Uncovering Hidden Themes in Audit Findings Through LDA-Based Topic Modeling

Yoyok Prastyo[1], Wiyli Yustanti[2], Yuni Yamasari[3]

[1,2,3] Informatics Program Study, Faculty of Engineering, Universitas Negeri Surabaya
[1]yoyokprastyo@unesa.ac.id
[2]wiyliyustanti@unesa.ac.id
[3]yuniyamasari@unesa.ac.id

Abstract— Academic audit reports play an important role in assessing and monitoring the quality of higher education. However, most of these reports are arranged in an unstructured narrative descriptive form, making it difficult to analyze systematically and consistently, especially if done manually. This poses a challenge for auditors and decision makers in identifying patterns of findings and quality issues efficiently. This study aims to apply and evaluate the Latent Dirichlet Allocation (LDA) method in extracting keywords and abstracting main topics from academic audit report texts. The dataset was obtained from the Quality Management System (SIMUTU) of Surabaya State University, which includes hundreds of audit finding descriptions from various faculties over the past three years. The methodology used includes text preprocessing stages using tokenization, stopword removal, and stemming techniques, followed by topic modeling using LDA. Evaluation was carried out quantitatively using a coherence score to assess topic quality, and qualitatively through visualization of results in the form of word clouds and pyLDAvis. The results showed that the LDA model was able to produce meaningful, representative, and relevant topics in the context of academic quality, such as document management, lecturer involvement, and implementation of learning evaluations. Manual validation by internal quality experts showed that the generated topics can help in understanding audit findings trends more quickly and objectively. Thus, LDA has proven to be effective as an approach to extracting important information from unstructured audit reports and has great potential to be integrated into data-driven quality dashboard systems to support more informed and evidence-based decision making.

Keywords: LDA, academic audit, topic modeling, keywords, NLP

I. INTRODUCTION

Internal audit plays a crucial role in ensuring compliance and operational quality of an institution, but the complexity and unstructured nature of audit reports often complicate the analysis process, requiring significant time and effort. Research shows that the application of artificial intelligence (AI) in the audit process can improve efficiency and effectiveness by automating routine tasks and assisting auditors in data analysis [1]. In addition, AI is able to identify risks faster and more accurately than traditional methods, which can improve the reliability of audit results [2]. However, the application of AI in auditing faces challenges, such as the lack of auditor understanding of AI technology, which can hinder the optimization of its benefits [2]. Nevertheless, AI still offers great potential in improving audit quality by providing more objective and systematic data-based analysis [1]. With proper utilization, AI can support auditors in identifying anomalies and increasing transparency and accuracy in audit decision making [2]. Therefore, developing auditor skills in understanding and utilizing AI is a strategic step in driving a more modern and effective internal audit transformation [1].

In this context, Natural Language Processing (NLP) is a technique that can be used to process text in audit reports automatically. With NLP, the system can recognize patterns in the description of findings and classify text according to predetermined categories. In previous studies, NLP has been used in various fields, including unstructured data analysis with topic modeling methods such as Latent Dirichlet Allocation (LDA), which allows efficient extraction of information from large amounts of text [3]. The application of NLP in audit report analysis can help overcome challenges in processing complex textual data, such as classifying findings based on certain categories, recognizing language patterns in audit descriptions, and drawing more objective conclusions.

In addition, the development of Natural Language Processing (NLP) technology has opened up great opportunities in processing unstructured data, especially in the form of narrative text such as audit reports. Previous studies have shown that NLP has been successfully applied in a number of fields to support automated decision-making processes. For example, in the business world, NLP is used to analyze customer feedback to identify sentiment and service issues in real time. In the financial sector, NLP techniques are used to detect anomalies or irregularities in financial reports that may indicate potential fraud or non-compliance with regulatory standards. Meanwhile, in the context of internal and external audits, NLP has been utilized to extract important information, extract key entities, and identify key issues hidden in text-based audit documents [4].

This NLP-based approach allows institutions to process large volumes of data efficiently and consistently, without having to rely entirely on subjective manual interpretations by auditors. Thus, critical findings can be recognized more quickly, so that the follow-up process for non-compliance or potential risks can be carried out proactively. The implementation of NLP in the audit process also supports the realization of a more

responsive, transparent, and data-based quality management system, in line with the demands of digital transformation in higher education and other professional sectors.

