EasyChair Preprint

№ 15272

Integrating BERTopic and Large Language Models for Thematic Identification of Indonesian Legal Documents

Moses Ananta, Rahayu Utari, Amany Akhyar and
Gusti Ayu Putri Saptawati

October 21, 2024

Integrating BERTopic and Large Language Models for Thematic Identification of Indonesian Legal Documents

Moses Ananta
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
23523016@std.stei.itb.ac.id

Rahayu Utari
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
23523009@std.stei.itb.ac.id

Amany Akhyar
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
33222305@std.stei.itb.ac.id

Gusti Ayu Putri Saptawati
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
putri@staff.stei.itb.ac.id

***Abstract*—The increasing complexity and volume of legal documents pose significant challenges for information retrieval and text analysis. Traditional text analysis methods are often inadequate, resulting in time-consuming and labor-intensive processes. This study applies advanced natural language processing (NLP) techniques, specifically BERTopic and large language models (LLMs), to cluster and identify themes within Indonesian legal paragraphs. The methodology includes data collection, preprocessing, BERTopic topic modeling, and LLM-based topic refinement. Results show that the "intfloat/multilingual-e5-large-instruct" embedding model, with a minimum cluster size of 40, achieves optimal performance with a Silhouette Score of 0.723 and a Davies-Bouldin Index of 0.340. Subsequent LLM refinement using Meta's LLaMA-3-8B-Instruct language model enhances the readability and relevance of the extracted topics. The approach enhances the organization and analysis of complex legal documents, with practical implications for improving legal information retrieval and management.**

Keywords— Natural Language Processing, BERTopic, Large Language Models, Topic Modeling, Indonesian Legal Documents

I. Introduction

The rapid expansion of legal documents demand advanced techniques for effective text analysis and information retrieval. The sheer volume and intricacy of these documents create significant challenges for legal professionals, making it difficult to efficiently manage, interpret, and extract relevant information. This issue is particularly critical in contexts where the legal system must handle large amounts of documents in a timely manner. Traditional methods of organizing and understanding legal documents are often inadequate, resulting in time-consuming and labor-intensive processes.

BERTopic [1], an innovative topic modeling method, utilizes pre-trained transformer language models along with a class-based TF-IDF process to generate coherent topic representations. This method has demonstrated superior performance across various domains by producing more meaningful and diverse topics, compared to classical models like Latent Dirichlet Allocation (LDA). Large language models such as ChatGPT, LLaMA, and Mistral further refine these topics by providing detailed and contextually relevant interpretations, enhancing the clarity and coherence of the identified themes.

This paper explores the application of advanced natural language processing (NLP) techniques—specifically BERTopic and large language models (LLMs)—for clustering and thematic identification in Indonesian legal documents. Our approach aims to improve the thematic structuring and comprehension of legal documents, thereby enhancing the efficiency of legal document management. The integration of these advanced NLP techniques presents a critical opportunity for the legal field to optimize document analysis and information retrieval, particularly in Indonesia, where the complexity of legal documents continues to rise. In this study, we apply BERTopic for initial topic modeling and clustering of Indonesian legal paragraphs and utilize LLMs for refining these topics. By leveraging these techniques, we aim to improve the organization, retrieval, and analysis of legal documents, contributing to more effective and timely legal decision-making.

II. Related Works

BERTopic, introduced by [1], extends traditional topic modeling by leveraging pre-trained transformer-based language models and a class-based TF-IDF procedure to generate coherent topic representations. This approach has demonstrated superior performance in various domains, including short text and multi-domain applications, by producing more meaningful and diverse topics compared to classical models like Latent Dirichlet Allocation (LDA) [2].

Recent advances in topic modeling and text classification have highlighted the versatility of BERTopic and transformer models in analyzing and understanding complex datasets from social media and beyond. [3] examined the performance of several topic modeling techniques including LDA, NMF, Top2Vec, and BERTopic on Twitter posts. Their study found that BERTopic and NMF were particularly effective in handling the short, unstructured nature of social media content, offering new insights into the dynamics of social interactions online. [4] focused on classifying Indonesian hoax news using a combination of a multilingual transformer model and BERTopic, demonstrating the potential of these models in improving classification accuracy in a low-resource language context. [5] explored the identification and evolution of interdisciplinary topics using BERTopic, providing a methodological framework for tracking topic

evolution across various disciplines. [6] analyzed the impact of ChatGPT on marketing strategies through the lens of early Twitter posts by using BERTopic to decode public sentiment and thematic clusters related to generative AI, illustrating its practical implications for marketing strategies and business practices. [7] proposed an approach to enhance the classification of Arabic cognitive distortions on Twitter using BERTopic, demonstrating the effectiveness of BERTopic in enriching text representation and improving classifier performance.

