Topic modeling of legal documents using NLP and bidirectional encoder representations from transformers

Amar Jeet Rawat1, Sunil Ghildiyal1, Anil Kumar Dixit2
1Department of Computer Science and Engineering, Uttaranchal University, Dehradun, India
2Law College Dehradun, Uttaranchal University, Dehradun, India

ABSTRACT
Modeling legal text is a difficult task because of its unique features, such as lengthy texts, complex language structures, and technical terms. During the last decade, there has been a big rise in the number of legislative documents, which makes it hard for law professionals to keep up with legislation like analyzing judgements and implementing acts. The relevancy of topics is heavily influenced by the processing and presentation of legal documents in some contexts. The objective of this work is to understand the legal judgement corpus related to cases under the Hindu Marriage Act of India. The study looked into various methods to generate sentence embeddings from the judgement. This paper employs the power of the BERTopic algorithm for generating significant topics.

Keywords:
BERTopic
Document clustering
Latent dirichlet allocation
Latent semantic analysis
Natural language processing
Topic modeling

This is an open access article under the CC BY-SA license.

Corresponding Author:
Amar Jeet Rawat
Department of Computer Science and Engineering, Uttaranchal University
Chandanwari, Prem Nagar, Dehradun, Uttarakhand, India
Email: aj.amar.rawat@gmail.com

1. INTRODUCTION
Legal texts are intended to be utilized in a specific manner. They have certain rules about how they should be organized and written. In legal texts, the smallest thing that makes sense is a clause. A group of words is called a clause that has a theme, a predicate, and is part of a compound or complex sentence [1]. Most of the law is written in natural languages like English. Therefore, natural language processing (NLP), along with machine learning (ML), is a crucial component for understanding, analyzing, topic modeling, and predicting laws. The recognition of words from the topics present in a corpus of data is called topic modelling [2]. Topic modelling can be applied to find topics that best describe a set of documents. The legal argumentation and judgement process is primarily reliant on textual information. Contract review, due diligence, understanding acts, and legal discovery are examples of time-consuming tasks that can benefit from NLP models and be automated, saving a significant amount of time. The goal of this paper is to obtain an abstract description of legal cases. The paper describes the approach to extracting topics from the judgement text of cases under the Hindu Marriage Act of India.

The process of extraction of collections of co-occurring words from a corpus is called topic modelling [3]. It is the most extensively used method in NLP for text mining. Some of the modelling techniques are latent semantic analysis (LSA), non-negative matrix-factorization (NMF), and latent dirichlet allocation (LDA). NMF is one of the factorization methods that ensures the non-negative elements of factorized matrices [4]. LSA is a statistical technique for representing and extracting the contextual sense of words from a text corpus [5]. The hidden concepts of a particular corpus are collected by LSA using singular value decomposition (SVD) [6]. It is also beneficial for information retrieval and filtering, and it works
effectively if the corpus is made up of documents that are meaningfully related [7]. LDA is a well-known topic model for identifying the set of hidden themes associated with a collection of documents [8]. In LDA, every file is modelled as a bag-of-words, with each topic modelled as a distribution of words [9].

There are many challenges in LDA and LSA topic modelling. Existing topic modelling models like LDA and LSA have many limitations. In LDA, the number of topics must be fixed. It also fails to demonstrate any relationship between the topics. It uses bag-of-words (BoW), which takes the assumption of word exchangeability without considering sentence structure. As LSA is a linear model, it is not suitable for datasets having non-linear dependencies. LSA uses SVD, which requires a lot of work and is challenging to update as new data becomes available.

2. BACKGROUND

Past studies show the implementation of ML and NLP techniques have been employed to analyze legal documents. To find a solution to unstructured data in Kadir and Aliman [10], the web-based text analytics and the R language are used to produce organized and summarized data. In Mangsor et al. [11], the traditional application of document clustering was combined with the topic modelling approach. With this integrated approach, it is possible to see the pattern. In Remmits and Kachergis [12], Araújo and Campos [13], to model legal corpus, LDA has been mostly used.

In Mohammed and Augby [14] compares the classification of scientific unstructured e-books using LDA and LSA. The work done in Neill et al. [15] focuses on making it easier to navigate and identify key legal topics and their associated collections of topic-specific terminology by evaluating the performance of topic-oriented models to summarize and display British statute. In Ravi [16], the researcher utilized LDA to model outstanding resources obtained by the Brazilian Supreme Court. The data set consists of a corpus of litigation that has been manually annotated with contextual labels by judicial professionals. Semantic analysis of the dataset shows that models have 10 or 30 topics that relate to the actual legal case discussed in court. The implementation of a model having 100 topics shows outstanding results.