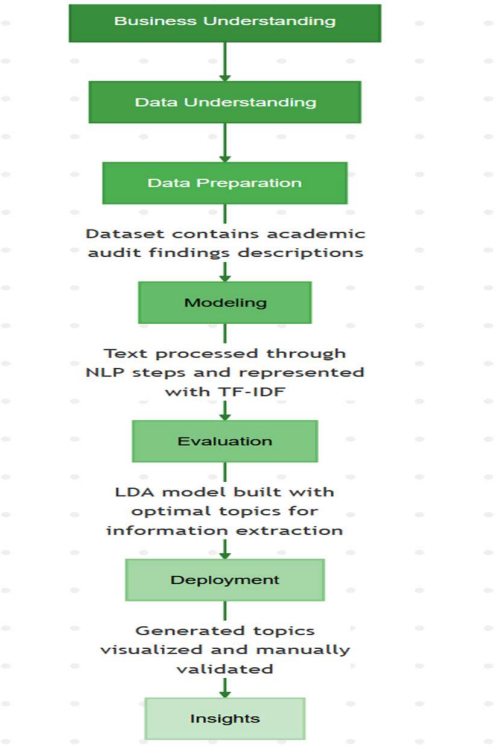

Fig. 1 CRISP-DM based research methodology

II. RESEARCH METHOD

A. Literature Study

This section will describe the literature review that is relevant to the research on Application and Evaluation of LDA for Keyword Extraction in Academic Audit Texts. This study discusses several key concepts that support the theoretical basis and methodology of the research.

1) Academic Audit and Quality Assurance

Academic audit is a systematic evaluation process of the implementation of academic policies to ensure that educational quality standards are met, thereby contributing to improving the quality of higher education institutions [5]. Quality assurance in higher education includes several main components, including identification of non-conformities (KTS) and observations (OB) in audit reports, as well as assessment of quality fulfillment indicators that reflect the level of conformity to academic standards [1].

2) Latent Dirichlet Allocation (LDA) for Topic Modeling

Latent Dirichlet Allocation (LDA) is an algorithm used to identify groups of words that frequently appear together in a document, thereby revealing hidden topics in audit reports and improving understanding of patterns in text [6]. The first step in LDA is to form a document-word matrix, which represents text in numerical form to facilitate analysis and identify relationships between words [7].

3) NLP Model Evaluation Techniques

Model evaluation in this study includes several techniques to ensure optimal performance and generalization, so that the analysis results can be relied on in various data conditions. First, Coherence Score is used to assess the quality of topics generated by the Latent Dirichlet Allocation (LDA) model, where a high coherence value indicates that the identified topics have clear and consistent meanings, which are essential in topic analysis [8].

4) Text Mining

Research in the field of unstructured text analysis continues to grow, along with the increasing need to manage and understand complex text-based information, such as social media conversations, audit reports, and academic documents.

Yustanti et al. [9] developed a probabilistic approach based on text clustering to optimize the extraction of mental health issues from social media. By utilizing probabilistic models such as LDA, this study shows that topic modeling techniques can reveal hidden public discussion patterns, as well as identify dominant topics related to mental health. The relevance of this method is very high in the context of academic audit reports which are also narrative and unstructured.

Yamasari et al. [10] showed that decision tree-based algorithms such as Random Forest and J48 can be used to build an optimal student stress detection system. This study emphasizes the importance of both supervised and unsupervised machine learning approaches in the context of predicting psychological conditions, which is parallel to efforts to classify audit findings based on thematics.

Buditjahjanto et al. [11] applied text preprocessing techniques and TF-IDF representation to an automated essay assessment system. This study also used neural networks to classify the competencies of test participants. Although the main methods are different, this study underlines the importance of strong text representation in supporting the accuracy of the classification model, which is very relevant for this study in the stage of document transformation to numeric format.

Putra and Suartana [12] studied the development of a digital-based interactive laboratory management system (SI-LMS). The main contribution of this study is how a web-based digital system can improve the efficiency of academic data management. This finding inspired the integration of the LDA topic labeling system into the academic quality dashboard.

Yohannes et al. [13] introduced business management training with a cryptocurrency-integrated web platform. Although the context is different, this study reinforces the urgency of integrating cutting-edge technologies (such as web-based systems and blockchain) in the

context of education and information management, which can be adapted for topic-based academic quality systems and NLP.

