Journal of Civil Engineering, Technology and Sciences
P–ISSN: 3090-0727 E-ISSN: 3090-0727
Vol. 2, No. 02, Juli 2026

COMPARISON OF LSA AND LDA ALGORITHMS IN TOPIC MODELING

Aris Subadi[1], Endang Kusnadi[2]
[1]STKIP Mutiara Banten, [2]Dinas Pekerjaan Umum dan Penataan Ruang Provinsi Banten
Corresponding: aris.subadi@gmail.com[1], ekusnadi77@gmail.com[2]

Abstract
This study aims to analyze topic modeling in incident ticket reports by comparing two methods, namely Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). The topic modeling method is used to uncover hidden topics in unstructured text. This study involved the initial stages of data preprocessing to convert unstructured text into structured text by changing the letters to lowercase, removing distracting characters, and removing stopwords. After the preprocessing stage was completed, topic modeling was carried out using LSA and LDA with variations in the number of topics. Performance evaluation was carried out using coherence scores and t-SNE visualization for each method. The evaluation results show that LSA has a higher coherence score for several topics, especially for topics 3 and 6, with a coherence score of around 0.57. On the other hand, LDA provides a better interpretation of the distribution of words in each topic, although the coherence score is slightly lower. The LDA coherence score is around 0.53 for topic 3, 0.43 for topic 6, and 0.46 for topic 10. In addition, a comparison between LDA and LSA can also assist researchers in choosing the most suitable method for topic analysis on a particular dataset.

Keywords— Topic Modelling, Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Tiket Insiden, Data analysist.

INTRODUCTION

Topic modeling is a branch of machine learning, specifically in the field of natural language processing, namely unsupervised learning, which means that the data used does not need to have identifiers. The topic modeling task is to scan documents and recognize word and sentence patterns (Amala et al., 2022). It then automatically groups together groups of words and similar subjects that best describe those groups of words or documents. Topic modeling generates automatic topic conclusions from a collection of texts, which is quite difficult to do manually with a large corpus. The basic task of topic modeling is to group words in a document by identifying words and patterns in the document in a way that describes a particular topic (Kamila et al., 2022) (Alghamdi, 2015).

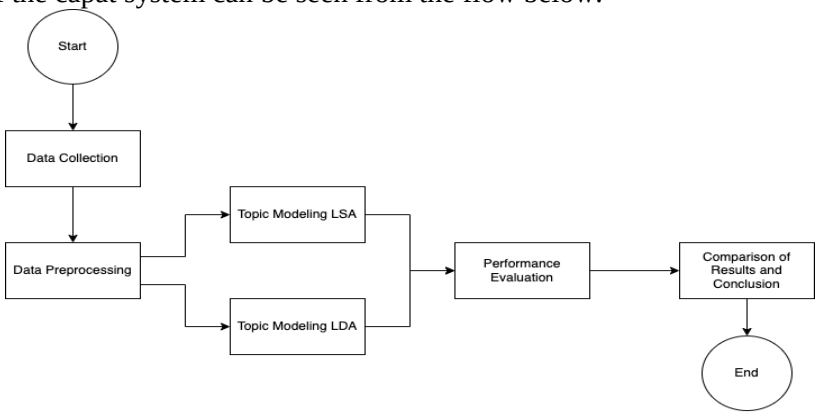
The purpose of this research is to compare two commonly used topic modeling methods, namely latent semantic analysis (LSA) and latent Dirichlet allocation (LDA), using a large abstract corpus of Incident Tickets. The purpose of this research is to identify the differences and similarities between the two methods by extracting patterns from the textual data of ticket transactions. In previous studies, LDA was the dominant method in topic modeling methodology, but the existence of the LSA method and the development of other topic modeling methods increased the need to compare their effectiveness in Incident Ticket analysis (Mohammed & Al-Augby, 2020)(Bellaouar et al., 2021). The results of this study can provide insight into the relative performance of LSA and LDA in analyzing abstract case tickets and can help analysts choose the topic modeling method that best suits their goals and analytical needs. This comparison aims to explain the advantages and disadvantages of each method and help analysts make better decisions when analyzing text data (Nufi, 2025).

METHOD

In this section, we provide detailed information about the materials and data used in this study and the methods used to collect and analyze the data. We explain in a systematic way how data is collected, processed, and interpreted so that the reader gets clarity. This step is very important in research because it determines the accuracy and validity of research results. All steps and procedures are explained transparently for other researchers to copy and review. The main topic of this section is a topic modeling comparison between Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), which are widely used methods for analyzing unstructured text. By comparing these two methods, our goal is to gain an understanding of how they work and to identify their strengths and weaknesses in answering research questions (Masan et al., 2023).

System Design

Systems design in research is critical because it helps organize and guide the way research is conducted and how information is collected, analyzed and interpreted. System planning in this research ensures that the research is carried out in a systematic and structured manner so that the results are reliable and relevant according to the research objectives. Good system design helps minimize bias and error and maximizes accuracy and interpretation of data. The design of the capat system can be seen from the flow below.



