Journal of Applied Data Sciences
Vol. 4, No. 3, September 2023, pp. 103-118

Unsupervised Learning Methods for Topic Extraction and Modeling in Large-scale Text Corpora using LSA and LDA

Henderi [1,*]; B Herawan Hayadi [2]; Sofa Sofiana [3]; Padeli [4]; Didik Setiyadi [5];

[1] Informatics, University of Raharja, Indonesia
[2] Information Technology Education, Bina Bangsa University, Indonesia
[3] Informatics, Pamulang University, Tangerang, Indonesia
[4] Information Technology Education, University of Raharja, Indonesia
[5] Informatics, Indonesia Mandiri University, Indonesia
[1] henderi@raharja.info*; [2] b.herawan.hayadi@gmail.com; [3] dosen00407@unpam.ac.id; [4] padeli@raharja.info; [5] ddk.setiyadi20@gmail.com
* corresponding author

(Received June 4, 2023; Revised July 17, 2023; Accepted July 17, 2023; Available online September 1, 2023)

Abstract

This research compares unsupervised learning methods in topic extraction and modeling in large-scale text corpora. The methods used are Singular Value Decomposition (SVD) and Latent Dirichlet Allocation (LDA). SVD is used to extract important features through term-document matrix decomposition, while LDA identifies hidden topics based on the probability distribution of words. The research involves data collection, data exploratory analysis (EDA), topic extraction using SVD, data preprocessing, and topic extraction using LDA. The data used were large-scale text corpora. Data explorative analysis was conducted to understand the characteristics and structure of text corpora before topic extraction was performed. SVD and LDA were used to identify the main topics in the text corpora. The results showed that SVD and LDA were successful in topic extraction and modeling of large-scale text corpora. SVD reveals cohesive patterns and thematically related topics. LDA identifies hidden topics based on the probability distribution of words. These findings have important implications in text processing and analysis. The resulting topic representations can be used for information mining, document categorization, and more in-depth text analysis. The use of SVD and LDA in topic extraction and modeling of large-scale text corpora provides valuable insights in text analysis. However, this research has limitations. The success of the methods depends on the quality and representativeness of the text corpora. Topic interpretation still requires further understanding and analysis. Future research can develop methods and techniques to improve the accuracy and efficiency of topic extraction and text corpora modeling.

Keywords: Topic Extraction, Topic Modeling, Large-Scale Text Corpora, SVD, LDA, Unsupervised Learning

1. Introduction

In the increasingly advanced digital information age, the amount of text generated every day reaches a tremendous scale. Large-scale text corpora, such as collections of news articles, e-books, social media, and academic documents, hold priceless information potential. [1], [2]. However, searching for relevant topics in such corpora has become an increasingly complicated and challenging task. It is not only related to the large volume of text, but also to the diversity of topics, the complex structure of the documents, and the interconnectedness of the documents. [3]-[5]. In this context, efficient and effective methods are needed to extract topics and model them in large-scale text corpora.

Topic search in large-scale text corpora faces several significant problems. First, the massive volume of text makes it difficult to manually identify key topics. [6], [7]. The process is time consuming and cumbersome, especially when dealing with millions or even billions of documents. Secondly, complex document structures and variations in writing styles can obscure topics. Different documents may contain related information, but with different writing styles, making topic identification difficult. Third, the interconnectedness of documents in large-scale text corpora implies that topic modeling must consider the context and information conveyed by related documents.

In this context, an automated and efficient approach for topic extraction and modeling in large-scale text corpora is required. The method should be able to overcome the challenges of large text volumes, identify related topics, and

cope with variations in writing styles and interrelationships between documents. In this study, an approach using unsupervised learning methods based on SVD and LDA is proposed to address these issues and produce accurate and comprehensive topic modeling in large-scale text corpora.

The goal of this research is to develop an efficient unsupervised learning method for topic extraction and modeling in large-scale text corpora. The method is expected to produce accurate topic representations and can be used to gain valuable insights from such text corpora. This research has the novelty of using SVD (Singular Value Decomposition) and LDA (Latent Dirichlet Allocation) techniques in the context of topic extraction and modeling in large-scale text corpora. SVD is used to reduce the dimensionality of the data and obtain a more compact representation, while LDA is used to model the topics present in the corpora by assuming a hidden topic distribution. The research questions to be answered in this study include:

1) How can the application of SVD and LDA methods help in topic extraction and modeling in large-scale text corpora?
2) How accurate and efficient is this method in identifying key topics in large-scale text corpora?
3) How can the resulting topic representations provide valuable insights from the text corpora?

By answering these questions, this research is expected to make significant contributions to the development of methods for topic extraction and modeling in large-scale text corpora, as well as improve our understanding of structure and content in complex text corpora.

2. Literature Review

2.1. Unsupervised Learning

Unsupervised Learning is a method in machine learning that allows computers to recognize patterns and structures in data without any predefined annotations or labels. [8], [9]. In unsupervised learning, the computer is given the task to explore the data and discover hidden patterns or relationships independently. In other words, this method allows the computer to learn on its own without clear instructions [10].

One of the common approaches in unsupervised learning is clustering. [10]. The goal is to group data based on similarities or patterns present in it. In clustering, the algorithm will look for similar patterns among a number of data and group them into appropriate categories. This helps to identify similar groups of data and understand the overall structure of the data.

Besides clustering, unsupervised learning also includes dimension-reduction methods, such as Singular Value Decomposition (SVD) and Principal Component Analysis (PCA). [10]. The goal is to reduce the dimensionality of the data while retaining important information. In this process, data that initially has many attributes or features is represented in a lower feature space, making it easier to understand and analyze the data.

Unsupervised learning approaches are often used in a variety of fields, including natural language processing, pattern recognition, data analysis, and intelligent computing. [11]. In natural language processing, these methods can be used for extracting information from text, such as topic identification, sentiment analysis, and language modeling. In pattern recognition, unsupervised learning can help in face recognition, speech recognition, and object classification. In data analysis, it can be used to discover hidden patterns in business or scientific data.

With unsupervised learning, computers can automatically and efficiently extract valuable information from data without the need for human supervision. These methods play an important role in big data processing and analysis, helping us understand patterns and relationships that exist in highly complex data and are useful in decision-making, modeling, and forecasting in various fields.

2.2. Singular Value Decomposition

Singular Value Decomposition (SVD) is a method in linear algebra to analyze and obtain useful information from matrices. [12]-[15]. SVD involves the decomposition of a matrix into three main components: the orthogonal matrix

U, the diagonal matrix containing the singular values, and the orthogonal matrix V. This method is widely used in various applications, including image processing, language processing, and natural language processing. This method is widely used in various applications, including image processing, natural language processing, data analysis, and topic modeling.