B. CRISP-DM

This study uses the CRISP-DM (Cross Industry Standard Process for Data Mining) approach, which is a cyclical framework commonly used in data analysis projects and information exploration from large data sets. This approach was chosen because it provides systematic and flexible stages in handling unstructured data such as narrative text in academic audit reports. The following are the details of the stages in CRISP-DM applied in this study:

1) Business Understanding

The Data Understanding stage is an important foundation in the data mining process because it determines the direction and strategy of further analysis. This stage aims to evaluate the structure, source, quality, and characteristics of the data content to be analyzed. In the context of this study, the data used is a collection of descriptions of academic audit findings sourced from the Quality Management System (SIMUTU) of Surabaya State University (UNESA).

2) Data Understanding

The dataset consists of more than 1,000 audit report entries collected over the past three years from various faculties and study programs. Each entry is a narrative text that describes the results of the auditor's observations related to the implementation of academic quality indicators. The recorded information includes actual conditions, compliance with standards, problem identification, and suggestions for improvement.

This data generally does not have a standard format, resulting in diversity in language style, text length, and use of terms between auditors. This characteristic makes the data fall into the category of unstructured data, which requires a Natural Language Processing (NLP)-based approach to be analyzed systematically.

To understand the characteristics of the data, a series of initial explorations were carried out, including:

a. Distribution of Text Length
   1. Counting the number of words and characters in each finding description to determine the distribution of text length.
   2. Initial findings show significant variation in length, from very short descriptions (<15 words) to very long (>100 words), which could potentially impact the effectiveness of the model in capturing context.

b. Vocabulary Diversity
   1) Analyze the number of unique words (vocabulary) in the entire text corpus, to determine the extent of the diversity of terms used.
   2) Vocabulary that is too homogeneous can hinder the formation of meaningful topics, while vocabulary that is too varied can cause noise.
   a) Identify the words that appear most frequently, such as “dosen”, “dokumen”, “evaluasi”, “pelaksanaan”.
   b) High frequencies of certain words may reflect the general focus of the audit findings, but also need to be filtered so that they do not disproportionately dominate the model.

c. Analysis of Specific Technical Terms and Acronyms
   1) Detecting the use of domain-specific terms such as “BKD”, “CPMK”, “RPS”, which frequently appear in academic reports.
   2) These terms require special handling so that they are not deleted in the stopword removal stage because they actually have important contextual meaning.

3) Data Preparation

The text obtained from the audit findings description is then processed through a series of initial processing stages based on Natural Language Processing (NLP) to ensure the data is ready to be used in the modeling process. This process includes:

a. Tokenization, which is breaking sentences into word units (tokens) to identify basic linguistic structures.
b. Stopword removal, which is eliminating common words that do not provide specific meaning in the context of analysis such as "and", "in", "yang".
c. Normalization, which includes standardizing spelling, removing punctuation, and converting capital letters to lowercase.
d. Stemming, which is changing words to their basic form, for example the word "peningkatan" becomes "tingkat". After preprocessing is complete, the text is represented using the Term Frequency-Inverse Document Frequency (TF-IDF) approach.
e. TF-IDF is used to assign weights to words based on their frequency of occurrence in a particular document relative to the entire corpus. This representation allows the model to emphasize contextually important words and lower the weight of words that are too common, thereby increasing accuracy in topic modeling and document classification.

4) Modelling

The Latent Dirichlet Allocation (LDA) model was built by determining the most optimal topic number configuration based on empirical performance and semantic evaluation. After conducting initial exploration of data characteristics and vocabulary distribution analysis results, a configuration with 10 topics was chosen because it provides a balance between information granularity and topic interpretability to determine the most optimal model in extracting information from audit documents. For each configuration, the model was trained using text data that had been processed and represented in TF-IDF form. The goal was to identify groups of words that frequently

appear together and semantically form a consistent topic. This process was carried out iteratively, and the results of each model were evaluated using the Coherence Score metric, a statistical measure that assesses the extent to which words in a topic have a strong semantic relationship. The coherence score was used because it has been proven to better represent the quality of topics in terms of human understanding than other probabilistic metrics such as perplexity. The higher the coherence value, the better the quality of the topic in terms of interpretability. The model with the highest coherence score was selected as the final LDA model used for further analysis and visualization.

