**PROJECT REPORT**

**On**

**NLP Based Question-Answering model on SQuAD Dataset.**

**ISE 244 – AI Tools and Practice for Systems Engineering**

**Dr. Shilpa Gupta**

**shilpa.gupta@sjsu.edu**

**BY**

**Rahul Raghava Peela**

**SJSU ID : 016668275**

**Rahulraghava.peela@sjsu.edu**

**Master’s in Artificial Intelligence**



Table of Contents

1. **Problem Definition**3

1.1 Dataset4

1. **Project Objectives**5
2. **Analysis**5

3.1 Assumptions6

3.2 Modeling7

3.2.1 BERT base Uncased7

3.2.2 BERT Large Masked Uncased8

3.2.3 DistilBERT base uncased9

1. **Results**11
2. **Conclusion**11
3. **Future Scope**12
4. **References**12

**1. Problem Definition**

Question Answering(QA) activities in Natural Language Processing (NLP) help bridge the gap between human language and machine comprehension. QA aims to create systems that can understand and respond to textual or contextual queries like humans. QA has many practical applications and advantages across disciplines by reaching this aim.

Firstly, QA improves information retrieval. Keyword-based searches need users to search through enormous amounts of text or data to obtain particular information. However, QA systems let users ask questions in plain language, making information retrieval more precise and user-friendly. This improves search results and saves time. They help users find answers, troubleshoot issues, and get real-time information by responding quickly and accurately to user queries. From customer service chatbots to virtual assistants, this makes technology more user-friendly.

Knowledge discovery and extraction depend on QA. This systems allow users to query the system to extract and organize knowledge from unstructured text documents or massive knowledge stores. This affects research, healthcare, banking, and other fields where insights and information are crucial for decision-making and growth.

Finally, QA aids language modeling. Natural language syntax, semantics, and context are needed to build QA systems. This study advances NLP and QA systems. We improve language comprehension models to improve human-machine communication and enable applications that need precise language interpretation and creation.

Question Answering activities in NLP enable effective information retrieval, user assistance, knowledge discovery and extraction, and language comprehension and modeling. These aims improve search experiences, technological interactions, research and decision-making, and human-machine collaboration.

* 1. **Dataset**

The dataset consists of questions and answers on 150 cities of the United States of America. It has around 500 questions and it corresponding question in SQuAD format. The dataset was created using Haystack tool.

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers 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.