While BERTopic provides an initial clustering and identification of themes within legal documents, large language models (LLMs) such as ChatGPT, LLaMA, and Mistral [8]–[10], play a crucial role in refining these topics. LLMs leverage extensive pre-training on diverse datasets to generate detailed and contextually relevant refinements of the topics identified by BERTopic. For example, ChatGPT can be used to further interpret and elaborate on the themes generated, providing deeper insights and more nuanced understandings of the legal paragraphs. LLaMA and Mistral, known for their robust performance in language generation tasks, can enhance the clarity and coherence of the topics, ensuring they are more precise and contextually accurate.

The use of domain-specific embeddings further enhances the effectiveness of topic modeling in specialized fields. [11] illustrated this by employing LEGAL-BERT embeddings with BERTopic to analyze legal documents, resulting in improved topic coherence and relevance. This method's ability to handle the unique characteristics of legal documents, such as complex language structures and technical vocabulary, makes it particularly suitable for legal document analysis.

The application of these advanced NLP techniques is not limited to Western legal systems. [12] utilized BERT-based models for topic modeling of legal documents under the Hindu Marriage Act, demonstrating the versatility and effectiveness of these models in diverse legal contexts. The success of such models underscores their potential applicability to the Indonesian legal system, which features its own unique linguistic and structural challenges.

III. Method

The method section outlines the process undertaken to apply BERTopic and large language models (LLMs) to the clustering and theme identification of Indonesian legal paragraphs. This includes data collection, preprocessing, BERTopic topic modeling, and evaluation techniques.

A. Data Collection

This study employs a legal knowledge graph sourced from the research by [13]. The "Lex2KG" model, detailed in their study, automatically converts legal documents into a structured knowledge graph, providing a well-organized dataset for our analysis. Lex2KG is a framework specifically designed for converting legal PDF documents into a knowledge graph (KG), which contains structured data such as metadata, document structures, textual content, and relationships between legal resources like amendments and citations. This structured format transforms otherwise unstructured legal documents into machine-readable data, facilitating various downstream applications like legal information retrieval and analysis.

B. Data Pre-Processing

Effective data preprocessing is crucial for ensuring the quality and consistency of the dataset used in topic modeling. The preprocessing stage includes several filtering techniques to prepare the legal paragraphs for analysis:

1) Removal of NaN Values: Any NaN (Not a Number) samples resulting from knowledge graph extraction were removed to ensure data integrity.

2) Extraction of Paragraphs: From the legal knowledge graph, we extracted the paragraph parts of each document, which typically contain the core legal arguments and references.

3) Stopword Removal: Stopwords, which do not carry significant meaning, were removed to eliminate unnecessary noise. The list of Indonesian stopwords used for filtering was acquired from the NLTK library [14].

4) Filtering Non-Alphabetic Characters: Non-alphabetic characters and alphabetic representations of integers were filtered out to focus on substantive text.

5) Word Length Filtering: Terms that were too short (less than 5 words) or too long (more than 99 words) to hold meaningful context were excluded. This step helps in removing irrelevant and noisy samples.

6) MinHash LSH Deduplication: Near-duplicate sentences within the dataset were identified and removed using MinHash and LSH Algorithm [15], [16]. This technique ensures the uniqueness and quality of the text corpus. The MinHashLSH algorithm was employed using the datasketch library [17] with a threshold parameter of 0.95 to detect near-duplicates. The threshold parameter determines the similarity level at which two sentences are considered duplicates.

C. BERTopic Topic Modelling

The following steps outline the topic modeling process of BERTopic and LLM to identify clusters and themes.

1) Create Paragraph Embeddings: Sentence embeddings were generated using pre-trained transformer-based language models. The following are the language models experimented for this research:

a) Alibaba-NLP/gte-Qwen1.5-7B-instruct [18]
b) intfloat/multilingual-e5-large-instruct [19]
c) intfloat/multilingual-e5-large [19]
d) distiluse-base-multilingual-cased-v2 [20]
e) denaya/indoSBERT-large [21]
f) indobenchmark/indobert-large-p1 [22]

The embeddings from each of the language models are clustered and evaluated using Silhouette Score and Davies-Bouldin Index and the embeddings with the best performance are chosen.