5) Evaluation
   a. The evaluation stage is carried out to assess the quality of the topic modeling results and ensure that the resulting topics can be interpreted well by end users. The best topics are selected based on the highest Coherence Score value, which indicates the extent to which words in a topic have strong and consistent semantic relationships. This coherence score provides a quantitative indicator of the quality of topic interpretability compared to other metrics such as perplexity which are more mathematical in nature.
   b. In addition to numerical evaluation, qualitative evaluation is also carried out through visualization techniques. The results of the LDA model are visualized in the form of a word cloud to display the dominant keywords in each topic visually, making it easier to quickly understand the contents of the topic.

III. RESULT AND DISCUSSION

In this chapter, the results of thematic analysis of various audit reports conducted on several important aspects in the academic environment will be presented. This analysis uses a topic modeling approach with the Latent Dirichlet Allocation (LDA) method to identify the main thematic patterns in various audit documents, ranging from study program research audits, community service, recognition of past learning, to audits of the implementation of study independent program. The results of this topic modeling not only provide an overview of the main focus in each type of audit, but also reveal consistent cross-audit thematic patterns, which are an important basis for developing an auto-labeling system for audit findings and building a dynamic and integrated data-based quality dashboard.

A. Research Audit

1) Based on the Distribution Table of Study Program Research Audit Topics and the Topic Probability Distribution Figure, it can be explained that the topics resulting from topic modeling describe various important aspects in the implementation of research audits.
2) The topic with the highest probability is Topic 2 (probability: 0.3119), which includes words such as "funds", "sources", "international", "study", and "program". This shows that this topic focuses on funding and internationalization aspects of research activities in the study program. This topic is the dominant theme in the audit report.
3) Furthermore, Topic 7 (probability: 0.1999) and Topic 6 (probability: 0.1766) also have significant weights. Topic 7 discusses things like "lecturers", "students", "evaluation", and "science", indicating a focus on academic evaluation of the involvement of lecturers and students in research activities. Meanwhile, Topic 6 highlights words such as "evaluation", "results", and "evidence", which can be interpreted as topics regarding the evaluation of research results and the appropriateness of supporting evidence.
4) Topic 5 (probability: 0.1210) is also relevant, because it raises words such as "research", "group", and "laboratory", so that they are associated with supporting facilities and collaboration in research.
5) On the other hand, topics with the lowest probability such as Topics 1, 4, 8, and 9 (each around 0.0063) have a narrower and more specific scope. For example, Topic 1 is related to coordinative activities such as "meetings", "roadmaps", and "study program coordination", while Topics 4 and 8 emphasize administrative and reporting aspects such as "documents", "uploads", and "assignments".
6) Meanwhile, Topic 10 (probability: 0.0590) mentions words such as "assignments", "percentages", and "titles", which can be associated with student data collection and reporting activities.
7) Overall, the main topics (Topics 2, 6, 7, and 5) illustrate that the implementation of research audits at the study program level focuses on financing, evaluation of research activities, involvement of human resources, and research support facilities. The tables and graphs provided support these findings by showing a clear probability distribution of the most dominant topics.

TABLE. 1
RESEARCH AUDIT TOPIC DISTRIBUTION

topic number	probability	words
1	0.006264186	['rapat', 'universitas', 'auditee', 'layar', 'sk', 'capai', 'roadmap', 'koorprodi', 'relevansi', 'in']
2	0.311886872	['dana', 'sumber', 'minimal', 'internasional', 'studi', 'program', 'nasional', 'milik', 'prodi', 'dosen']

topic number	probability	words
3	0.106596471	['capai', 'ajar', 'dukung', 'prasarana', 'sarana', 'cakup', 'peta', 'jalan', 'agenda', 'relevan']
4	0.006264186	['dokumen', 'peta', 'sesuai', 'mahasiswa', 'prodi', 'evaluasi', 'unggah', 'jalan', 'website', 'sosialisasi']
5	0.120967191	['riset', 'kelompok', 'laboratorium', 'fungsional', 'dukung', 'program', 'studi', 'mohon', 'tabel', 'prodi']
6	0.176559982	['evaluasi', 'sesuai', 'jalan', 'peta', 'hasil', 'relevansi', 'baik', 'bukti', 'ps', 'guna']
7	0.199955396	['dosen', 'mahasiswa', 'jalan', 'peta', 'ps', 'sesuai', 'evaluasi', 'ilmu', 'agenda', 'kembang']
8	0.006264186	['evaluasi', 'sesuai', 'dosen', 'jalan', 'ps', 'mahasiswa', 'peta', 'hasil', 'koorprodi', 'tugas']
9	0.006264187	['dana', 'sumber', 'studi', 'program', 'nasional', 'minimal', 'internasional', 'milik', 'laboratorium', 'riset']
10	0.058977342	['mahasiswa', 'tugas', 'line', 'in', 'agenda', 'minimal', 'persentase', 'dosen', 'judul', 'data']

