# A Machine Learning Approach used for Hindi Question Answering System

*** Pooja Kumari**
**** Prof. Rakesh Shivhare**

## ABSTRACT

When a user types in a question in natural language, a Question Answering (QA) system—which is really an Information Retrieval (IR) system—finds the most relevant or near-matching results. This is a result of NLIDB, or the Natural Language Interface to Database. The study delves into the implementation of a Machine Learning-based Hindi Language QA system. There are three stages to the QA system that has been put into place: The first step is to access the natural language query, which involves reading, preprocessing, and tokenizing the input query. Then comes the feature extraction phase, which involves identifying specific feature vectors from the results of the previous phase. Lastly comes the classification phase, which involves using the Naïve Baye's classifier and the system's stored knowledge base. To define the overall accuracy of discovering the relevant answers of the user's specific inquiries, this study indicates that the ideas of categorization and similarity give better results than the usage of the 'equals' concept.

**Keywords:-** Question Answering System; Natural Language; Information Retrieval; Classification; Machine Learning**.**

*Pooja Kumari, Research Scholar, Department of Computer Science & Engineering, Radharaman Engineering College, Bhopal , India, poojarec6@gmail.com
**Rakesh Shivhare, Professor, Department of Computer Science & Engineering, Radharaman Engineering College, Bhopal, India

## I. INTRODUCTION

In light of the rapid increase in the quantity of readily accessible knowledge resources or papers that can be found through search engines, we have reached a point in time where an efficient quality assurance system will become an indispensable component of our day-to-day lives in order to gain access to information. A Knowledge-based Quality Assurance (QA) system is a specialised kind of information retrieval activity in which information is specified by information needs that are relatively articulated as questions or queries in natural language. There are many other ways that people interact with computers, but this is one of the most common. The quality assurance system gives the user the ability to access information resources in a natural way by communicating their questions in natural language and receiving a response that is both brief and pertinent as a consequence. Artificial intelligence, natural language processing, information retrieval, information extraction, and machine learning are all combined in order to create a more effective quality assurance system. The definition of a natural language statement includes both a question or query and the answers or output that it produces.

The concept of classification is well-known and can be defined as the process of selecting the appropriate "Class Label" from the input (Query) that is provided. Generally speaking, in classification tasks, a set of labels is clarified in advance, and each input is evaluated separately from all of the other inputs that are provided. The QA system considers the questions that are posed by the user in NL, searches for the correct responses from a collection of catalogues, and then provides the user with the answer that is both perfect and quick. As can be seen in Figure 1, a major provocation is noticed in knowledge-based quality assurance systems. This provocation is the creation of a magnificent knowledge base that contains empirical and correct information and domains.

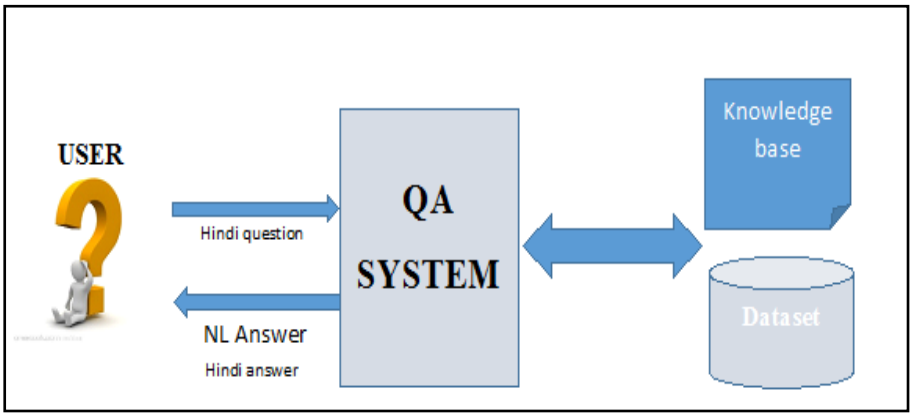


Figure 1: Question Answering (QA) System

## II. LITERATURE REVIEW

In an effort to increase performance while decreasing development costs, Question Answering Systems use a wide range of tactics. To acquire precise responses, people use a wide range of natural language queries. Quality assurance systems have been the focus of researchers for a long time, and they come from various linguistic backgrounds.

Research by Gupta and Khade [1] on BERT-based bilingual machine comprehension in Hindi and English shows that transformer-based models can capture semantic similarities across languages and make cross-lingual knowledge transfer easier.

Their results demonstrated that pre-trained embeddings could comprehend closely related languages, which led Artetxe et al. [2] to study the feasibility of translating monolingual representations to other languages.

Conneau et al. [3] suggested unsupervised cross-lingual representation learning at scale to extract embeddings independent of language from big multilingual datasets. Their research shed light on

the difficulties and potential benefits of cross-lingual knowledge transfer, paving the way for unsupervised methods to cross-lingual natural language processing tasks.

The multilingual representation model MuRIL was presented by Hanuja et al. [4] and is specifically designed for Indian languages. To achieve linguistic inclusion in natural language processing research, their study focused on places with a high level of linguistic diversity and the need for language-specific representations and resources.

Modern healthcare organisations are striving to employ cutting-eorganizationsy like fog computing and the Internet of Things (IoT) to improve patient outcomes, healthcare delivery, and disease prediction. This integration is already attracting a lot of attention. Recent developments in healthcare-related areas such analytics, fog computing, predictive analytics, the Internet of Things (IoT), machine reading comprehension (MRC) and question-answering (QA) systems are included in this literature review.