Picture 1. System design

Data Collection

The data collection process is carried out to retrieve the necessary text data from incident ticket reports to compare the performance of Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) in topic modeling by:

1. Identification of Data Sources: Researchers identify sources from incident ticket reports that will be used in this study. The data source can be an internal database, a ticketing system, or some other platform where incident ticket reports are recorded and stored.
2. Data Access and Permissions: Access to data sources was obtained, and necessary permissions and consents were obtained to use the data for research purposes. Researchers may need to comply with privacy and data protection regulations, especially if the data contains sensitive information.
3. Data Extraction: Once access is granted, the researcher extracts the incident ticket report from the selected data source. This process involves extracting relevant textual data, which may include details about the incident, descriptions, and other relevant information.

Data Pre-Processing: As mentioned in the abstract, data pre-processing is done to convert unstructured text into structured text. This involves steps such as converting text to lowercase, removing unwanted characters, and removing stopwords (common words that don't make a significant contribution to meaning). Data division: The dataset is divided into a training set and testing set. The training set is used to build the LSA and LDA models, while the test set is used to evaluate the performance of these models

Created	Short description	Description
2023-07-20 17:13:33	RITM0700364: Request to create	Hi Server Team
2023-07-20 10:33:43	Tidak bisa login ke terminal server 1,	Tidak bisa login ke terminal server
2023-07-20 09:45:20	Profile Login masuk temporary	Minta tolong bantu check di
2023-07-20 09:38:17	Tidak bisa login (T0007)	Pada saat login loading terus
2023-07-20 09:13:35	IP 10.89.8.46 Virtual Machine slow	Virtual Machine IP 10.89.8.46 is
2023-07-20 08:40:02	APP-NAPP-SVR29 Sonar Update	Hi Team,
2023-07-19 12:32:14	Please help, I cannot login to	Please help, I cannot login to
2023-07-19 09:20:37	id server nggak bisa connect	Id Server saya 722341$(\\prwfs02)
2023-07-19 01:05:48	SAP IKP link slowness.	Team Server Bantuk Check untuk
2023-07-18 17:51:37	RITM0699267: Request to create	Request to create folder & grant
2023-07-18 16:41:14	tidak bisa login ke server billing	tidak bisa login ke server billing

Picture 2. Data Preprocessing

Data Preprocessing is a stage for converting unstructured text data into structured text so as to provide better results when the dataset is processed by the model (Amala et al., 2022). There are several text pre-processing processes:

1) Lowercase: In this stage, all text that will be used in model processing is converted to lowercase. This aims to provide consistency to the words in the text, so that there is no difference in words written in capital or lowercase letters.
2) Noise Removal: Noise removal is a stage where characters that can interfere with data processing in the model are removed from the dataset. These characters can be punctuation marks, hash marks, URLs, and other characters that do not contribute to the meaning of the text.
3) Stopwords Removal: Stopwords removal is the stage where words that have high frequency in the document but do not have important meaning information for analysis are removed from the text. Examples of stopwords are "and", "or", "in", "of", etc.
4) Stemming (word reduction): Stemming is the process of changing the words that have been formed into their basic forms or basic words. In this stage, affixes (prefixes or suffixes) to words are removed so that these words have the same root word.
5) Tokenizing (Text splitting): Tokenization is the process of breaking text into the smallest units in sentences, namely tokens. Tokenization aims to identify words or smaller units in text that can be used for further analysis.

Topic Modelling

In this study, two topic modeling methods, namely LSA (Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation), are used to explore patterns and relationships between words in incident ticket reports. Topic modeling is a technique in text analysis that aims to identify hidden topics that are contained in unstructured text (Devi et al., 2022). Described as follows:

LSA : LSA (Latent Semantic Analysis) uses SVD (Singular Value Decomposition) to reduce the size of the term matrix. This reduction is done to reduce matrices that are sparse, noisy, and redundant in many dimensions. Dimensional reduction can be done using finite SVD (Truncated Singular Value Decomposition). SVD captures the underlying semantic structure of a set of document representation matrices. Singular Value Decomposition (SVD) is a method that determines the latent structure of words, sentences or text based on a combination of statistical and algebraic methods. The basic idea of SVD is to find the most valuable information and use smaller dimensions to represent valuable information (Mohammed & Al-Augby, 2020).

LDA : LDA (Latent Dirichlet Allocation) is a topic probability modeling method that represents documents as a combination of topics. Unlike LSA, LDA does not use single-value decomposition. Instead, this method uses a generative probabilistic model to estimate the probability distribution of topics in each document and the probability distribution of words in each topic. In LDA, each document is considered as the result of a latent topic combination, and each topic is characterized by a word distribution. This model iteratively decreases the distribution of topics in each document and the distribution of words in each topic until a stable solution is reached (Devi et al., 2022).

Two methods were used to evaluate model performance, namely subject cohesion scores and 2D t-SNE plots. First, the topic cohesion value method is used to get an idea of the optimal number of topics according to human interpretation. Subject cohesion provides a cohesion score for each number of subjects tested to determine the number of subjects with the highest cohesion score, which shows the most consistent interpretation (Devi et al., 2022).

Next, 2D t-SNE plots (t-distributed stochastic neighbor embedding) are used to visualize the thematic modeling results. 2D t-SNE charts help to visually observe the distribution and grouping of subjects in two-dimensional space. This makes it possible to see whether the LSA and LDA models produce significantly different themes or whether there is overlap between the themes of the two models. By using these two evaluation methods, a more comprehensive insight into the functionality and effectiveness of the LSA and LDA models in event ticket report topic modeling should be obtained (Nurlayli & Nasichuddin, 2019).