In the context of text processing and data analysis, SVD can be used to reduce the dimensionality of term-document matrices of text corpora. In such matrices, rows represent words or terms, while columns represent documents. [16]-[19]. SVD allows for a more compact matrix representation by identifying key patterns or significant singular values. Thus, SVD can help in eliminating information redundancy and obtaining a more efficient representation.

The main advantage of SVD is its ability to reveal hidden data structures. By looking at the largest singular values, we can identify the most important dimensions in the data. For example, in text processing, the largest singular values can represent key topics in text corpora. Therefore, SVD is often used in topic analysis, document classification, and text-based recommendation systems. [17].

In addition, SVD can also be used for data reconstruction. By using some of the largest singular values, we can approximate the original matrix with controllable precision. This approach is useful in noise or outlier removal, data compression, and recovery of missing or incomplete information.

In conclusion, SVD is a powerful and flexible method in data analysis. In the context of text processing, SVD can help in topic extraction, modeling, and efficient text processing. By understanding the singular values and patterns present in data, we can uncover hidden structures and utilize them for various applications in data science and artificial intelligence.

2.3. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a method in natural language processing used to reveal semantic relations between words and documents in a text corpus. It is based on the mathematical representation of words and documents in the form of a term-document matrix, where rows represent words and columns represent documents. [20]-[24]. LSA aims to extract hidden or latent patterns that reflect the semantic relationships between words and documents.

In LSA, the term-document matrix is then decomposed using the Singular Value Decomposition (SVD) method. SVD decomposes the matrix into three new matrices: the singular value matrix, the words matrix, and the documents matrix. The main components of the singular value matrix describe the main information in the data, while the words matrix and document matrix describe the relationship between words and documents in the latent space. [25], [26].

By using LSA, we can calculate the similarity between words or documents based on the cosine distance between their representation vectors in latent space. LSA allows us to find words that have similar meanings or documents that have similar topics based on the similarity of semantic patterns found through SVD decomposition.

The advantage of LSA is its ability to overcome the problems of synonymy, polysemy, and ambiguity in language processing. By combining context information and the distribution of words, LSA can reveal semantic relationships that are not directly visible in the text. However, LSA also has some limitations, such as the lack of intuitive interpretation and difficulty in handling very large text corpora.

LSA has been used in various applications, including information retrieval, document categorization, and recommendation systems. In the context of information retrieval, LSA can assist in finding relevant documents based on topic similarity, even if the words used in the query or search are imprecise. In document categorization, LSA can help in grouping documents based on similar topics, even if the words used in the documents are different. Overall, LSA is a useful approach in text analysis and processing to reveal semantic structures in text corpora.

2.4. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative model used in natural language processing and topic modeling. It enables the identification of topics hidden in text corpora by assuming a distribution of topics hidden in documents. [27], [28]. LDA provides a powerful approach to understanding topic structure in complex text data.

In LDA, each document is considered as a linear combination of hidden topics. These topics are considered as probability distributions over the words in the corpus. LDA assumes that there is a document-level Dirichlet distribution that describes the proportion of topics in each document, as well as a word-level Dirichlet distribution that describes the occurrence of words in each topic. By using modeling algorithms, LDA can estimate the probability distributions of topics and words that best fit the data. [29].

The main goal of LDA is to identify topics present in a text corpus without any predefined annotations or labels. Using this approach, LDA can uncover the topic structure hidden in the corpus and identify the words that are most related to each topic.

LDA has been used in various applications, such as information retrieval, recommendation systems, and text analysis. In information retrieval, LDA helps in identifying topics that are most relevant to a user's search query. In recommender systems, LDA is used to understand users' interests and provide relevant recommendations based on their topics of interest. In text analysis, LDA can help in document classification, sentiment analysis, and understanding issues that arise in text corpora. [30].

In conclusion, LDA is a powerful and effective model in topic modeling in text corpus. By using topic and word probability distributions, LDA can uncover the topic structure hidden in text data. This method paves the way for a deeper understanding of text content, document clustering, and topic analysis in various applications that require rich understanding of large and complex texts.

2.5. Corpora Text

Text corpora are collections of text documents used as data sources for natural language analysis and processing. Text corpora can be collections of news articles, academic documents, web texts, social media posts, or other text collections. Text corpora hold large and diverse potential information, which can be explored to understand patterns, relationships, and topics present in the text.

Text corpora are often used in various applications such as natural language processing, text analysis, and information mining. In natural language processing, text corpora are used to train language models, understand grammar, and build text processing algorithms. Text analysis utilizes text corpora to identify patterns and trends, perform sentiment analysis, and perform information extraction. Information mining utilizes text corpora to search, extract, and present relevant information based on a specified query or topic.

It is important to have representative and diverse text corpora to make the results of natural language analysis and processing more accurate and meaningful. The size of the text corpora also plays an important role, the larger the text corpora, the more information available and the more varied topics and issues that can be identified. However, managing and processing large-scale text corpora is also challenging, as it requires adequate storage and efficient algorithms.

In addition, text corpora can also present several challenges such as the diversity of writing styles, the complexity of document structures, and the presence of noise or irrelevant information. Therefore, processing text corpora requires effective methods and algorithms for information extraction, topic modeling, and accurate analysis.

2.6. Past Related Study

Many previous studies similar to the current study have been conducted to explore methods for extracting and modeling topics in large-scale text corpora using unsupervised learning methods such as SVD and LDA. Previous studies have proven the effectiveness and usefulness of these methods in understanding the topic structure in text.

Some previous studies [27]-[29] have adopted SVD as an approach for text corpora processing. For example, research conducted by [31] used SVD to reduce the dimension of term-document matrix and identify key topics in news corpora. The results showed that SVD was successful in extracting cohesive topics and eliminating redundant information.

In addition, some previous studies have also implemented LDA for topic modeling in text corpora. For example, research by [32] applied LDA to research article corpora and successfully identified topics hidden in these documents. This research shows that LDA can be used to extract useful and in-depth information in text corpora.

Previous research has also compared the two methods, SVD and LDA, in the context of topic modeling. For example, research by Wang et al. (2014) compared the performance of SVD and LDA in identifying topics in text corpora. The results showed that LDA produced better and more cohesive topic modeling compared to SVD.

However, although previous research has provided valuable insights into the use of unsupervised learning methods for topic extraction and modeling in text corpora, there are still some shortcomings that need to be addressed. For example, the challenge of dealing with noise or irrelevant information in text corpora, as well as the complexity of handling large-scale text corpora that require high computational efficiency. Therefore, the current research aims to address these shortcomings and develop more advanced and effective methods for topic extraction and modeling in large-scale text corpora.

