Legal Question Answering System using LLaMA, RoBERTa and Cosine Similarity

Shivangi Singh
Symbiosis Institute of Technology
Department of Artificial Intelligence
Pune,India
shivangi.singh.btech2021@sitpune.edu.in

Sachit Lele
Symbiosis Institute of Technology
Department of Artificial Intelligence
Pune,India
sachit.lele.btech2021@sitpune.edu.in

Sahil Shenoy
Symbiosis Institute of Technology
Department of Artificial Intelligence
Pune,India
sahil.shenoy.btech2021@sitpune.edu.in

Mayur Gaikwad
Symbiosis Institute of Technology
Department of Artificial Intelligence
Pune,India
mayur.gaikwad@sitpune.edu.in

Dr. Shruti Patil
Symbiosis Institute of Technology
Department of Artificial Intelligence
Pune,India
shruti.patil@sitpune.edu.in

***Abstract*—In information retrieval, a Question Answering (QA) system's job is to automatically provide accurate responses in natural language to questions posed by humans. This can be done by either a pre-organized database or a set of documents records. It displays just the data that has been requested rather than looking up entire documents using a search engine. According to data in daily life is getting busier, so in order to find the precise piece of Even a basic query requires a large and costly amount of information supplies. The document that explains the various methodology and specifics of the question-answering implementation for legal documents to get more accurate responses through the use of NLP methods.**

***Index Terms*—Legal Documents, LLaMA, Cosine-Similarity,RoBERTa.**

I. Introduction

In India, legal documents are essential to the operation of the judicial system and the government. Statutes, case law, agreements, rules, and other legal documents are among them. However, it can be difficult and time-consuming to understand and extract useful information from these documents [1]. Nowadays, people—including legal professionals—find it difficult to sift through the enormous amount of legal materials, which leads to mistakes, misunderstanding, and time-consuming work. This study suggest creating an Indian legal document question-answering paradigm as a solution to this problem [1].

The objective is to develop a model that can accurately answer questions based on the content of legal documents especially indian legal docs, making it easier for users to find and utilize the information they need. Models such as RoBERTa, LLaMA and naive approaches are applied to get the result.

The RoBERTa model and the BERT model have the same architecture. It is a reimplementation of BERT with a few minor embedding tweaks and adjustments to the important hyperparameters [8]. A vast dataset comprising more than 160GB of uncompressed text is used to train RoBERTa. The English Wikipedia and Books Corpus (16GB) used in BERT are included in the dataset for RoBERTa. The Web text corpus (38 GB), the CommonCrawl News dataset (63 million articles, 76 GB). RoBERTa was pre-trained using this dataset and 1024 V100 Tesla GPUs that were operating for a day [10].

whereas, Llama 2 is a collection of pre-trained and fine-tuned large language models (LLMs) ranging in scale from 7B to 70B parameters from Meta, Facebook's parent company [16].

II. Data Preparation

From the website "indiacode.nic.in," we gathered an extensive dataset of legal documents covering a variety of legal topics, such as health law, education law, and more [17].

A. Preprocessing

In order to get the data ready for use in an answering model, we extensively extracted text from PDF files and underwent a thorough preprocessing process such as tokenization, which divides text into meaningful units, sentence segmentation, which handles special characters and formatting, stop word removal, stemming and lemmatization, text normalization, which reduces noise, data cleaning, which balances data for class distribution, data splitting, which evaluates performance, and data augmentation, which increases the diversity of training data [3]. In the end, these preprocessing procedures enable

