Project Report

Type of NLP Problem: Question-Answering

By

Aishwarya Paruchuri : 015017639

Shubhangi Manoj Kumar : 015932826

Priyank Thakkar : 015944318

Vineet Kiragi : 015924701

Guided by

Jorjeta Jetcheva

1. INTRODUCTION

Both academics and industry have long been interested in question-and-answer (QA) systems, which are rooted in the development of artificial intelligence, specifically the well-known Turing test. Over the past few years, developments in the field of natural language processing have continuously improved technologies related to QA (NLP). Due to limited computing resources, early research on QA mainly depended on manually designed syntactic rules to answer straightforward questions. Recent years have seen significant progress in the QA sector due to the availability of multiple large-scale benchmark datasets and the rapid development of deep-learning algorithms.

2. PROBLEM STATEMENT

Natural language Processing focuses on three variants of question-answering (QA) based on the inputs and outputs:

- **Extractive QA :** The model **extracts** the answer from a context. The context could be a text that has been supplied, a table, or even HTML. BERT-like models are typically used to resolve this.
- **Open Generative QA :** The process of creating new text is known as a generation. For instance, these models can paraphrase or complete missing text.
- **Closed Generative QA :** In this case, no context is provided. The answer is completely generated by a model

In our case, we are going with option one, extractive QA. According to the internet, multiple-choice test items are less prone to guessing than true or false questions, giving them a more reliable method of assessment. One-word replies are frequently used in the curriculum, from elementary school through universities like SJSU. Now, if we can automate the process of responding to the same questions from a paragraph randomly chosen from a textbook, internet, or research paper, it will drastically minimize the amount of effort. Moreover, we can apply this to some company's FAQ section where customers or visitors of the website can ask custom questions, once we feed the informational context to the model. And we can also do the same thing for some products' reviews. Additionally, it automates extracting information from unknown text; this seems quite intriguing to our team. Hence, we decided to work on Extractive QA NLP models. For which, we tried two different ways: First we tried hugging face fine tuning models on our own custom dataset. Second way we tried was Haystack Farmreader API, we tried a total of 6 models. We will go over the models and draw comparisons among them under the models utilized section below.

3. DATA DESCRIPTION

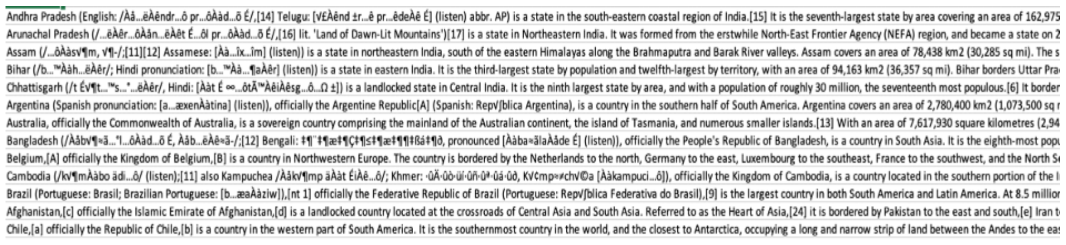
Stanford **Q**uestion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowd workers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. This dataset is available on the internet for Extractive QA and they follow a specific annotation design. We choose to create a dataset using documents fetched from the internet about various nations and states while using the same annotation style as SQuAD.

Our dataset comprises distinct questions posed by each team member on a collection of around

200 articles that are relevant to various nations and/or states, with each question's response being a segment of the article.

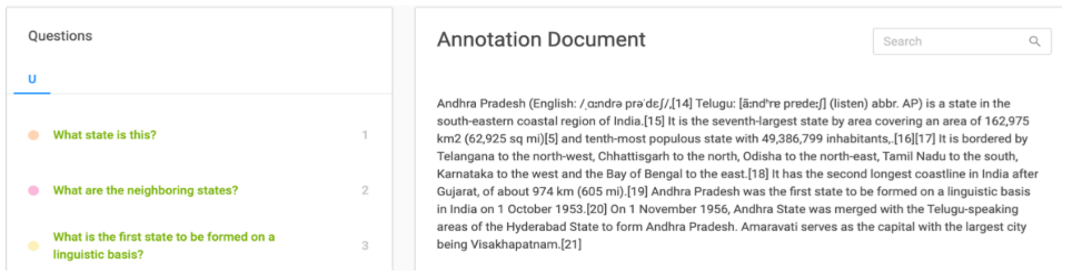
To develop the Q&A

- **Step1:** We first collected the 200 documents related to various nations and states in an excel sheet as shown below.