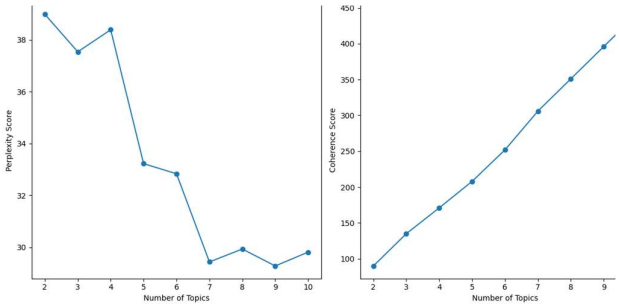

Fig. 2 Perplexity and Coherence values of the LDA model Study Program Research Audit

The distribution of documents showed that topic 2 (research funding) dominated (33.1%), followed by topic 7 (20.4%) and topic 6 (17.6%). This shows that aspects of funding, evaluation and availability of roadmaps are the main issues in research audits.

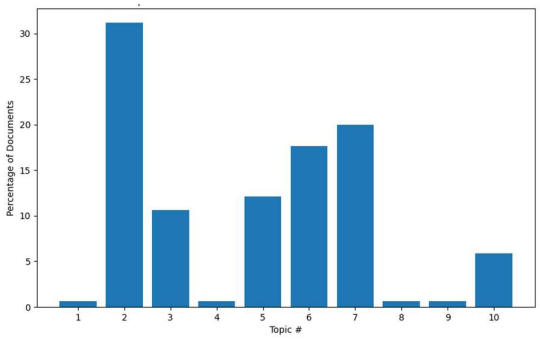

Fig. 3 distribution of topics in a document Research Audit

B. Community Service Audit

1) Based on the Community Service Audit Topic Distribution Table and the Topic Probability Distribution Figure, it can be explained that the results of topic modeling show a diversity of themes related to the implementation and evaluation of community service activities.
2) The topics with the highest probability are Topic 7 (probability: 0.2102) and Topic 10 (probability: 0.2023). Topic 7 highlights keywords such as "implement", "facilities", "institutions", and "documents", indicating a focus on providing facilities and infrastructure and institutional support in the implementation of community service. Topic 10, on the other hand, includes words such as "activity", "teaching", "implementation", "lectures", and "facilities", indicating that the implementation of community service activities is integrated with the learning process and academic facilities.
3) Topic 3 (probability: 0.1321) and Topic 6 (probability: 0.0988) also stand out. Topic 3 contains words such as "prodi", "giat", "jalan", "hasil", and "standar", describing the study program's efforts in planning, implementing, and evaluating the results of community service in a structured manner according to standards. Meanwhile, Topic 6 contains words such as "implementation", "process", "industry", and "ilmu", which show collaboration between community service and the industrial sector as well as the application of science.
4) Topic 5 (probability: 0.0968) is also relevant because it discusses social aspects such as "society", "region", "poor", "entas", and "solution", so it can be concluded that this topic emphasizes the program's contribution to community empowerment in marginal or underdeveloped areas.
5) Topics with medium probability such as Topic 2 (0.0741) and Topic 8 (0.0807) show a focus on community comfort and the use of technology and science in community service activities.
6) In contrast, Topics 1, 4, and 9 have the lowest probabilities (each below 0.05). Topic 1 relates to student involvement in activity reporting ("folder",

"report", "data"), while Topic 4 refers to "schedule", "plan", and "hook", indicating the administrative focus and activity planning. Topic 9 contains words such as "cv", "qualification", and "upload", which are more directed at the administrative requirements and data collection of activity implementers. 7. Thus, from the distribution graph and table, it can be concluded that the main topics of the community service audit are centered on implementation integrated with academics, support for facilities and institutions, and social impacts on the community, which is reflected in the high probability in Topics 7, 10, and 3.