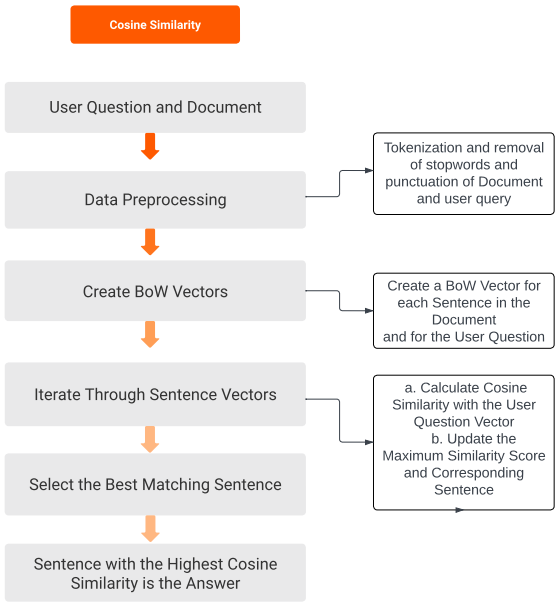
more accurate and effective legal information retrieval and analysis by ensuring the data is clean, structured, and suitable for training and testing the question-answering model [4].

III. Model Architecture

For the purpose of this project we have implemented question answering systems using three approaches Naive approach (Cosine Similarity), LLaMA and RoBERTa.

A. Cosine Similarity

Cosine similarity is an approach used in question answering to measure the similarity between two vectors, often representing a question and a document (e.g., a passage of text) [5] .The process involves the following steps[6].

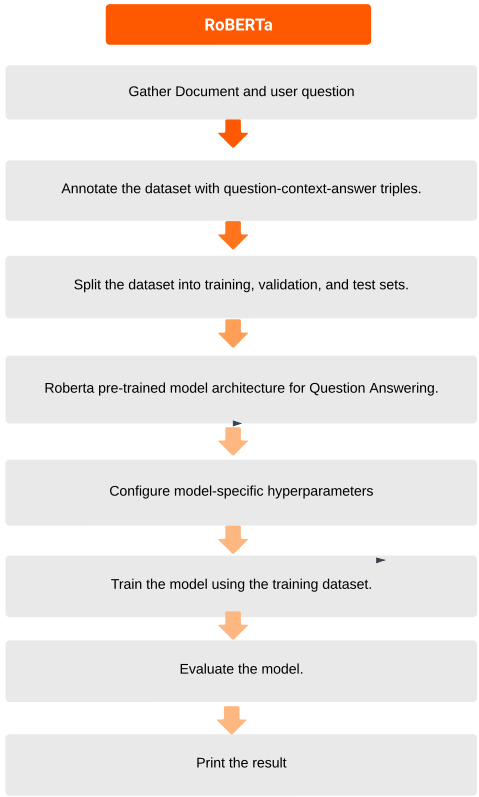


- Cosine Similarity Calculation: To compare the similarity between a question vector and a document vector, the cosine similarity is computed. Cosine similarity measures the cosine of the angle between two vectors in a multi-dimensional space. A value of 1 means the vectors are identical, and 0 means they are orthogonal (completely dissimilar)[6].
- Ranking: For a given question, the documents are ranked based on their cosine similarity scores [5]. Documents with higher similarity scores are considered more relevant to the question.
- Answer Extraction: The answer to the question is often extracted from the top-ranked document(s) [6].

Because cosine similarity cannot understand word context, it is less useful in solving the question-answering problem.

B. RoBERTa

RoBERTa is a strong transformer-based language model whose powerful language understanding abilities and extensive pretraining have revolutionized the field of natural language processing. Given their renowned complexity, legal texts provide special difficulties. This study investigate the use of RoBERTa's flexible architecture to create a legal document-specific question-answering model [7].



Following preprocessing, each document was methodically separated into contexts, which were distinguished by the legal act's subsections. A set of questions and answers was carefully created within each act context, following the format of the Stanford Question Answering Dataset (SQuAD) [4]. In order to ensure compatibility and structured representation for future analysis and utilization, the resultant data was arranged and saved in JSON format [10].

After creating this large dataset of context-question-answer pairs, RoBERTa model was used to perform question answering tasks. The user's query and the document's legal

context were tokenized into subword tokens by means of RoBERTa's sophisticated language understanding capabilities. These tokenized inputs were then encoded and used as the RoBERTa model's input [9]. We used unique tokens like [CLS] (classification token) and [SEP] (separator token) to organize the input for RoBERTa in order to speed up the question-answering process.

The RoBERTa model, which was optimized for question answering on this particular legal dataset, played a crucial role in extracting accurate responses to user inquiries from the context passages[1].