Andhra Pradesh (English: /Àå...ëÀêndr...ô pr...ôÀàd...õ É/,[14] Telugu: [√£Àênd ±r...ê pr...êdeÀê É] (listen) abbr. AP) is a state in the south-eastern coastal region of India.[15] It is the seventh-largest state by area covering an area of 162,975
Arunachal Pradesh (/...ëÀêr...ôÀån...ëÀêt É...ôl pr...ôÀàd...õ É/,[16] lit. 'Land of Dawn-Lit Mountains')[17] is a state in Northeastern India. It was formed from the erstwhile North-East Frontier Agency (NEFA) region, and became a state on 2
Assam (/...ôÀàs√¶m, √¶-/;[11][12] Assamese: [Àà...ĩx...ĩm] (listen)) is a state in northeastern India, south of the eastern Himalayas along the Brahmaputra and Barak River valleys. Assam covers an area of 78,438 km2 (30,285 sq mi). The s
Bihar (/b...™Ààh...ëÀêr/; Hindi pronunciation: [b...™Àà...¶aÀêr] (listen)) is a state in eastern India. It is the third-largest state by population and twelfth-largest by territory, with an area of 94,163 km2 (36,357 sq mi). Bihar borders Uttar Pra
Chhattisgarh (/t É√¶t...™s...*...ëÀêr/, Hindi: [Ààt É ∞...ôtÃ™ÀêiÀêsg...ô...Ω ±]) is a landlocked state in Central India. It is the ninth largest state by area, and with a population of roughly 30 million, the seventeenth most populous.[6] It border
Argentina (Spanish pronunciation: [a...æxenÀàtina] (listen)), officially the Argentine Republic[A] (Spanish: Rep√∫blica Argentina), is a country in the southern half of South America. Argentina covers an area of 2,780,400 km2 (1,073,500 sq r
Australia, officially the Commonwealth of Australia, is a sovereign country comprising the mainland of the Australian continent, the island of Tasmania, and numerous smaller islands.[13] With an area of 7,617,930 square kilometres (2,94
Bangladesh (/Ààb√¶≈ã..."l...ôÀàd...õ É, Ààb...ëÀê≈ã-/;[12] Bengali: ‡¶¨‡¶æ‡¶Ç‡¶≤‡¶æ‡¶¶‡ßá‡¶ð, pronounced [Ààba≈ãlaÀåde É] (listen)), officially the People's Republic of Bangladesh, is a country in South Asia. It is the eighth-most popu
Belgium,[A] officially the Kingdom of Belgium,[B] is a country in Northwestern Europe. The country is bordered by the Netherlands to the north, Germany to the east, Luxembourg to the southeast, France to the southwest, and the North Se
Cambodia (/k√¶mÀàbo ädi...ô/ (listen);[11] also Kampuchea /Ààk√¶mp äÀàt ÉiÀê...ô/; Khmer: ·ûÄ·ûò·ûí·ûñ·ûª·ûá·ûð, K√¢mp≈≠ch√©a [Ààkampuci...ô]), officially the Kingdom of Cambodia, is a country located in the southern portion of the I
Brazil (Portuguese: Brasil; Brazilian Portuguese: [b...æaÀàziw]),[nt 1] officially the Federative Republic of Brazil (Portuguese: Rep√∫blica Federativa do Brasil),[9] is the largest country in both South America and Latin America. At 8.5 millior
Afghanistan,[c] officially the Islamic Emirate of Afghanistan,[d] is a landlocked country located at the crossroads of Central Asia and South Asia. Referred to as the Heart of Asia,[24] it is bordered by Pakistan to the east and south,[e] Iran t
Chile,[a] officially the Republic of Chile,[b] is a country in the western part of South America. It is the southernmost country in the world, and the closest to Antarctica, occupying a long and narrow strip of land between the Andes to the eas

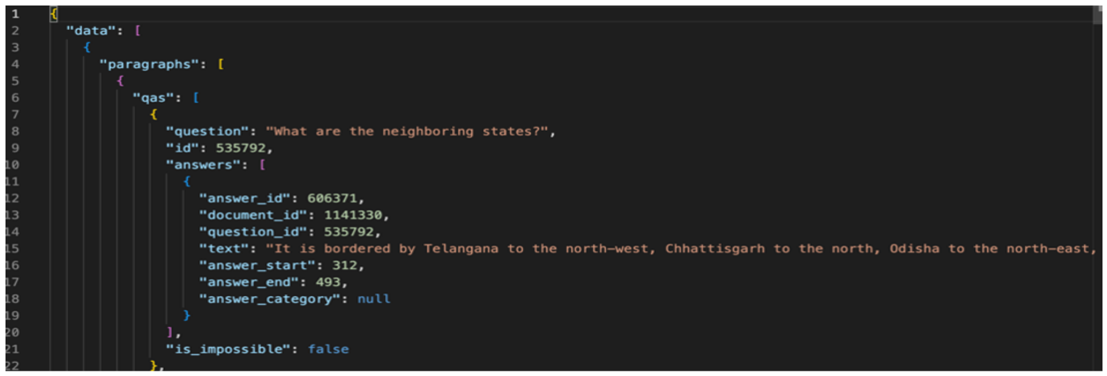
- **Step 2:** We used the Haystack annotation tool to annotate the text within the document to serve as the responses to the questions we had formulated, minimum 5 each approximately.