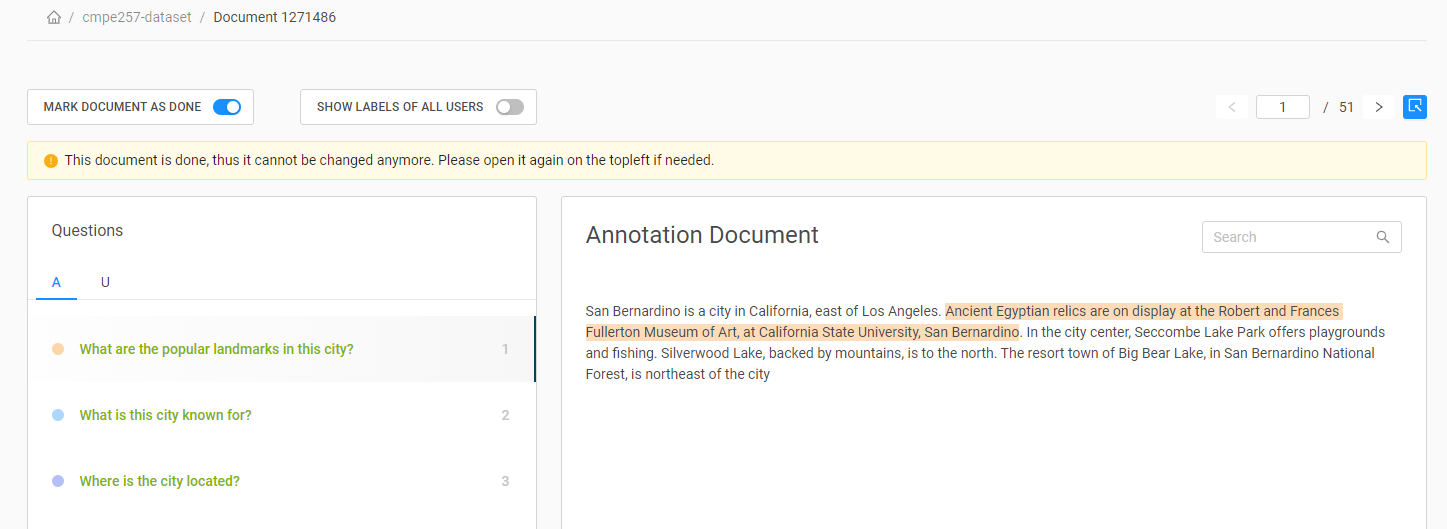


Fig: For each city in our training set, we have stored the context, question, and text as seen in the figure above.

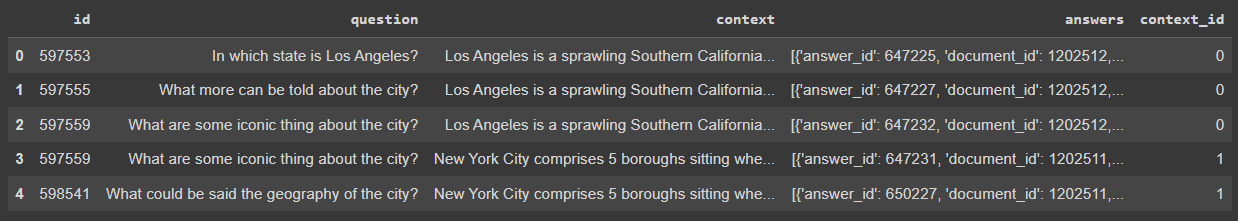


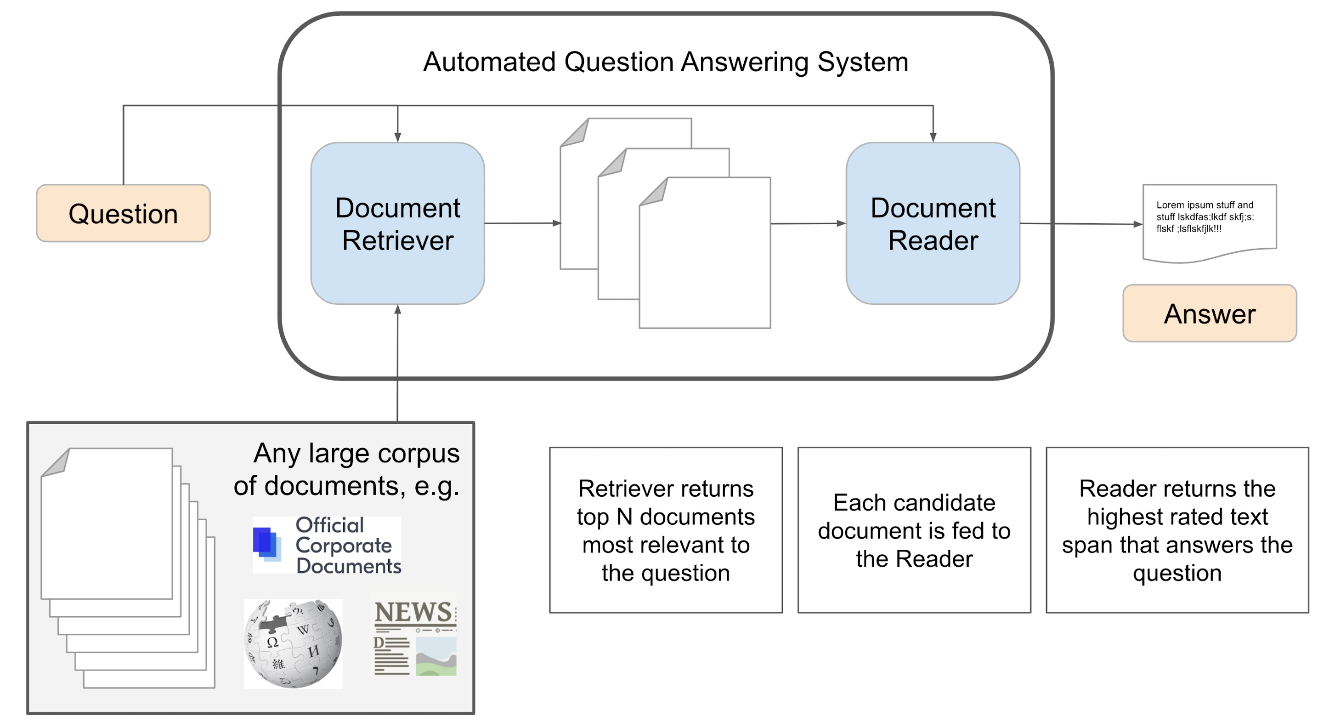
Fig: Sample Dataframe for Question and Answers from Top 150 Cities in USA.