To anticipate cases of COVID-19 early on, Singh and Kaur [5] created a method based on fog computing. Their groundbreaking work demonstrated the potential of fog computing to assess healthcare data in real-time, paving the way for quicker response times and better pandemic control.

The Internet of Things (IoT) has the potential to enhance healthcare service delivery, medication adherence, and patient monitoring, according to research by Singh et al. [6]. Internet of Things (IoT) devices have the ability to improve healthcare accessibility and efficiency, according to their research.

To enhance insurance risk assessment and pricing approaches, Singh et al. [7] presented an ensemble-based strategy to predict health insurance rates. Healthcare decision-makers can benefit from data-driven insights based on their results.

The software-based architecture proposed by Singh and Kaur [8] for the construction of smart healthcare systems using fog computing offers a complete platform for the management, analysis, and support of choices pertaining to healthcare data. By streamlining integration with existing systems and infrastructure, this method promotes the development of fog computing applications in healthcare.

Natural language processing, medical record retrieval, and quality assurance system improvement have all been the subject of recent research. Ramesh et al. [8] introduced Samanantar, the most

publicly accessible collection of parallel corpora for eleven Indic languages. The ease of this has greatly facilitated studies of multilingual NLP and cross-lingual understanding.

Due to developments in deep learning and natural language processing, there has been a surge in research on MRC and QA systems in the past few years. This research review covers a wide range of subjects, including advancements in QA dataset extension, semantic matching in MRC, the current status of answering complicated questions in the open domain, and the development of QA systems based on deep learning.

To better understand how different semantic matching strategies could improve comprehension and response retrieval accuracy, Liu et al. [9] performed an empirical study on MRC. By filling in gaps in our understanding of semantic modelling approaches and how they impact MRC performance, their work lays the groundwork for better QA systems.

Among the many datasets available for training and testing QA models, Rogers et al.'s [10] exhaustive taxonomy of NLP tools for question answering and reading comprehension stands out. Researchers and practitioners can enhance decision-making by comparing their taxonomy to quality assurance databases.

By examining current methods in open-domain complicated question answering, Etezadi and Shamsfard [11] shed light on the difficulties, approaches, and outcomes of handling questions of varying levels of complexity. Their survey covers a lot of ground and covers a lot of different approaches used in open-domain QA systems.

In their study, Abdel-Nabi et al. [12] followed the development of deep learning-based question-answering systems and how they were used for quality assurance tasks. They provide insightful commentary on current practices and potential future improvements by compiling important research results and trends in QA systems based on deep learning.

When taken as a whole, these studies and surveys enhance our knowledge of MRC semantic matching, the current situation of QA datasets, the difficulties and potential solutions associated with answering complicated questions in the open domain, and the use of deep learning in QA systems. By assisting researchers and industry professionals in enhancing MRC and QA, they promote growth and development in NLP.

The increasing demand for effective access to legal information and knowledge has piqued the interest of many in legal question-answering systems. Martinez-Gil [13] uncovered a wealth of

information about the many techniques, methodologies, and difficulties related to legal question-answering systems after delving deeply into the subject. The study emphasises the importance of deep learning andemphasizesearning approaches in tackling challenges connected to answering legal questions in its discussion of present and future trends in the field.

One way that communities can work together and share what they know is through community question-answering (CQA) platforms. Roy et al. [14] recently evaluated CQA difficulties with an emphasis on deep learning and machine learning techniques. Their research sheds light on the difficulties and potential benefits of CQA systems in areas such as community involvement, evaluation of answer quality, and question comprehension.

In order to train robots to understand clinical practice recommendations, Mahbub et al. [15] created the CPGQA benchmark dataset. Their research aids continuing efforts to address the requirement for healthcare-specific datasets in order to apply transfer learning to CQA problems connected to clinical practice.

By merging a statistical scoring method with a vectorization methodology, Manjunath et al. [1vectorizationan intelligent system that can answer queries. By enhancing question comprehension and answer retrieval, their strategy improves CQA systems overall through the application of cutting-edge vectorization techniques and statistical modelvectorizationress the needs of Persian-speaking communities, Darvishi et al. [17] proposed PQuAD, a dataset for Persian question-answering. Through their contributions, the existing dataset becomes more diverse, which in turn aids studies of Persian language processing and the development of CQA systems tailored to Persian speakers.

Using a deep learning method, Qiu et al. [18] created a system that can generate mineral exploration ontologies and utilise them to answer questions. Their study utilizefies the use of deep learning techniques in specialised domains by improving mineral explospecializedy response using domain knowledge.

Abedissa, Usbeck, and Assabie [19] released AMQA, an Amharic question-answering dataset, to fill the need for resources in underrepresented languages. The wealth of data provided by their research enhances natural language processing, which in turn allows for the development of Amharic-specific question-answering systems.

A financial named entity recognition system called FinBERT-MRC was introduced by Zhang and Zhang [20]. It uses BERT within the machine reading comprehension paradigm. Their main goal has been to enhance algorithms that extract financial entities from text. When it comes to financial data extraction, their cutting-edge deep learning techniques are light years ahead of the competition.

A memory-aware attentive control system for responding community inquiries was presented by Wu et al. [21] using knowledge-based dual refinement approaches. They hope to improve community question-answering systems by leveraging domain-specific knowledge and attention processes to provide more accurate and context-relevant answers.

To analyse needs expressed in natural language, Eanalyzet al. [22] presented an AI-assisted question-answering method. Their goal is to make requirement analysis in software engineering more efficient and accurate by using AI approaches to understand needs expressed in plain language.

