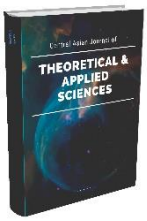


# CENTRAL ASIAN JOURNAL OF THEORETICAL AND APPLIED SCIENCES

**Volume: 04 Issue: 09 | Sep 2023 ISSN: 2660-5317**
**https://cajotas.centralasianstudies.org**


# Twitter Topic Modelling Using Latent Dirichlet Allocation Approach


**Uce Indahyanti, Yulian Findawati, Achmad Ariansyah**
Faculty of Science and Technology, Universitas Muhammadiyah Sidoarjo, Indonesia

**Endah Asmawati**
Faculty of Engineering, Universitas Surabaya, Indonesia



*Received 4th Jul 2023, Accepted 6th Aug 2023, Online 8th Sep 2023*



***Abstract:*** *This study aims to apply topic modeling from Twitter data about the Kanjuruhan tragedy, one of the trending topics due to a fatal incident that occurred after a football match at Kanjuruhan Stadium, in Malang, Indonesia. The research was conducted using the Latent Dirichlet Allocation (LDA), namely a text mining method to find certain patterns in a document by producing several different kinds of topics. The data used consists of 1480 tweets in the Indonesia language that had been pre-processed. This modeling has produced 5 main topics related to the Kanjuruhan tragedy such as the PSSI (Indonesian Football Association) investigation, suspects, the Itaewon tragedy, Korean netizens (Knetz), and tear gas. The implication of this research is not only to provide information about the comments and expectations of Twitter users regarding the Kanjuruhan tragedy but also to provide considerations for the stakeholder.*



***Keywords:*** *topic modeling, Twitter data, the Kanjuruhan tragedy, LDA, text mining.*


## 1. Introduction

Social media Twitter is a forum that is widely used by the public to express opinions and comments on popular issues. Twitter provides an online social networking and microblogging service, which enables its users to send and read text-based messages. public comments on social media Twitter is a huge body of text data, which can be mined and analyzed. To obtain hidden topics from the corpus (collection of natural texts), topic modeling can be applied using the most popular topic modeling approaches are LDA (Latent Dirichlet Allocation) and Latent Semantic Analysis (LSA). Each topic will represent a variety of comments discussing the same context.

Several studies related to topic modeling have been applied in various fields, such as bioinformatics [1] and transportation [2]. Twitter data-based topic modeling using the LDA method has also been carried out by several previous researchers [3][4][5]. Another study aims to make topic modeling to determine the topic of tweets about football news in Indonesian, using the LDA method, which has produced several topics such as pre-match analysis, live match updates, and football club achievements [6].

This study applies topic modeling based on tweet data about the tragedy at the Kanjuruhan Stadium, a fatal incident that caused hundreds of football spectators to die. The data used was taken from Twitter for the period of October 2022. We use Latent Dirichlet Allocation (LDA) as a topic modeling method to determine what topics appear on Twitter. The remainder of this paper consists of Section 2 describing the

materials and methods, section 3 describing the results and discussion, and section 4 explaining the conclusions.

## 2. Materials and Methods

### 2.1 Twitter Datasets

Twitter is a website owned and operated by Twitter, Inc. which offers a social network in the form of a microblog. This site allows users to send and read blog messages as usual but is limited to only 140 characters displayed on the user's profile page. Twitter has unique characteristics and writing formats with special symbols or rules. Messages on Twitter are known as Tweets. Twitter as one of the popular social media makes it very easy for its users to access a lot of information and channel their opinions [7].

The use of Twitter soars when an event that attracts public attention occurs, such as the Kanjuruhan tragedy. Tweet data related to the tragedy reached thousands from various user accounts. This study retrieved 2000 data from the Kanjuruhan tragedy tweets during the period of October 2022. After cleaning up the duplicate data, 1480 tweet data remained.

### 2.2. Topic Modelling

The concept of topic modeling consists of entities namely "word", "document", and "corpora". “Word” is considered the basic unit of discrete data in a document, defined as an item of vocabulary that is indexed for each unique word in the document. “Document” is an arrangement of N words. A corpus is a collection of M documents and corpora is the plural form of corpus. While "topic" is the distribution of some fixed vocabulary. each document in the corpus contains its own proportion of the topics discussed according to the words contained therein. Topic modeling has been of interest to most authors from the fields of Text Mining, Natural Language Processing, and Machine Learning [1].

