

# Extractive Question Answering System for SEBI Documents

Deepthi Saravanan

*Abstract*—Securities and Exchange Board of India (SEBI) is the regulatory body for securities and commodity market in India under the jurisdiction of Ministry of Finance, Government of India. SEBI has framed a set of regulations that companies are expected to abide by. In addition, it also regularly publishes additional documents that clarifies company's queries, amplifies the regulations with the purpose and exceptions and also discusses the financial market activities. These lengthy and interconnected documents might be hard for legal experts to search for information given the situation and issues. Advances in NLP, coupled with the vast availability of digital text data have yielded impressive results in the legal domain. In this paper, we develop an extractive question-answering system that could answer SEBI-related queries from a knowledge base of diverse SEBI documents of varying degree of legal language ranging from regulations to legal cases. The proposed model can output answers for a wide range of queries with an accuracy of 79 percent.

## I. INTRODUCTION

Text Documents in legal domain are constantly increasing in volume and complexity. The information in such documents is usually in an unstructured text format for which a number of systems have been proposed and developed. The information made available by such legal information systems, however, is often accessed with simple, keyword-based search interfaces and presented as a simple list of hits. This makes the process of information retrieval time consuming with poor accuracy, especially when dealing with large volume of information. Moreover, the usefulness of such information varies widely and depends on its structure and its representation and the context in question. Thus, although the information may be available, users and legal professionals may find the information vague and noisy when interested in specific circumstances or investigating a particular case. Such issues have led to a need for improving ways to search and structure large amounts of legal information.

While the task of ranking candidate answers in QA is similar to document-retrieval, there are some key differences. While search datasets also tend to have a high ratio of irrelevant to relevant documents, in the case of QA, there is typically only one correct answer. Hence, it is imperative to develop a model that can accurately rank the potential answer passages for given queries. In this paper, we propose an extractive question answering working model that could potentially overcome the drawbacks of the classical approaches [1] available. In addition, we also propose an enhanced attention-based deep neural model for ranking task that could improve the performance of the QA model. Our framework that facilitates the processing of semantics of

these documents not only improves the efficiency, but could also be used at ease by the end-users.

The novel properties of the proposed deep ranking model and our contributions can be summarized as follows:

- **Dynamic denoising of irrelevant passages with a siamese layer**

We incorporate a pre-processing layer that employs a siamese network that was trained to send in only the passages relevant to the query into the attention network with an accuracy of 0.89. This inturn reduces the number of computations while improving the quality of results.

- **Combining Feature-based engineering with attention layers**

The normalized frequency of query terms are incorporated into the attention layer as an additional information. This boosted the performance of the ranking model by 5 percent.

- **Internal re-ranking procedure**

We introduce a novel mechanism of internal ranking and re-ranking procedure as opposed to the existing single iteration of ranking models that employ an external re-ranking model in addition for better performance. The proposed idea reduces the complexity and effort of training two individual models for the two tasks.

**Roadmap.** The paper is divided into the following section - A literature survey discussing a few existing solutions, Knowledge Base used for the model, our proposed model for an efficient information retrieval, an enhanced neural architecture for ranking answer passages and finally the conclusion derived from this exploration along with the future scope.

## II. LITERATURE REVIEW

Owing to the overwhelming amount of data available on the internet, many methods have been devised to automate various mundane works and analysis in the legal domain, such as retrieving relevant legal documents and law statements. Most of the legal related data are in the text format and hence is an active region of research in the field of Natural Language Processing. There have been various attempts of developing AI systems that could perform various higher level useful legal tasks efficiently, which otherwise would take a lot of legal hours by lawyers and other legal experts.

Owing to large volumes of data, one of the important tasks where an automated system would be of great help is information retrieval. [2] developed a question-answering system for legal documents in Portuguese language. The

system consists of two components - preliminary analysis of documents (information extraction) and query processing (information retrieval), based on computational linguistic theories: syntactical analysis (constraint grammars); followed by semantic analysis using the discourse representation theory; and, finally, a semantic/pragmatic interpretation using ontologies and logical inference.

QA models could also be developed by leveraging various language models, as discussed in [3]. The author constructs a simple Convolutional Neural Network architecture with a Multi-Head Attention mechanism using BERT pre-trained embeddings to represent the asked question dynamically in multiple aspects and achieved enhanced retrieval performance.