3. Methodology

This research follows a multi-stage flow to analyze and model topics in text corpora using unsupervised learning methods such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).

3.1. Data Collection

The first stage is Data Collection, where text data is collected from relevant sources such as articles, academic documents, or web texts. This data collection aims to build a representative and diverse text corpora that will inform the research.

3.2. EDA (Exploratory Data Analysis)

The second stage is EDA (Exploratory Data Analysis), where text data is analyzed descriptively to understand its characteristics. EDA involves descriptive statistics, data visualization, and an initial introduction to the distribution of words, document length, and common patterns in text corpora. The purpose of this stage is to gain initial insight into the data and identify interesting patterns for further analysis.

3.3. Topic Modeling

After EDA, the next stage is Topic Modeling, which is the core of this research. In this stage, methods such as LSA and LDA are used to identify topics hidden in text corpora. LSA uses SVD to reduce the dimension of the term-document matrix and extract topic information, while LDA uses generative models to estimate the probability distribution of topics and words. The result of this stage is topic modeling that provides insight into the topics present in text corpora.

3.4. Preprocessing

The next stage is Preprocessing, where the subject text is cleaned up and prepared before further analysis. Preprocessing involves steps such as stop words removal, stemming, punctuation removal, and text normalization. The purpose of this stage is to improve the quality of the data and prepare it to suit the analysis method to be used.

3.5. LSA, LDA, and SVD

Finally, the Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) stages are used to analyze and model topics in text corpora. LSA identifies key dimensions in the text data and reveals hidden semantic patterns, while LDA identifies hidden topics and the distribution of words within each topic. Through these analyses, this research aims to gain a deeper understanding of the topics present in text corpora and the relationship between words within each topic.

With this research flow, it is expected to generate a better understanding of topic structure in text corpora and provide valuable insights for applications such as information retrieval, text analysis, and topic modeling. Figure 1 is the flow of this research.

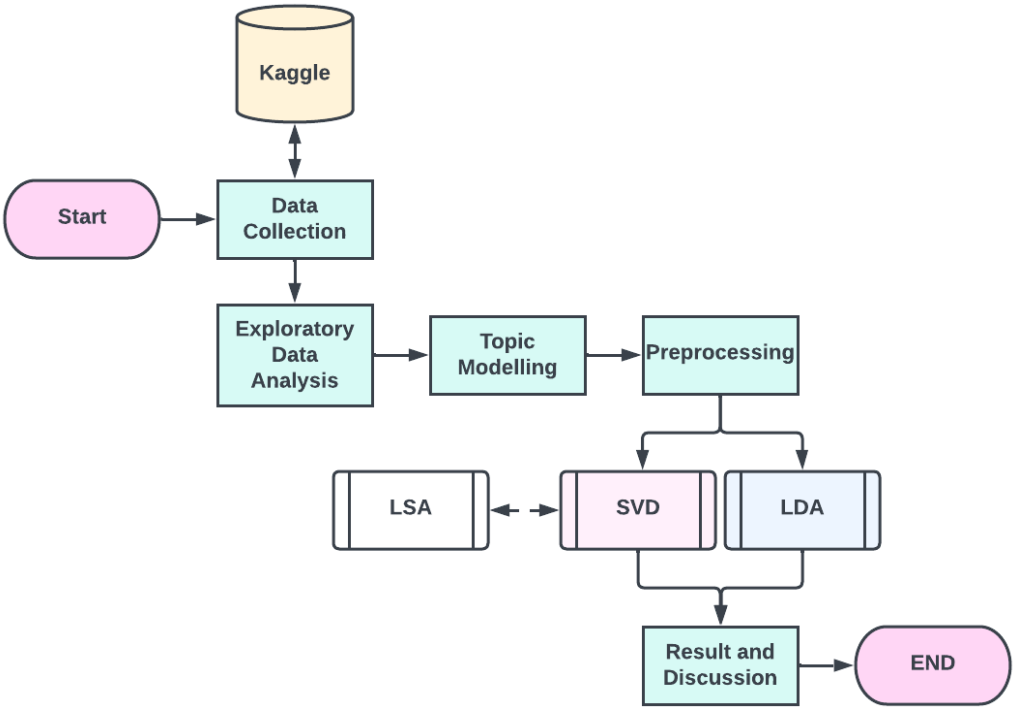

Figure. 1. Research Steps

4. Results and Discussion

4.1. Exploratory Data Analysis

At the EDA stage, we developed a list of the top words used in all one million news headlines, providing an overview of the core vocabulary of the source data. Stop words are removed here to avoid unnecessary conjunctions, prepositions, and so on. Table 1 below is a sample of the data used.

Table. 1. Sample dataset

	publish_date	headline_text
0	2003-02-19	aba decides against community broadcasting lic...
1	2003-02-19	act fire witnesses must be aware of defamation
2	2003-02-19	a g calls for infrastructure protection summit
3	2003-02-19	air nz staff in aust strike for pay rise
4	2003-02-19	air nz strike to affect australian travellers

Next, we generate a histogram of the length of the title words and use part-of-speech tagging to understand the types of words used in the corpus (presented in figure 2 below). This is done by converting all headline strings into TextBlobs and calling the pos_tags method on each headline, resulting in a list of tagged words for each headline. Table 2 below shows the Mean and Total words per headline.