TABLE. 2

COMMUNITY SERVICE AUDIT TOPIC DISTRIBUTION

topic number	probability	words
1	0.046779248	['libat', 'mahasiswa', 'folder', 'data', 'laksana', 'proposal', 'file', 'lapor', 'kait', 'capai']
2	0.074063267	['unsur', 'sarana', 'prasarana', 'nyaman', 'masyarakat', 'abdi', 'penuh', 'aman', 'selamat', 'sehat']
3	0.132065007	['prodi', 'giat', 'jalan', 'peta', 'hasil', 'capai', 'standar', 'analis', 'dokumen', 'libat']
4	0.036455653	['laksana', 'sesuai', 'sasar', 'jadwal', 'aspek', 'rencana', 'folder', 'lapor', 'giat', 'kait']
5	0.096752913	['masyarakat', 'aspek', 'wilayah', 'kembang', 'miskin', 'entas', 'laksana', 'daya', 'solusi', 'alih']
6	0.098809386	['hasil', 'giat', 'sesuai', 'implementasi', 'proses', 'industri', 'bidang', 'ilmu', 'dokumen', 'kembang']
7	0.210187122	['laksana', 'sarana', 'lembaga', 'fasilitas', 'harga', 'prestasi', 'dokumen', 'prodi', 'parasarana', 'dukung']
8	0.080653555	['laksana', 'ilmu', 'teknologi', 'giat', 'kembang', 'pendayagunaan', 'tahu', 'rangka', 'manfaat', 'tim']
9	0.021925403	['cv', 'laksana', 'muat', 'hasil', 'kualifikasi', 'akademik', 'ilmu', 'lapor', 'unggah', 'dokumen']
10	0.202308446	['giat', 'ajar', 'hasil', 'implementasi', 'kait', 'fasilitas', 'terap', 'ijin', 'aspek', 'kuliah']

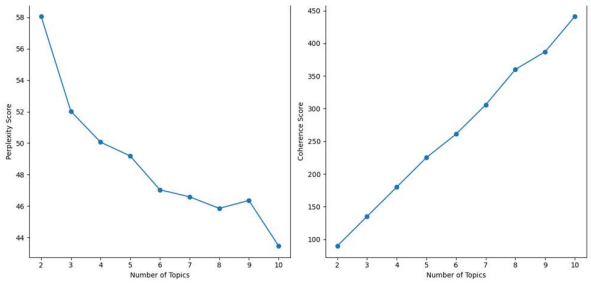

Fig. 4 Perplexity and Coherence values of the LDA model Community Service Audit

The distribution of documents shows an even distribution, but topic 7 (21.9%) and topic 10 (21.1%) dominate. This shows that the implementation of activities and the involvement of institutions in supporting community service are the main concerns.

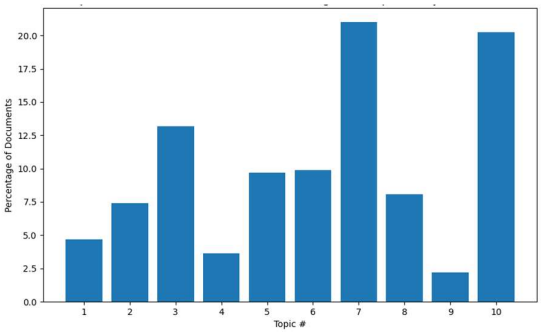

Fig. 5 distribution of topics in a document Community Service Audit

C. Prior Learning Recognition Audit

1) Based on the MBKM Audit Topic Distribution Table and Topic Probability Distribution Figure, the results of topic modeling show that most of the report content or narratives are very focused on Topic 4, which has a dominant probability of 0.7828.
2) Topic 4 has keywords such as "guidelines", "satisfied", "survey", "students", "sahahkan", and "bkp", which explicitly illustrate that the document is very focused on measuring the level of satisfaction of MBKM students based on official guidelines approved by the institution or partner. This shows that evaluation and accountability for the implementation of MBKM are the main concerns.
3) Other topics that have quite significant contributions are Topic 2 (probability: 0.0860) and Topic 6 (probability: 0.0855).
4) Topic 2 contains words such as "students", "survey", "validity", "instrument", and "kembang", indicating a focus on validating survey instruments and developing services for students.
5) Topic 6 with keywords such as "work", "program", "activity", "sub", and "plan", reflects the activities of the MBKM work program under a certain section or directorate and the achievement of the study program's strategic plan.