A question-answering model focused on Regulatory Documents is proposed by [4] where they employ a two-step approach which first selects relevant paragraphs given a question; and then compares the selected paragraph with user query to predict a span in the paragraph as the answer. The model achieved a precision of 67 percent.

The state-of-the-art research in the ranking task of information retrieval domain involves employment of attention layers. One such model is proposed in [5] where a decoder mechanism is used to learn the ranks of the search results in a listwise fashion, using embedding values from trained convolutional neural networks.

Siamese networks are an active part of the current research for a diverse range of tasks such as sentence similarity. [6] evaluates several variants of Siamese recurrent architectures which are used to measure Semantic Textual Similarity.

[7] presents a comprehensive simulation study examining the performance characteristics of a collection of existing Rank Aggregation methods that are suitable for genomic applications under simulations that mimic practical situations. In addition, the key factors that affect the performance of the different methods are also analysed.

[8] presents three neural architectures built on top of BERT for question generation tasks. The first one is a straightforward BERT employment, which reveals the defects of directly using BERT for text generation. The other two models were hence proposed by restructuring our BERT employment into a sequential manner for taking information from previous decoded results. They successfully advanced the BLEU 4 score of the existing best models.

Existing literature explores various use cases in the legal domain where the tasks could be automated for improved efficiency. In this paper, we propose an enhanced neural network architecture employing attention layers to improve the ranking accuracy in the information retrieval domain, which could be incorporated into the extractive question-answering model we built for the Securities and Exchange Board of India (SEBI) documents.

### III. KNOWLEDGE BASE SEMANTICS

#### A. Documents involved

The Knowledge Base used for the question-answering model is composed of four kinds of documents:

- **Regulatory Documents**  
Composed of SEBI regulations and definitions of certain selective legal terms. There are 48 regulatory documents in total. Each document varies in size, ranging from 20 to around 120 regulations.
- **Informal Guidance**  
Presents a detailed answer given by SEBI for common queries raised by companies. The answers are provided based on the SEBI regulations as facts.
- **Conceptual Papers**  
Provides a detailed discussion on a topic with respect to a regulatory document, ranging from amendment proposal to expanded semantics of the concise regulations for better and uniform understanding.
- **Legal Cases**  
These documents present a detailed analysis of cases between companies and SEBI. Allegations made, facts, company background, decision and penalty are a few subjects covered predominantly in these documents.

#### B. Semantics Extraction

The documents in the Knowledge Base are of unstructured format. The SEBI regulatory documents consists of both regulatory and non-regulatory sentences (for instance, additional explanations, footnotes and schedule clauses). Hence, it is important to identify regulatory sentences and extract them. Analysing the content across documents, it was observed that the words 'shall' and 'may', tagged as rule-relevant words, could be found majorly in sentences that are regulations. On this basis, the regulations were extracted and saved in the Knowledge Base.

The other three documents namely Informal Guidance, Conceptual Papers and Legal cases are divided into individual passages or paragraph points and stored in the Knowledge Base documentwise.

These extracted semantics are then applied in the Question-Answering Model for ranking and answer extraction, as explained in the following section.

### IV. EXTRACTIVE QUESTION ANSWERING MODEL

Question Answering Systems are gaining increasing popularity due to the ability of the model to promptly answer the queries put forth by humans with accurate understanding of the context specified. [2] introduces a question answering system for Portuguese Legal Documents that works based on the logical representation of the language. The proposed extractive question answering model was developed for SEBI domain to capture the context from input query and output relevant answers and also, an expanded context of other answers around the main one via interactive network visualization.

It has four major components as shown in Figure 1.

#### A. Query Pre-processing Module

In this pre-processing module, the input query is conditionally expanded based on the presence of SEBI definition phrases and LDA topic words in the query. In addition,