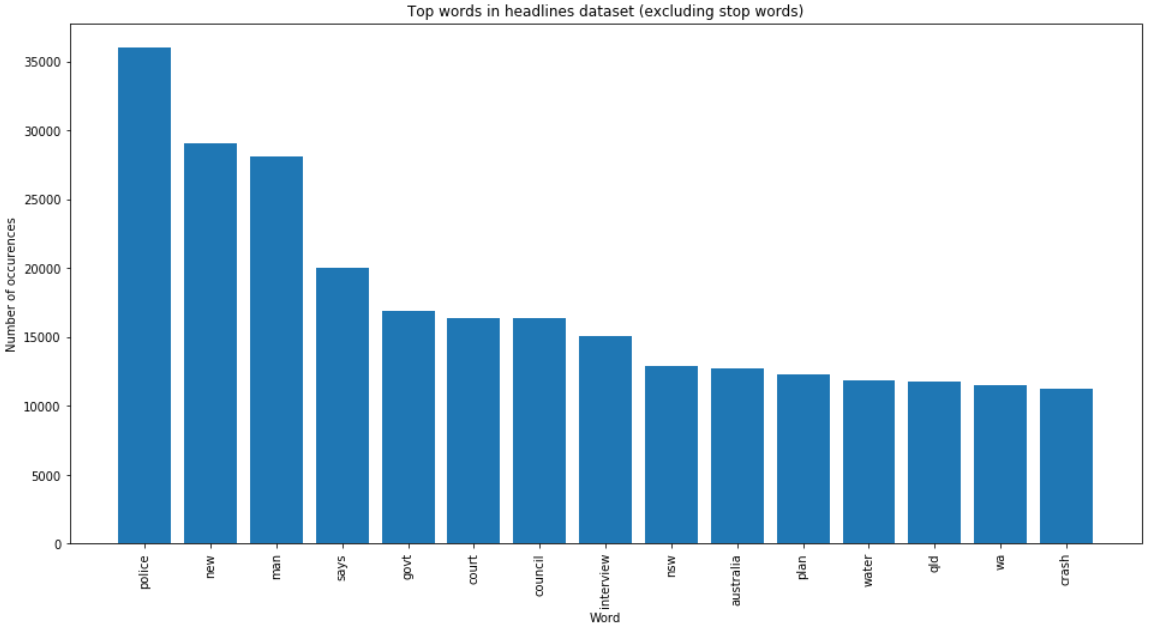

Figure. 2. Histogram of headline word lengths

Table. 2. Mean and total number of words per headline

Total number of words	Mean number of words per headline
7179551	6.295543927301008

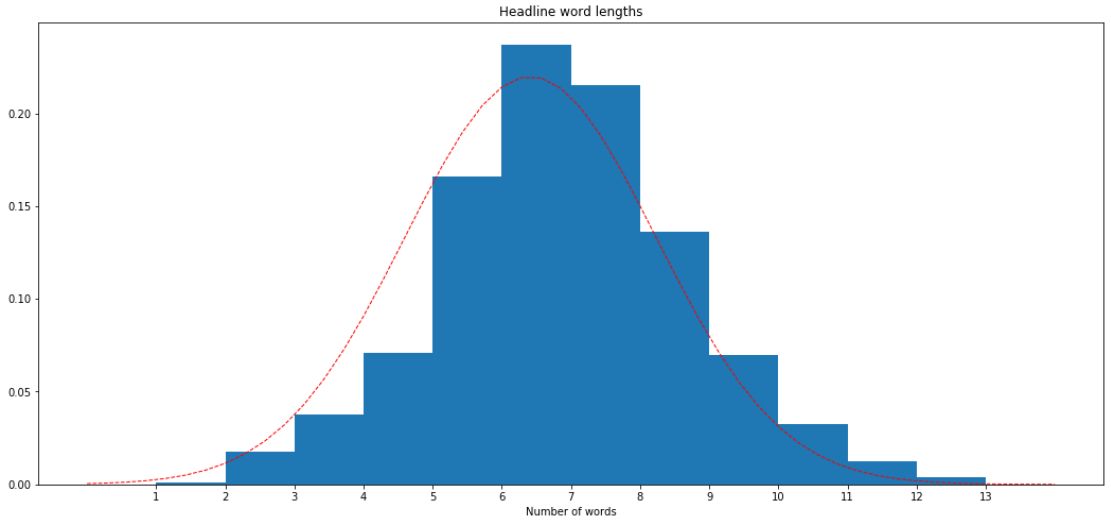

Figure. 3. Headline word lengths

By plotting the number of titles published per day, per month, and per year (attached below in figure 4), we can also get an idea of the sample density. These plots provide information about the publication patterns of titles in a given time period, helping us understand the crowding and trends in the corpus data.

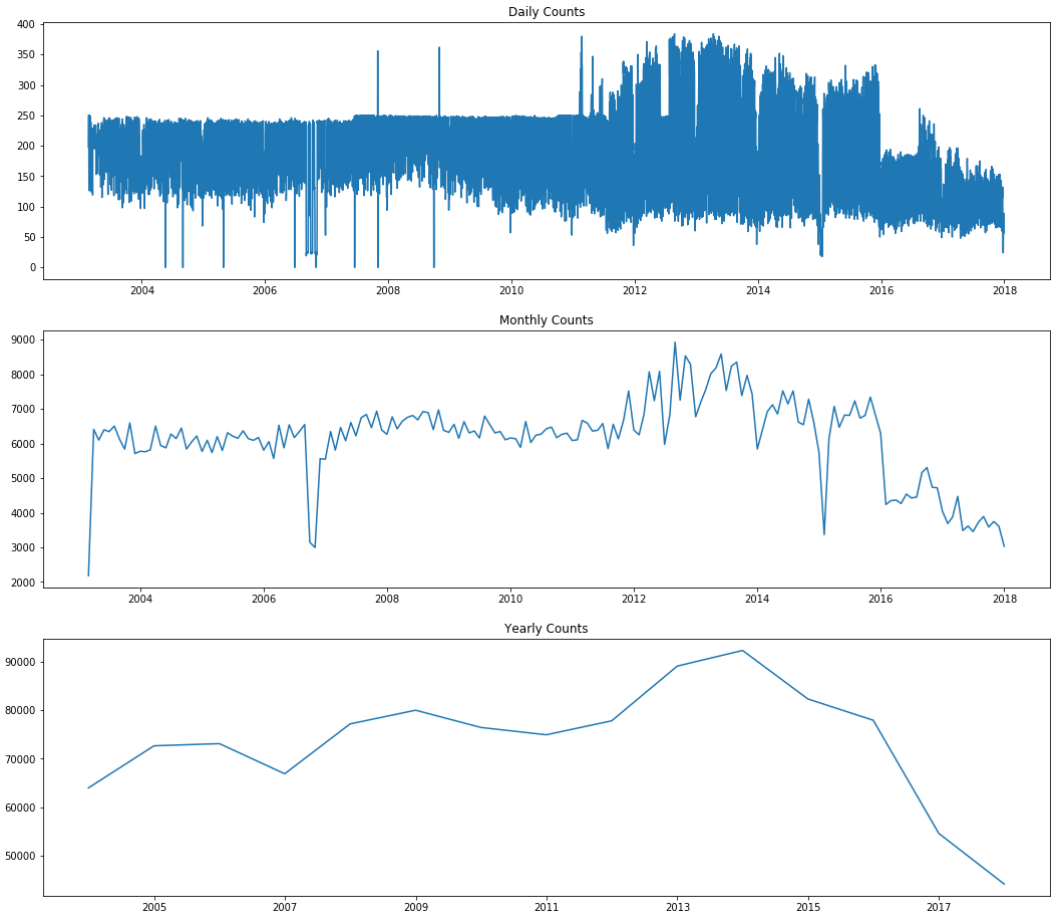

Figure. 4. The frequency of headlines published on a daily, monthly, and yearly basis

Through this stage of EDA, we gained an initial understanding of the characteristics of the text data that we would analyze further. We looked at the patterns of the most commonly used words in the headlines, gained insight into the sentence structure, and identified publication patterns in the corpus. This information will help us direct the subsequent analysis in this study, as well as provide the necessary context to understand the corpus of texts we are exploring.