2) Reduce Embeddings Dimension: The dimensionality of the embeddings was reduced using UMAP [23] to facilitate clustering.

3) Cluster Embeddings: The reduced embeddings were clustered using the HDBSCAN algorithm [24] to identify dense regions in the embedding space. Several values of minimum cluster size parameters of HDBSCAN were searched and the value that gives the best performance according to Silhouette Score and Davies-Bouldin Index was chosen.

4) Extract Topics: Topics were extracted from the clusters using the mixture of bag-of-words and class-based TF-IDF (c-TF-IDF) procedure. Each term in each cluster is counted using the CountVectorizer function similar to the sklearn implementation [25], before applying the c-TF-IDF algorithm to measure the importance of each term for each cluster, top 4 topics for each cluster are designated as labels for the said cluster.

The importance of each term in a cluster is calculated using the following equation:

$$W_{x,c} = \|tf_{x,c}\| \times log(1 + \frac{A}{f_x}) \quad (1)$$

Where, is the importance score of term x in cluster c, $tf_{x,c}$ is the term frequency of term x in cluster c, A is the average number of terms per cluster, f_x is the frequency of term x across all clusters.

Additionally, other configurations are experimented beside the default configuration:

- *Modify CountVectorizer's ngram_range value*: we modify the range of n-values for different word n-grams or char n-grams to be extracted.
- *Use c-TF-IDF with BM-25 weighting [26], & square root TF*: changing the c-TF-IDF calculation from equation (1) to:

$$W_{x,c} = \sqrt{\|tf_{x,c}\|} \times log(1 + \frac{A - f_x + 0.5}{f_x + 0.5}) \quad (2)$$

  This process further improves the robustness of the model to the unregulated stopwords.
- *MaximalMarginalRelevance* [27]: decrease term redundancy that represents the same information such as term "car" or "cars".

D. LLM Topic Refinement

Extracted topics were refined using large language models. Meta's LLaMA-3-8B-Instruct language model [28] was employed to enhance the coherence and relevance of the identified topics. *W*e utilized a specific prompt structure designed to guide the LLM in producing coherent and contextually relevant topic labels. This prompt structure included a system message, a one-shot example of user-assistant interaction, and a user query.

The system message established the LLM's role as a helpful assistant for labeling topics, setting the context for its responses. The one-shot example demonstrated an interaction between the user and assistant, providing a clear template for the model to follow. The user query contained the necessary context, including documents and keywords related to the topic to be refined. The user query included placeholders [DOCUMENTS] and [KEYWORDS] for the actual documents and keywords specific to each topic. Italicized sentences are provided as English translations for readers but are not included in the original prompt.

The prompt structure is detailed as follows:

LLaMA-3-8B Prompt

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a helpful, respectful and honest assistant for labeling topics.<|eot_id|><|start_header_id|>user<|end_header_id|>

I have a topic that contains the following documents:
- Perjanjian kontrak harus memenuhi syarat-syarat sah, termasuk kesepakatan bersama, kemampuan hukum, dan objek yang halal.
- Pengingkaran kontrak adalah ketika salah satu pihak gagal memenuhi kewajibannya sesuai perjanjian.
- Undang-Undang Perlindungan Konsumen melindungi hak konsumen dalam transaksi bisnis.
- A contract agreement must meet the legal requirements, including mutual agreement, legal capacity, and lawful object.
- Breach of contract occurs when one party fails to fulfill their obligations under the agreement.
- The Consumer Protection Act safeguards consumer rights in business transactions.

The topic is described by the following keywords: 'kontrak, perjanjian, hukum, perlindungan konsumen, kewajiban, pelanggaran, kesepakatan, transaksi'. *'contract, agreement, law, consumer protection, obligation, breach, settlement, transaction'*

Based on the information about the topic above, please create a short label of this topic. Make sure you to only return indonesian label and nothing more.<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Pelanggaran Kontrak dan Perlindungan Konsumen
Breach of Contract and Consumer Protection
<|eot_id|><|start_header_id|>user<|end_header_id|>

I have a topic that contains the following documents:
[DOCUMENTS]

The topic is described by the following keywords: '[KEYWORDS]'.