RESULT AND DISCUSSION

This experiment uses a special database containing incident ticket reports with a total of 7,148 reports. This text data has a total of 262,704 words, with an average number of words per report of 36,932939687895406 words. These incident ticket reports cover various events and incidents that occurred and were downloaded from an automated reporting system application to form the dataset used in this study. The selection of incident ticket reports was carried out to evaluate the performance of topic modeling techniques with a focus on topics with high and low correlation (Novarian et al., 2023).

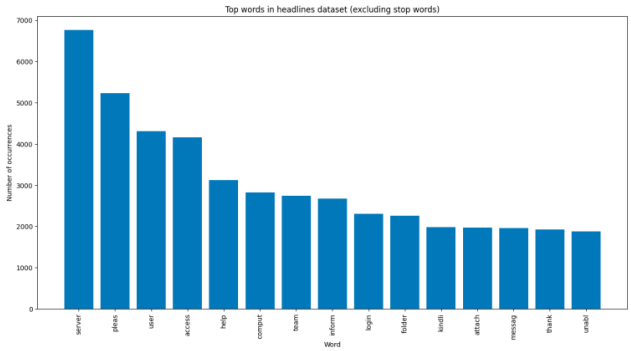
Preprocessing Data

Pre-processing operations are performed, which include clearing the text of characters and numbers, removing stop signs and special symbols, removing unnecessary words, then reconfiguring the data for unstructured data representation in "bag of words" (corpus) form. The second step is to train the data using two topic modeling methods namely LSA and LDA, with various numbers of topics (3, 6, 10, 15) to see which method gives the best performance with our database. Keywords for each topic are used as features to classify the books. The next step is to use the topic cohesion measure to evaluate the topic modeling method.

TABLE 1. NUMBER OF WORD

Headline	Number Of Word		
	After removing numbers & characters	After removing punctuation & stop words	After combining duplicates for repetitive words
1	30	30	22
2	20	20	8
3	20	20	11
4	9	9	6
5	31	31	12

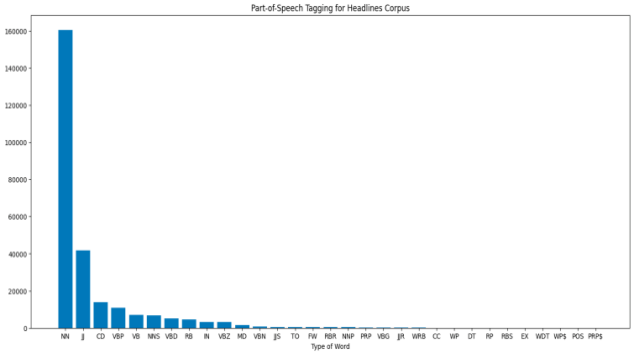
After the incident ticket report data has been processed and pre-processed, the next step is to visualize the top words in the report using the CountVectorizer. CountVectorizer is a method of Natural Language Processing (NLP) that is used to transform text into feature vectors based on the frequency of words appearing in the text. In this visualization, we will display the top 15 words and their number of occurrences in an incident ticket report. Bar graphs will be used to depict these words in order of frequency. The taller the bar graph, the more often the word appears in the report.



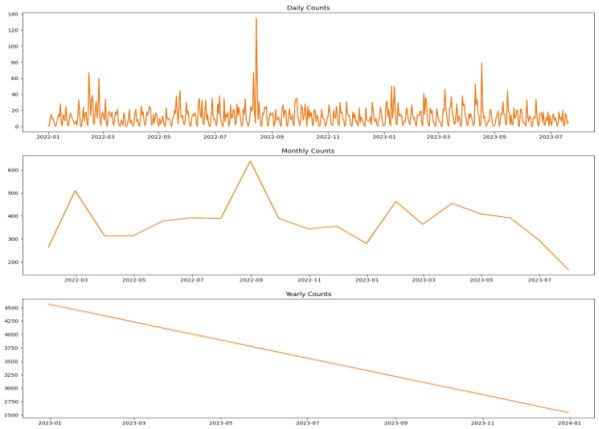
Picture 3. Display the top 15 words

The graphic above shows the top 15 words in an incident ticket report after going through the CountVectorizer process. These words are the words that appear most often in the dataset and provide important information about the dominant topics in the incident ticket report. By looking at this graph, we can identify keywords that may relate to frequently reported events and incidents.

After visualizing the top words in the incident ticket report, the next step is to do POS tagging to identify the types of words in the text. POS tagging will mark each word, such as 'NN' for a noun, 'JJ' for an adjective, and so on. This helps in analyzing sentence structure and word composition in the report. After performing POS tagging, visualization of the word length in the incident ticket report is performed using a histogram graph. This histogram graph will show you the distribution of the most common word lengths in the report. This information is important for understanding the language patterns used and the characteristics of the text in incident ticket reports. The histogram graph can be seen in picture 4.



Picture 4. Part of Speech Tagging



Picture 5. Displays the frequency of incident ticket reports

Figure 4 displays the frequency of incident ticket reports by time period: daily, monthly, and yearly. The first graph shows daily changes, the second shows monthly changes, and the third shows annual changes.