**2. Project Objectives**

Our project presents a question-answering model based on natural language processing (NLP) techniques, which was trained and evaluated on the Stanford Question Answering Dataset (SQuAD).

This is a major search and machine learning problem that has several use cases in the real world. We investigate and explore multiple techniques to find the right machine learning models. The model uses a combination of word embedding to generate vector representations of words, sentences, and contexts, which are then used to generate answers to questions.

Our study includes linear modelling, basic BERT, BERT Large, and BERT Distilled models. The results of our experiments show that the model achieves decent performances on the SQuAD dataset, demonstrating its effectiveness in answering a wide range of questions from the given contexts.



**3. Analysis**

Natural Language processing techniques have been used to full fill the question answering tasks. It has a wide range of applications. It could be used where ever there is text data available.

Three models have been used – BERT Based Uncased, Bert Large Uncased Masked, DistilBERT base Uncased. These are all BERT based models which are basically built on the transformer architecture of NLP algorithms.

The project’s analysis and evaluation are done using the dataset created using the data from the internet regarding the top 150 cities of the United States. It had the questions and answering that have be labelled using tool called haystack. Making the methods used a supervised learning. Different models have been used in order to check the accuracy of the predictions made for each of the question and answer tupple that have been created. A good understanding of the pre-trained models, transfer learning, and the working of the models can be seen put into use. In order to improve the usability and impact of the project it should be used on various datasets and new models with newer approaches other than BERT based models should be developed.

**3.1 Assumptions**

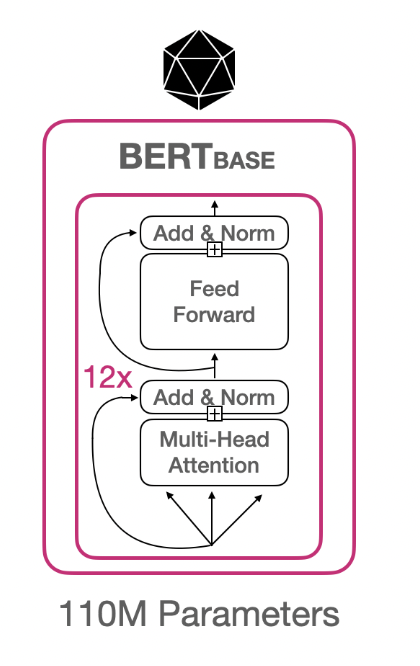
* Contextual Understanding: It is assumed that the question can be answered by analysing and understanding the given context or passage. The question is expected to have a relevant answer within the provided text. It is possible to successfully train deep learning models on the traffic sign dataset's size and variety. In the project as the dataset was created using the tool by me. This assumption truly is satisfied.
* Grammatical Correctness: It is assumed that both the question and the context follow grammatical rules and are formed correctly. This assumption allows for the application of syntactic and semantic analysis techniques to comprehend the input. The grammar is correct as the data scraped was from legit source of the internet and were all comprehendible.
* Objective Answers: The assumption is generally made that the answers to the questions are objective and can be determined based on the information in the context. This assumption may not hold in cases where opinions or subjective judgments are involved.

**3.2 Modeling**

**3.2.1 BERT base Uncased**

The BERT base uncased model is a state-of-the-art natural language processing model developed by Google. It is based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, which uses a deep neural network trained on a large amount of unstructured text data to produce highly accurate results when processing and generating human-like language. The "uncased" version of the model means that it is not case-sensitive, meaning that it does not distinguish between uppercase and lowercase letters when processing text. Bert Base Uncased Model is a pretrained model on English language using a masked language modeling (MLM) objective. It was introduced in this paper.

