UNIVERSITY PARTNER

UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

6CS007
PROJECT AND PROFESSIONALISM

LITERATURE REVIEW REPORT

NYAYAMITRA: RAG based AI for Nepali legal Q&A

University Id : 2408239

Group : L6CG15

Reader : Mr. Akash Adhikari

Supervisor : Mr. Yamu Poudel

Student Name : Nikisha Shrestha

Course : BSc (Hons) Computer Science

Submitted on : 11/14/2025

Contents

1. Introduction 1
2. Literature Review 2
   2.1. Transformer-based models 3
   2.2. Evolution of multilingual and monolingual models 4
   2.3. Monolingual models in context of Nepal 5
   2.4. LLM models in context of Legal domain 6
   2.5. NepKanun 7
3. Conclusion 8
4. Future Work 9
5. References 9

1. Introduction

Lawyers and legal professionals are often burdened by the need to manually analyze large volumes of complex documents such as case laws, contracts, statutes, and regulations - a process that is both time-consuming and prone to inefficiency. Existing AI systems like ChatGPT and Claude, while powerful in general domains, lack the legal domain training required to deliver reliable, citation-grounded outputs for specialized use cases. In Nepal, these challenges are amplified by the scarcity of Nepali legal corpora and complex syntactic structures. Therefore, there is a clear need for AI-powered legal tools capable of supporting accurate, efficient, and context-aware legal analysis.

To address this gap, NYAYAMITRA is proposed as a Retrieval-Augmented Generation (RAG)-based legal question-answering system designed to provide accurate, citation-based responses to Nepali legal queries by integrating user-uploaded documents with Nepali law. To build such a system, it is crucial to understand how modern Natural Language Processing (NLP) techniques support reliable reasoning and context-aware generation. Therefore, this literature review examines the evolution of NLP – from early statistical approaches to advanced transformer-based architecture, as well as the development of multilingual and monolingual Nepali language models and domain-specific legal LLMs. By analyzing existing research on Nepali NLP, legal-domain modeling, and RAG methodologies, this section establishes the theoretical foundation for NYAYAMITRA and identifies the gaps that this system aims to address.

2. Literature Review

There have been many advancements in the field of Natural Language Processing (NLP) with initial approaches including n-grams and rule-based systems. These systems laid the foundation for understanding natural languages but faced significant limitations in terms of more complex nuances. After this, models such as conventional Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory networks (LSTMs) were introduced, allowing models to process sequential data more effectively. Word embeddings such as Word2Vec, GloVe, and fastText were used as input vectors for these models. These systems performed better in tasks like language modelling and sequence prediction (Thapa et al., 2025). However, since the embeddings were static, they could not capture semantic relationship between words and faced challenges in terms of long-range dependencies and computational efficiency (Pudasaini et al., 2023).

The next step in the evolution of NLP was the introduction of pre-trained models, starting with approaches like ULMFit (which uses RNN as its core) and transformer-based models such as BERT, RoBERTa, XL-NET, and ALBERT. These transformer-based models integrate attention mechanisms which differ from RNNs and CNNs. Through the attention mechanism, transformers can discern salient features from input context, disregarding noise, thus enabling nuanced assessments of word dependencies regardless of their sequential proximity. Transformer framework uses two types of attention modalities: self-attention and encoder-decoder (or cross) attention. In self - attention modality, each word in the input sequence looks at all the other words (including itself), to decide which ones are important for understanding its meaning. This modality encodes both syntactic and semantic relationships. After self-attention builds rich representations of the input (in the encoder), the model then uses cross-attention to generate output (in the decoder). The decoder looks at the encoder's output (the processed input sequence) and uses weighted attention to decide which parts of the input are most relevant for predicting the next word in the output in an autoregressive

fashion (predicts one word at a time - each prediction depends on all the previous predicted words) (Siino et al., 2025).

The field of NLP was further enhanced by the concept of self-supervised model pre-training, where models like ElMo, BERT, and GPT use advantage of pre-training on large volumes of unlabeled text dataset to learn general language representations (Timilsina et al., n.d.). Further advancements include instruction tuning, a specific kind of fine tuning in which models are trained on instruction-output pairs to improve their ability to follow user commands (Thapa et al., 2025).

2.1. Transformer-based models

Among the wide variety of transformer-based models, BERT, RoBERTa, and GPT-2 are the most notable ones.

1. BERT

Bidirectional Encoder Representations from Transformers (BERT), developed by Google in 2018 introduced bidirectional contextual embeddings, which predict masked words by considering both right and left context, and laid the foundation for many subsequent NLP models. BERT was trained on 3.3 B words from the English Wikipedia and Book corpus. The model was pre-trained on Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM task, certain words in a sentence are intentionally masked and the model's goal is to predict these words by using the context provided by the surrounding words whereas in NSP task, sentence pairs are fed to the model, and the model has to predict whether the second sentence logically follows the first or if these sentences are random and unrelated (Thapa et al., 2025).