4.2. Topic Modeling

This research then applies clustering algorithms to a corpus of news headlines to learn about ABC News' topic focus, as well as how that focus has evolved over time. To do so, the research first conducted experiments with a small sample of the dataset to determine the most suitable clustering algorithm. Once that was confirmed, the research then extended the analysis to a larger subset of the available data.

Clustering is a technique that groups similar data based on their characteristics. By applying this approach to a corpus of news headlines, this research aims to identify different clusters that represent different topic categories or themes. This allows this study to find patterns and trends in ABC News' news coverage over the analyzed time period.

Through clustering analysis, this research was able to see how the topics covered by ABC News have evolved over time. By examining the composition of each group and analyzing the headlines within it, this research gained a comprehensive understanding of the dominant themes and the changes that have occurred. This analysis provides valuable insight into the editorial focus of ABC News and reveals the changing patterns of news coverage.

By extending the analysis to a larger portion of the available data, this research was able to increase the reliability and representativeness of the findings. This expansion allowed this research to conduct a more comprehensive exploration of the topic landscape in the ABC News corpus and strengthen the validity of conclusions about topic focus and its evolution over time.

4.2. Preprocessing

In the Preprocessing stage, the only step required in our case is feature building, where we sample text titles and represent them in a manageable feature space. In practical terms, this means converting each string into a numeric vector. This can be done using SKLearn's CountVectorizer object, which produces a document-word matrix of size n × K where K is the number of distinct words across the n text titles in our sample (with stop word removal and max_features constraints).

Thus, we have our very high-rank and sparse training data, i.e. small_document_term_matrix, and can now implement a clustering algorithm. Our choice will fall on either Latent Semantic Analysis or Latent Dirichlet Allocation. Both will take our document-word matrix as input and produce a topic matrix of size n × N as output, where N is the number of topic categories (which we provide as a parameter). For now, we will take the value of N as 8.

In the preprocessing process, we focus on feature construction, where we convert text titles into a numerical representation that can be processed. We use SKLearn's CountVectorizer object to convert each string into a numeric vector. The result is a document-word matrix with dimensions n × K, where n is the number of text headings in our sample, and K is the number of distinct words across text headings. We also perform stop word removal and limit the number of features by using the max_features parameter.

With the training data that we have built, i.e. small_document_term_matrix which is very high dimensional and sparse, we can proceed to the next step, which is the implementation of the clustering algorithm. Our choice is to use the Latent Semantic Analysis (LSA) or Latent Dirichlet Allocation (LDA) method. Both methods will use the document-word matrix as input and produce a topic matrix with dimensions n × N as output, where N is the number of topic categories we specify. For this study, we chose an N value of 8.

Thus, through the preprocessing stage, we managed to construct the necessary features for topic analysis and modeling. Through text conversion to numerical representation and stop word removal, we generate a document-word matrix that will be used as input in the subsequent clustering algorithm. In this study, we chose to use LSA or LDA algorithms to analyze the matrix and generate a topic matrix that will provide insight into the topics present in the text corpus.

4.3. Latent Semantic Analysis and Latent Dirichilet Allocation

The first stage in the LSA (Latent Semantic Analysis) phase is to perform a truncated singular value decomposition (SVD) of the document-term matrix. This decomposition reduces the highly ranked and sparse matrix to a lower ranked representation, retaining only the r largest values. By doing this, we effectively capture the hidden semantic structure in the corpus.

After obtaining the truncated SVD, we can proceed to determine the predicted topic for each headline in the sample. This is done by taking the maximum value (argmax) of each headline in the topic matrix, which assigns the headline to the topic with the highest score. By sorting and counting those assigned topics, we can gain insight into the topic distribution in the corpus.

However, these topic categories may not have a clear interpretation. To better understand and describe the topics, it is helpful to identify the words that appear most frequently within each topic. By examining the top words in each topic, we can gain a better understanding of the underlying themes and concepts.

For example, the given sample output shows eight topics, where each topic is represented by a set of frequently occurring words. These topics cover a range of subjects such as police investigations, court proceedings, government reports, health, and local council issues. This division allows us to understand the main themes present in the corpus.

Topic 1: police death crash probe car drug missing woman attack fatal

Topic 2: man charged with murder dies jailed court accused guilty bail arrested

Topic 3: new laws abc year years sets weather cancer trial queensland

Topic 4: says wa report government helped australian group interview iraq claims

Topic 5: court interview face high accused charges trial told faces ban

Topic 6: govt water qld urged work act vic closer funds considers

Topic 7: council plan election backs center takes fears business park lake

Topic 8: australia health nsw coast world win hospital south cup

To analyze and compare the results with other clustering algorithms, a dimensionality reduction technique called t-SNE (t-distributed Stochastic Neighbor Embedding) is used. t-SNE visualizes the relationship between data points in a lower dimensionality space, providing insight into the success of the clustering process and possibly uncovering patterns or clusters that may not be visible in the original higher dimensionality space. Figure 5 below is the calculated LSA topic.

