

# Machine Learning Engineer Nanodegree

## Capstone Project

Wong Teck Meng  
November 5th, 2021

## Automatic Text Summarization Application

### I. Definition

#### Project Overview

In this Information Age, the abundance of text content has created value for summarized text content. Summarization consist of condensing text to a concise summary of the initial text while retaining its most essential points and overall context.

Automatic Summarization is not new. First notable attempt dates back to 1958 made by IBM<sup>1</sup>. Since then, there are two main approaches to summarize text content:

1. Extractive summarization <sup>2</sup>– Essential sentence selection from a/set of documents.
2. Abstractive summarization <sup>3</sup>– Rewriting the text content in a more precise manner.

There are many datasets available for summarization. The datasets domain ranges from News articles, Reddit and Wikipedia. This project will choose a dataset with less than 3000 documents/summary pairs. The is constrained because the training time of huge datasets will go up to several weeks on my local machine setup.

#### Problem Statement

The goal of this project is to create an end-to-end application that automatically summarize text input while retaining its most essential points to a text output.

This project is to build and deploy:

---

<sup>1</sup> Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. 2(2), 159–165 (1958)

<sup>2</sup> [Sinha et al.](#) : Extractive Text Summarization using Neural Networks (2018)

<sup>3</sup> [Rush et al.](#): A Neural Attention Model for Abstractive Sentence Summarization (2015)

- A REACT web application that input text and output summarized text. The application will provide evaluation metrics when reference summary is provided.
- A model to perform summarization task.
- A backend application to initialize and serve the summarization model.

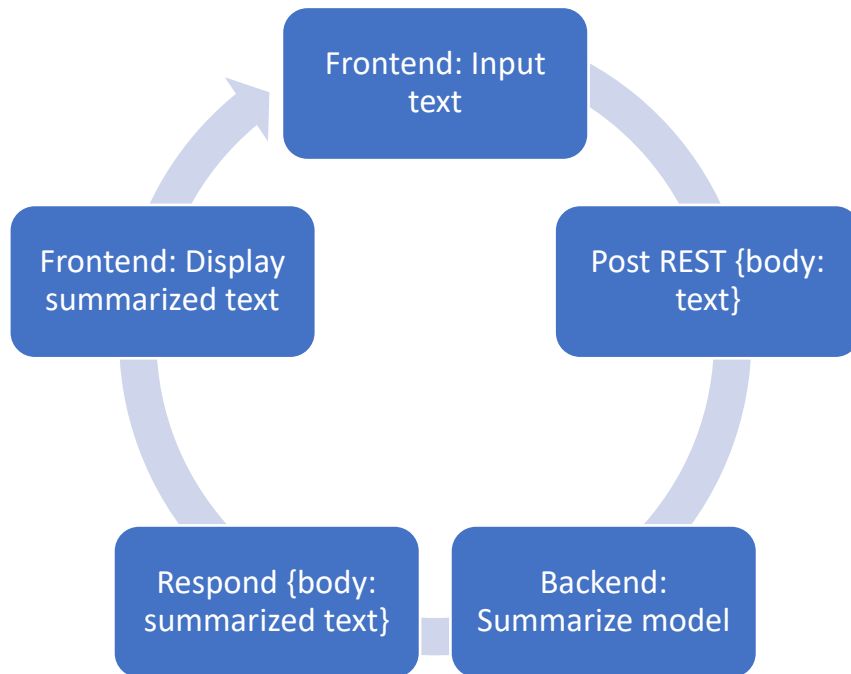


Figure 1 End to end process flow diagram

## Metrics

Recall-Oriented Understudy for Gisting Evaluation aka **ROUGE**<sup>4</sup> is a commonly used metric for summarization and translation in natural language processing. The metrics are based on comparison between a **reference summary** (usually human produced) and an automatically produced **system summary**. However, ROUGE does not take into account the sentence structure or grammatical correctness.

ROUGE comprises of various metrics. Of the following evaluation metrics available, ROUGE-N is the main one.

### ROUGE-N

Overlap of N-grams between system and reference summaries. **ROUGE-1** will be measuring unigrams overlap. **ROUGE-2** and **ROUGE-3** would use bigrams and trigrams respectively.

### ROUGE-L

Measure the longest matching sequence of words using **LCS**.