1) Evaluation-Metric: To evaluate the answers generated by the the RoBERTa generative model we used the metric of perplexity score.

2) Evaluation Definition: Perplexity is a measurement often used in natural language processing to assess the quality of language models. It provides a way to quantify how well a language model predicts a sequence of words. Language models assign probabilities to sequences of words. The perplexity score is the inverse of the probability of the sequence. Lower perplexity indicates that the language model is less "perplexed" and is better at predicting the text, while higher perplexity suggests the model is less accurate in predicting the sequence [8]. Perplexity is often calculated using the following formula:

$$\text{Perplexity}(W) = P(w_1, w_2, \ldots, w_n)^{-\frac{1}{n}} \quad (1)$$

P(w_1, w_2, ..., w_n) is the probability assigned by the language model to the sequence of words w_1, w_2, ..., w_n. n is the number of words in the sequence.

3) Evaluating the model: For evaluating the model we first fine tuned the model on a set of question-answer-context SQuAD styled dataset based on Indian Legal Documents, the context asked for the evaluation for the model was as follows: Context:-"An Act to provide for the better prevention of the spread of Dangerous Epidemic Diseases. (1) This Act may be called the Epidemic Diseases Act, 1897. Definitions (a) 'healthcare service personnel' means a person who while carrying out his duties in relation to epidemic related responsibilities, may come in direct contact with affected patients and thereby is at the risk of being impacted by such disease, and includes- (i) any public and clinical healthcare provider such as doctor, nurse, paramedical worker and community health worker; (ii) any other person empowered under the Act to take measures to prevent the outbreak of the disease or spread thereof; and (c) "property" includes- (i) any facility identified for quarantine and isolation of patients during an epidemic; (ii) a mobile medical unit; and (iii) any other property in which a healthcare service personnel has direct interest in relating to the epidemic;.

The question asked to extract information from the above and to evaluate the model was "How is 'healthcare service personnel' defined under the Act?". A set of 20 answers of were generated by the the RoBERTa model having perplexity scores between the range of 1.003 to 1.185 with an average perplexity score of 1.005,a sample of the answers generated and there respective perplexity scores has been shown in table 1 ,the perplexity scores are plotted in the fig1, the frequency of the perplexity scores of the answers generated by the model are plotted in the fig2

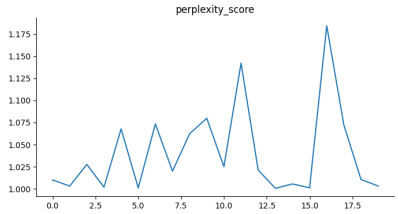


Fig. 1. Perplexity Scores for every Answer generated by the model

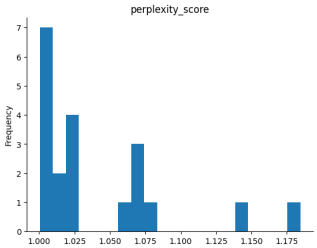


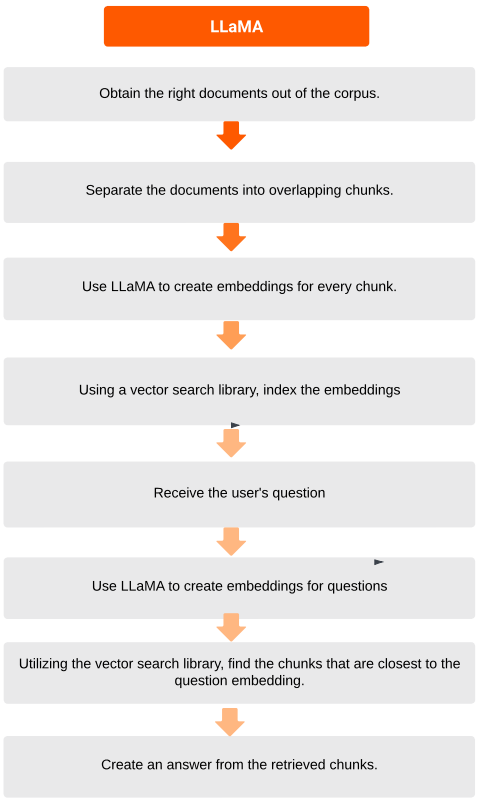
Fig. 2. Frequency of Perplexity Scores for Answers Generated by the model