A technique for query response about semi-structured heterogeneous genealogical knowledge trees using deep neural networks was proposed by Suissa et al. [23]. Through the application of deep learning techniques, they tackle the difficulties of deducing answers from complex knowledge networks, leading to enhanced genealogy data retrieval and comprehension.

## III. ARCHITECTURE AND PROPOSED WORK

As can be seen in Figure 2, the architecture of the Hindi Quality Assurance System is composed of three phases:

A. Accessing NL Query phase.
B. Feature Extraction Phase
C. Classification
D. Knowledge Base

### A. Accessing NL Query Phase

In the beginning, the user sends a query in Hindi to the QA system. In this case, Hindi is employed as the language of inquiry. The Query that was asked is subsequently passed on to the initial step of the system, which is known as the Accessing NL query step. The query from the NL is read, accessible, and preprocessed in this section. After passing through the preprocessing

stage, the tokenization procedure is carried out, which ultimately results in the elimination of stop words and the creation of tokens.

Example: भारत की राजधानी क्या है

Tokens: भारत राजधानी क्या

**B. Feature Extraction Phase**

The level at which entity prediction is located is this one. Following the reading of the question that was provided by the user as input, the preprocessed query is fed, which ultimately outcomes in tokens. The tokens that were generated in the previous stage are transferred to the subsequent phase, which then divides into two distinct subphases.

1) Entity detection: The use of similarity measures is one method that can be utilised for entity discovery. When searching for entities, one can make use of a variety of similarity functions. In this case, the Smith-Waterman algorithm is utilised [5].Taking into consideration the entire sequence, this method analyses and evaluates segments of every conceivable length, and then optimises the similarity measure.

2) Feature Vectors: If the data that is sent to an algorithm or any system that refers to NL systems is enormous and it is determined that it is not vital, then it is necessary to transform it into a limited set of features, which is simply referred to as a features vector. For the expression "term frequency," which is written as tf(t,d), the most straightforward option is to take the raw frequency of a term in a document. This refers to the number of times that the term t appears in the document, where t is the term and d is the document.

**C. Classification Phase**

The feature vector that was generated as a consequence of the previous step is then delivered to the subsequent phase of classification, which is where the task of determining the appropriate label for the class based on the query that was provided is completed. In this quality assurance system, the Naive Bayes Classifier is included. For it to be trained, it must be provided with some data. Following the completion of the training, some testing of the data is required, and lastly, an evaluation of the performance is taken into consideration. Should the classifier be based on training, it is imperative that it incorporates the appropriate label for every input. Some notations for probabilities are utilised in the process of evaluating the performance of the NB classifier.

These notations include p(c), which represents the prior probability of class c, and p'(c), which represents the posterior probability of class c that is returned by the particular classifier.

**D. Knowledge Base**

What the Knowledge Base of the established Quality Assurance System is made up of:

• SWD, which stands for "Stop Words Database," is a database of stop words that has been kept in the past.

• Entities: In a similar fashion, the entities dataset is made up of entities, which are the words that are already known and the outcomes of the Entity Detection process.

• Trained Data: An example of the dataset that contains trained data is the data that has been trained manually and is stored in this location.