The work done in Angelov [17] examines the usage of the LDA in obtaining accurate and meaningful topics in case law documents to discover the possibility of discovering subjects in the documents related to case law documents. The LDA has remained the favored model for modelling issues until now. Despite its ubiquity, LDA has a number of flaws. To get the optimized results from the LDA model, there should be a good number of topics. Furthermore, the LDA method uses a bag-of-words model of words, which ignores word order and semantics.

In (Chakravarty et al.) [18], the authors employ LDA to cluster Indian court decisions, with cosine similarity as the distance metric between documents. However, their assessment does not include a legal expert's prior knowledge to determine whether the clusters correspond to legal knowledge on the topic. The potential of distributed representations to capture the semantics of words and texts is gaining prominence Silveira et al. [19]. Google introduced bidirectional encoder representations from transformers (BERT), a sophisticated sentence embedding method Radford et al. [20].

In the family of BERT models, LEGAL-BERT is designed to aid NLP-based research in the law domain, application of legal technology, and computational law. The LEGAL-BERT model family is released in Devlin et al. [21], which benefits NLP-based research. It is pre-trained with legislation based on the UK and EU. To have token level context-specific word embedding, authors used generic context-specific language models like GPT-2 [22], BERT Gunjan et al. [23], and RoBERTa. BERTopic is a topic modelling technique that employs transformer-based models to achieve reliable word representation Okazaki et al. [24].

3. MATERIAL AND METHOD

This section explains the methodology used for building topic models and setting configurations that are used for analysis. In the first place, this paper describes the process of dataset acquisition. The second phase includes the procedure of preparing datasets and the implementation of BERTopic for topic modelling. In this section, the brief architecture of BERTopic is described. Lastly, it describes the topic representation and document clustering using term frequency–inverse document frequency (TF-IDF).

3.1. Data collection

For this work, we extracted data from the "LegalCrystal" website. Since the source data is not in text or csv format, we employ web scraping with Python's BeautifulSoup package. BeautifulSoup uses regular expressions to parse elements on an HTML page and generates a parse tree for easy searching, navigation, and editing. Legal case data is organized into three sections, namely: case details, case description, and judgment. Case details include subject, court name, decision date, case id, case name, acts, and names of
judges. For this work, case number, case name, acts, and case description are extracted from 1200 cases into a csv file. 8–10 paragraphs are found in each case judgement.

3.2. Topic modelling

An unsupervised learning approach that determines the distribution of themes in a corpus is referred to as topic modelling, where topics are known as a recurring pattern of terms [25]. The goal of topic modelling is to extract the words that convey the document’s concept. The extracted case dataset includes cases under the Hindu Marriage Act (HMA) and the algorithm used for topic modelling is aimed at identifying the words like “divorce, maintenance, custody, compromise, and settlement.” from case judgement. Figure 1 depicts the process of topic modeling. The python spacy package is used for data preprocessing. The selection of only those paragraphs that have some previous case citations or act related information is part of data processing.

![Figure 1. Topic modeling process](image)

3.3. BERTopic modelling

Bidirectional encoder representations from transformers (BERT) is a transformer-based pre-trained model, which has generated remarkable results for NLP based problems. Pre-trained models are especially useful because they are believed to have more accurate word and phrase representations. The approach discussed in this work uses BERTopic to identify document topics. BerTopic is a topic-modelling technique that forms condensed collections using transformers (BERT embedding) and class-based TF-IDF. In Figure 2, the architecture of BERTopic is shown. This algorithm consists of three steps. In the first step, it uses embedding techniques like BERT to excerpt document embeddings. The second step deals with the forming of clusters. It uses uniform manifold approximation and projection (UMAP) to decrease embedding dimensionality and hdbscan package to cluster reduced embeddings and construct semantically comparable document clusters. The final step is to use class-based TF-IDF to extract and reduce topics, and then use Maximal Marginal Relevance (MMR) to improve word coherence.