Topic Modelling LSA

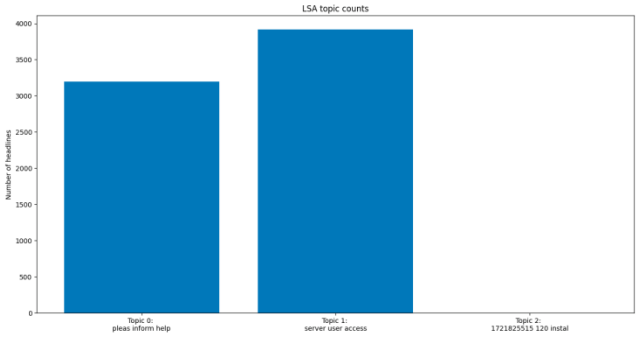
Topic modeling using the LSA method with various numbers of topics (3, 6, 10, 15) is used to identify the main topics in incident ticket reports (Williams, 2018). The selection of the number of topics is adjusted to the complexity of the data and the purpose of the analysis. The results of this modeling help classify incident ticket reports into relevant topics, providing important insights for more effective decision-making and system improvement. Below are some keywords that represent each topic in the modeling results table:

TABLE 2. TOPIC MODELLING LSA

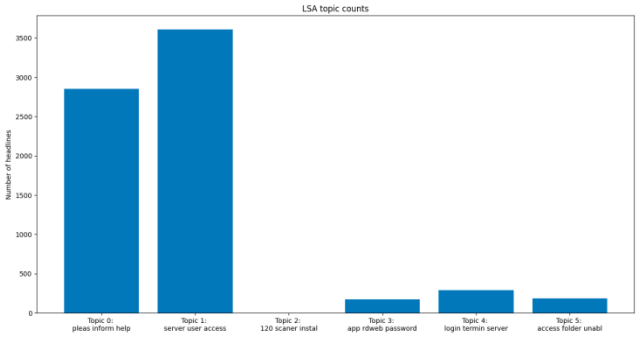
Topic (n)	Keyword
3	Please inform help message access attach computer user use email, server user access please login team terminal unable computer IP, 1721825515 120 install scanner amd17218128129 zypper cause category categori cater.
6	Please inform help message attached access computer using user email, server user access please team login computer IP folder unable, amd17218128129 120 scanner install zypper category cater caturbenni cause caution, application rdweb reset password account please forest security wfh access, login server terminal window farm help farm01 ask log 01, access folder unable share tkrdweb file new previous ID joiner.
10	please inform message help attached access use user computer email, server user access please team IP folder computer login drive, 120 scanner install zypper caution category categori cater caturbenni cause, app rdweb password reset account please forest security access user, terminal server login help ask farm01 farm enter window log, access folder unable share tkrdweb new ID joiner previous request, NA window impact adlan product server globalnetlcl 11162022 ID number, team dfm sent chin regard aspiro account service server business, login unable pc computer log farm terminal help enter number, remove team pua file please check manual access snapshot require.
15	Please inform message attach help access use email user sender, server user access please team folder IP drive request create, 120 scanner install zypper caution category categori cater caturbenni cause, app rdweb account password reset please forest security access email, server terminal help login ask farm01 enter farm application 01, access folder share unable previous drive request urgent ticket new, NA impact window adlan product server globalnetlcl ID 11162022 number, team dfm chin sent regard aspiro service account server business, PC help number please login enter unable printer password ID, remove team pua file please check manual snapshot require cleanup, password reset wfh please help IP tkrdweb PC update user, access tkrdweb user unable joiner new qlikview ID application web, log computer PC folder share help login ts printer please, login farm please terminal check user computer enter help profile, login unable computer terminal log farm server 01 user access.

In figure 5, each bar represents the number of incident ticket reports related to a particular topic identified by the LSA (Latent Semantic Analysis) method. The x-axis represents the topics, labeled as "Topic 1," "Topic 2," and so on, along with the top 3 words for each topic. The y-axis represents the number of incident ticket reports included in each topic. This

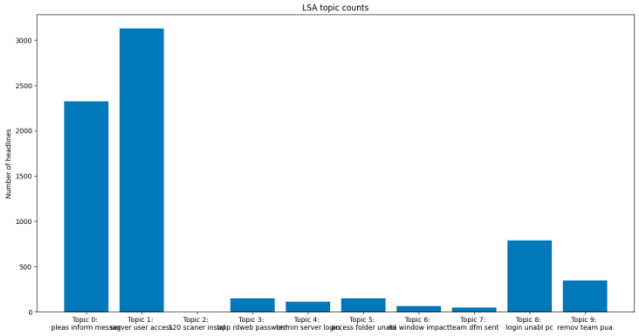
visualization gives an idea of the distribution of topics and their respective frequencies in the dataset.