Based on the information about the topic above, please create a short label of this topic. Make sure you to only return indonesian label and nothing more.<|eot_id|><|start_header_id|>assistant<|end_header_id|>

E. Evaluation

The effectiveness of the topic modeling and refinement process was evaluated using clustering and topic evaluation metrics:

1) Silhouette Score [29]: This metric measures the quality of the clusters formed by our model. It calculates how similar an object is to its own cluster compared to other clusters. The Silhouette Score ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. The equation for the Silhouette Score, s, for each data point is given by:

$$s = \frac{b-a}{ma(a,b)} \quad (3)$$

Where a is the mean distance to the other points in the same cluster (cohesion), and b is the mean distance to the points in the nearest cluster that the data point is not a part of (separation) .

2) Davies-Bouldin Index [30]: This index is an internal evaluation scheme for clustering algorithms. It is defined as the average 'similarity' between each cluster and its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. Thus, lower values of the Davies-Bouldin index indicate better

clustering. The Davies-Bouldin Index, DB, can be formulated as:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right) \quad (4)$$

where n is the number of clusters, i is the average distance of all points in cluster i to the centroid c_i , and $d(c_i,c_j)$ is the distance between centroids c_i and c_j .

3) Cv Topic Coherence [31]: This metric measures the degree of semantic similarity between high scoring words in the topic. It provides a numerical value that helps in assessing how interpretable the topics are. Coherence values typically range between 0 and 1, where higher values correspond to more semantically coherent topics .

4) Topic Diversity [32]: This metric quantifies the uniqueness of topics by calculating the percentage of unique words across the top words of each topic, which helps to gauge the overlap between topics. If each topic consistently uses distinct sets of words, the topics are considered more diverse. The equation for calculating Topic Diversity, TD, is given by:

$$TD = \frac{|unique\ words\ in\ top\ N\ words\ of\ all\ topics|}{N\times K} \quad (5)$$

where N is the number of top words considered per topic, K is the total number of topics, and |unique words in top N words of all topics| is the count of unique words appearing in the top N words across all topics .

IV. Results and Discussion

The Results and Discussion section provides a comprehensive analysis of the findings from the application of BERTopic and LLM-based topic refinement to the dataset of Indonesian legal paragraphs. This section is structured to present the results of the various methodological steps, interpret the significance of these findings, and discuss their implications.

A. Data Collection and Data Pre-Processing Results

Table 1 summarizes the preprocessing and the number of documents retained after each step.

TABLE I. The number of documents retained after each preprocessing step

Pre-processing Step	Number of Documents
Initial	1,163,051
Drop NaN Row	1,161,481
Extract Paragraphs	37,453
Non-Alphabet Filtering	37,453
Alphabet Integer Filtering	37,453
Stopword Filtering	37,453
Length Filtering	36,689
MinHash LSH Deduplication	34,038

The initial dataset comprised 1,163,051 documents. After dropping rows with NaN values, 1,161,481 documents were retained. Subsequent steps, including extracting paragraphs and filtering out non-alphabetic characters, further refined the dataset to 37,453 documents. Filtering for alphabetic characters and integers, stopword removal, and length filtering reduced the dataset to 36,689 documents. Finally, MinHash LSH deduplication, which eliminates near-duplicate entries, resulted in a final dataset of 34,038 documents.

B. Embeddings Selection Results

The embeddings generated by multiple pre-trained transformer-based language models were evaluated to determine the best performing model for topic modeling. The embeddings were reduced to 100 dimensions using UMAP and clustered using the HDBSCAN algorithm with a minimal cluster size of 2. The performance of each embedding model was assessed using the Silhouette Score and Davies-Bouldin Index. The embeddings generated by the model "multilingual-e5-large-instruct" achieved the highest Silhouette Score (0.571) and the lowest Davies-Bouldin Index (0.530), indicating the best cluster quality among the models evaluated.