The purpose of topic modeling is to determine topics automatically from a set of documents that have a hidden structure in the form of topics, distribution of topics per document, and determination of topics per word in each document. Topic modeling uses these documents to infer hidden topic structures. The number of topics to be generated has been determined before the topic modeling process is carried out. [2].

### 2.3. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of discrete data collections such as a set of documents (text corpus). In the context of text modeling, topic probabilities provide an explicit representation of a document. The basic idea of LDA is that a document consists of several topics. The LDA process is generative through an imaginary random process in a model that assumes that documents originate from a certain topic, and each topic consists of a distribution of words. The LDA concept is shown in Figure 1 [8].

 

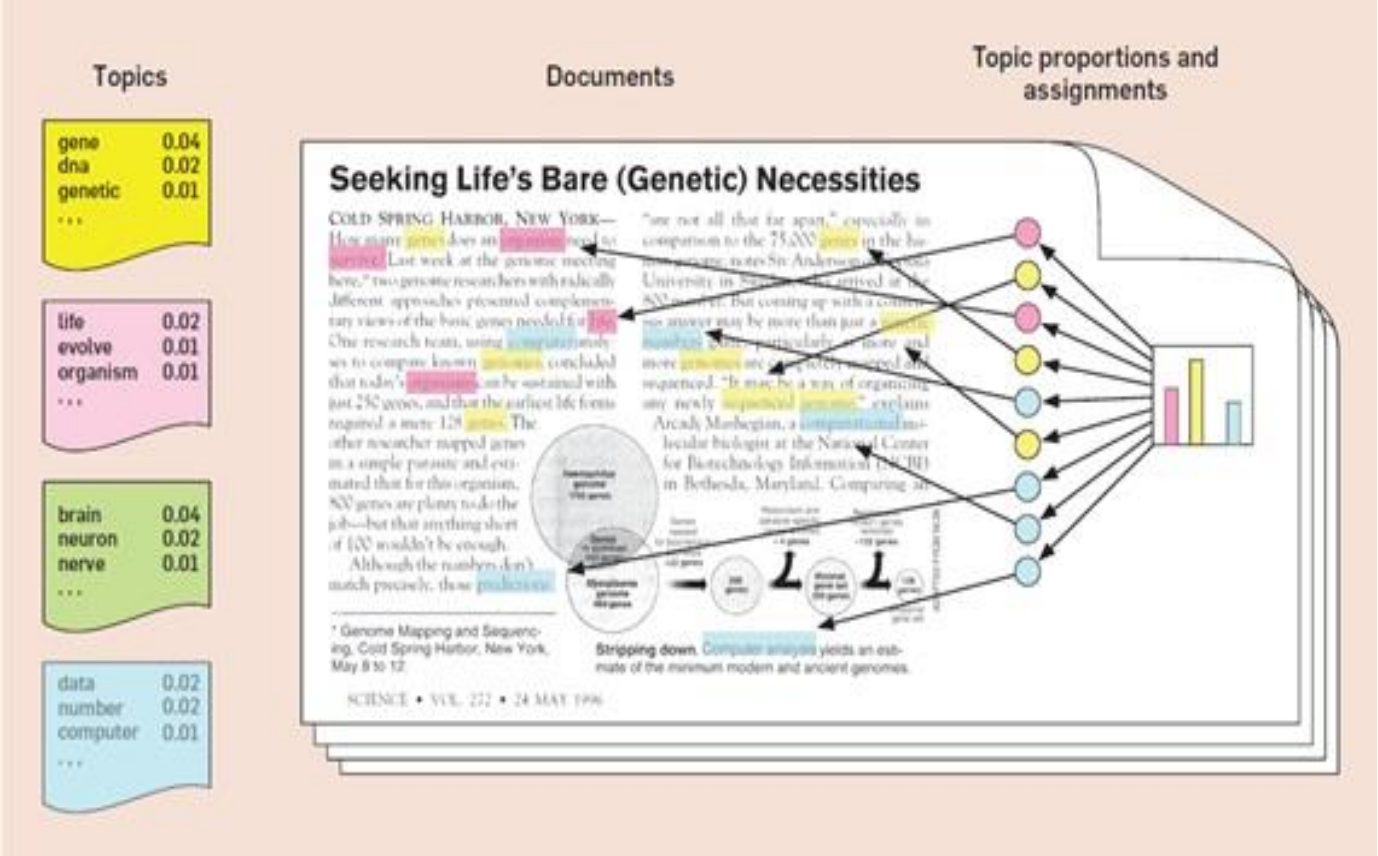


Fig. 1. LDA concept [9]

LDA is also called a text mining method for finding certain patterns in a document by generating several different types of topics [10]. LDA was chosen because it can analyze large data and documents. LDA uses the bag of words method to identify hidden topic information in large sets of documents [11].