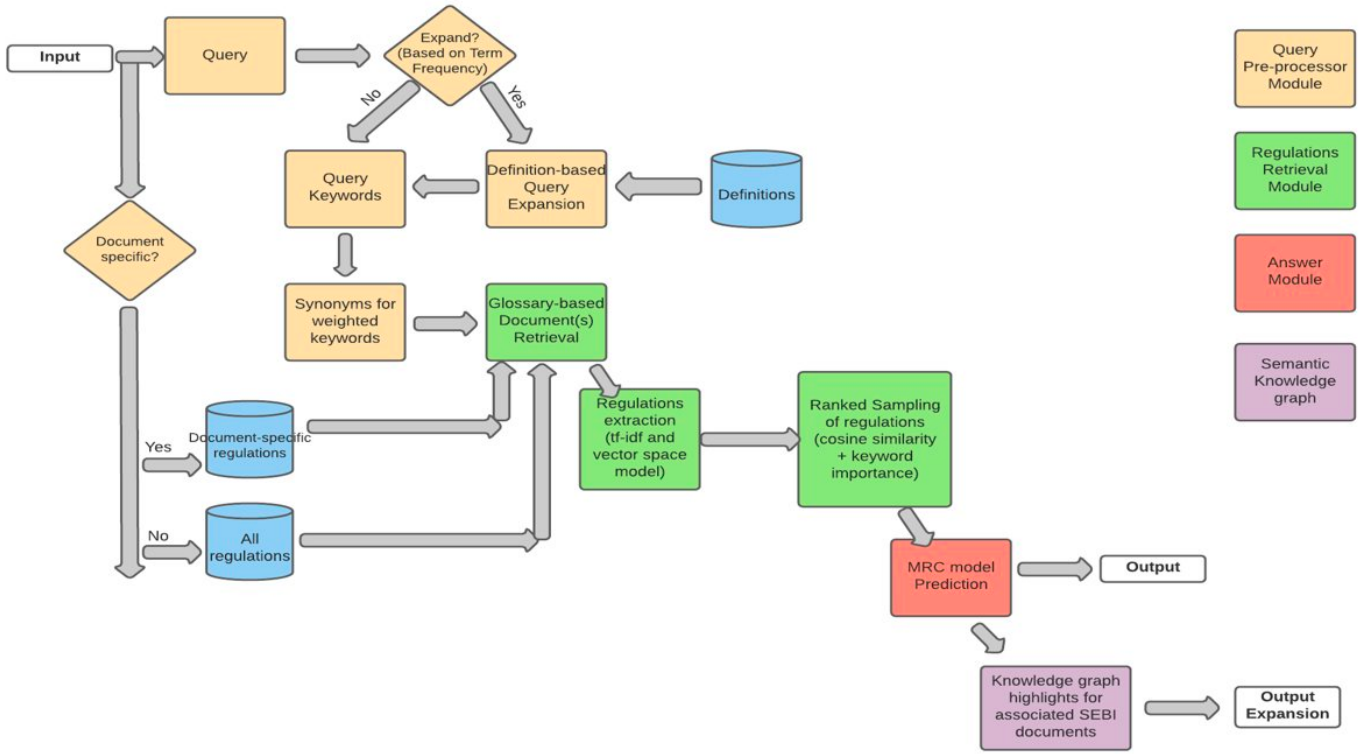


Fig. 1: Query Answering Model - Flow Chart

synonyms and hypernyms (collective words that describe a category) are added from combination of results from a pre-trained legal model embeddings, definitions similarity and WordNet similarity functions. Hypernyms are helpful to generalise specific queries.

### B. Regulations Retrieval Module

Based on the expanded query from the previous module, relevant documents are retrieved based on the glossary of unique terms representing the individual documents built by human expert. The passages from the retrieved documents and the expanded query are embed in the normalized tf-idf based vector model space and the passages are ranked based on the cosine similarity values.

### C. Answer Extraction Module

The top ten ranked passages from the previous module and the query are fed into a pre-trained Machine Reading Comprehension Model built on BERT, called cdQa model. The model extracts the answer span based on maximum of the product of the start and end probabilities of a token pair. The output consists of the answer span and the corresponding passage from which it was extracted.

### D. Knowledge Graph Module

Based on the answer passage extracted, the corresponding node in the SEBI Knowledge graph is highlighted along with the first degree neighbours representing the related answers to the user's query of interest. The user is free to navigate around and also view the original document as required.

TABLE I: Model Outputs

Query	Answer	Rationale
How much amount will be provided as reward for informants in insider trading cases?	ten percent of the monetary sanctions collected or recovered and shall not exceed Rupees One crore	insider trading document identified from the terms 'insider trading' and 'informants'. The term 'how much amount' helped rank the answer.
What legal formalities are to be followed for sweat equity issuance?	Section 79A of Companies Act, 1956 and these Regulations to its – (a) Employees (b) Directors Special Resolution	sweat equity document identified from the term 'sweat equity'. The term 'legal formalities' helped rank the answer.

### E. Sample Outputs

Table I shows a few outputs from the question-answering model. The model answers matched the ground truth from legal expert.