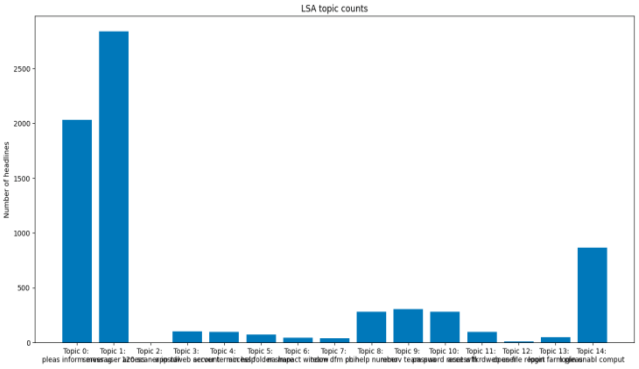
Picture 6. distribution of 3 topics



Picture 7. LSA of 6 topic



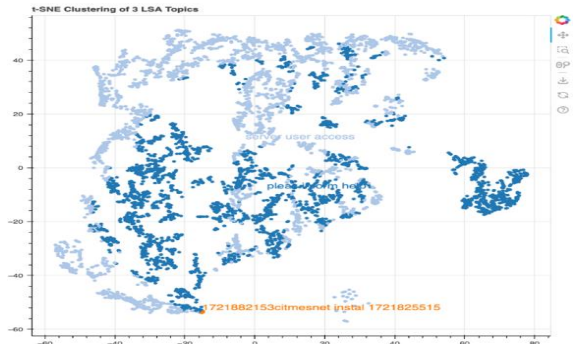
Picture 8. LSA of 10 topic



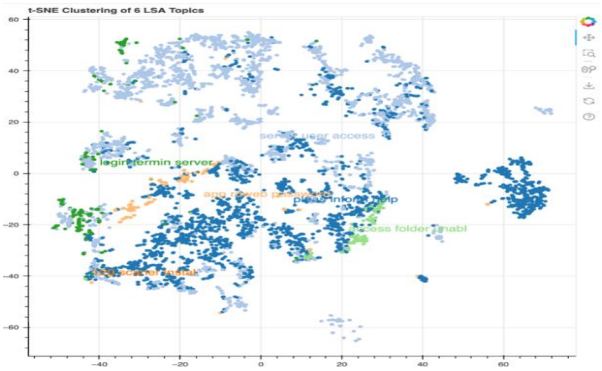
Picture 9. LSA of 15 topic

After that, t-SNE (t-distributed Stochastic Neighbor Embedding) mapping was carried out to group LSA topics into two-dimensional space so that they could be visualized more clearly. The t-SNE method is used to reduce the dimensions of the LSA topic matrix into two dimensions so that they can be represented in graphs with different colors for each topic. The results of this t-SNE mapping are displayed in a graph that shows the distribution of LSA topics in two-dimensional space. Each dot in the graph represents an incident ticket report and is

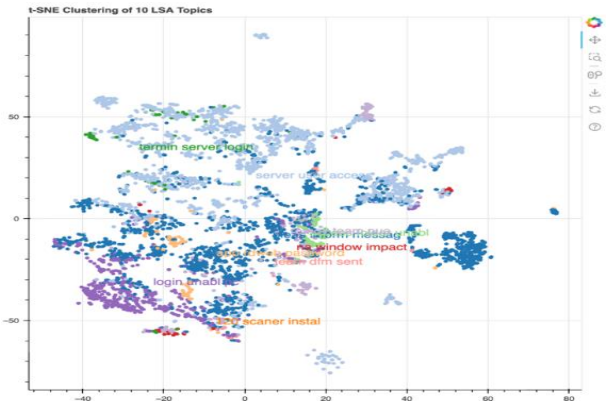
assigned a different color according to the topic. The results can be seen in Figures 8 – 11 below:



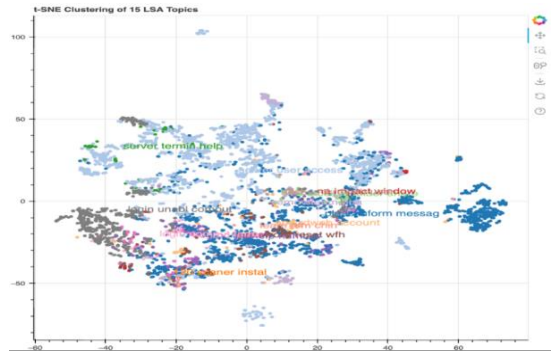
Picture 10. t-SNE Clustering of 3 LSA Topics



Picture 11. t-SNE Clustering of 6 LSA Topics



Picture 12. t-SNE Clustering of 10 LSA Topics



Picture 13. t-SNE Clustering of 15 LSA Topics

From this visualization, it can be seen that there is a relationship or group pattern between topics 3 and 6, indicating a similarity in the distribution of words. This information can assist in identifying similar problem patterns in incident ticket reports. However, the relationship between topics 10 and 15 is still seen separately in the t-SNE visualization due to the complexity of the data and variations in the distribution of words on each topic. T-SNE attempts to depict relationships between data based on distance, but sometimes the dots representing related topics can be spread out quite a distance due to differences in the representation of words.

TABLE 3. TOPIC MODELLING LSA

Topic (n)	Coherence Score
3	0.5705067299519792
6	0.5708585978243623
10	0.49092609251967184
15	0.2981242486270825

Topics 3 and 6 had a high coherence score of around 0.57, indicating strong word connections and consistent patterns. Topic 10 has a coherence score of 0.49, slightly lower but still quite good. Meanwhile, Topic 15 has the lowest coherence score, which is 0.30, indicating weaker word associations and less consistent patterns. A coherence score describes how organized and interrelated the words in a topic are, and a high score indicates a clearer and more cohesive topic.