<sup>4</sup> [Chin-Yew Lin: Text Summarization Branches Out](#) (2004)

The following metrics are provided by **ROUGE**:

### Recall

Words in the **reference summary** that have been captured by the **system summary**. This measure how many information are captured in the **system summary**.

$$recall = \frac{\text{Number\_of\_overlapping\_words}}{\text{Total\_words\_in\_reference\_summary}}$$

### Precision

It measures relevance of **system summary**. This measures the number of relevant words.

$$recall = \frac{\text{Number\_of\_overlapping\_words}}{\text{Total\_words\_in\_system\_summary}}$$

### F-measure

F-measure is a harmonic mean of the **precision** and **recall**, where F-measure reaches its best score at 1 and worst score at 0. The relative contribution of precision and recall to F-measure score are equal.

$$fmeasure = 2 * \frac{(precision * recall)}{(precision + recall)}$$














## II. Analysis

### Data Exploration

BBC News Summary<sup>5</sup> will be used for this project dataset.

#### Data Explorer

7.27 MB

- ▼  BBC News Summary
  - ▼  News Articles
    - ▶  business
    - ▶  entertainment
    - ▶  politics
    - ▶  sport
    - ▶  tech
  - ▼  Summaries
    - ▶  business
    - ▶  entertainment
    - ▶  politics
    - ▶  sport
    - ▶  tech

This dataset has 2225 news articles of BBC news website from 2004 to 2005. The dataset comprised of 5 topics, business, entertainment, politics, sport and technology. Each news article has summary provided in the Summaries folder. The first line of the text of articles is the respective title. It was used in the paper of D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006; whose all rights, including copyright, in the content of the original articles are owned by the BBC. More at <http://mlg.ucd.ie/datasets/bbc.html>.

This project will use the news articles as text inputs to the automatic text summarization model and the corresponding summaries folder used as reference summary for evaluation purpose.

---

<sup>5</sup> <https://www.kaggle.com/pariza/bbc-news-summary>

	title	article	summary
0	Ad sales boost Time Warner profit	Quarterly profits at US media giant TimeWarner jumped 76% to \$1.13bn (£600m) for the three months to December, from \$639m year-earlier. The firm...	TimeWarner said fourth quarter sales rose 2% to \$11.1bn from \$10.9bn. For the full-year, TimeWarner posted a profit of \$3.36bn, up 27% from its 200...
1	Dollar gains on Greenspan speech	The dollar has hit its highest level against the euro in almost three months after the Federal Reserve head said the US trade deficit is set to s...	The dollar has hit its highest level against the euro in almost three months after the Federal Reserve head said the US trade deficit is set to st...
2	Yukos unit buyer faces loan claim	The owners of embattled Russian oil giant Yukos are to ask the buyer of its former production unit to pay back a \$900m (£479m) loan. State-owned...	Yukos' owner Menatep Group says it will ask Rosneft to repay a loan that Yugansk had secured on its assets. State-owned Rosneft bought the Yugansk ...
3	High fuel prices hit BA's profits	British Airways has blamed high fuel prices for a 40% drop in profits. Reporting its results for the three months to 31 December 2004, the airli...	Rod Eddington, BA's chief executive, said the results were "respectable" in a third quarter when fuel costs rose by £106m or 47.3%. To help offset ...
4	Pernod takeover talk lifts Domecq	Shares in UK drinks and food firm Allied Domecq have risen on speculation that it could be the target of a takeover by France's Pernod Ricard. R...	Pernod has reduced the debt it took on to fund the Seagram purchase to just 1.8bn euros, while Allied has improved the performance of its fast-foo...

Figure 2 BBC news dataset header

The dataset consists of the following fields:

- **title**: the title for the respective article.
- **article**: the news article itself.
- **summary**: the provided summary of the news article. This will be used as the golden reference in evaluation metrics.

	count	mean	std	min	25%	50%	75%	max
article_length	2225.00	2232.79	1364.25	471.00	1414.00	1936.00	2774.00	25454.00
headline_length	2225.00	31.37	2.61	16.00	31.00	32.00	33.00	52.00
summary_length	2225.00	1000.56	638.43	227.00	624.00	862.00	1235.00	12344.00
summary_ratio	2225.00	0.45	0.05	0.19	0.42	0.45	0.48	0.68

Figure 3 Summary Statistics for Columns

Summary statistics of the BBC dataset as show above. Summary ratio are calculate using summary length divided by article length.

Article length (characters counts) is between roughly 1414 and 2774 characters, with the median at about 2232 and a long tail with outliers to the right.