![Figure 2. BERTopic architecture](image)

3.4. Creating topic representation

To generate topics, we change the TF-IDF so that interesting words can be found in clusters of documents rather than per document. C-TF-IDF is a TF-IDF formula that has been applied to multiple classes by joining all documents in each class. As a result, instead of a set of documents, each class is converted into a single document. For each class I the frequency of words t is calculated and divided by the number of total words, ‘W’.

$$ W_{x,c} = \frac{TF_{x,c} \times \log (1 + \frac{A}{fx})}{W} \quad (1) $$

Where $TF_{x,c}$ denotes the frequency of word x in class c, $fx$ denotes the frequency of word x across all classes. $A$ stands for average number of words per class.

---

*Topic modeling of legal documents using NLP and bidirectional encoder ... (Amar Jeet Rawat)*
4. RESULT AND DISCUSSION

The model is initialized with the parameter verbose set to true so that the model's stages can be tracked. By running the model, we found 24 topics in each class. Figure 3 depicts the 2D representation of intertopic distance of legal document paragraphs. Figure 3(a) and Figure 3(b) depict the intertopic distance map without topic reduction and with topic reduction, respectively. By putting "nr_topics=15" in the model_reduce_topic function, we tried to cut down on topics that overlapped. Figure 4 and Figure 5 show the top eight most frequent topics with five words per topic before the topic reduction process, and after topic reduction, respectively. In Figure 6, a heat map depicting the similarity between topics is created based on the cosine similarity matrix between topic embeddings. In Figure 3(a), judgement paragraphs are clustered into 24 topics, and topic T1 has a maximum of 58 words. After applying topic reduction in Figure 3(b), paragraphs were clustered into 15 topics and topic id T0 has the highest 79 words.

![Intertopic Distance Map](a)

![Intertopic Distance Map](b)

Figure 3. Representation of Intertopic Distance of legal document (a) without topic reduction and (b) with topic reduction

![Topic Word Scores](c)

Figure 4. Top 8 most frequent topics with five words per topic (before topic reduction)
4.1. Application of BERTopic

BERTopic has a number of distinguishing advantages over the other topic models. The results show that, independent of the language model used to embed the documents, BERTopic maintains its competitiveness and that, in some cases, and performance may even improve when using cutting-edge language models. This shows that even if traditional language models are utilized, it can scale performance to keep up with new advancements in the field of language models and still be competitive. The usage and fine-tuning of BERTopic are greatly facilitated by the separation of the procedure of embedding documents from presenting topics.
4.2. Evaluation

The two most widely used metrics, topic diversity and topic coherence, serve as indicators of the effectiveness of the topic models in this study. The topic coherence of each topic model was assessed using normalized pointwise mutual information (NPMI). In this matrix, the measure scale goes from \([-1, 1]\), with 1 denoting the strongest connotation. The work of [26] defines topic variety as the proportion of unique words across all themes. The scale goes from \([0, 1]\), with 0 denoting superfluous topics and 1 denoting topics with more variety. Topic coherence and topic variety are examples of validation metrics that serve as proxies for what is a subjective assessment. Different users may have different opinions about a topic's coherence and diversity. Because of this, these metrics can be used to gain an idea of how well a model is performing.

5. CONCLUSION

In this work, we have shown the implementation of the BERTopic algorithm for topic modelling in Indian legal case judgement text. In terms of qualitative evaluation, the approach yields positive results, revealing topics that are consistent with the theme of the document. This paper can be taken as an initial approach for future studies. Furthermore, the performance of BERTopic can be compared with other topic modelling techniques. Different embedding models can be compared to construct a BERTopic model.

REFERENCES

BIOGRAPHIES OF AUTHORS

Amar Jeet Rawat is a research scholar at Uttaranchal University, Dehradun, Uttarakhand, India. He has 10 years of teaching experience. He holds a Master degree in Computer Science and Engineering from Graphic Era University and Bachelor degree from Hinacl Pradesh University. His research area includes recommendation system, natural language processing, machine learning and deep learning. He can be contacted at email: aj.amar.rawat@gmail.com.

Dr. Sunil Ghildiyal is an Associate Processor, Department of Computer Science and Engineering at Uttaranchal University, Dehradun, Uttarakhand, India. He has been awarded Ph.D. from Venkateshwra University, Gajraula, Amroga, Uttar Pradesh. His research interests include Computer Science Information Systems, machine learning, and intelligent systems. His has published many research papers in good SCOPUS and SCI journals. He can be contacted at email: Sg124ddn@gmail.com.

Prof. Dr. Anil Kumar Dixit is a Professor at Law College Dehradun, Dehradun, Uttarakhand, Indian. His has published many research papers in good SCOPUS and SCI journals. He can be contacted at anil@uttaranchaluniversity.ac.in.