Topic Modelling LDA

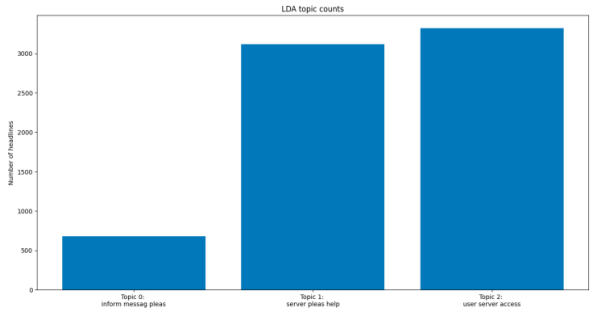
Topic modeling using the LDA (Latent Dirichlet Allocation) method with various numbers of topics (3, 6, 10, 15) is used to identify the main topics in incident ticket reports. The selection of the number of topics is adjusted to the complexity of the data and the purpose of the analysis. The results of this modeling help classify incident ticket reports into relevant topics, providing important insights for more effective decision-making and system improvement. Here are some keywords that represent each topic in the modeling results:

TABLE 4. TOPIC MODELLING LDA

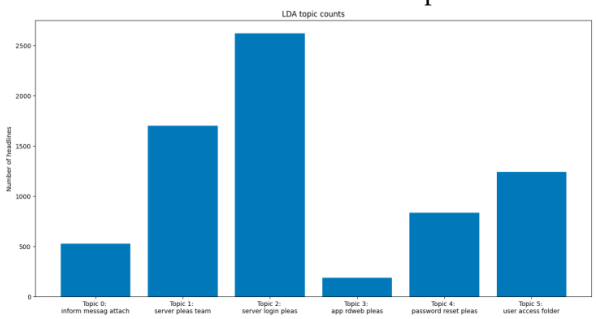
Topic (n)	Keyword
3	inform messag pleas attach sender origin email use contain commun, server pleas help team access check comput login ip user, user server access pleas folder login comput help unabl termin
6	inform messag attach sender pleas use origin email contain commun, server pleas team help check access kindli printer comput hi, server login pleas comput unabl termin ip help log user, app rdweb pleas access inform attach email account messag user, password reset pleas help wfh server remov snapshot user check, user access folder drive request creat grant server home provid
10	server pleas access inform email data open sender team accentur, server pleas team check help access file kindli hi remov, server login comput pleas unabl termin ip user help log, inform messag attach pleas sender origin use email contain commun, password reset pleas wfh help user team server thank account, user access folder creat request drive grant home server provid, server access pleas help web urgent slow unabl error user, access server tkrdweb window 11162022 globalnetlcl user pleas standard 2012, connect server pleas access ip load pi help comput login, printer pleas pc help ip comput server epson thank instal

15	pleas access inform server email sender attach origin messag open, server team pleas check kindli remov hi sopho file updat, server login pleas comput unabl termin help ip user log, app rdweb pleas access inform account user email messag attach, server connect remot pleas desktop access help cpu pc network, user access folder creat drive grant request home provid quota, access help server unabl pleas login user problem folder deni, farm ts log server user access 01 pc pleas login, password reset pleas wfh help tkrdweb server access user set, printer pleas pc help ip epson instal comput scan thank, server pleas help team printer devic thank access add idprwvpifs01, server pleas zhu thank email regard ssc inform lin hi, inform messag attach use sender origin pleas contain email commun, user account lock log got kindli busi attach host failur, access server pleas window user aks report unabl pc login

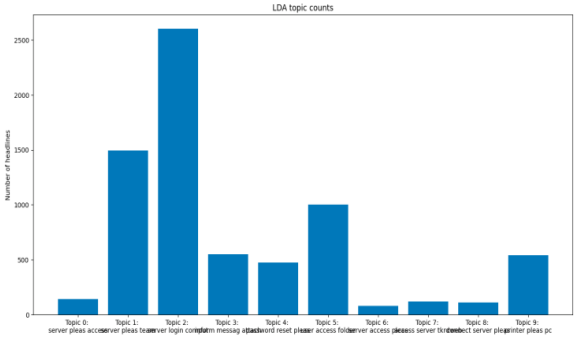
In Figure 5, each bar represents the number of incident ticket reports related to a particular topic identified by the LDA (Latent Dirichlet Allocation) method. The x-axis represents the topics, labeled as "Topic 1," "Topic 2," and so on, along with 3 keywords for each topic. The y-axis represents the number of incident ticket reports included in each topic. This visualization gives an idea of the distribution of topics and their respective frequencies in the dataset.