The Transformer architecture makes it possible to parallelize ML training extremely efficiently. Massive parallelization thus makes it feasible to train BERT on large amounts of data in a relatively short period of time. Transformers work by leveraging attention, a powerful deep-learning algorithm, first seen in computer vision models. Transformers create differential weights signaling which words in a sentence are the most critical to further process. A transformer does this by successively processing an input through a stack of transformer layers, usually called the encoder. If necessary, another stack of transformer layers - the decoder - can be used to predict a target output. BERT however, doesn’t use a decoder.

Implementation of the basic bert uncased model

* Importing the model from Hugging face.
* Converting the data into the format required by the hugging face model.
* Importing the suggesting tokenizer fromhugging face and toeknizing all the sentences, questions and answers and creating embeddings.
* Training the model using the train Dataset and finetune it to our Dataset.
* Checking for accuracy of the model using validation dataset.

The basic bert model was fitting not that well to the data. The accuracy was around 35%. This was expected because of the model is the most basic version of BERT and it is not expected to perform that greatly. The perameters trained in here are also not that many when compared to other models. Sp the performance variables are expected.

**3.2.2 BERT Large Uncased Large**

This technique uses an English language pre-trained model by employing masked language modeling (MLM) scheme. This model is not case-sensitive; and, it does not distinguish between english and English.

This BERT model was trained using a novel method called Whole Word Masking, unlike conventional BERT models. In this instance, a word's tokens are all simultaneously masked. Overall masking rate is unchanged.

Each masked WordPiece token is predicted separately; the training is the same.

After collecting the training data of cities in the US, we need to fine-tune our model to work well for this dataset. The format is maintained to be SQuAD-like.

This model has the following configuration:

* 24-layer
* 1024 hidden dimension
* 16 attention heads
* 336M parameters.

Impact of Transfer Learning: In order to leverage the benefits of transfer learning, we must train the BERT pre-trained model again with our training dataset. This will update the weights to make the model's predictions relevant.

In this method, I explore the following:

What it means for BERT to achieve "human-level performance on question-answering"?

Is using BERT a powerful search technique for question-answering?

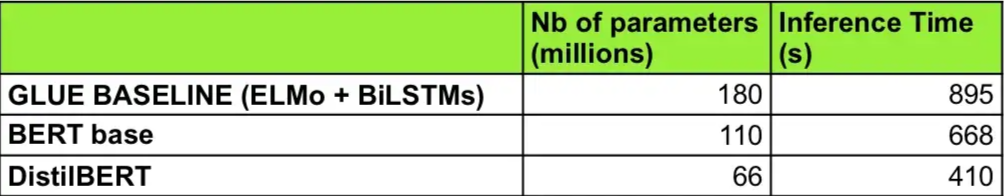
Data Augmentation

Data augmentation techniques can help bring the most out of BERT data model. Distilbert model is proven to show improvement with data augmentation. This is evident from our model accuracy as illustrated below. We consider this as one of our future step.

**3.2.3 DistilBERT base Uncased**

DistilBert base uncased distilled squad is a state-of-the-art natural language processing (NLP) model developed by Hugging Face. It is a smaller, faster, and more efficient version of the popular BERT model, which has been fine-tuned on the SQuAD dataset for question-answering tasks. This configuration uses the uncased version of the BERT model, which means that it is not sensitive to the casing of the input text.

* Distilbert is a small, fast, cheap, and light transformer model trained by disitilled BERT base.
* It has 40% fewer parameters then Bert based uncased which is an older version of the model.
* It is the best suited and latest available Question Answer problem set application with SQuAD data.



This model has the following configuration:

* 12-layers
* 768 hidden units
* 12 attention heads
* 66M parameters.