TABLE I
TABLE OF COMPARING ROBERTA GENERATED ANSWERS WITH THEIR RESPECTIVE PERPLEXITY SCORES (SMALLEST, AVERAGE, LARGEST)

Text	Perplexity Score
'healthcare service personnel' means a person who while carrying out his duties in relation to epidemic related responsibilities, may come in direct contact with affected patients and thereby is at the risk of being impacted by such disease	1.003
any facility identified for quarantine and isolation of patients during an epidemic; (iii) a mobile medical unit; and (iv) any other property in which a healthcare service personnel has a direct interest in relating to the epidemic.	1.015
a mobile medical unit; in contradiction to any other property in which a personnel has direct interest in relating to the pandemic;	1.185

C. LLaMA

To begin, Llama converts all of the data it receives from text files, PDFs, and other sources into word embeddings. These embeddings are kept in a database by it. Subsequently, it extracts these embeddings by comparing their similarity scores to the queries the user poses. Subsequently, refined responses are produced by utilizing the extracted similarity scores [16].



1) Evaluating The model: To evaluate the model we calculated the perplexity score of the generated answers.We asked the model a series of three questions and calculated the perplexity scores of the answers.The questions asked to the model were:
1)What powers does the government have to stop the spread of diseases?
2)How does government control use of electronic cigarettes?
3)What are the penalties for violence against healthcare service personnel and damage to property during an epidemic?
The perplexity scores of the generated model ranged between 1.775 to 2.56 and were plotted in the plot given in fig3

IV. Related Works

In the field of question answering on legal documents, there are renowned research studies emphasizing on the extraction and interpretation of legal information from textual documents. The research emphasis is in the application of advanced NLP methods such as Cosine Similarity,RoBERTa and LLaMA, to facilitate precise question answering within the legal domain. This section further describes some of the prominent research work in the field of interest.

Gil Martinez et. al in [1] have employed employs both quantitative and qualitative research methods to survey

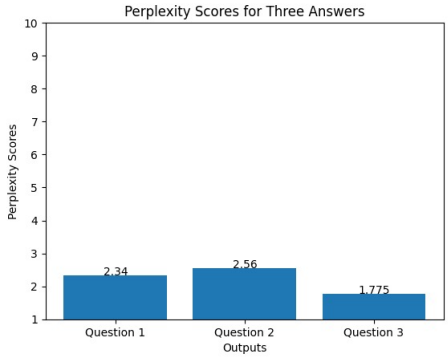


Fig. 3. Perplexity Scores for the three answers generated by the LLaMA model

existing solutions for legal question-answering systems. This approach ensures a comprehensive assessment of the current landscape. The paper may not cover every existing legal question-answering system, as the field is rapidly evolving, and new solutions may have emerged since the paper's publication.

Phong-Khac et al. in [2] made Legal Question Answering using Ranking SVM and Deep Convolutional Neural Networks, the paper introduces a combination of features (LSI, Manhattan, Jaccard) and focuses on single-paragraph articles, achieving improved results.

In paper [5] Vold et. all compares the performance of a RoBERTa Base classifier with a traditional linear SVM on legal domain question answering tasks. RoBERTa significantly outperforms the SVM in terms of F1-score (31 percent improvement) and Mean Reciprocal Rank (41.4 percent improvement) in legal question answering. While RoBERTa achieves better performance, the use of traditional ML models like SVM may still be appropriate for cases where data quality and redundancy pose challenges.

The author of paper [11] created a geographical domain question-answering system that responds to user inquiries about different cities. In order to design the system, the author first created a knowledge base document and undertook document pre-processing using a Named entity tagger, parser, and word net tool. This included noise removal, tokenization, sentence splitting, and document tagging. The three primary components of this system are document processing, answer processing, and question processing. The subclassification and reformulation of questions are dealt with in question processing. The question is classified using a simple matching pattern algorithm. Following that, passage retrieval was implemented, using an indexed and pre-processed corpus. The candidate answer generated by the retrieval module is used as input for the answer extraction process, where a WordNet tool was used to perform semantic relation ranking. Following ranking, the final response with the maximum rank is displayed.