What state is this? 1
What are the neighboring states? 2
What is the first state to be formed on a linguistic basis? 3

Andhra Pradesh (English: /ˌɑːndrə prəˈdɛʃ/,[14] Telugu: [ãːndʱrɐ prdeːʃ] (listen) abbr. AP) is a state in the south-eastern coastal region of India.[15] It is the seventh-largest state by area covering an area of 162,975 km2 (62,925 sq mi)[5] and tenth-most populous state with 49,386,799 inhabitants,.[16][17] It is bordered by Telangana to the north-west, Chhattisgarh to the north, Odisha to the north-east, Tamil Nadu to the south, Karnataka to the west and the Bay of Bengal to the east.[18] It has the second longest coastline in India after Gujarat, of about 974 km (605 mi).[19] Andhra Pradesh was the first state to be formed on a linguistic basis in India on 1 October 1953.[20] On 1 November 1956, Andhra State was merged with the Telugu-speaking areas of the Hyderabad State to form Andhra Pradesh. Amaravati serves as the capital with the largest city being Visakhapatnam.[21]

- **Step 3:** After the annotations across all the documents were done we exported them into a SQuAD format.



```
{
  "data": [
    {
      "paragraphs": [
        {
          "qas": [
            {
              "question": "What are the neighboring states?",
              "id": 535792,
              "answers": [
                {
                  "answer_id": 606371,
                  "document_id": 1141330,
                  "question_id": 535792,
                  "text": "It is bordered by Telangana to the north-west, Chhattisgarh to the north, Odisha to the north-east,
                  "answer_start": 312,
                  "answer_end": 493,
                  "answer_category": null
                }
              ],
              "is_impossible": false
            },
```

Since we created the dataset from scratch, It was easy for us to do preprocessing steps. In addition, it was in a format of SQuAD, but not the exact parameters are SQuAD, we had to change a few parameters on the time of parsing the dataset.

4. NLP MODELS AND TECHNIQUES UTILIZED

We tried a structured approach to find the best suitable solution for our particular problem statement. First we tried to start with the most basic model, **the BERT base uncased mode**l. It's a state-of-the-art natural language processing model developed by Google. As the parameters in this basic models are less, its a good starting point for us to see how NLP tasks work on our Dataset. After fine tuning the pretrained model imported from haystack. We got decent enough results to see that, the data we collected is enough and the model is actually learning.

After basic bert, we tried experimenting with **distilbert-base-cased-distilled-squad model**. The distilbert-base-cased-distilled-squad model is a 40% lighter model than the original BERT model, by the virtue of which it is faster as well. The distillation process enabled model to achieve comparable performance with faster training times, improved model interpretability and enhanced memory utilization.

Furthermore, we tried experimenting with 4 different models from the German company deepset.

- **deepset/roberta-base-squad2:** It is based on the RoBERTa (Robustly Optimized BERT Pre-training Approach) model, which is a variant of the popular BERT.
- **deepset_roberta-large-squad2:** The models is very similar to roberta base in terms of the underlying transformer, but only differs in terms of the amount of data used for training and the no of parameters.
- **Deepset_tinyroberta-squad2:** It is based on the tiny-RoBERTa architecture, It is trained using unsupervised learning.This allowed us to generalize to new questions and answers.
- **deepset/minilm-uncased-squad2: It** is an open-source transformer based language model pre-trained on SQUAD 2.0. Compared to other existing language models, minilm-uncased-squad2 provides a higher level of accuracy and efficiency. Its main advantages include better understanding of context and entity information, faster inference time, and enhanced performance