TABLE II. Evaluation of different embedding models using Silhouette Score and Davies-Bouldin Index

Embedding Model	Silhouette Score[a]	Davies-Bouldin Index[b]
gte-Gwen 1.5-7B-instruct	0.537	0.570
multilingual-e5-large-instruct	**0.571**	**0.530**
multilingual-e5-large	0.539	0.576
distiluse-base-multilingual-cased-v2	0.544	0.560
indoSBERT-large	0.517	0.590
indobert-large-p1	0.521	0.602

a. Higher is better.

b. Lower is better

C. Minimum Cluster Size Search Results

To further optimize the clustering performance, the minimum cluster size parameter in HDBSCAN was varied and evaluated using the embeddings from the "intfloat/multilingual-e5-large-instruct" model. The embeddings were reduced to 100 dimensions using UMAP, and various minimum cluster sizes were tested. The minimum cluster size of 40 was chosen because it provided a good balance between cluster quality and the number of clusters formed. This setting achieved a Silhouette Score of 0.723, which is the highest score achieved during the experiment, and a Davies-Bouldin Score of 0.340, with 189 clusters and 7,142 outliers. Table III presents the evaluation results:

TABLE III. Evaluation of minimum cluster size variations using Silhouette Score and Davies-Bouldin Index

Minimum Cluster Size	Silhouette Score[a]	Davies-Bouldin Index[b]	Number of Clusters	Number of Outliers
2	0.571	0.530	3,340	9,088
5	0.678	0.405	1,169	8,453
15	0.719	0.351	467	7,637
20	0.698	0.341	340	6,354
25	0.706	**0.327**	276	6,336

Minimum Cluster Size	Silhouette Score[a]	Davies-Bouldin Index[b]	Number of Clusters	Number of Outliers
30	0.697	0.334	233	6,540
40	**0.723**	0.340	189	7,142
50	0.710	0.343	149	6,659
75	0.705	0.332	96	7,042
100	0.685	0.342	65	7,519
150	0.714	0.349	50	8,132
200	0.709	0.360	40	9,325
500	0.660	0.455	15	14,872

a. Higher is better.

b. Lower is better.

D. Topic Extraction Results

After selecting the optimal embeddings and determining the best minimum cluster size, the next step was to extract topics from the clusters. Several different extraction techniques were applied to determine the most effective method for topic extraction. The performance of each technique was evaluated using the Cv Coherence Score and Topic Diversity Score. Table IV presents the evaluation results:

TABLE IV. EVALUATION OF DIFFERENT EXTRACTION CONFIGURATIONS USING CV COHERENCE AND TOPIC DIVERSITY

BERTopic Configuration	Cv Coherence Score[a]	Topic Diversity[a]
Default Settings	**0.469**	0.589
CountVectorizer ngram_range(1,5)	0.384	0.296
CountVectorizer ngram_range(1,3)	0.390	0.301
c-TF-IDF with BM-25 weighting & square root TF normalization	0.401	**0.774**
MaximalMarginalRelevance(diversity=0.3)	**0.469**	0.589
MaximalMarginalRelevance(diversity=0.5)	**0.469**	0.589

a. Higher is better

The default settings, CountVectorizer ngram_range(1,3 and 1,5), and MaximalMarginalRelevance (diversity=0.3 and 0.5) achieved the highest Cv Coherence Score of 0.469. However, the configuration using c-TF-IDF with BM-25 weighting and square root TF normalization achieved the highest Topic Diversity Score of 0.774.

Based on these findings, the c-TF-IDF configuration with BM25 weighting and square root TF normalization emerges as the most promising choice for our topic modeling task. While other configurations, particularly the default settings, demonstrated high coherence scores, the c-TF-IDF approach offers a superior balance between topic coherence and diversity.

E. LLM Topic Refinement Results

After extracting the initial topics using BERTopic, the topics were further refined using a large language model (LLM) to enhance their relevance and readability. The refinement process aimed to improve the semantic clarity of the topics, making them more interpretable and meaningful. Below are the 5 samples demonstrating the results of the LLM refinement:

TABLE V. COMPARISON OF EXTRACTED TOPICS BEFORE AND AFTER LLM REFINEMENT, SHOWING THE IMPROVEMENT IN SEMANTIC CLARITY

Extracted Topics	Topics After LLM Refinement
buruh, pekerja, pengusaha, serikat, pengaturan, pengurus, pengadilan, ilmu, mekanisme *workers, employees, entrepreneurs, unions, regulations, management, court, science, mechanism*	Hukum Hubungan Industrial (HRI) dan Pengaturan Pemutusan Hubungan Kerja *Industrial Relations Law (HRI) and Employment Termination Regulations*
kepolisian, republik, warga, indonesia, migran, asing, penyidik, kesatuan, tentara, bahasa *police, republic, citizens, Indonesia, migrants, foreigners, investigator, unit, military, language*	Penyidikan dan Wewenang Penyidik dalam Hukum Acara Pidana *Investigation and Investigator Authority in Criminal Procedure Law*
partai, politik, perolehan, gabungan, peserta, rekapitulasi, penghitungan, suara, lokal, pasangan *party, politics, acquisition, coalition, participants, recapitulation, vote counting, local, candidate pair*	Proses Penghitungan dan Rekapitulasi Hasil Pemilu *Vote Counting and Recapitulation Process in Elections*
presiden, wakil, bekas, terpilih, ketua, mpr, usul, diberhentikan, sungguh, jabatannya *president, vice president, former, elected, chairperson, MPR, proposal, dismissed, truly, office*	Proses Pengangkatan dan Pemberhentian Presiden dan Wakil Presiden *Process of Appointment and Dismissal of the President and Vice President*
produk, produksi, hortikultura, produktivitas, halal, produktif, bpjph, peredaran, pemasaran, proses *products, production, horticulture, productivity, halal, productive, BPJPH, circulation, marketing, process*	Regulasi dan Pengawasan Produk Pangan dan Hortikultura *Regulation and Supervision of Food and Horticultural Products*

Italicized sentences are provided as English translations for readers but are not part of model output. Correlation between the extracted topics and the refined topics are represented by the words highlighted in the same color.

The topics refined using the LLM showed notable improvement in semantic clarity compared to the initial extracted topics. For example, the topic "buruh, pekerja, pengusaha" (workers, employees, entrepreneurs) was refined into "Hukum Hubungan Industrial (HRI) dan Pengaturan Pemutusan Hubungan Kerja" (Industrial Relations Law and Employment Termination Regulations), offering a more precise and interpretable label relevant to legal discourse.

This refinement process significantly enhanced the coherence and relevance of the topics, making them clearer and more meaningful for legal analysis. By applying transformer-based topic modeling in combination with LLM refinement, the methodology produced topics that were not only more descriptive but also more aligned with legal terminology. For instance, the topic "kepolisian, republik, warga" (police, republic, citizens) was refined into "Penyidikan dan Wewenang Penyidik dalam Hukum Acara Pidana" (Investigation and Investigator Authority in Criminal Procedure Law), providing a more contextually appropriate representation.

Overall, this approach improves the interpretability of extracted topics from legal documents, making the analysis more valuable for research and practice while offering a robust methodology for future studies in legal text analysis.

V. Limitations & Future Work

This study faced several limitations that should be considered when interpreting the results. First, the dataset used was limited to a specific set of Indonesian legal documents, which may affect the generalizability of the findings to other legal systems or broader datasets. Additionally, while the LLM refinement process significantly improved the coherence and interpretability of the topics, the computational resources required to run large models like LLaMA can be a constraint, particularly for organizations with limited resources.

Looking forward, future research could address these limitations by applying the methodology to a larger and more diverse corpus, including legal documents from different jurisdictions and languages, to better assess the model's versatility. Additionally, exploring the use of alternative models, including smaller or more efficient LLMs, could help balance the need for high performance with computational feasibility. Fine-tuning the LLMs on more domain-specific legal data or incorporating expert feedback during the refinement process could further improve the accuracy and relevance of the extracted topics. Finally, practical applications of this method could be explored, such as its use in automating legal case management or assisting legal professionals in quickly identifying relevant case law, which would provide valuable insights into its real-world utility.

VI. Conclusion

This research demonstrates the effectiveness of combining BERTopic and large language models (LLM) for clustering and thematically identifying Indonesian legal paragraphs, significantly improving the coherence and relevance of extracted topics. Its application to Indonesian legal documents shows how these tools can enhance the organization and analysis of complex legal documents. The results have practical implications for improving the classification and retrieval of legal information, which can benefit legal research and document management. Future work can expand on this by applying the methodology to other legal systems and exploring its use in real-world legal settings.

References

[1] Maarten Grootendorst, "BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics," *Zenodo*, 2020.

[2] D. M. Blei et al., "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, no. 4–5, 2003, doi: 10.7551/mitpress/1120.003.0082.

[3] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," Front. Sociol., vol. 7, May 2022, doi: 10.3389/fsoc.2022.886498..