### 2.4. Python dan Google Colab

Google colab is an executable document that can be used to store, write, and share programs that have been written via Google Drive. Google colab is a coding environment in a notebook format that is user-friendly and can support all needs related to data science and machine learning. This software is similar to Jupyter Notebook in the form of a cloud that runs using the Google Chrome browser. Meanwhile, Python is a popular programming language used in the Google Colab environment. Python is an open-source programming language, easy to use and has many supporting libraries for data science and machine learning needs. for example, text pre-processing used for topic modeling using the Python programming language was carried out by [6] [12].

### 2.5. Modelling Stages

The stages of topic modeling in this study are shown in Figure 2, starting from the data collection stage, data pre-processing, topic modeling, visualization, until the results analysis.

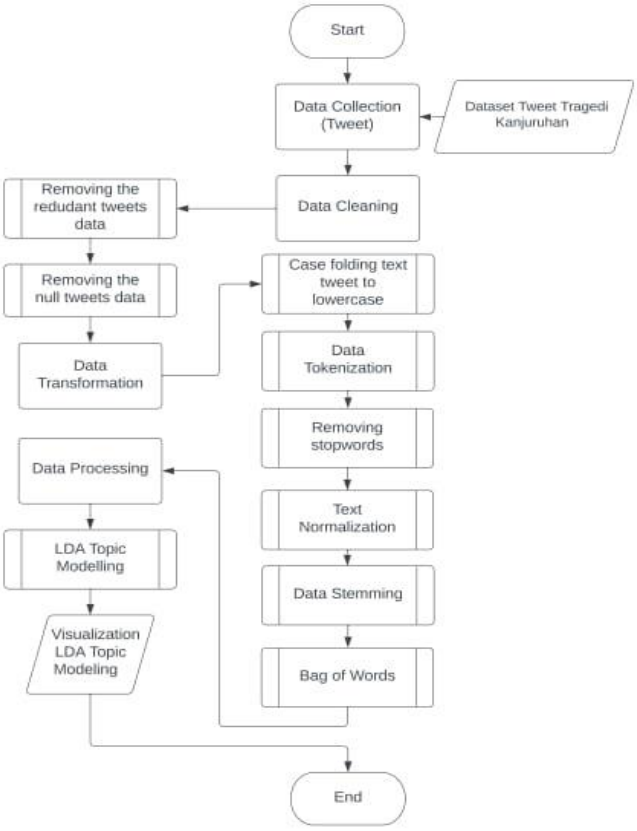


Fig. 2. Modelling stages

**A.** Data Collection

Data retrieval using the Twint library with the keywords "Tragedi Kanjuruhan", language "id", limit 2000, period 1 October to 31 October 2022, from several accounts, including official accounts such as detik.com, jawapos.com , and hariankompas.com.

**B.** Data Pre-processing

The preprocessing stage consists of cleaning data, selecting attributes, case folding (changing into lowercase), tokenizing (removing unnecessary characters or symbols), stopwords (cleaning text from words that have no meaning), normalizing (replacing certain words with more appropriate words such as jatim to Jawa Timur), stemming (cutting affixes to text using the Sastrawi and Swifter packages).

**C.** LDA Topic Modelling

LDA topic modeling in this study uses the LdaModel library provided by the Gensim library with Python [6]. We determine five topics as parameters, and the following is a modeling code snippet.

```
import gensim
from gensim import corpora
Lda = gensim.models.LdaModel
dictionary = corpora.Dictionary(doc_clean)
bow_corpus = [dictionary.doc2bow(doc) for doc in doc_clean]
```

```
total_topics = 5
number_words = 8
```

**D.** Visualization

The results of LDA topic modeling are visualized using the Gensim library and pyLDAvis in Python. PyLDAvis is a web-based interactive topic model visualization using LDA built from LDAvis using a combination of R and D3 [13]. The pyLDAvis library, used for browsing relationships between topics and terms to understand LDA model. PyLDAvis has two panels, the distribution map each topic and the most representative intensity graph terms frequently found in the corpus. The following is a visualization code snippet.

```
import pyLDAvis.gensim
import pickle
import pyLDAvis
import os

# Visualize the topics
pyLDAvis.enable_notebook()
LDAvis_data_filepath =
os.path.join('ldavis_prepared_'+str(total_topics))
corpus = [dictionary.doc2bow(text) for text in doc_clean]
if 1 == 1:
LDAvis_prepared= pyLDAvis.gensim.prepare(lda_model, corpus,
dictionary)
with open(LDAvis_data_filepath, 'wb') as f:
pickle.dump(LDAvis_prepared, f)
```

