|  |  |
| --- | --- |
|  | Tech Immersion and Placement Programme  Applied Artificial Intelligence  6 January – 7 April 2020  Project Capstone Report  Submitted By:  Eugin Lee  Koay Seng Tian |

Table of Contents

[1 Background 2](#_Toc35934095)

[1.1 Background (1.1) 2](#_Toc35934096)

[2 Methodology and Design 2](#_Toc35934097)

[2.1 Software Data Flow 2](#_Toc35934098)

[3 Findings 2](#_Toc35934099)

[3.1 Findings (3.1) 2](#_Toc35934100)

[4 Evaluation and Analysis 2](#_Toc35934101)

[4.1 Evaluation and Analysis (4.1) 3](#_Toc35934102)

[5 Results 3](#_Toc35934103)

[5.1 Results (5.1) 3](#_Toc35934104)

[6 Conclusion 3](#_Toc35934105)

[6.1 Conclusion (6.1) 3](#_Toc35934106)

[7 Recommendations 3](#_Toc35934107)

[7.1 Recommendations (7.1) 3](#_Toc35934108)

[8 Appendices 3](#_Toc35934109)

[8.1 Appendices (8.1) 3](#_Toc35934110)

[9 References 3](#_Toc35934111)

[9.1 References (9.1) 3](#_Toc35934112)

# Background

This project is a collaboration between Nvidia and RP as a capstone project for the TIPP programme. The project aim to provide a real world work environment for the student to apply artificial intelligence techniques taught in the course. Supervisors in return receive the intellectual property of the final product.

## Background (1.1)

The goal of this project is to provide some kind of an annotation tool, to differentiate key takeaways in a corpus of text. The summarisation technique will be powered by BERT models.

# Methodology and Design

Extractive Summariser using BERT transformer model

Using a webpage as an input, this BERT model summariser aim to extract key feature sentences of the main corpus.

These extracted features will then be presented as an annotated form, together with the main corpus as an output document.

Using a client-server model, the web application provide seamless transition between server(flask), UI (streamlit), and the underlying python code.

For evaluation, we will be using ROUGE scoring system which is optimised to calculate distances/ similarities between summarised articles. By using ROUGE, and comparing between BERT generated summaries against human generated summaries, we are able to calculate the similarity between the text and recommend ways to improve the results.

## Software Data Flow

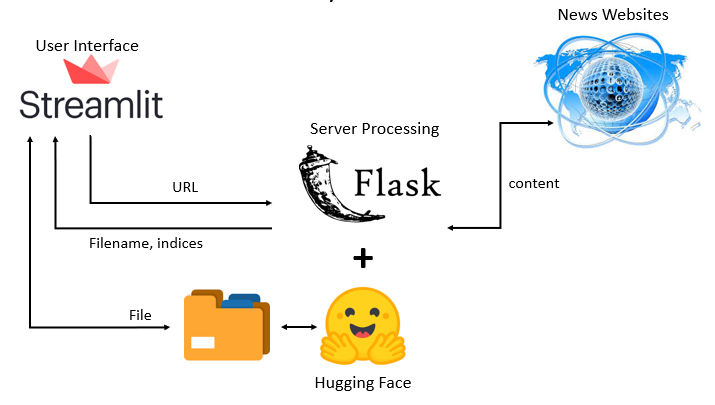


Figure 1 Data Flow

The software is developed using available open-source application frameworks (Streamlit, Flask) and Hugging Face BERT/Transformer model. The software is developed using Python scripting language.

Streamlit architecture is based on the ability to write web application the same way a plain Python scripts is written. Streamlit applications have a unique data flow. Any time something must be updated on the screen (for example, the application responding to a button is pressed), Streamlit will just rerun the entire Python script from top to bottom.

This will pose a challenge for the application developer because it is not implemented as a 'callback', like most web applications will perform. Some of these quirks can be modified using Streamlit's cache decorator (streamlit@cache) which allows developers to skip certain costly computations when the application reruns. However, such modification, as we have observed, may create stability issues.