## V. ATTENTION-BASED NEURAL ARCHITECTURE FOR RANKING PASSAGES

The version one of the QnA model described in section IV gave accurate results for a few queries but failed for a few. For instance, given a query about provisions for transfer of shares, instead of the provision given under Section 109 of Companies Act 1956, the model just mentioned 'Companies Act'. On analysis, it was found that the retrieval module performed quite poorly for certain queries which affected the overall performance. Owing to the fact that the tf-idf based vector space model is outdated, several deep learning alternatives were explored. A lot of state-of-the-art models employed attention mechanism, which is currently in research too. Referring to the architecture proposed in [9], we propose an enhanced neural architecture that improved the performance of the original model from the paper for the ranking task.

TABLE II: Model Hyperparameters

Parameter	Value
Optimizer	adam
Activation function	relu
Dropout rate	0.1
Loss function	Triplet loss
Batch size	128
Epochs	100

### A. Proposed Architecture

The proposed architecture consists of a siamese classifier followed by a two-layered attention network. The input query and passages from the database are sent into the siamese model that classifies the passage as relevant or not relevant with respect to the query. The relevant passages are fed into the attention network, as shown in figure 2.

The attention network is classified into two layers - answer-focused and query-focused. These layers work in a manner such that the second layer re-ranks the list passages based on query importance values that was initially ranked based on passage context.

The answer-focused layer forms a matrix of cosine similarity values between the individual query and passage tokens using the embedding values from SEBI-BERT, which was trained on SEBI documents. The answer vector is formed where each value represents the importance value assigned to the individual passage token. This importance value is calculated as the sum of the cosine values of the individual passage token with all query tokens. To this answer vector, a frequency feature vector is concatenated where each value maps to the normalized frequency of the individual query token in the passage. This final concatenated vector is fed into a dense layer which generates a single scalar value. This value is treated as the score assigned to the passage by the answer-focused attention layer. The passages are ranked based on these score values.

The query-focused layer follows a similar approach. Instead of the answer vector, it forms the query vector where each value corresponds to the sum of cosine values of each query token with all the passage tokens. The sum of cosine values provides the importance score for each query term. Higher the sum value, more relevant and important is the query token in context. This vector is concatenated with the score vector from the previous layer and fed into the dense layer. The output from this dense layer serves as the final score value assigned to the passage. The passages are then re-ranked based on these values.

The hyperparameters and loss functions employed by the model is described in table II.

TABLE III: Model Evaluation

Model	Mean Average Precision	Mean Reciprocal Rank
Proposed Model	0.79	0.85
Original Base Model	0.75	0.79

### B. Training Data Generation

The training data comprises of both real and synthetic queries. Around 350 real queries were scraped from SEBI FAQ documents. While around 8500 synthetic queries were generated from SEBI regulations using a pre-trained query generation model built on the t5 language model. Procuring expert-labelled ranked data was a major challenge given the volume of data. Hence, weakly ranked passages were generated for these queries to serve as the ground truth for the evaluation of the proposed model. This was achieved by aggregating the rank results from tf-idf based vector space model and BM25 model via the majority voting algorithm.

### C. Evaluation Metrics

For evaluation, we rank answer sentences with the predicted score of each method and compare the rank list with the ground truth to compute metrics. We choose Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR), which are commonly used in information retrieval and question answering, as the metric to evaluate our model. The definition of MRR is as follows:

$$MRR = \frac{1}{|Q|} * \sum_{q \in Q} \frac{1}{rank(fa)} \quad (1)$$

where rank(fa) is the position of the first correct answer in the rank list for the question q.

Thus MRR is only based on the rank of the first correct answer. It is more suitable for the cases where the rank of the first correct answer is emphasized or each question only have one correct answer. On the other hand, MAP looks at the ranks of all correct answers with respect to the ground truth rank labels given. It is computed as following:

$$MAP = \frac{1}{|Q|} * \sum_{q \in Q} AP(q) \quad (2)$$

where AP(q) is the average precision for each query.

Thus MAP is the average performance on all correct answers.

### D. Model Learning Results

The **Siamese network** trained on the generated queries achieved an **accuracy of 0.892, Validation Accuracy of 0.893 and Validation loss of 0.25.**

Table III provides the comparative evaluation metric values obtained by training and testing the proposed and original ranking neural architectures on the SEBI training data generated. As can be inferred, the proposed model outperformed the reference base model in both the evaluation metrics.