V. Conclusion

This study solves a significant challenge in the legal domain by utilizing multiple techniques, such as Cosine Similarity, RoBERTa, and LLaMA, in the context of question-answering for legal documents. Legal documents frequently use complex, highly technical language that calls for a thorough grasp of context and semantics. Better legal research, analysis, and decision-making could be facilitated by these models.In this study perplexity score as a metric was used for primary evaluation of the models, with the performance of the RoBERTa model having answers between the score range of 1.003 to 1.185, while the LLaMa model had answers generated ranged between 1.775 to 2.56 for a set of three different questions.

While discussing the scores its also important to mention that RoBERTa's answers were realtively more extractive in nature and were reliant on the context provided while the answers generated by the LLaMA were more generative in nature and had the influence of additional information added by the model's pretrained influence which resulted better quality of answers but at the cost of higher computational cost.Hence a hybrid approach involving both techniques has a potential for future work for optimum information extraction at reduced computational cost.

VI. References

1.Martinez-Gil, Jorge. "A survey on legal questions–answering systems." Computer Science Review 48 (2023): 100552.

2.Do, Phong-Khac, et al. "Legal question answering using ranking SVM and deep convolutional neural.

3.Kim, Mi-Young, Ying Xu, and Randy Goebel. "Applying a convolutional neural network to legal question answering." New Frontiers in Artificial Intelligence, Japan, November 16-18, 2015, Revised Selected Papers. Springer International Publishing, 2017.

4.Khazaeli, Soha, et al. "A free format legal question answering system." Proceedings of the Natural Legal Language Processing Workshop 2021. 2021.

5. F Meng, W Wang, J Wang. 2021. "Research on Short Text similarity calculation Method for Power Intelligent Question Answering".In 13th International Conference on Computational Intelligence 2021.

6.M Ahmed, HU Khan, S Iqbal, Q Althebyan. 2022" Automated Question Answering based on Improved TF-IDF and Cosine Similarity". In Ninth International Conference on Social Networks Analysis.

7.Xiao, Chaojun, et al. "Lawformer: A pre-trained language model for chinese legal long documents." AI Open 2 (2021): 79-84.

8.S. Abualhaija, C. Arora, A. Sleimi and L. C. Briand, "Automated Question Answering for Improved Understanding of Compliance Requirements: A Multi-Document Study," 2022 IEEE 30th International Requirements Engineering Conference (RE), Melbourne, Australia, 2022, pp. 39-50.

9. Van, Hieu Nguyen, et al. "Miko team: Deep learning approach for legal question answering in alqac 2022." 2022 14th International Conference on Knowledge and Systems Engineering (KSE). IEEE, 2022.

10.Jha, Raj, and V. Susheela Devi. "Extractive Question Answering Using Transformer-Based LM." International Conference on Neural Information Processing. Singapore: Springer Nature Singapore, 2022.

11. Amit Mishra, Nidhi Mishra and Anupam Agrawal, "Context Aware Restricted Geographical Domain Question Answering System", In 2010 International Conference on Computational Intelligence and Communication Networks.

12. Zeng-Jian Liu, Xiao-Long Wang and Qing-Cai Chen "A Question answering system on web search" International conference on machine learning 2014.

13. Pum-Mo Ryu, Myung-Gil Jang and Hyun-Ki Kim. 2014. "Open domain question answering using Wikipedia-based knowledge model." In Information Processing and Management 50 (2014) 683– 692, Elsevier

14. Vold, Andrew, and Jack G. Conrad. "Using transformers to improve answer retrieval for legal questions." Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. 2021.

15. Caballero, Ernesto Quevedo, et al. "Study of Question Answering on Legal Software Document using BERT based models." LatinX in Natural Language Processing Research Workshop. 2022.

16.H Touvron, T Lavril, G Izacard, X Martinet.2023." Llama: Open and efficient foundation language models".

17. https://www.indiacode.nic.in/