# Applications

In this project we utilised these libraries:

|  |  |
| --- | --- |
| Python Library | Use |
| Newspaper3k | Scraping tool |
| Pandas | Data manipulation |
| summarizer | BERT summariser |
| datetime | timestamp |
| rouge | ROUGE scoring |
| NLTK | Tokenising words |
| Flask | Server |
| streamlit | User Interface |

# Insights

In this project we discovered 2 major takeaways that arise due to our experiments and research. We discovered that (1) model selection plays a big part of in operational efficiency, (2) and

## BERT models are not equal (4.1)

When we first started out, we utilised the “stock” BERT model (bert-base-uncased) that was embedded in the summariser library. While testing, we quickly realised that we may face into some usability issues as the waiting time for the extractive summary takes around 17 seconds to complete.

According to recent research on attention spans, current attention spans range between 8 to 12 seconds. This means that users might regard a 17 second wait to be “too long” and may not even use the product even if it was a perfect product. We need to find a compromise between speed and performance.

To achieve this, we look at the different BERT models available in the market. We eliminated the “large” BERT models at first cut, because they contain more parameters which will mean longer processing time. The narrows our focus to the base BERT model. To evaluate performance, we run a loop test, controlling all parameters except a model change. In this experiment, we recorded the performance in the form of processing time, and text length.

At this point, we are using just quantitative measures, such as (1) processing speed and (2) text length as our selection.

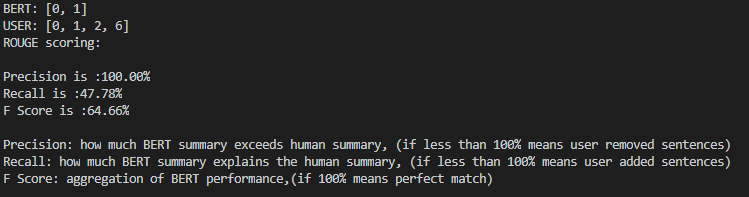
Within the BERT models, we chose DistillBERT as its processing speed falls within the sweet spot between 8-12 seconds and the summarised text length seems to be within an acceptable ballpark area. In addition, to artificially inprove on the run time, we created a time illusion (to make time seem to pass faster),by inserting loading animation and stimulating an active program during the processing runtime.

|  |  |
| --- | --- |
|  |  |

## ROUGE and not BLEU

We selected ROUGE as our main technique to measure model accuracy. The main reason is because ROUGE provides more insights into the scoring mechanism. In ROUGE, the precision and recall score is also provided which gives an understanding in the event of a low score, whether the machine generated text, is too short or too long, or is it not relevant. With this knowledge, we could either tweak the parameters correctly, or if the summary is poor, we can fine-tune it with a longer relevant corpus. With BLEU, another popular technique, only the final BLEU score (between 0 to 1) is given. This gives us no direction to improve.

ROUGE and BLEU are both popular methods in summaries evaluation. They use similar calculation methods surrounding n-grams and variations of recall and precision. When testing, we discovered that BLEU does not seem to work very well when there are differences in corpus length (between BERT and human generated). This could be due to the brevity penalty effect. The documentation recommended a smoothing technique to overcome this but we did not find the value to be as intuitively accurate as ROUGE.



# Results and conclusions

We successfully created and deployed a web application that allows a user to (1) enter a URL and received a full text with annotated key points (2) user can provide feedback, which can be used to provide a quantitative scoring (3) we could improve the BERT model based on the scoring.

Our average runtime for the model is around 8-10 seconds. The model for some reason, seems to run slightly faster after a few iterations.

Project managers from Nvidia are extremely satisfied with the progress we made.

# Recommendations

Given that the code is designed to be modular based and robust, the engineer will be able to switch out part of the code and be able to run. For example, as our parser is optimised for newspaper articles, it may not work as well for PDF. In which case, the engineer could switch in a PDF parser instead.

The streamlit application works perfectly for a proof-of-concept user interface; to quickly unlock the power of the underlying model, and test out concept feasibility. However, the UI content some limitations such as feedback functionality like a HTML, and also limited fine tuning options, which hinders our ability to improve the product.

In NLP literature, the computer generated summary is always compared against a “gold standard” summary to determine its accuracy. However, in our case, the BERT summary is compared against to a random user for feedback. This user may have its own biasness, and may have an inconsistent proficiency of the subject. This could all affect our evaluation score of the model. To avoid this issue, it might be more useful to collect an aggregation of user responses, before tweaking the model.

# Appendices/ Github

Blah blah blah

Share our github here, and which are the main .py to run. Maybe ST can write on how to run the program

## Appendices (8.1)

Blah blah

# References

Blah blah blah Stick out workflow and research list in here