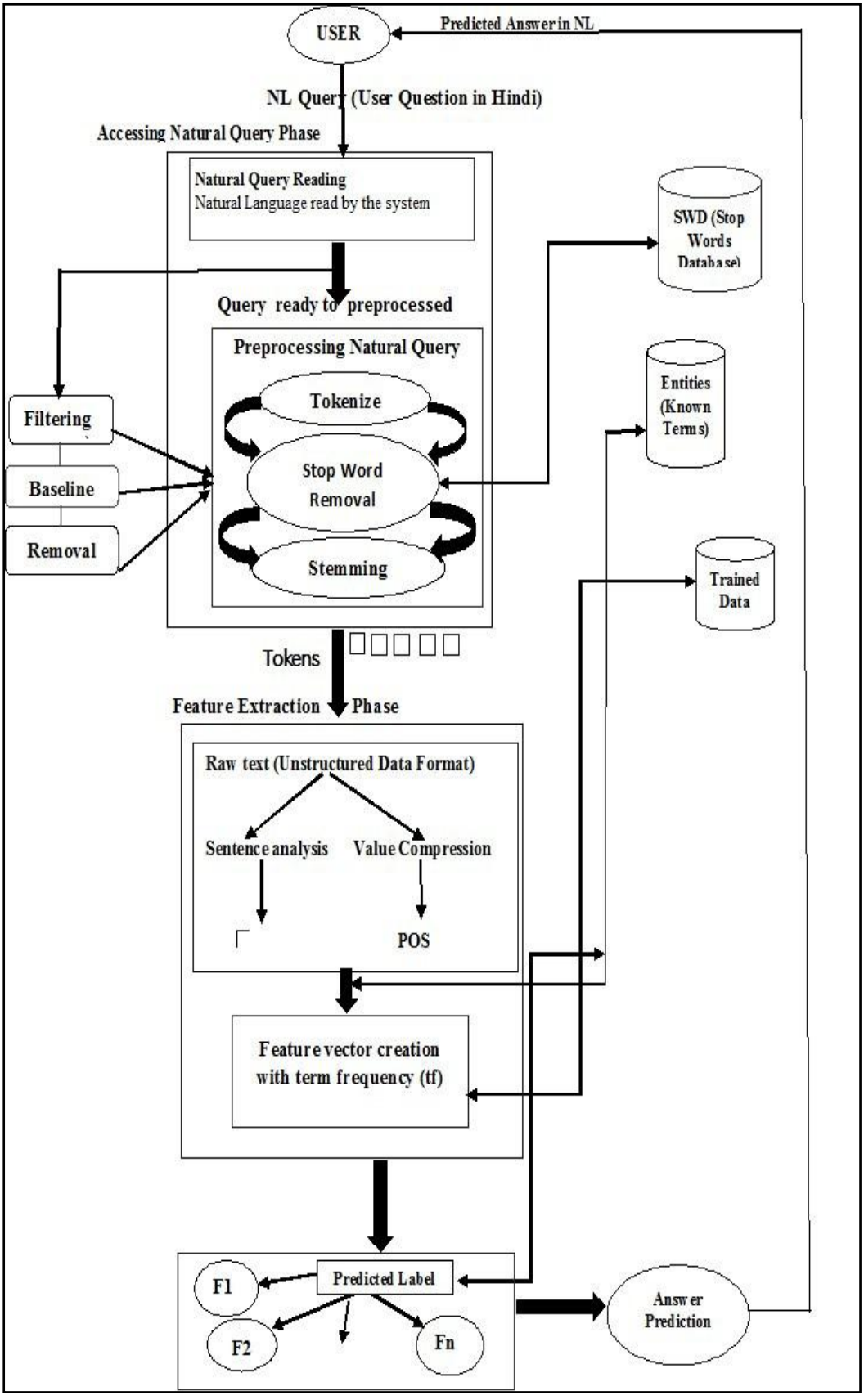


Figure 2: Architecture Module of QA System

## IV. EXPERIMENTS AND RESULTS

It does not matter which platform you use to implement the QA System; it may be done on either Windows or Linux computer systems. The QA system that was built features a graphical user interface that is easy to use. In the experiment of the system, the Naïve Bayes Classifier was utilized as the algorithm for classification. Several different similarity functions are utilized in the testing process, which is carried out on the text-based dataset. It is shown here in the article that the result for the same is shown, and the article concludes with a similarity graph.

### A. Evaluation metrics

To ensure a thorough performance review, evaluation metrics are essential. We fed the QAS a series of user queries and then analysed the answers. Precision (P), Recall (R), and F-measure (F) were some of the metrics used to assess QAS.

To measure the accuracy of the predictions made by the Question Answering System, a unique scoring system known as a Confusion Matrix was employed. Based on the actual label and the predicted label provided by a prediction algorithm, a confusion matrix assigned predictions to different categories.

The normalized nTF-nIDF vectorization approach is used for feature extractionis given in Eqs. (1.1 and 1.2)..

$$nTF_{t,d} = \frac{\text{freq}_{t,d}}{\max freq_{t,d}} \qquad (1.1)$$

$$IDF_t = 1 + \log \frac{P+1}{df_t + 1} \qquad (1.2)$$

A highly accurate system can accurately forecast True-Positives, True-Negatives, or both, since accuracy gives equal weight to the two.

Preciseness quantifies the accuracy of predictions. Calculates the proportion of forecasts that come true.

Positive recall Future expectations.

Accuracy is given in Eq. (1.3) is often used to measure the fraction of correct predictions among all predictions

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (1.3)$$

The main screen is the first shot of the QA system. It displays a dialogue box for the user to enter a query, which includes an answer button. When the user clicks on the answer button, they will receive the response to the particular query they entered.

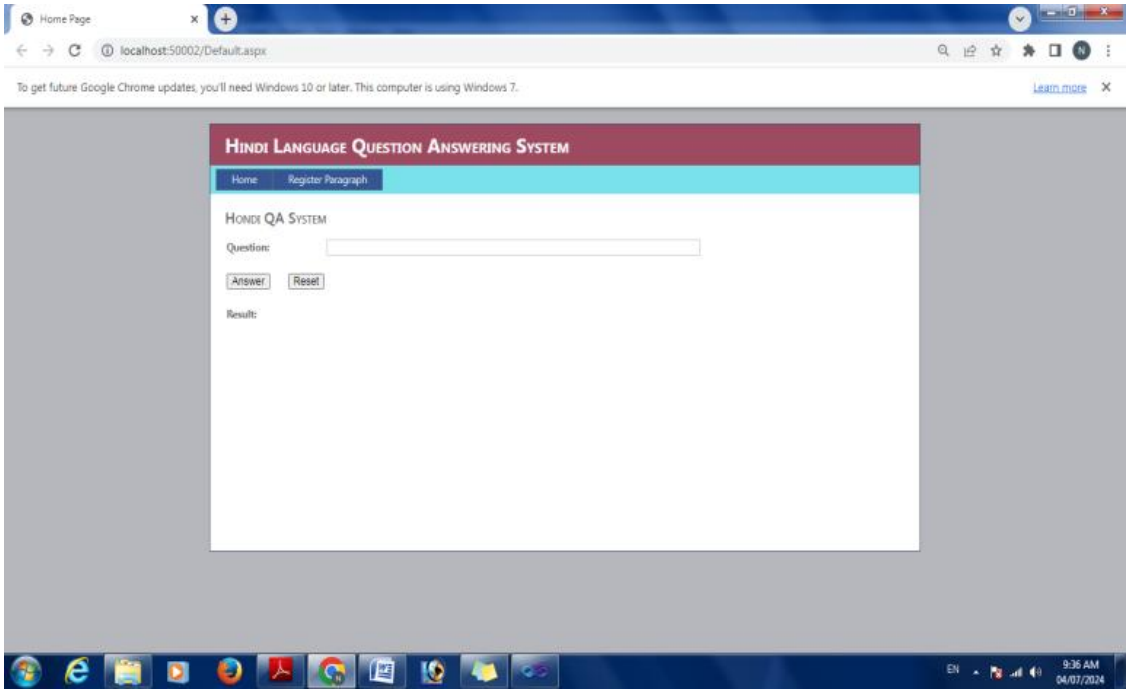


Figure 3: Main Screen

This system is supporting the function of auto complete also, in which the earlier searched questions will be displayed while the user is typing the query.