6) Topic 1 emphasizes "teaching", "assistance", and "education guidelines", which may be related to the learning and guidance process during MBKM activities.
7) Topic 3 contains words such as "lecture", "student", and "independent", most likely reflecting independent learning activities and credit recognition.
8) Topic 5 leads to "entrepreneurs", "projects", and "wise", indicating the existence of entrepreneurship programs and work projects based on central policies.
9) Thus, from the distribution table and graph, it can be concluded that the MBKM audit narrative is dominated by the evaluation of student satisfaction with the implementation of the program (Topic 4), supported by the development of survey instruments (Topic 2), and institutional work program activities (Topic 6). Other topics serve as a complement to the context of the learning process, student independence, and entrepreneurship policies that are part of the MBKM framework.

TABLE. 3

PRIOR LEARNING RECOGNITION AUDIT TOPIC DISTRIBUTION

topic number	probability	words
1	0.027625819	['ajar', 'satu', 'asistensi', 'didik', 'mana', 'pedoman', 'puas', 'survey', 'sahkan', 'bkp']
2	0.085988246	['mahasiswa', 'survey', 'validitas', 'layan', 'kembang', 'hasil', 'subdirektorat', 'puasa', 'lanjut', 'instrumen']
3	0.009060455	['hasil', 'sop', 'kuliah', 'mahasiswa', 'lanjut', 'rektor', 'inggris', 'studi', 'aktivitas', 'independen']
4	0.782805589	['pedoman', 'puas', 'survey', 'mahasiiswa', 'sahkan', 'mitra', 'bkp', 'rektor', 'atur', 'inggris']
5	0.009060342	['konsisten', 'hasil', 'tetap', 'pusat', 'bijak', 'wirausaha', 'proyek', 'kerja', 'independen', 'subdirektorat']
6	0.085459549	['kerja', 'program', 'giat', 'seksi', 'direktorat', 'capai', 'sub', 'mbkm', 'rencana', 'prodi']

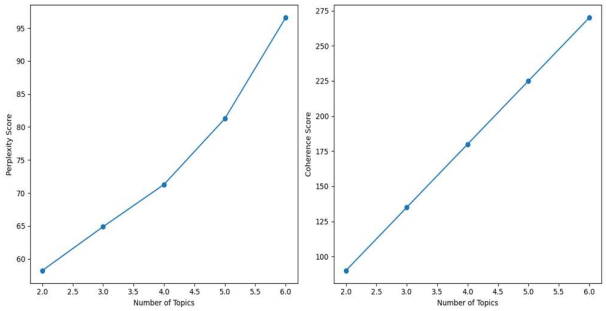

Fig. 6 Perplexity and Coherence values of the LDA model Prior Learning Recognition Audit

There are 12 documents in this type of audit. Words such as “survey”, “bkp”, and “student” are dominant. From testing 2–6 topics, the optimal configuration is 6 topics, where:

1. Topic 4 (83.3%) emphasizes the MBKM implementation guidelines and student satisfaction,
2. Other topics are more spread out on aspects of program implementation and partner collaboration.
3. The high dominance of one topic indicates that most reports discuss the consistency of guidelines and the involvement of external partners in the MBKM program.

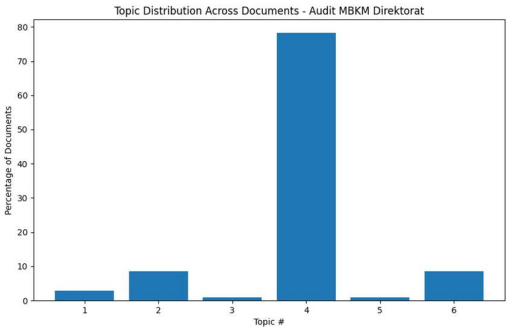

Fig. 7 distribution of topics in a document Prior Learning Recognition Audit

IV. CONCLUSION AND FUTURE WORKS

This study shows that the Latent Dirichlet Allocation (LDA) method is effective in extracting key topics from unstructured academic audit reports. With a CRISP-DM-based approach, the study successfully processed and analyzed more than 1,000 descriptions of audit findings from various types of internal audits, such as research audits, community service, MBKM, and recognition of prior learning (RPL).

The application of LDA supported by text preprocessing stages and TF-IDF representation allows the identification of dominant themes in each type of audit, such as research funding, student engagement, learning achievement assessment, and implementation of MBKM policies. Model evaluation using coherence scores shows that the optimal number of topics

varies, and the quality of the resulting topics has been validated through visualization and expert interpretation.