2. RoBERTa

Following BERT, models such as RoBERTa and XLNET were introduced with larger pre-training data and increased model parameters (Timilsina et al., n.d.). Robustly Optimized BERT Pretraining Approach (RoBERTa), released by Facebook AI in 2019, in contrast to BERT, focuses solely on MLM task, completely disregarding NSP task. This modification has been proven to improve performance in various benchmarks (Thapa et al., 2025).

3. GPT-2

Generative Pre-trained Transformer (GPT-2) on the other hand, uses Casual Language Modeling (CLM), which predicts the next word in a sequence given the preceding context in an autoregressive, left-to-right manner. (Thapa et al., 2025).

2.2. Evolution of multilingual and monolingual models

Initially, only two versions of BERT were available in English and Chinese. Later, multilingual BERT (m-BERT) was developed, trained in 104 languages to support cross-lingual understanding. This created a benchmark for many multilingual tasks, demonstrating that a single model can acquire shared representations across multiple languages. Since then, many multilingual models such as XLM and XLM-RoBERTa have been introduced. These models made significant advances in terms of size and performance. Furthermore, inspired by m-BERT, NLP communities developed their own language-specific BERT models. Some of the popular BERT models include Russian, Dutch, Arabic, French, and Portuguese (Timilsina et al., n.d.). Research has shown that multilingual models often underperform compared to monolingual models. Due to this, other language specific models such as FinBERT for Finnish, BERTje and RoBERT for Dutch, FlauBERT for French, NorBERT for Nordic languages, Chinese BERT for

Chinese, Her-BERT for Polish languages, GBERT for German, IndicBERT for Indian languages etc were introduced (Thapa et al., 2025).

2.3. Monolingual models in context of Nepal

Nepali is a low-resource language, with less than 1 GB of text data available in Wikipedia, highlighting the challenges of data scarcity for NLP tasks (Timilsina et al., n.d.). Additionally, the syntactic structure of Nepali language which follows Subject-Object-Verb (SOV) differs from English language which follows Subject-Verb-Object (SVO) order. This factor poses additional challenges for developing an effective monolingual NLP models for Nepali (Thapa et al., 2025).

IndicBERT, a monolingual model for Indian languages, also supports Nepali language and has demonstrated that language specific models can outperform their multilingual counterparts on specialized tasks. Furthermore, models such as NepBERTa, NepaliBERT, BERT-Nepali, NepBERT, etc. have introduced monolingual BERT models for the Nepali language (Thapa et al., 2025).

NepBERTa released by (Timilsina et al., n.d.) is a variant of BERT model developed specifically for Nepali languages. The data used to train this model was collected through scrapping the top 36 news sites in the Nepali language. A total of 12.4 GB data was scrapped during this process, and the model was trained on 0.8 B words. Similarly, other models such as NepBERT and NepaliBERT were trained in text corpuses made available by the OSCAR dataset (Timilsina et al., n.d.)

2.4. LLM models in context of Legal domain

Transformer-based models and Large Language Models (LLMs) such as GPT-4, and Llama series have greatly advanced NLP tasks. However, in context of specific domains like law, these models are prone to hallucinations. So, to generate more context aware answers, domain-based models must be introduced. One such model is LEGAL-BERT which is trained in large legal corpora from US, UK, and European countries. This model has greatly improved performance on legal tasks like Named Entity Recognition (NER) and classification.

Despite the prominence of domain-specific models, they may still be prone to hallucinations as these LLMs are still statistical models – they predict text based on patterns in training data. To address this, Retrieval-Augmented Generation (RAG) has surfaced which combines LLMs with external retrieval steps and grounds responses in reliable sources prior to generation. RAG is particularly valuable in scenarios where users need to upload their own documents, allowing the model to provide answers that are directly informed by the uploaded content. Thus, a hybrid approach involving a combination of RAG with Nepali law-specific LLMs is proposed by (“NepKanun: A RAG-Based Nepali Legal Assistant,” 2025).

Another notable research in the field of Nepali NLP includes development of transformer-based bidirectional neural Machine Translation system for English-Nepali legal texts with a corpus of 125,000 sentences. However, this model also presents challenges in terms of data scarcity for question answering and simplification for general public (Poudel et al., 2024).

2.5. NepKanun

NepKaunun is a RAG based system combined with a LLaMa 3.2 3B model fine-tuned on Nepali legal QA dataset. It includes over 16,000 entries from Supreme Court's website, news sources as well as processed legal documents. The text data were transformed into question answering pairs designed for instruction-tuning using prompt engineering. The resulting dataset was then fine-tuned on Llama 3.2 3B model employing a Parameter-Efficient Fine-Tuning (PEFT), specifically Low-Rank Adaptation (LoRA) and Quantized LoRA (QLoRA) ("NepKanun: A RAG-Based Nepali Legal Assistant," 2025).