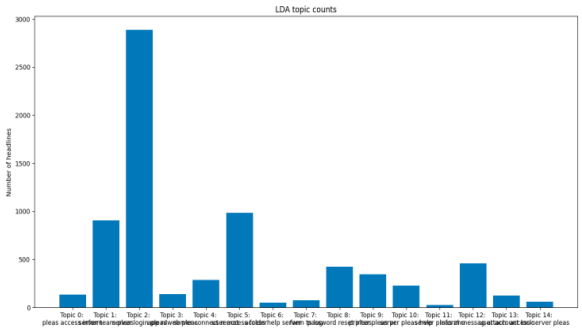
Picture 14. LDA of 3 topic



Picture 15. LDA of 6 topic

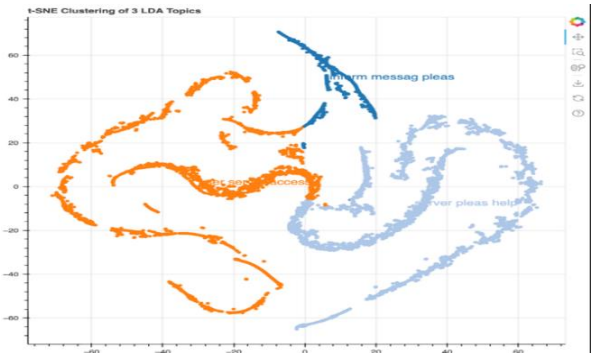


Picture 16. LDA of 10 topic

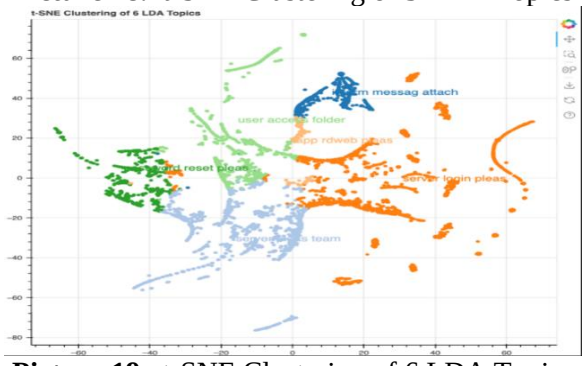


Picture 17. LDA of 15 topic

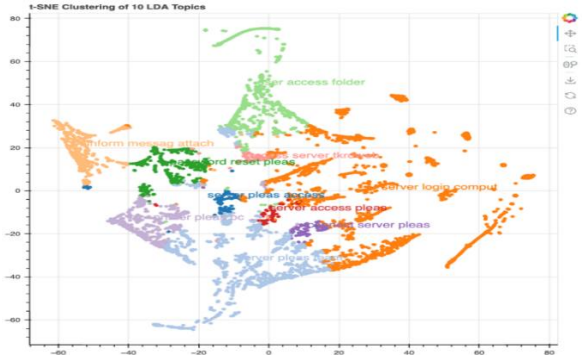
After that, t-SNE (t-distributed Stochastic Neighbor Embedding) mapping was carried out to group LDA topics into two-dimensional space so that they could be visualized more clearly. The t-SNE method is used to reduce the dimensions of the LDA topic matrix into two dimensions so that they can be represented in graphs with different colors for each topic. The results of this t-SNE mapping are displayed in a graph showing the distribution of LDA topics in two-dimensional space. Each dot in the graph represents an incident ticket report and is assigned a different color according to the topic. The results can be seen in Figure 16 - 19 below:



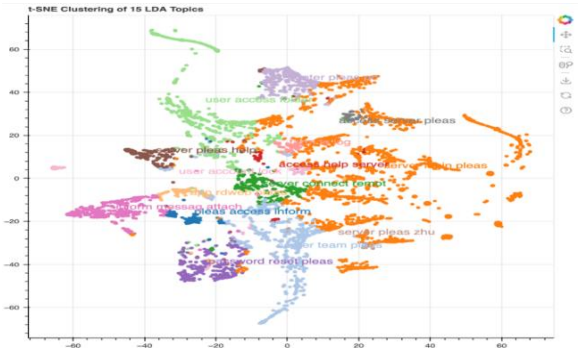
Picture 18. t-SNE Clustering of 3 LDA Topics



Picture 19. t-SNE Clustering of 6 LDA Topics



Picture 20. t-SNE Clustering of 10 LDA Topics



Picture 21. t-SNE Clustering of 15 LDA Topics

From this visualization, it can be seen that there is a relationship or group pattern between topics 3, 6 and 10, indicating similarities in the distribution of words. This could indicate that these topics share some similarities in content or characteristics that appear in incident ticket reports. This information can help in identifying similar or related problem patterns among these topics. However, the linkages between the 15 topics are still seen separately in the t-SNE visualization. This may be due to the complexity of the data and the greater variation in the distribution of words in topic 15, so that the points representing this topic are spread far enough from other topic points. Nonetheless, information about the interrelationships between topics 3, 6, and 10 can still provide valuable insights in the analysis of incident ticket reports.

TABLE 5. SUMMARY COHERENCE SCORE

Topic (n)	Coherence Score
3	0.5292519539139374
6	0.42769862356824445
10	0.45836772867840114
15	0.25997918635203104

The results of the LDA topic modeling show that there are four topics with different coherence scores. Topic 3 has a coherence score of 0.53, indicating a strong word connection and a consistent pattern. Topic 10 also has a good coherence score, which is around 0.46, although slightly lower than Topic 3. However, Topic 6 has a coherence score of 0.43, indicating less consistent word connections compared to Topics 3 and 10. Topic 15 has the lowest coherence score, only 0.26, indicating a less consistent pattern and weak word connections within the topic. The coherence score is an important indicator for measuring the quality of topics in describing patterns and relationships in incident ticket data, so topics with higher scores are perceived as more cohesive and meaningful. *Comparison LSA & LDA*

In a comparison between LSA (Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation) in topic modeling for incident ticket reports, both methods demonstrate the ability to identify key topics and describe the distribution of relevant words. From the t-SNE visualization, it can be seen that LSA succeeded in grouping topics 3 and 6 better, indicating a relationship or group pattern between the two topics. However, LDA also succeeded in grouping topics 3, 6, and 10 quite well, indicating a similarity in the distribution of words among the three topics. In terms of coherence scores, LSA had higher scores for topics 3 and 6, indicating stronger relationships among words and more consistent patterns in the topics. However, the LDA also produced a good coherence score for topic 10, even though it was slightly lower than the LSA. Meanwhile, LDA had a lower coherence score for topic 15, indicating weaker word associations and less consistent patterns within the topic. Overall, LSA and LDA have their respective advantages and disadvantages in topic modeling. LSA tends to produce clearer links between topics in the t-SNE visualization, while LDA provides a better interpretation of the distribution of words in each topic