[4] L. B. Hutama and D. Suhartono, "Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic," Inform., vol. 46, no. 8, pp. 81–90, 2022, doi: 10.31449/inf.v46i8.4336.

[5] Z. Wang et al., "Identifying interdisciplinary topics and their evolution based on BERTopic," Scientometrics, 2023, doi: 10.1007/s11192-023-04776-5.

[6] W. Zhou et al., "ChatGPT and marketing: Analyzing public discourse in early Twitter posts," J. Mark. Anal., vol. 11, no. 4, pp. 693–706, Dec. 2023, doi: 10.1057/s41270-023-00250-6.

[7] F. Alhaj et al., "Improving Arabic Cognitive Distortion Classification in Twitter using BERTopic," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 1, 2022, doi: 10.14569/IJACSA.2022.0130199.

[8] OpenAi, "ChatGPT: Optimizing Language Models for Dialogue," OpenAi Blog, 2022.

[9] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023.

[10] A. Q. Jiang et al., "Mistral 7B," Oct. 2023.

[11] R. Silveira et al., "Topic modeling of legal documents via LEGAL-BERT," in CEUR Workshop Proceedings, 2021, doi: 10.2139/ssrn.4539091.

[12] A. J. Rawat et al., "Topic modeling of legal documents using NLP and bidirectional encoder representations from transformers," Indones. J. Electr. Eng. Comput. Sci., vol. 28, no. 3, 2022, doi: 10.11591/ijeecs.v28.i3.pp1749-1755.

[13] M. Abdurahman et al., "Lex2KG: Automatic Conversion of Legal Documents to Knowledge Graph," in 2021 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2021, 2021, doi: 10.1109/ICACSIS53237.2021.9631310.

[14] S. Bird, "NLTK: The natural language toolkit," in COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Interactive Presentation Sessions, 2006.

[15] A. Z. Broder et al., "Min-wise independent permutations," J. Comput. Syst. Sci., vol. 60, no. 3, 2000, doi: 10.1006/jcss.1999.1690.

[16] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in Conference Proceedings of the Annual ACM Symposium on Theory of Computing, 1998.

[17] Eric Zhu, "ekzhu/datasketch: v1.6.4," Zenodo, Oct. 03, 2023, doi: 10.5281/zenodo.8402527.

[18] Z. Li et al., "Towards General Text Embeddings with Multi-stage Contrastive Learning," Aug. 2023, doi: 10.48550/arXiv.2308.03281.

[19] L. Wang et al., "Multilingual E5 Text Embeddings: A Technical Report," Feb. 2024, doi: 10.48550/arXiv.2402.05672.

[20] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 2019, doi: 10.18653/v1/d19-1410.

[21] K. D. Rahadika Diana and M. L. Khodra, "IndoSBERT: Enhancing Indonesian Sentence Embeddings with Siamese Networks Fine-tuning," in 2023 10th International Conference on Advanced Informatics: Concept, Theory and Application, ICAICTA 2023, 2023, doi: 10.1109/ICAICTA59291.2023.10390469.

[22] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," Sep. 2020, doi: 10.48550/arXiv.2009.05387.

[23] L. McInnes et al., "UMAP: Uniform Manifold Approximation and Projection," J. Open Source Softw., vol. 3, no. 29, 2018, doi: 10.21105/joss.00861.

[24] R. J. G. B. Campello et al., "Density-based clustering based on hierarchical density estimates," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013, doi: 10.1007/978-3-642-37456-2_14.

[25] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, 2011.

[26] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Found. Trends Inf. Retr., vol. 3, no. 4, 2009, doi: 10.1561/1500000019.

[27] J. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," in SIGIR 1998 - Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, doi: 10.1145/290941.291025.

[28] AI@Meta, "Llama 3 Model Card," 2024..

[29] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., vol. 20, no. C, 1987, doi: 10.1016/0377-0427(87)90125-7.

[30] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-1, no. 2, 1979, doi: 10.1109/TPAMI.1979.4766909.

[31] M. Röder et al., "Exploring the space of topic coherence measures," in WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining, 2015, doi: 10.1145/2684822.2685324.

[32] A. B. Dieng et al., "Topic modeling in embedding spaces," Trans. Assoc. Comput. Linguist., vol. 8, 2020, doi: 10.1162/tacl_a_00325