Use of RAG Methodology in NepKanun

NepKanun combines LLM that produces the final response, with a retrieval component that finds relevant data chunks from a knowledge base.

Steps employed to combine RAG methodology with LLMs

1. Knowledge base construction and preprocessing

For the construction of knowledge base, important legal texts such as Constitution of Nepal 2072 & portions of important legislation (such as Environmental Act & Muluki Ain) were retrieved.

2. Chunking

Structure-aware chunking technique was used that divided the texts logically to maintain the semantic structure present in legal documents, frequently arranged hierarchically (e.g., into Parts, Chapter, and Articles).

3. Vector Embedding & Indexing

The divided chunks were then transformed into dense vector embeddings using a multilingual Sentence Transformer model, based on architecture like Sentence-BERT. This model was used as they can capture semantic linkages across languages including Nepali. Then, vector embeddings of 384 dimensions were used to represent each piece. As for storing and indexing the vector embeddings, ChromaDB, an open-source vector database was used as it is designed for similarity search.

4. Retrieval and Generation

The user's query is embedded into the same 384-dimensional vector space. The retrieval component then does a similarity search to find the most relevant ChromaDB pieces. Maximal Marginal Relevance (MMR) was used to choose the top k=9 document chunks. In the generating component, the fine-tuned Llama 3.2 3B model receives the retrieved document chunks and concatenates it with the initial user query and finally generates a context-grounded legal answer ("NepKanun: A RAG-Based Nepali Legal Assistant," 2025).

3. Conclusion

This literature review highlights substantial progress in NLP, from transformer-based architectures to the emergence of Nepali monolingual models and domain-specific legal language systems. While these developments establish a strong foundation, they also reveal critical gaps – particularly in addressing domain-specific hallucinations, and the absence of robust Nepali legal LLMs capable of delivering grounded, citation-first answers. Recently, NepKanun proposed a hybrid solution that combines RAG with a Nepali legal dataset fine-tuned model; however, the associated research paper is still in the process of formal publication ("NepKanun: A RAG-Based Nepali Legal Assistant,"

2025). Building upon this emerging work, the present project aims to replicate and extend the same architectural direction by integrating a Nepali legal language model, NyayaLM, fine-tuned on Nepali legal datasets with a structured RAG pipeline.

4. Future Work

While the first version of NYAYAMITRA will rely on a RAG-based architecture combined with a Nepali legal language model (NyayaLM) fine-tuned on domain-specific datasets, the long-term vision for the system extends beyond question answering. Future iterations will integrate additional NLP capabilities such as Named Entity Recognition (NER) for extracting legal entities and semantic similarity estimation for tasks like case law comparison. These enhancements aim to transform NYAYAMITRA from a legal Q&A assistant into a more comprehensive legal intelligence platform capable of assisting lawyers, students, and citizens in deeper legal research and document understanding.

5. References

NepKanun: A RAG-Based Nepali Legal Assistant, 2025. , in: Proceedings of the 2025 ACL Rolling Review May Submission. Presented at the ACL (Association for Computational Linguistics) Rolling Review.

Poudel, S., Bal, B.K., Acharya, P., 2024. Bidirectional English-Nepali Machine Translation(MT) System for Legal Domain, in: Melero, M., Sakti, S., Soria, C. (Eds.), Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-Resourced Languages @ LREC-COLING 2024. ELRA and ICCL, Torino, Italia, pp. 53–58.

Pudasaini, S., Shakya, S., Tamang, A., Adhikari, S., Thapa, S., Lamichhane, S., 2023. NepaliBERT: Pre-training of Masked Language Model in Nepali Corpus, in: 2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). Presented at the 2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 325–330. https://doi.org/10.1109/I-SMAC58438.2023.10290690

Siino, M., Falco, M., Croce, D., Rosso, P., 2025. Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. IEEE Access 13, 18253–18276. https://doi.org/10.1109/ACCESS.2025.3533217

Thapa, P., Nyachhyon, J., Sharma, M., Bal, B.K., 2025. Development of Pre-Trained Transformer-based Models for the Nepali Language, in: Sarveswaran, K., Vaidya, A., Krishna Bal, B., Shams, S., Thapa, S. (Eds.), Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025). International Committee on Computational Linguistics, Abu Dhabi, UAE, pp. 9–16.

Timilsina, S., Gautam, M., Bhattarai, B., n.d. NepBERTa : Nepali Language Model Trained in a Large Corpus.

NyayaLM, https://huggingface.co/chhatramani/NyayaLM_v0.5_gemma3n4B