## 3. Results and Discussion

This section discusses the results of topic modelling. This modeling has produced 5 main topics related to the Kanjuruhan tragedy such as the Itaewon tragedy (topic #1), the PSSI investigation (topic #2), suspects (topic #3), Korean netizens/Knetz (topic #4), and tear gas (topic #5). Table 1 shows the results of the bag of words weighting. We determine eight words as parameters (K = kata = word) and translate in English for common words.

Table 1. The results of the bag of words weighting

| Topic | K1 (Word1) | K2 | K3 | K4 | K5 | K6 | K7 | K8 |
|---|---|---|---|---|---|---|---|---|
| Topic #1 | itaewon (63%) | october (40%) | victim (31%) | halloween (16%) | people (13%) | dead (12%) | stadium (8%) | closed (8%) |
| Topic #2 | indonesia (23%) | investigate (15%) | pssi (14%) | ball (14%) | thoroughly (14%) | suporter (12%) | stay (12%) | soccer (9%) |
| Topic | suspect | kapolri | pssi | permanent | lib (19%) | pt | malang | director |

 

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| #3 | (57%) | (26%) | (23%) | (20%) | | (16%) | (14%) | (12%) |
| Topik #4 | Country (22%) | knetz (15%) | lu (10%) | indo (10%) | people (10%) | gw (10%) | itaewon (8%) | yesterday (8%) |
| Topik #5 | eye (19%) | people (17%) | itaewon (12%) | pas (12%) | gas (11%) | water (11%) | hope (10%) | victim (10%) |

The distance map visualization between topics from this model and the top 30 most prominent words in the corpus is shown in Figure 3, which is one of the results of modeling visualization. The bar chart in Figure 3 shows the 30 most prominent words in the corpus on the topic "Itaewon". Figure 3 shows five topic clusters that can be grouped independently. These clusters cover topics that can be seen from a distance between clusters, and explain that the distribution and frequency of words within these topics is very unique. The word "Itaewon" appeared at the top because of the many comments by Indonesian netizens who replied to comments by Korean netizens. Previously, many Korean netizens commented on the Kanjuruhan tragedy. Examples of other topic visualizations are shown in Figure 4.

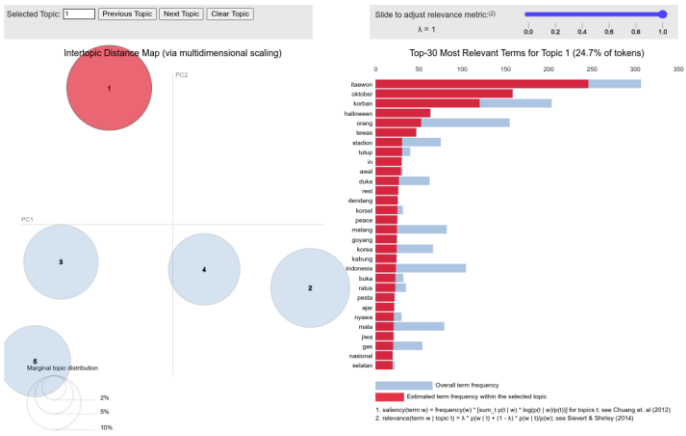


Fig.3. LDA visualization (topic #1 - the Itaewon tragedy)

Figure 4 shows the 30 most prominent words in the corpus on the topic "suspects". The bar chart in Figure 4 illustrates the distribution of words that refer to the topic of the suspect in the Kanjuruhan tragedy.