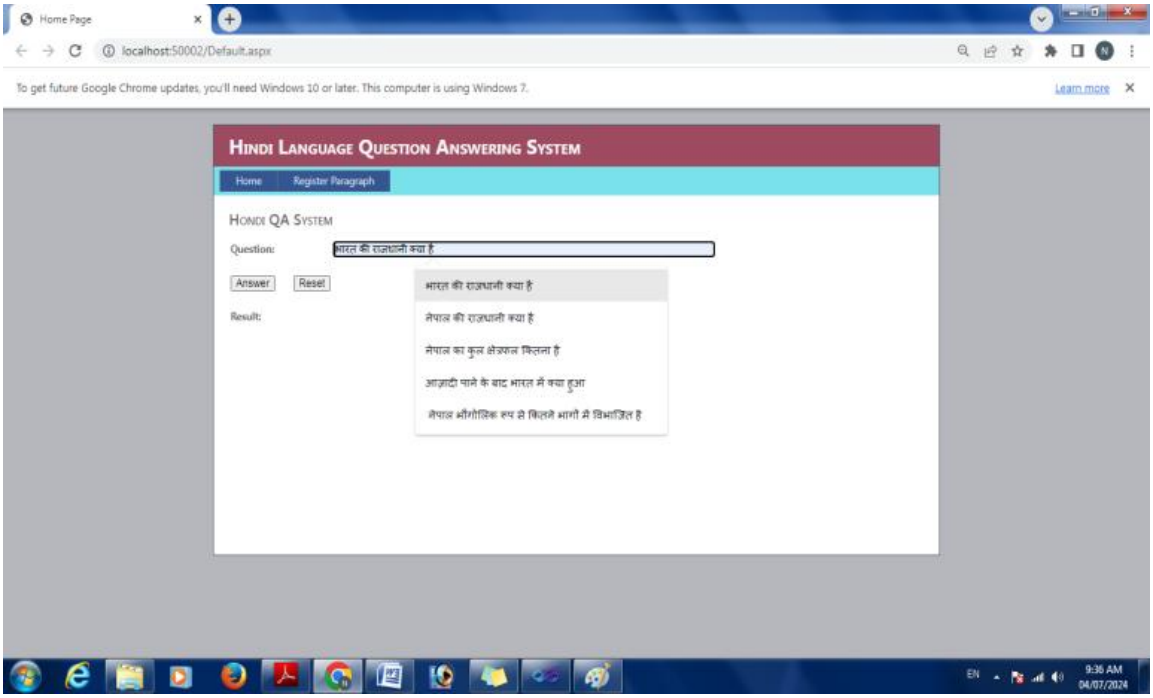


Figure 4: Auto Complete

When the question is given to the system, it will be represented.

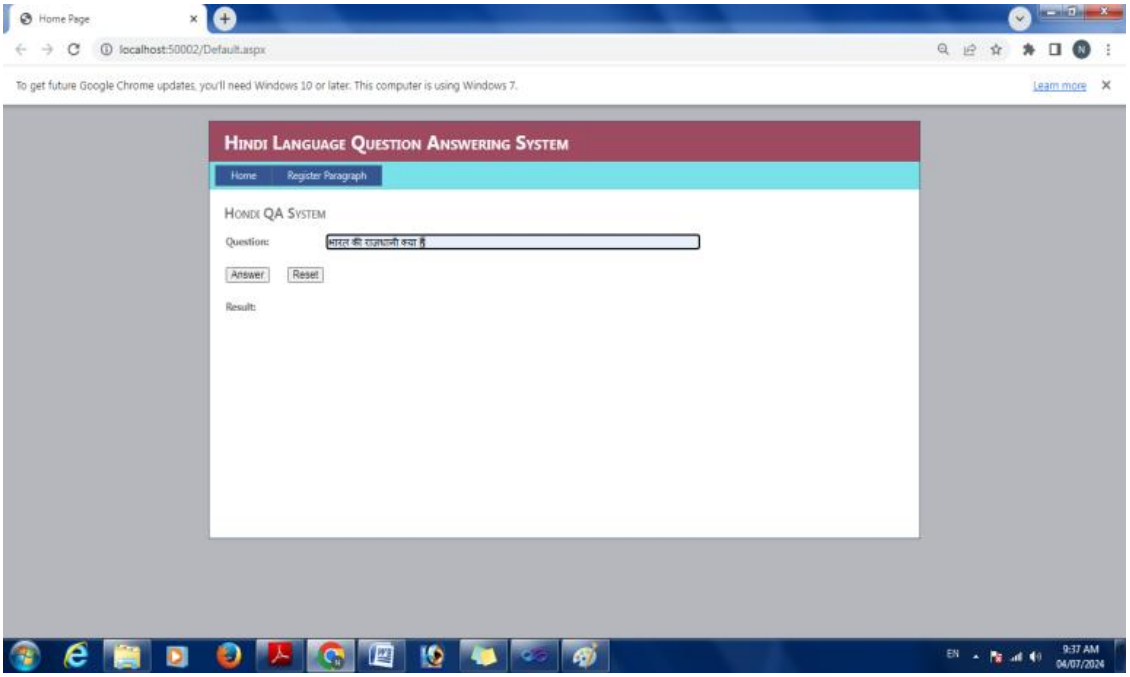


Figure 5: Question Input

Now, finally when question is given, answer button is clicked and the answer to that specific query is displayed.

