Question Aswering System for Draft Red Herring Prospectus

Arnav Sharma[1], Arun Sar[2], Kaushal Kishore[3], and Shuchismit Jha[4]

[1,2,3,4]Plaksha University
Email ids: [1]arnav.sharma@plaksha.edu.in,
[2]arun.sar@plaksha.edu.in, [3]kaushal.kishore@plaksha.edu.in,
[4]shuchismit.jha@plaksha.edu.in

April 8, 2022

1 Introduction

The year of 2021 saw 63 companies float their Initial Public Offering (IPO) combined on the National Stock Exchange (NSE) & Bombay Stock Exchange (BSE). Each of these companies had to file and submit a Draft Red Herring Prospectus (DRHP) with the market regulator which is also made available to the public. The DRHP enables potential investors to analyze the company's financial health, market valuation, issuance objective and other important information to decide on purchase. Each report is around 500 pages long, which makes it extremely complicated for people to understand, often even find the information they are seeking. Natural Language Processing on these documents can help in extracting desired information which can be presented as output in multiple forms be it a summarizer to show only certain information which are considered important or a Question Answering System, even a chatbot.

2 Aim

The scope of this project entails creating a Question Answering System to answer all sorts of Questions based on the company's performance as mentioned in their DRHP.

3 Dataset

The dataset being used for fine-tuning is the Stanford Questioning Answering Dataset (SQuAD). SQuAD is a is a collection of question-answer pairs derived

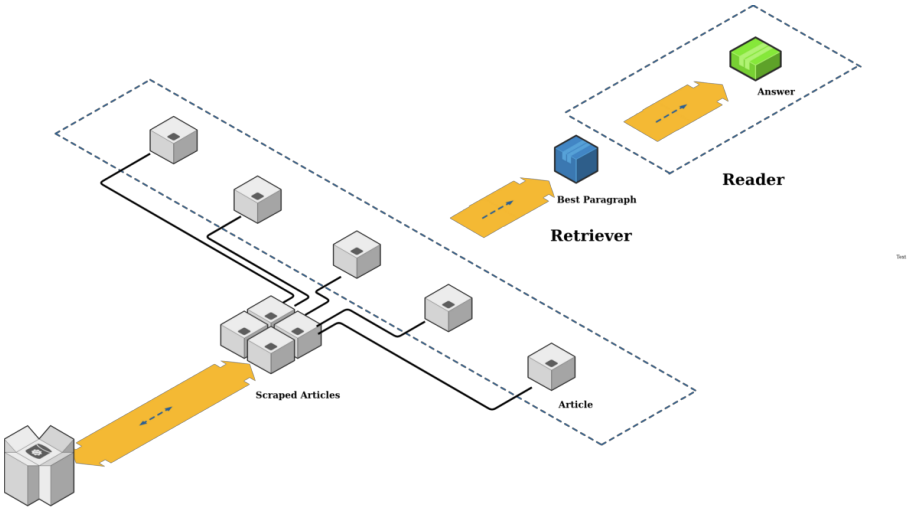

Figure 1: Q/A Flow

from Wikipedia articles. Since the questions and answers have been produced by humans, they include the extent of subjectivity and ambiguity which would be expected from questions on the deployed model. SQuAD contains 100,000 questions which have been answered by the crowd along with 50,000 unanswerable questions written adversarially by people in forms that are similar to the other answerable ones.

4 Model Framework

4.1 Document Store

The DRHPs are readily available in PDF format. To store the text for further processing (eg. Tokenize) they need to be parsed. Base Document Store creates a base class to store the PDF content from where they can be efficiently processed.

4.2 Model

The model being used is DistilBERT. DistilBERT is uses a distillation technique for making Bert models smaller and faster. Distillation is a technique used for compressing a large model, called the teacher, into a smaller model, called the student to reproduce the same behavior. Its breakdown consists of Retriever, Ranker & Reader. The Retriever is a lightweight filter that can quickly go through the full document store and pass on a set of candidate documents that are relevant to the query. BM25, a bag-of-words retrieval function ranks a set of

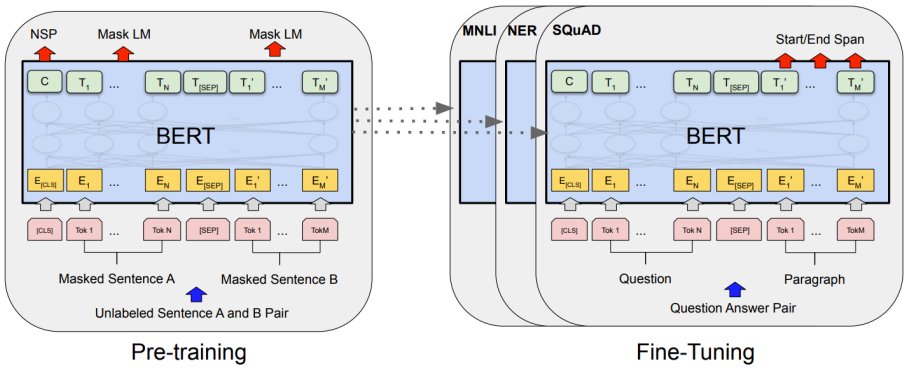

Figure 2: BERT/DistilBERT Architecture

documents based on the query terms appearing in each document, regardless of their proximity within the document. When used in combination with a Reader, the retriever is a tool for sifting out irrelevant documents, saving the Reader from doing more work than it needs to and speeding up the querying process.

5 Challenges

- The dataset being used is the SQuAD dataset which although includes human subjectivity, still does not provide fully accurate answers. Since the output in demand requires often basic yet specific financial information which is extremely tough to map from a generic dataset which has not been tagged and trained for financial output, the answer received is ambiguous most of the time. To handle this limitation, other datasets like FinQA can be used which have been specifically trained for numerical reasoning over financial data. An attempt to use this dataset for training failed due to lack of expertise in handling 'Json' file type, in which the dataset was available.

- The size of the input texts is also a limiting factor – both to processing time and answer accuracy. Each report is 500 pages long, which is practically impossible for DistilBERT to parse. One solution to this problem is using Longformers which is specifically designed to summarize long texts. Longformers are capable of parsing upto thirty-two thousand characters at a time. The attention mechanism in longformers is a drop-in replacement for the standard self-attention and combines a local windowed attention with a task motivated global attention. The attention patterns include sliding window attention, dilated sliding window attention and global plus sliding window. The reports can be broken down into subsets of thirty-two thousand characters and parsed separately. This provides better performance

over DistilBERT.

- FinBERT is a Sentiment Analysis model trained on Financial Opinion Mining and Question Answering (FiQA) dataset which uses pretrained BERT models. This can potentially provide more accurate results however it has not been used for modelling due to inability to install certain required packages which are incompatible with current versions of PyTorch.