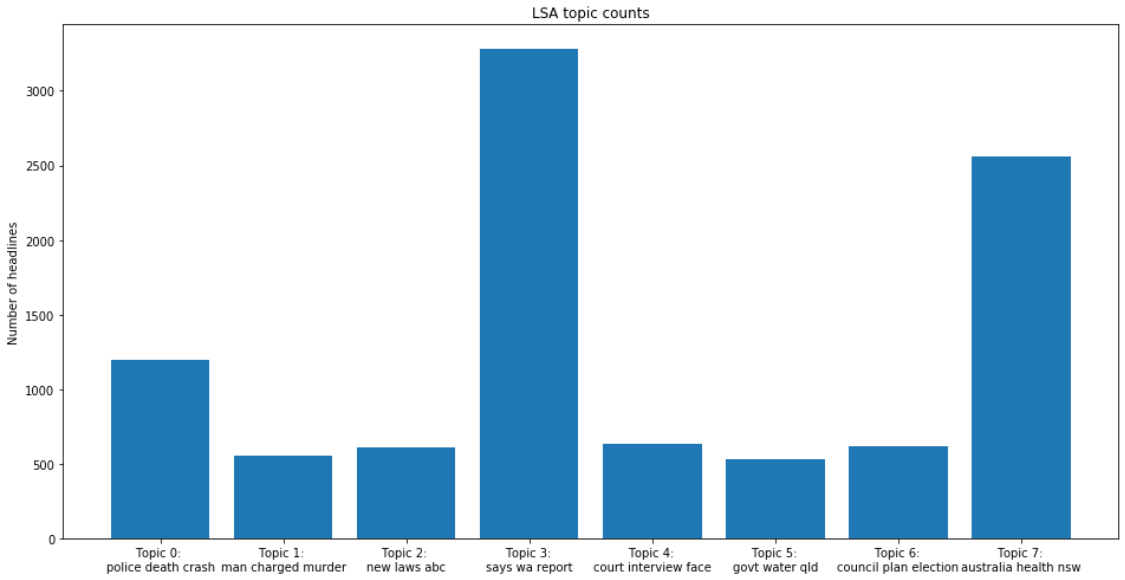

Figure. 5. LSA topic counts

By using LSA, examining the most frequently occurring words in each topic, and applying t-SNE for visualization, we can effectively explore and interpret the hidden semantic structures in the corpus, gaining valuable insights into the topics and their relationships in the text data.

Next, the last step in the LSA stage is to plot the grouped titles. In this plot, the top three words in each group are also included, which are placed at the centroid for that topic. This provides a visual representation of the topics found through LSA. However, the results obtained were not satisfactory. Apparently, we did not manage to achieve a significant degree of separation between the topic categories, and it is difficult to determine whether this is due to the LSA decomposition or the t-SNE dimension reduction process. In this context, we need to go further and try other clustering techniques. Figure 6 below is the t-SNE clustering of the 8 LSA topics.

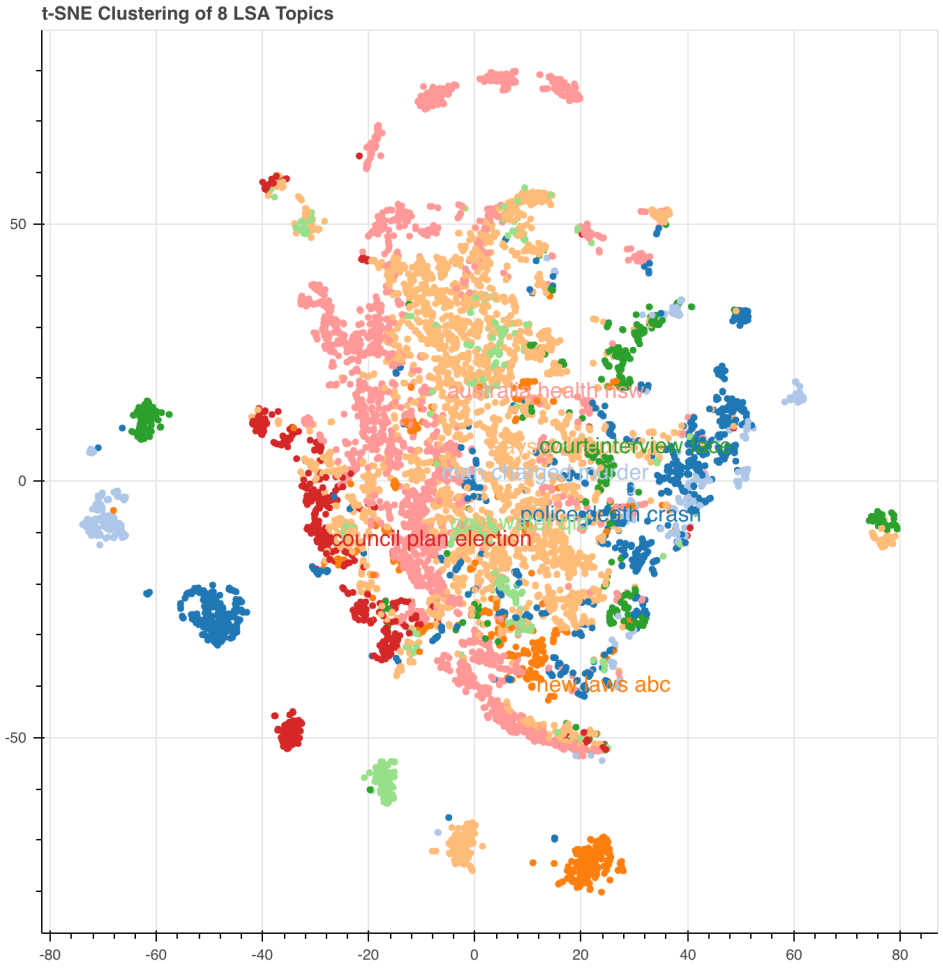

Figure. 6. t-SNE clustering of 8 LSA topics

Although the LSA results are not yet satisfactory, this can be an opportunity for further exploration. One alternative that can be tried is to use different clustering methods such as k-means clustering or hierarchical clustering. By trying other clustering techniques, it is expected that there will be an improvement in the separation between topic categories. The selection of an appropriate dimensionality reduction method is also an important consideration in this process. If t-SNE does not produce adequate separation, other alternatives such as Principal Component Analysis (PCA) or Non-Negative Matrix Factorization (NMF) can be tried to see if they produce better results in separating topic categories. By evaluating and improving the steps performed in the LSA stage, it is expected that better results will be obtained in topic modeling and separation between topic categories. This research will continue by trying alternative clustering techniques and other dimensionality reduction methods to achieve more satisfactory results.

At a later stage in this analysis, the process is repeated using the LDA (Latent Dirichlet Allocation) method as an alternative to LSA (Latent Semantic Analysis). Unlike LSA which focuses on dimensionality reduction, LDA is a generative probabilistic process specifically designed to uncover hidden topic structures in text corpora. As in the previous stage, an argmax operation is performed on each entry of the topic matrix obtained from LDA. This allows determining the predicted topic category for each headline based on the highest probability. By looking at the most frequently occurring words associated with these topic categories, an understanding of the characteristics and themes of each identified topic can be obtained.

To fairly compare LDA with LSA, the topic matrix obtained from LDA is projected into two dimensions using t-SNE (t-Distributed Stochastic Neighbor Embedding). This dimensionality reduction technique helps visualize and

compare the separation of topic categories in a two-dimensional space. Figure 7 below is the t-SNE clustering of 8 LDA topics.

Figure. 7. t-SNE clustering of 8 LDA topics

The results of the t-SNE visualization show a significant improvement when using LDA compared to LSA. The topic categories identified by LDA show better separation and cohesion. This suggests that LDA is more effective than LSA in uncovering and distinguishing topic structures hidden in text corpora.

