Why Distil-BART?

To find the best text summarizer, we went throught multiple models on Hugging Face and multiple other places on the internet where we could find pre-trained models. Our observations led us to believe that Distil-BART text summarizer by Facebook was the best for our use case.

To go through with this task of narrowing down the text summarizers, we made a dataset using CNN articles as it was a very general dataset and tested all the models on the same 10 articles. We tested 8 models on these 10 articles and noted down the results with a word count in the same excel sheet. These results helped us compare the performance of these models on similar datasets and helped us narrow down the model a lot.

There are two types of text summarizers on the internet -

1. Extractive Text Summarizers - The process of extractive text summarization looks for important sentences and adds them to the summary, which includes exact sentences from the source text.

2. Abstractive Text Summarization - In this method, it makes an effort to recognize key passages, evaluate the context, and deftly provide a summary.

We found that the abstractive paradigm of summarizing was extremely useful for our use case as our inputs might have meaning in every sentence and we need the summarizer to understand the whole input and give a summary such that a small kid can also understand it. Taking out the most important sentences might not accomplish this task, hence we decided to go ahead with abstractive summarization engines.

We tested multiple etxractive and abstractive summarizers to compare their performances.

We tested out sumy, BERTSUM and summa which are extractive models. As expected these models just took sentences from the beginning, middle and end of the input data to model a summary. This

did not convey the important points of the input paragraph and often led to nonsensical and illogical outputs. Thus we eliminated extractive models from our options.

Other abstractive models we tested included BART, T5 and Pegasus. These models often resulted in extractive outputs. In some models like Pegasus we found it rather challenging to modify the parameters based on which the summarization is performed. Some of them also performed well only on constrained input text types and lengths. Thus amongst all abstractive models we found that distil-BART gave us more reliable and correct outputs.

After testing and comparing the performance and results of these models, we decided to go ahead with BART because -

1. It has been trained on large datasets using unsupervised learning methods, which allows it to generate high-quality summaries that are faithful to the original text. As a result, DistilBART has been shown to outperform other text summarizers on various benchmark datasets.
2. It is a much smaller and faster model compared to the others mentioned above because of which it runs much faster and has a much lower CPU requirement.This allows DistilBART to achieve similar levels of accuracy as the larger models while using fewer computational resources and taking less time to generate summaries.
3. It is open-source and free unlike a few models mentioned above. This also means that it is constantly being updated with the latest techniques and improvements in natural language processing.

All these reasons motivated us to pick DistilBART as the text summarizer for our application.