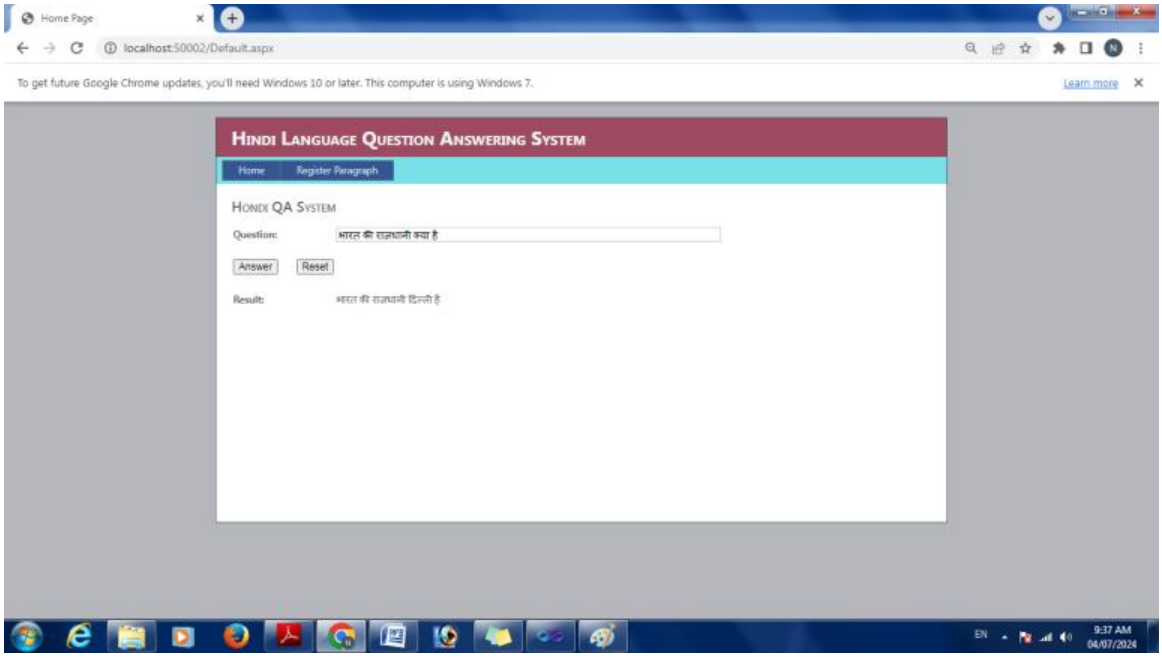


Figure 6: Result Screen

The system show that paragraph input for token generation.