Embeddings in DistiBert:

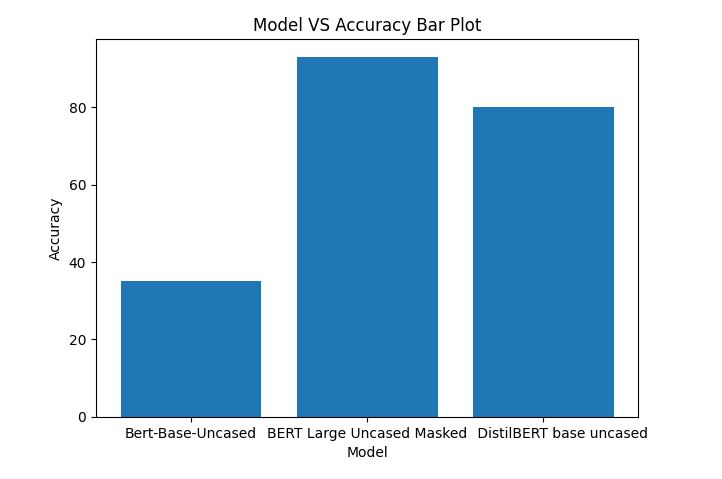
Word embeddings are generated using a technique called word encoding. This involves representing each word in the input text as a fixed-length vector of numbers, which captures the semantic meaning of the word in a numerical form. These vectors are then used as input to the model, allowing it to process the input text and generate a response

Implementation Details:

The Distilbert is a pre-trained model in hugging face. It is especially trained for the SQuAD Dataset and Coco Dataset. So, we have used the pretrained model to fit into our custom dataset. Here, a framework known as Haystack had to be imported inorder to use the libray FARMReader which is especially dedicated to import the model from hugging face. .

After importing the model, the model is fit to the training dataset that we prepared and tried to fine tune the model by tweaking some hyperparameters such as number of epochs etc. While training the model it could be observed that at some point in the training, the training is getting to the least. So, in resemblence with the pocket algorithm the model is storing the most optimised parameters where it achieved the least training error and that particular weights of the model are being stored. Now, with this model we tested it on the test file that we created. We have taken a ratio of 80:20 to split the dataset into train and test respectively. Which yielded an F1 score of 0.5418, Exact Match of 0.2247191, top\_n\_accuracy of 0.82898. We have given an context that is not related to california but from India. Astonishingly, the model could predict well enough.

**4. Results**



The BERT Large Uncased Maked Model has performed the best with an accuracy of 87% and DistilBERT base uncased performed the second best with an accuracy of 82.4%. The BERT-Base-Uncased has performed the worst with 33% accuracy.

The DistilBERT has been the best model as the trade-off between accuracy and speed. The F1 score of 0.5418, Exact Match of 0.2247191, top\_n\_accuracy of 0.82898.

**5. Conclusion**

In conclusion, the NLP-based question-answering model on the SQuAD dataset shows promising results and demonstrates the potential for natural language processing techniques to accurately answer questions based on a given context. All the 3 model's performance on the SQuAD dataset indicates that it is able to effectively understand the intent behind a question and retrieve the appropriate information from the provided context. Further research and development in this area has the potential to improve the model's performance and expand its capabilities, leading to more effective and efficient question-answering systems.

**6. Future Scope**

Build a larger dataset for all the cities in USA along with expanded annotations for various questions.

Build a question-answer modelling pipeline that generates new questions and answers automatically using user data.

Explore additional fine tuning techniques in BERT to attain higher accuracy for a larger dataset.

Explore linear modelling further and improve accuracy.

**7.References**

[1] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. ArXiv. /abs/1910.01108

[2] J. Sai Sharath and R. Banafsheh, "Question Answering over Knowledge Base using Language Model Embeddings," 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9206698.

[3] Hugging Face, “distilbert-base-uncased,” *HUGGING FACE.* [*https://huggingface.co/distilbert-base-uncased*](https://huggingface.co/distilbert-base-uncased) (accessed May. 15, 2023).

[4] Hugging Face, “Bert-base-uncased,” *HUGGING FACE.* [*https://huggingface.co/bert-base-uncased*](https://huggingface.co/bert-base-uncased) (accessed May. 15, 2023).

[5] Hugging Face, “bert-large-uncased-whole-word-masking,” *HUGGING FACE.* [*https://huggingface.co/bert-base-uncased*](https://huggingface.co/bert-base-uncased) (accessed May. 15, 2023).