Based on these results, it can be concluded that LDA is a more suitable algorithm to develop the clustering process in the following sections. Its ability to capture hidden topic structures and provide cohesive topic categories makes it a more promising approach for analyzing large-scale text corpora. The improved performance of LDA in separating topic categories increases its potential in applications such as information retrieval, text analysis, and topic modeling.

4.4. Discussion

This research uses SVD (Singular Value Decomposition) and LDA (Latent Dirichlet Allocation) methods in topic extraction and modeling in large-scale text corpora. The application of both methods provides significant benefits in understanding the topic structure in large and complex text corpora. The SVD method is used to reduce the dimensionality of the term-document matrix in text corpora. By identifying the largest singular value, SVD can extract the most important features in the text data, which can then be linked to the main topics. In this research, SVD helps uncover cohesive patterns in text corpora and identify interrelated topics.

On the other hand, the LDA method models text corpora as a probabilistic generative process. LDA assumes that each document consists of a probability distribution of topics and each topic consists of a probability distribution of

words. By using LDA, this research can identify topics hidden in text corpora and analyze the relationship between words and topics. Both methods have high accuracy in identifying key topics in large-scale text corpora. The SVD method reveals cohesive patterns and reveals thematically related topics, while the LDA method can identify hidden topics by estimating the probability distribution of words within each topic. This accuracy allows researchers to gain a deeper understanding of the topics present in text corpora.

Moreover, both methods are also efficient in dealing with large-scale text corpora. The SVD method reduces the dimension of the term-document matrix, thus speeding up the topic extraction process. The LDA method uses an efficient probability approach in modeling text corpora, so it can be applied to large data with affordable computation time. The topic representations generated by SVD and LDA provide valuable insights from the text corpora. They help organize the scattered information in text corpora and allow researchers to see emerging patterns. These topic representations can be used for information mining, document categorization, and deeper text analysis. Thus, this study shows that the application of SVD and LDA methods in topic extraction and modeling in large-scale text corpora has great potential to provide valuable insights in text analysis.

5. Conclusion

This research uses SVD and LDA methods in topic extraction and modeling in large-scale text corpora. Through the application of both methods, significant results were obtained in understanding the topic structure in complex text corpora. The SVD method reveals cohesive patterns and thematically related topics, while the LDA method can identify hidden topics by estimating the probability distribution of words within each topic. The accuracy and efficiency of these two methods enable researchers to identify and analyze key topics in large-scale text corpora.

This research has important benefits in text processing and analysis. The topic extraction and modeling results obtained provide valuable insights into the topics present in text corpora. The resulting topic representations can be used for information mining, document categorization, and more in-depth text analysis. This can aid in decision-making, new knowledge discovery, and further understanding of large text corpora.

This research has the novelty of applying SVD and LDA methods to large-scale text corpora. In the context of growing text corpora in number and complexity, this research makes an important contribution to topic modeling and extraction of relevant information from text corpora.

Although this study provides significant results, there are some limitations that need to be noted. First, the success of the SVD and LDA methods is highly dependent on the quality and representativeness of the text corpora used. In addition, the resulting topic interpretation still requires further understanding and analysis by the researcher. In addition, this research only compares SVD and LDA in the context of topic extraction, and there is potential for the use of other methods that can be investigated for topic modeling in text corpora.

For future research, further development can be done in topic modeling by considering contextual factors such as time and location. In addition, further research can be conducted to improve topic interpretation by involving domain knowledge or more advanced visualization methods. Research can also involve the use of more complex machine learning techniques and ensemble methods to improve the performance of topic extraction and modeling in larger and heterogeneous text corpora.

References

[1] G. Ignatow, N. Evangelopoulos, and K. Zougris, "Sentiment Analysis of Polarizing Topics in Social Media: News Site Readers' Comments on the Trayvon Martin Controversy," in *Communication and Information Technologies Annual*, in Studies in Media and Communications, vol. 11. Emerald Group Publishing Limited, 2016, pp. 259-284. doi: 10.1108/S2050-206020160000011021.

[2] F. Petroni *et al.*, "Language models as knowledge bases?" *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, vol. 1, no. 1, pp. 2463-2473, 2020, doi: 10.18653/v1/d19-1250.

[3] A. S. Hosseini, "Sentence-level emotion mining based on combination of adaptive Meta-level features and sentence