These findings confirm that LDA not only helps understand recurring issues in the implementation of academic quality audits but also provides a basis for the development of automated systems for topic labeling, monitoring finding trends, and supporting data-based decision making. Further research can expand this approach by integrating deep learning-based topic models and combining them with classification systems for predictive audits.

REFERENSI

[1] R. Silaen and T. Dewayanto, "Penggunaan berbagai artificial intelligence pada proses audit: A systematic literature review," Diponegoro J. Account., vol. 13, no. 2, pp. 112–125, 2024. [Online]. Available: https://ejournal3.undip.ac.id/index.php/accounting/article/view/43916

[2] H. H. Rumahorbo and T. Dewayanto, "Pengaruh Transformasi Digital: Kecerdasan Buatan Dan Internet Of Things Terhadap Peran Dan Praktik Audit Internal: Systematic Literature Review," Diponegoro J. Account., vol. 12, no. 4, pp. 1–15, 2023. [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/accounting

[3] D. Yu and B. Xiang, "Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling," Expert Syst. Appl., vol. 225, 2023. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.120114

[4] L. Zheng, Z. He, and S. He, "A topic model-based knowledge graph to detect product defects from social media data," Expert Syst. Appl., vol. 268, 2025. [Online]. Available: https://doi.org/10.1016/j.eswa.2024.126313

[5] V. S. Patil and V. B. Shinde, "The impact of an effective academic audit on accreditation performance," Qual. Assur. Educ., vol. 33, no. 1, pp. 14–27, 2025. [Online]. Available: https://doi.org/10.1108/QAE-12-2024-0275

[6] M. L. C. Chilmi, "Latent Dirichlet Allocation (LDA) untuk Mengetahui Topik Pembicaraan Publik tentang Omnibus Law," Jurnal Informatika, vol. 15, no. 1, pp. 45–58, 2021. [Online]. Available: https://repository.uinjkt.ac.id/dspace/bitstream/123456789/56724/1/M.%20LUVIAN%20CHISNI%20CHILMI-FST.pdf

[7] Y. S. Wardhana and A. Kesumawati, "Analisis Topik Skripsi Menerapkan Pemodelan Latent Dirichlet Allocation," J. Teknol. Inform. Sistem Informasi, vol. 4, no. 2, pp. 98–110, 2023. [Online]. Available: https://ojs.stmik-banjarbaru.ac.id/index.php/jutisi/article/download/2271/1196

[8] R. Gautam and M. Sharma, "Improving SVM performance for type II diabetes prediction with an integrated kernel function," Materials Today: Proceedings, vol. 66, pp. 1727–1731, 2023.

[9] W. Yustanti, A. W. Utami, G. S. Palupi and P. S. Nautika, "Probabilistic-based Text Clustering for Optimizing Mental Health Issues Extraction on Social Media," 2024 Seventh International Conference on Vocational Education and Electrical Engineering (ICVEE), Malang, Indonesia, 2024, pp. 70-75, doi: 10.1109/ICVEE63912.2024.10823692.

[10] Yamasari, Y., Qoiriah, A., Rochmawati, N., Prapanca, A., Prihanto, A., Suartana, I. M., & Ahmad, T. (2024). Exploring the tree algorithms to generate the optimal detection system of students' stress levels. Indonesian Journal of Electrical Engineering and Computer Science, 36(1), 548–558. https://doi.org/10.11591/ijeecs.v36.i1.pp548-558

[11] Buditjahjanto, I. G. P. A., Idhom, M., Munoto, M., & Samani, M. (2022). An Automated Essay Scoring Based on Neural Networks to Predict and Classify Competence of Examinees in Community Academy. TEM Journal, 11(4), 1694–1701. https://doi.org/10.18421/TEM114-34

[12] R. E. Putra and I. Made Suartana, "Development of Smart and Interactive Laboratory Management System (SI-LMS)," 2021 Fourth International Conference on Vocational Education and Electrical Engineering (ICVEE), Surabaya, Indonesia, 2021, pp. 1-5, doi: 10.1109/ICVEE54186.2021.9649702.

[13] E. Yohannes et al., "Educational Training on Business Management Using Web-Based Applications with Cryptocurrency Integration," 2024 Seventh International Conference on Vocational Education and Electrical Engineering (ICVEE), Malang, Indonesia, 2024, pp. 157-162, doi: 10.1109/ICVEE63912.2024.10823784.