CONCLUSION

In topic modeling for incident ticket reports, a comparison has been made between the two methods, namely Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), with reference to the results of the previous t-SNE performance evaluation and visualization. The results of the performance evaluation show that LSA has a higher coherence score for topics 3 and 6, which is around 0.57, indicating a strong word association and a consistent pattern in the two topics. Meanwhile, LDA had lower coherence scores for topics 3, 6, and 10, which were around 0.53, 0.43, and 0.46, respectively, indicating less consistent word relationships compared to LSA (Latif et al., 2021)(Fatima-azzahrae et al., 2024).

From the t-SNE visualization, it can be seen that LSA succeeded in grouping topics 3 and 6 better, indicating a relationship or group pattern between the two topics. However, LDA also succeeded in grouping topics 3, 6, and 10 quite well, showing similarities in the distribution of words among the three topics. Even so, there is a more disaggregated spread between the 15 topics in the t-SNE visualization of the LDA method. Distribution of words in each topic. Therefore, the selection of the method depends on the purpose of the analysis and the desired interpretation in the incident ticket report topic modeling. If the main goal is to more clearly identify the linkages between topics, then LSA may be a better choice. However, if a more in-depth interpretation of the distribution of words within each topic becomes more important, then LDA may be a more suitable choice (Farkhod et al., 2021).

ACKNOWLEDGMENT

We sincerely thank our friends and lecturers who have provided support, assistance, and motivation in this research. Their contributions have helped us achieve good results. Thank you for your time, support, and inspiration that has been given so that this research can be carried out properly. Without their help, this research would not have been successful.

REFERENCES

Alghamdi, R. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, *6*(1), 147–153.

Amala, I. A., Richasdy, D., & Purbolaksono, M. D. (2022). Telkom University News Topic Modeling Using Latent Semantic Analysis (LSA) Method on Online News Portal. *Building of Informatics, Technology and Science (BITS)*, *4*(1), 110–115. https://doi.org/10.47065/bits.v4i1.1584

Bellaouar, S., Bellaouar, M. M., & Ghada, I. E. (2021). Topic modeling: Comparison of LSA and LDA on scientific publications. *ACM International Conference Proceeding Series*, 59–64. https://doi.org/10.1145/3456146.3456156

Devi, S., Dhavale, C., Moharkar, L., & Khanvilkar, S. (2022). Impact of Online Education and Sentiment Analysis from Twitter Data using Topic Modeling Algorithms. *International Journal of Applied Sciences and Smart Technologies*, *4*(1), 21–34. https://doi.org/10.24071/ijasst.v4i1.4637

Farkhod, A., Abdusalomov, A., Makhmudov, F., & Cho, Y. I. (2021). applied sciences LDA-Based Topic Modeling Sentiment Analysis Using Topic / Document / Sentence (TDS) Model. *Applied Sciences*, *11*, 1–15.

Fatima-azzahrae, A., Mehdi, A., & Lily, L. (2024). ScienceDirect NLP NLP and and Topic Topic Modeling Modeling with with LDA , LDA , LSA , and NMF NMF for for Monitoring Monitoring Psychosocial in Monthly Surveys Psychosocial Well-being in Monthly Surveys. *Procedia Computer Science*, *251*(2018), 398–405. https://doi.org/10.1016/j.procs.2024.11.126

Latif, S., Member, S., Shafait, F., & Latif, R. (2021). Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling. *IEEE Access*, *9*, 127531–127547. https://doi.org/10.1109/ACCESS.2021.3112620

Mohammed, S. H., & Al-Augby, S. (2020). LSA & LDA topic modeling classification: Comparison study on E-books. *Indonesian Journal of Electrical Engineering and*

Computer Science, *19*(1), 353–362. https://doi.org/10.11591/ijeecs.v19.i1.pp353-362

Novarian, N., Khomsah, S., & Arifa, A. B. (2023). *Topic Modeling Tugas Akhir Mahasiswa Fakultas Informatika Institut Teknologi Telkom Purwokerto Menggunakan Metode Latent Dirichlet Allocation*. *8798*.

Nufi, O. S. (2025). Topic Modelling Skripsi Manajemen Dakwah PTKIN menggunakan Latent Dirichlet Allocation (LDA). *Remik: Riset Dan E-Jurnal Manajemen Informatika Komputer*, *9*(2), 539–549.

Nurlayli, A., & Nasichuddin, M. A. (2019). Topic Modeling Penelitian Dosen JPTEI UNY pada Google Scholar Menggunakan Latent Dirichlet Allocation. *ELINVO (Electronics, Informatics, and Vocational Education)*, *4*(November), 154–161. https://doi.org/10.21831/elinvo.v4i2.

Williams, T. (2018). ScienceDirect ScienceDirect ScienceDirect Comparison of of LSA LSA and and LDA LDA for for the the Analysis Analysis of of Railroad Railroad Accident Accident Text Text. *Procedia Computer Science*, *130*, 98–102. https://doi.org/10.1016/j.procs.2018.04.017