syntactic features," *Eng. Appl. Artif. Intell.*, vol. 65, no.1, pp. 361-374, 2017, doi: 10.1016/j.engappai.2017.08.006.
[4] M. Yu and C. Guo, "Using news to predict Chinese medicinal material price index movements," *Ind. Manag. Data Syst.*, vol. 118, no. 5, pp. 998-1017, Jan. 2018, doi: 10.1108/IMDS-06-2017-0287.
[5] P. Eachempati and P. R. Srivastava, "Accounting for unadjusted news sentiment for asset pricing," *Qual. Res. Financ. Mark.*, vol. 13, no. 3, pp. 383-422, Jan. 2021, doi: 10.1108/QRFM-11-2019-0130.
[6] A. Gupta and R. Katarya, "Social media based surveillance systems for healthcare using machine learning: A systematic review," *J. Biomed. Inform.*, vol. 108, no. 1, p. 103500, 2020, doi: https://doi.org/10.1016/j.jbi.2020.103500.
[7] F. Rahutomo and A. Hafidh Ayatullah, "Indonesian Dataset Expansion of Microsoft Research Video Description Corpus and Its Similarity Analysis," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, vol. 3, no. 4 SE-, pp. 319-326, Oct. 2018, doi: 10.22219/kinetik.v3i4.680.
[8] D. Singh and C. K. Mohan, "Deep Spatio-Temporal Representation for Detection of Road Accidents Using Stacked Autoencoder," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 879-887, 2019, doi: 10.1109/TITS.2018.2835308.
[9] K. Nam and F. Wang, "The performance of using an autoencoder for prediction and susceptibility assessment of landslides: A case study on landslides triggered by the 2018 Hokkaido Eastern Iburi earthquake in Japan," *Geoenvironmental Disasters*, vol. 6, no. 1, pp. 1-14, 2019, doi: 10.1186/s40677-019-0137-5.
[10] G. Czibula, A. Mihai, and L. M. Crivei, "S PRAR: A novel relational association rule mining classification model applied for academic performance prediction," *Procedia Comput. Sci.*, vol. 159, no.1, pp. 20-29, 2019, doi: https://doi.org/10.1016/j.procs.2019.09.156.
[11] S. Sinha *et al.*, "Variational Autoencoder Anomaly-Detection of Avalanche Deposits in Satellite SAR Imagery," *ACM Int. Conf. Proceeding Ser.*, vol. 1, no.1, pp. 113-119, 2020, doi: 10.1145/3429309.3429326.
[12] B. Hawashin, S. Alzubi, T. Kanan, and A. Mansour, "An efficient semantic recommender method for Arabic text," *Electron. Libr.*, vol. 37, no. 2, pp. 263-280, 2019, doi: 10.1108/EL-12-2018-0245.
[13] G. Sudeepa and P. Jagadeesh, "Foreground Detection in Dynamic Scenes using Singular Value Decomposition Algorithm in Comparison with Gaussian Mixture Model to measure F-score," *J. Pharm. Negat. Results*, vol. 13, no. 1, pp. 1772-1785, 2022, doi: 10.47750/pnr.2022.13.S04.214.
[14] M. Arif, N. Qaisar, and S. Kanwal, "Factors affecting students' knowledge sharing over social media and individual creativity: An empirical investigation in Pakistan," Int. J. Manag. Educ., vol. 20, no. 1, p. 100598, 2022, doi: https://doi.org/10.1016/j.ijme.2021.100598.
[15] I. Armawan, S. Sudarmiatin, A. Hermawan, and W. Rahayu, "The effect of social media marketing, SerQual, eWOM on purchase intention mediated by brand image and brand trust: Evidence from black sweet coffee shop," Int. J. Data Netw. Sci., vol. 7, no. 1, pp. 141–152, 2023.
[16] T. Vu, E. Chunikhina, and R. Raich, "Perturbation expansions and error bounds for the truncated singular value decomposition," *Linear Algebra Appl.*, vol. 627, no.1, pp. 94-139, 2021, doi: 10.1016/j.laa.2021.05.020.
[17] Z. Chen, L. Deng, and X. Kong, "Modified truncated singular value decomposition method for moving force identification," *Adv. Struct. Eng.*, vol. 25, no. 12, pp. 2609-2623, 2022, doi: 10.1177/13694332221104278.
[18] M. S. Alomari, "The Legal System for the Conversion of Commercial Companies in the Light of the Rules of the Saudi Corporate System," *Int. J. Appl. Inf. Manag.*, vol. 2, no. 4, pp. 106-111, Aug. 2022, doi: 10.47738/ijaim.v2i4.43.
[19] D. Vazquez, X. Wu, B. Nguyen, A. Kent, A. Gutierrez, and T. Chen, "Investigating narrative involvement, parasocial interactions, and impulse buying behaviours within a second screen social commerce context," Int. J. Inf. Manage., vol. 53, no. April, p. 102135, 2020, doi: 10.1016/j.ijinfomgt.2020.102135.
[20] A. A. P. Ratna, B. Budiardjo, and D. Hartanto, "Simple: An Automated Essay Grading System for Grading Exams in Indonesian," *MAKARA Technol. Ser.*, vol. 11, no. 1, pp. 5-11, 2010, doi: 10.7454/mst.v11i1.435.
[21] D. D. Gaikar, B. Marakarkandy, and C. Dasgupta, "Using Twitter data to predict the performance of Bollywood movies," *Ind. Manag. Data Syst.*, vol. 115, no. 9, pp. 1604-1621, Jan. 2015, doi: 10.1108/IMDS-04-2015-0145.
[22] S. Jain, K. R. Seeja, and R. Jindal, "Computing semantic relatedness using latent semantic analysis and fuzzy formal concept analysis," *Int. J. Reason. Intell. Syst.*, vol. 13, no. 2, pp. 92-100, 2021, doi: 10.1504/IJRIS.2021.114635.
[23] N. A. Prabowo, "Social Network Analysis for User Interaction Analysis on Social Media Regarding E-Commerce Business," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 4, no. 2, pp. 95-102, 2021.
[23] V. Morozov, O. Mezentseva, A. Kolomiiets, and M. Proskurin, "Predicting Customer Churn Using Machine Learning in IT Startups," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 77, no. 1, pp. 645-664, 2022. doi: 10.1007/978-3-030-82014-5_45.
[25] J. G. Borade, A. W. Kiwelekar, and L. D. Netak, "Automated Grading of PowerPoint Presentations Using Latent Semantic Analysis," *Rev. d'Intelligence Artif.*, vol. 36, no. 2, pp. 305-311, 2022, doi: 10.18280/ria.360215.
[26] L. C. Q. Lam, T. K. Toai, and S. Vaclav, "A latent semantic analysis method for ranking the results of human disease search engines," *Bull. Electr. Eng. Informatics*, vol. 12, no. 2, pp. 1189-1195, 2023, doi: 10.11591/eei.v12i2.4602.
[27] M. K. Lim, Y. Li, and X. Song, "Exploring customer satisfaction in cold chain logistics using a text mining approach," *Ind. Manag. Data Syst.*, vol. 121, no. 12, pp. 2426-2449, Jan. 2021, doi: 10.1108/IMDS-05-2021-0283.
[28] J. R. Saura, D. Ribeiro-Soriano, and P. Zegarra Saldaña, "Exploring the challenges of remote work on Twitter users'

sentiments: From digital technology development to a post-pandemic era," *J. Bus. Res.*, vol. 142, no.1, pp. 242-254, 2022, doi: https://doi.org/10.1016/j.jbusres.2021.12.052.

[29] M. A. N. Febriansyach, F. Rashif, G. I. P. Nirvana, and N. A. Rakhmawati, "Implementation of LDA for Topic Grouping of Twitter Bot Account Tweets with #covid-19," *CogITo Smart J.*, vol. 7, no. 1, pp. 170-181, 2021, doi: 10.31154/cogito.v7i1.299.170-181.

[30] N. Miao, F. Xue, and R. Hong, "Multimodal Semantics-Based Supervised Latent Dirichlet Allocation for Event Classification," *IEEE Multimed.,* vol. 28, no. 4, pp. 8-17, 2021, doi: 10.1109/MMUL.2021.3077915.

[31] Y. Luo, Z. Yang, Y. Liang, X. Zhang, and H. Xiao, "Exploring energy-saving refrigerators through online e-commerce reviews: an augmented mining model based on machine learning methods," *Kybernetes*, vol. 51, no. 9, pp.2768-2794. Jan. 2021, doi: 10.1108/K-11-2020-0788.

[32] S. Zhou, P. Kan, Q. Huang, and J. Silbernagel, "A guided latent Dirichlet allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura," *J. Inf. Sci.*, vol. 49, no. 2, pp. 465-479, 2023, doi: 10.1177/01655515211007724.