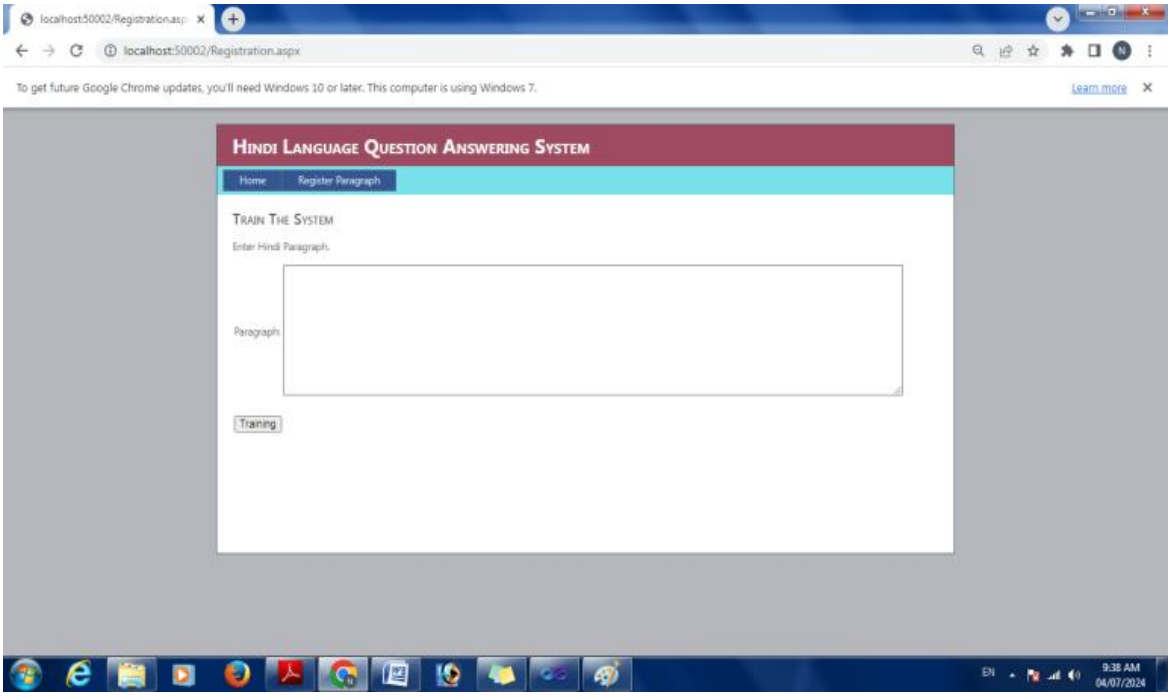


Figure 7: Paragraph Input Screen

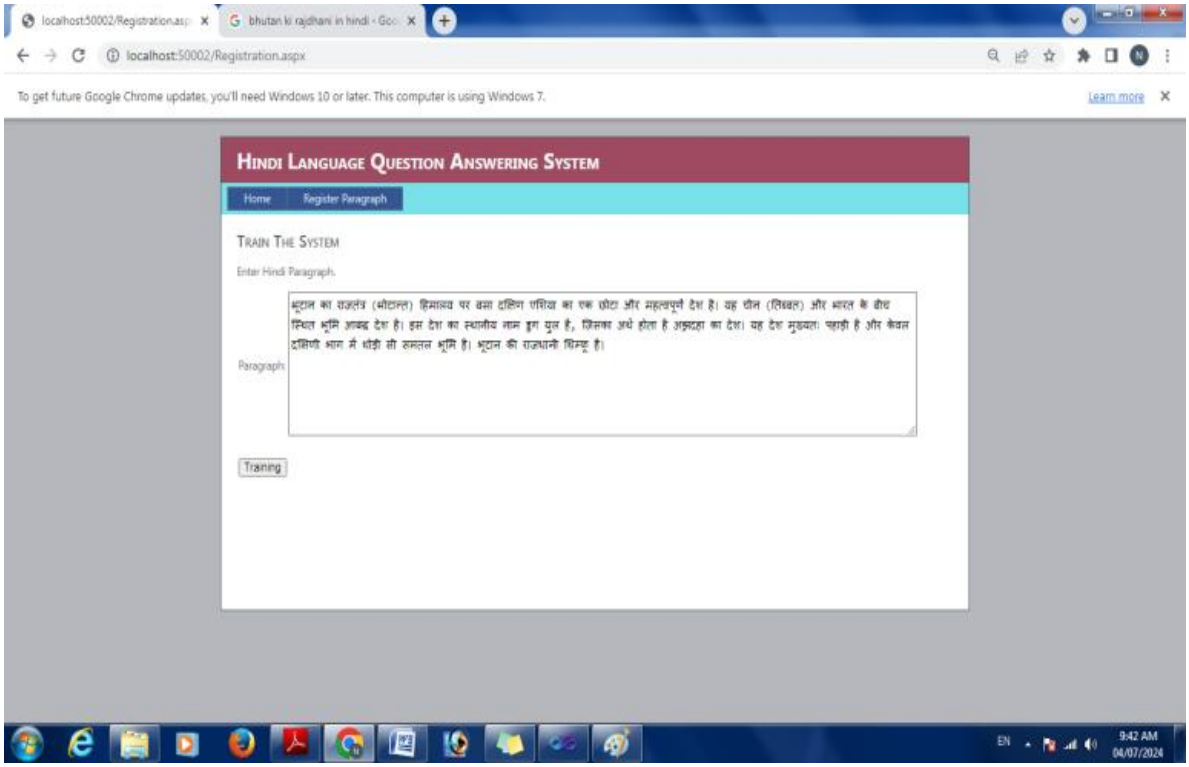


Figure 8: Mining Screen

Following is a Table showing the overall accuracy when a complete test set is given as input to the QA system.

Table 1: Overall Accuracy Table

| Test Set | Total | Correct | Overall Accuracy % |
|---|---|---|---|
| TS1 | 25 | 23 | 92 |
| TS2 | 50 | 44 | 88 |

The test sets TS1 and TS2 represent two distinct test sets. TS1 contains questions that are posed by a user who is already familiar with the domain, whereas TS2 contains questions that are posed by a user who is not familiar with the domain. Therefore, the accuracy percentage is a reflection of which of the two test sets is performing better than the other.

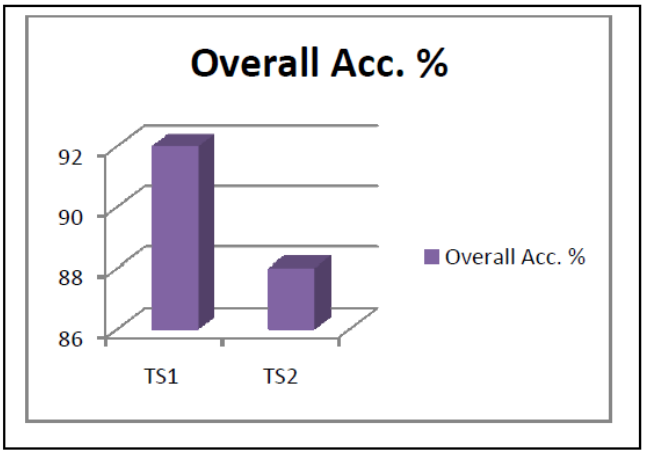


Figure 9: Graph of Overall Accuracy

The overall accuracy graph can be found above, and it represents the proportion of the data that was entered into the system that was accurate. This graph compares the data to the table that is displayed.