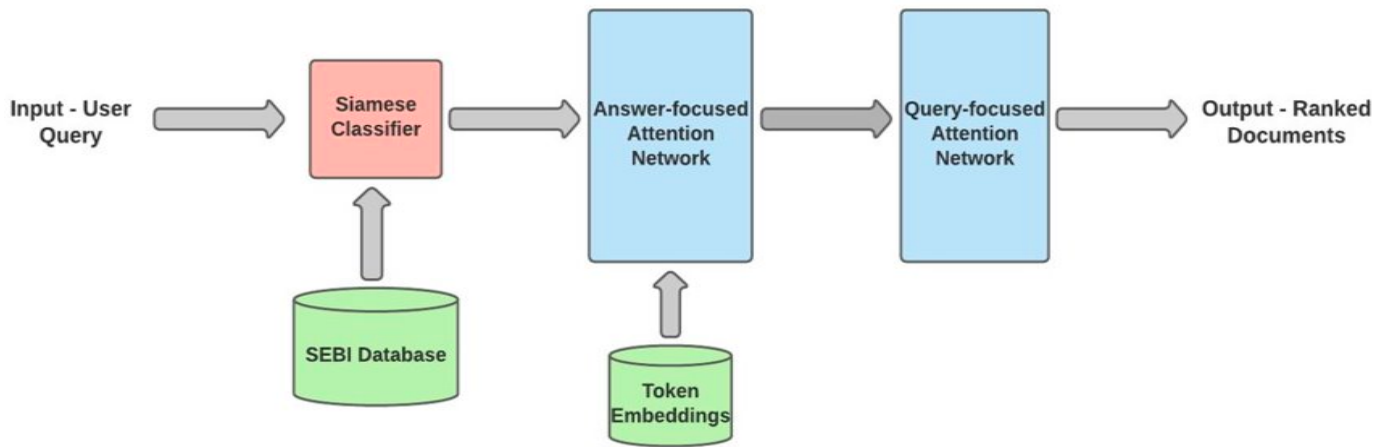


Fig. 2: Proposed Ranking Network Architecture

## VI. CONCLUSION

In this paper, we proposed an enhanced neural network architecture that employs attention mechanism with additional feature engineering for an efficient ranking which performed marginally better than the baseline architecture. The proposed model achieved Mean Average Precision of 0.79 and Mean Reciprocal Rank of 0.85. This could replace the existing retrieval module in the proposed question-answering system. For future work, we will improve the neural net performance for answer ranking by exploring deeper and extend it to an open-domain answering task. We will also expand the knowledge base to provide answers to a given query from the perspectives of different countries.

## REFERENCES

- [1] S. Jabri, A. Dahbi, T. Gadi, and A. Bassir, "Ranking of text documents using tf-idf weighting and association rules mining," in *2018 4th International Conference on Optimization and Applications (ICOA)*, 2018, pp. 1–6.
- [2] P. Quaresma and I. Rodrigues, "A question answer system for legal information retrieval," 01 2005, pp. 91–100.
- [3] J. Sai Sharath and R. Banafsheh, "Question answering over knowledge base using language model embeddings," *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul 2020. [Online]. Available: <http://dx.doi.org/10.1109/IJCNN48605.2020.9206698>
- [4] D. Collarana, T. Heuss, J. Lehmann, I. Lytra, G. Maheshwari, R. Nedelchev, and P. Trivedi, "A question answering system on regulatory documents," 12 2018.
- [5] B. Wang and D. Klabjan, "An attention-based deep net for learning to rank," 2017.
- [6] T. Ranasinghe, C. Orasan, and R. Mitkov, "Semantic textual similarity with Siamese neural networks," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulgaria: INCOMA Ltd., Sep. 2019, pp. 1004–1011. [Online]. Available: <https://www.aclweb.org/anthology/R19-1116>
- [7] X. Li, X. Wang, and G. Xiao, "A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications," *Briefings in Bioinformatics*, vol. 20, no. 1, pp. 178–189, 08 2017. [Online]. Available: <https://doi.org/10.1093/bib/bbx101>
- [8] Y.-H. Chan and Y.-C. Fan, "A recurrent BERT-based model for question generation," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 154–162. [Online]. Available: <https://www.aclweb.org/anthology/D19-5821>
- [9] L. Yang, Q. Ai, J. Guo, and W. B. Croft, "anmm: Ranking short answer texts with attention-based neural matching model," 2019.