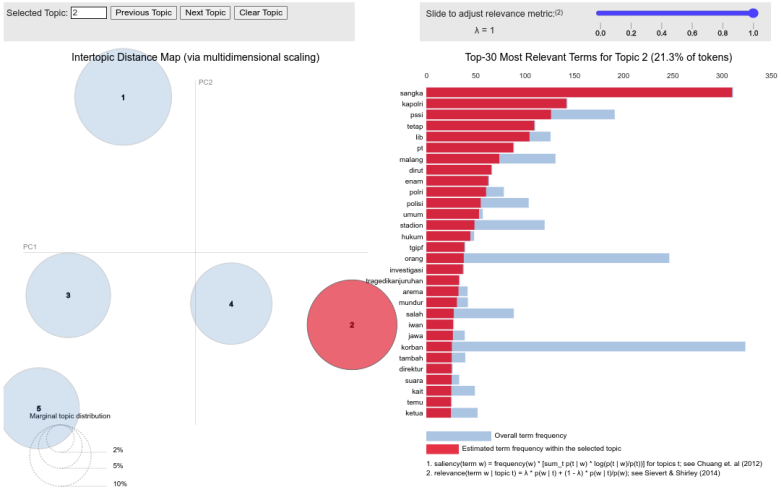


Fig.4. LDA visualization (topic #2 - the PSSI investigation)

 

## 4. Conclusion

This modeling has produced 5 main topics related to the Kanjuruhan tragedy such as the PSSI (Indonesian Football Association) investigation, suspects, the Itaewon tragedy, Korean netizens (Knetz), and tear gas. The word "Itaewon" appeared at the top because of the many comments by Indonesian netizens who replied to comments by Korean netizens. Previously, many Korean netizens commented on the Kanjuruhan tragedy.

The implication of this research is not only to provide information about the comments and expectations of Twitter users regarding the Kanjuruhan tragedy but also to provide considerations for the stakeholder. Meanwhile, this study still needs to be improved such as the use of metric coherence scores in determining the number of topics. To find out more about the performance of the LDA method in extracting topics from Bahasa Indonesian text documents, by comparing this method with other non-topic based methods.

## 5 Acknowledgements

This research is supported by Universitas Muhammadiyah Sidoarjo (UMSIDA).

## References

1. L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, “An overview of topic modeling and its current applications in bioinformatics,” *Springerplus*, vol. 5, no. 1, 2016, doi: 10.1186/s40064-016-3252-8.
2. L. Sun and Y. Yin, “Discovering themes and trends in transportation research using topic modeling,” *Transp. Res. Part C Emerg. Technol.*, vol. 77, no. April, pp. 49–66, 2017, doi: 10.1016/j.trc.2017.01.013.
3. G. Lansley and P. A. Longley, “The geography of Twitter topics in London,” *Comput. Environ. Urban Syst.*, vol. 58, pp. 85–96, 2016, doi: 10.1016/j.compenvurbsys.2016.04.002.
4. A. F. Hidayatullah and and M. R. Ma’arif, “Pre-processing Tasks in Indonesian Twitter Messages,” *J. Phys. Conf. Ser.*, 2017, doi: 10.1088/1742-6596/755/1/011001.
5. H. Jelodar *et al.*, “Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey,” *Multimed. Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, 2019, doi: 10.1007/s11042-018-6894-4.
6. A. F. Hidayatullah, E. C. Pembrani, W. Kurniawan, G. Akbar, and R. Pranata, “Twitter Topic Modeling on Football News,” *2018 3rd Int. Conf. Comput. Commun. Syst. ICCCS 2018*, pp. 94–98, 2018, doi: 10.1109/CCOMS.2018.8463231.
7. A. A. Amrullah, A. Tantoni, N. Hamdani, R. T. R. L. Bau, and E. U. Ahsan, Muhammad Rafiqudin, “Review Atas Analisis Sentimen Pada Twitter Sebagai Representasi Opini Publik Terhadap Bakal Calon Pemimpin. Prosiding Seminar Nasional Multi Disiplin Ilmu & Call For Papers Unisbank,” 2016.
8. A. Y. N. I. J. David M. Blei, “Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
9. D. Blei, L. Carin, and D. Dunson, “Probabilistic topic models,” *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55–65, 2010, doi: 10.1109/MSP.2010.938079.
10. I. M. K. B. Putra and R. P. Kusumawardani, “Analisis Topik Informasi Publik Media Sosial Di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation ( LDA ),” *J. Tek. Its*, vol. 6, no. 2, pp. 2–7, 2017.

11. Y. Sahria, "Analisis Topik Penelitian Kesehatan di Indonesia Menggunakan Metode Topic Modeling LDA (Latent Dirichlet Allocation)," *Resti*, vol. 4, no. 2, pp. 336–344, 2020.
12. M. Cendana and S. D. H. Permana, "Pra-Pemrosesan Teks Pada Grup Whatsapp Untuk Pemodelan Topik," *Junal Mantik Penusa*, vol. 3, no. 3, pp. 107–116, 2019.
13. C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," no. September, pp. 63–70, 2015, doi: 10.3115/v1/w14-3110.

 