## V. CONCLUSION

With an overall accuracy of 0.9 and a threshold of 0.9, the Question Answering System for Hindi Natural Language provides a comprehensive understanding of the QA System. The ideas of overall accuracy and similarity are utilised in this context, which provides a user with a tremendous platform on which they may ask a question using natural language and receive the answer in the same language. This is a significant improvement over the notions that were made use of in earlier systems. With regard to the work that will be done in the future, this system may be implemented in a variety of languages, and it may employ a variety of classification strategies in conjunction with a variety of datasets.

## REFERENCES

1. Artetxe, Mikel, Sebastian Ruder, and Dani Yogatama. (2020) "On the Cross-lingual Transfer ability of Monolingual Representations". Proceed-ings of the 58th Annual Meeting of the Association for Computational Linguistics: 4623–4637.
2. Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzm´an, Edouard Grave, Myle Ott, Luke Zettlemoyer and Veselin Stoyanov. (2020) "Unsupervised Cross-lingual Representation Learning at Scale". Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: 8440–8451.
3. Hanuja, Simran, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. (2021) "MuRIL: Multilingual Representations for Indian Languages". CoRRabs/2103.10730.
4. P. Singh, and R. Kaur, "Implementation of the QoS framework using fog computing to predict COVID-19 disease at early stage", World Journal of Engineering, vol. 12, pp. 345-356, 2021.
5. P. D. Singh, G. Dhiman, and R. Sharma, "Internet of things for sustaining a smart and secure healthcare system", Sustainable Computing: Informatics And Systems, vol. 33, pp. 622-634, 2022.
6. P. Singh, K. D. Singh, V. Tripathi, and V. Chaudhari, "Use of ensemble based approach to predict health insurance premium at early stage", International Conference on Computational Intelligence and Sustainable Engineering Solutions, pp. 566-569, IEEE, 2022.
7. P. Singh, and R. Kaur, "A software-based framework for the development of smart healthcare systems using fog computing", IET Software, vol. 34, pp. 145-159, 2022.
8. Ramesh, Gowtham, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, MahalakshmiJ, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. (2022) "Samanantar: The Largest Publicly Available Parallel Corpora Collec-tion for 11 Indic Languages". Transactions of the Association for Computational Linguistics 10:145–162.
9. Liu, Qian, Rui Mao, Xiubo Geng, and Erik Cambria. "Semantic matching in machine reading comprehension: An empirical study". Information Processing & Management 60, no. 2 (2023): 103145.
10. Rogers, Anna, Matt Gardner, and Isabelle Augenstein. "Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension". ACM Computing Surveys 55, no. 10 (2023): 1-45.
11. Etezadi, Romina, and Mehrnoush Shamsfard. "The state of the art in open domain complex question answering: a survey". Applied Intelligence 53, no. 4 (2023): 4124-4144.
12. Abdel-Nabi, Heba, Arafat Awajan, and Mostafa Z. Ali. "Deep learning-based question answering: a survey". Knowledge and Information Systems 65, no. 4 (2023): 1399-1485.
13. Martinez-Gil, Jorge. "A survey on legal question–answering systems". *Computer Science Review* 48 (2023): 100552.
14. Roy, Pradeep Kumar, Sunil Saumya, Jyoti Prakash Singh, Snehasish Banerjee, and Adnan Gutub. "Analysis of community question-answering issues via machine learning and deep learning: State-of-the-art review". CAAI Transactions on Intelligence Technology 8, no. 1 (2023): 95-117.
15. Mahbub, Maria, Edmon Begoli, Susana Martins, Alina Peluso, Suzanne Tamang, and Gregory Peterson. "cpgqa: A benchmark dataset for machine reading comprehension tasks on clinical

practice guidelines and a case study using transfer learning". IEEE Access 11 (2023): 3691-3705.
16. Manjunath, T. N., Deepa Yogish, S. Mahalakshmi, and H. K. Yogish. "Smart question answering system using vectorization approach and statistical scoring method." Materials Today: Proceedings 80 (2023): 3719-3725.
17. Darvishi, Kasra, Newsha Shahbodaghkhan, Zahra Abbasiantaeb, and Saeedeh Momtazi. "PQuAD: A Persian question answering dataset". Computer Speech & Language 80 (2023): 101486.
18. Qiu, Qinjun, Miao Tian, Kai Ma, Yong Jian Tan, Liufeng Tao, and Zhong Xie. "A question answering system based on mineral exploration ontology generation: A deep learning methodology". Ore Geology Reviews (2023): 105294.
19. Abedissa, Tilahun, Ricardo Usbeck, and Yaregal Assabie. "Amqa: Amharic question answering dataset". arXiv preprint arXiv:2303.03290 (2023).
20. Zhang, Yuzhe, and Hong Zhang. "FinBERT–MRC: Financial Named Entity Recognition Using BERT Under the Machine Reading Comprehension Paradigm". Neural Processing Letters (2023): 1-21.
21. Wu, Jinmeng, Tingting Mu, Jeyan Thiyagalingam, and John Y. Goulermas. "Memory-Aware Attentive Control for Community Question Answering With Knowledge-Based Dual Refinement". IEEE Transactions on Systems, Man, and Cybernetics: Systems (2023).
22. Ezzini, Saad, Sallam Abualhaija, Chetan Arora, and Mehrdad Sabetzadeh. "Ai-based question answering assistance for analyzing natural-language requirements". arXiv preprint arXiv:2302.04793 (2023).
23. Suissa, Omri, Maayan Zhitomirsky-Geffet, and Avshalom Elmalech. "Question answering with deep neural networks for semi-structured heterogeneous genealogical knowledge graphs". Semantic Web 14, no. 2 (2023): 209-237.