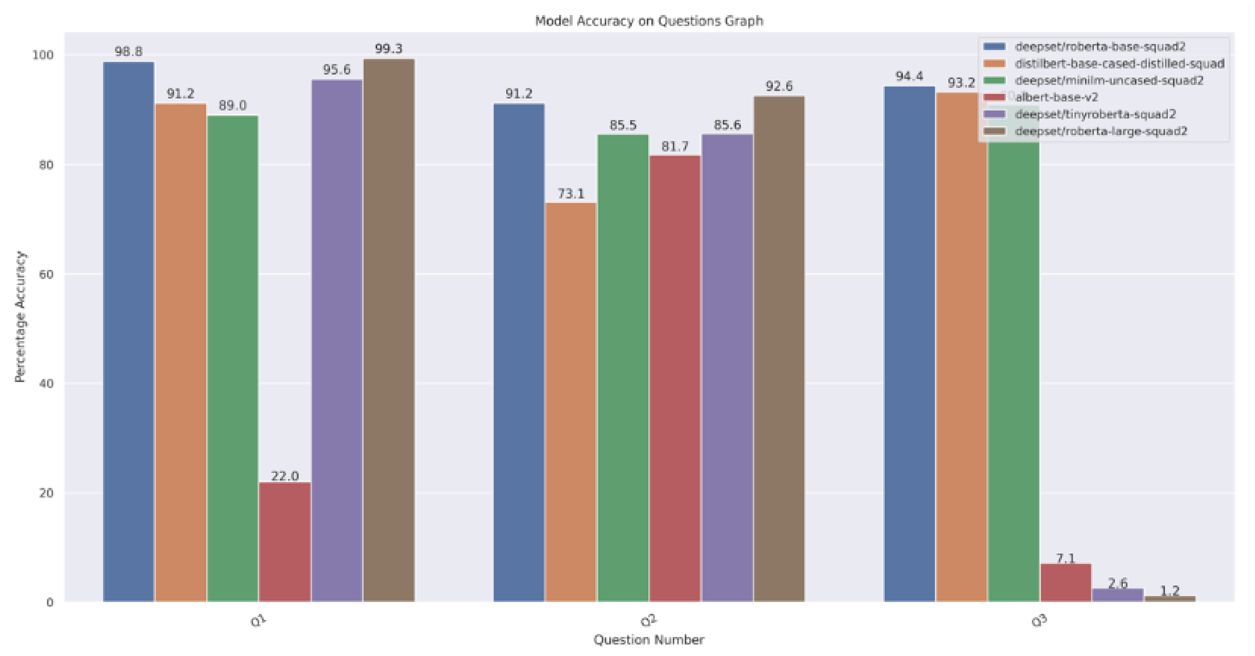
Generally all deepset models performed well with high ranking on all the verification parameters, among them Roberta large model had the best results.

Finally we tried applying the albert-base-v2 model, which is a pre-trained language model developed by Google AI. ALBERT is to improve the efficiency of natural language processing models by reducing the number of parameters and sharing parameters. But we got similar results from the Alberta model as from deep set models.

UI using Streamlit: We built a basic UI to show the working of the models we built. The stream lit is a python based GUI application. Once a model is built on google collab and fine tuned with our custom dataset. The model is saved and downloaded. Then that saved model can be imported into the UI code and verified on a random sample.

5. CONCLUSIONS AND FUTURE PLANS

Upon performing the analysis between the various models implemented and their F1 scores, accuracy, and top 4 match, we were able to conclude that DistilBERT was giving us the best results. We performed the random sample context test on the models, which included 3 Questions which the models had not been trained on, hence creating a benchmark to draw conclusions about the performance of each model.



To improve on our approach, future work could involve including support for languages apart from English. We could also improve our system by including models like GPT3 instead of focusing on a BERT based approach. Since GPT3 has the advantage of performing better with unseen data, it would be a great option to explore. Additionally, we can improve on the performance of the models in our approach by increasing train data size or by further fine tuning the models.

6. INDIVIDUAL CONTRIBUTIONS

Names	Data Preparation	Model Training	Extra
Aishwarya Paruchuri	Collection and Annotation of Paragraphs : 50	Models: 1. deepset/roberta-base 2. deepset/tinyroberta 3. deepset/roberta-large-squad2 4. Fine-tuned Roberta-base model	1. Worked extensively on Project Report and Project Presentation. 2. Transferred the contents of our presentation slides to Canva to make it more interactive.
Shubhangi Manoj Kumar	Collection and Annotation of Paragraphs : 50	Models, via huggingface and haystack FARMreader. 1. albert-base-v2 - Fine tuned 2. deepset/deberta-v3-base-squad2 3. Fine tuned Alberta.	1. Wrote content for various Project Report sections and sections of the Project Presentation as well.
Priyank Thakkar	1. Collection of Paragraphs : 51 2. Total questions annotated: 420	HayStack based FARMreader Models. 1. deepset/roberta-base-squad2 2.distilbert-base-cased-distilled-squad 3. deepset/minilm-uncased-squad2 Comparison graphs for all 6 models evaluation & benchmarking on our custom dataset using above models.	1. Exploratory Data Analysis (EDA) to know more about custom dataset. 2. Information Retrival Methods (TFIDF). 3. Compared all 6 models on three questions taken from random context available on the internet. 4. Project Proposal, PPT, Report
Vineet Kiragi	Collection and Annotation of Paragraphs : 50	1. Built a bert base uncased model using hugging face and fine tuned it to fit our dataset with a decent accuracy. 2. Built an UI to deploy the model using streamlit and ran all the implemented models on the UI.	1. Wrote explanations for the models in the report.