By checking all text columns for null values by using `df.isna()` and compute a summary of the result:

```

title      0
article    0
summary    0

```

Figure 4 Checking for Missing Data

There are no abnormalities detected.

## Exploratory Visualization

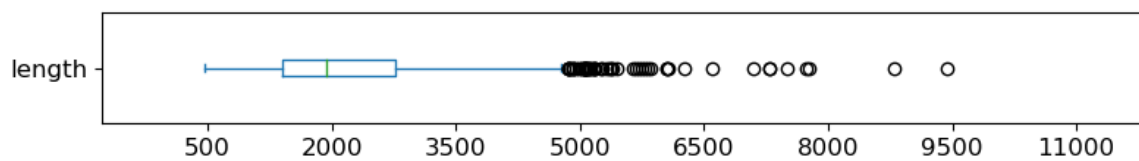


Figure 5 Plotting Value Distributions of articles words count

50% percent of the articles (the box in the middle) have a length between roughly 1414 and 2774 characters, with the median at about 2232 and a long tail with outliers to the right. The distribution is left-skewed.

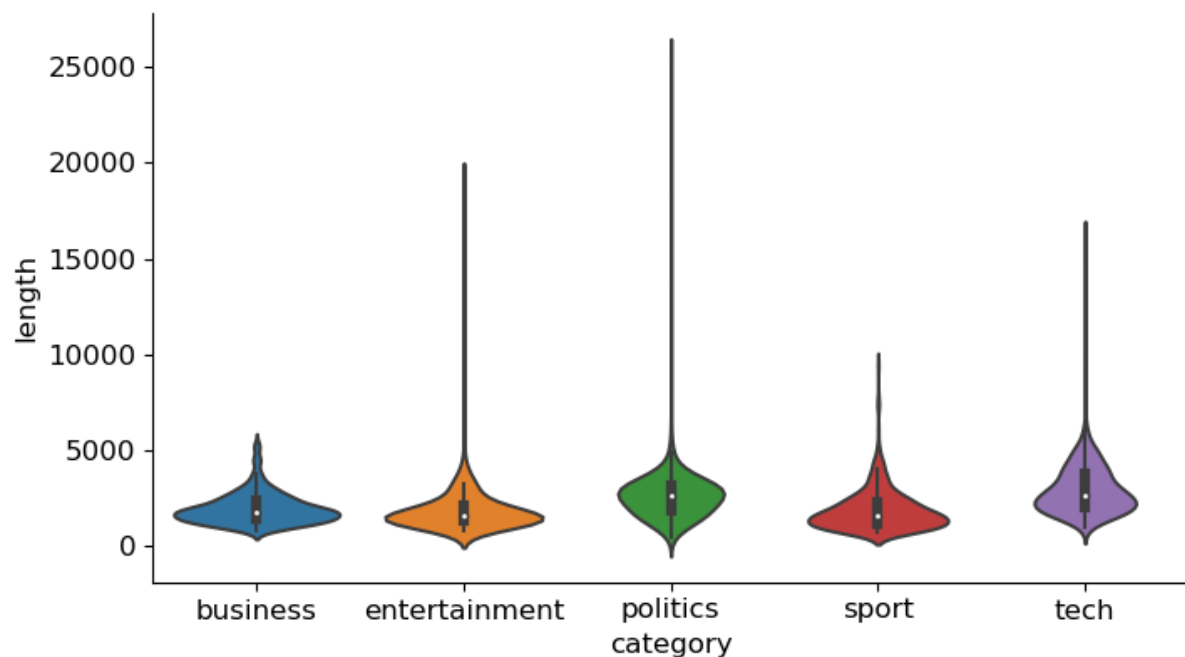


Figure 6 Comparing Value Distributions Across Categories

Peculiarities in the data often become visible when different subsets of the data are examined. The violin plot is the “smoothed” version of a box plot. Frequencies are visualized by the width of the violin body, while the box is still visible inside the violin. No significant peculiarities are noticed across the different categories, except for business where there are fewer outliers than the rest of the categories.

## Algorithms and Techniques

Create a new model by using **transfer learning** from a **pretrained T5 model** and **fine-tuning** with BBC News summary dataset.

The following parameters are used for the tokenizer:

- max\_input\_length: 1024
- max\_target\_length: 128

Let's discuss the algorithms and techniques used as below.

## Transfer Learning and Fine-Tuning

Transfer learning is the fastest way to perform common NLP tasks use pretrained language models. Transfer learning is possible because pretrained language models are neural networks. The first several layers of the neural network have already learned some useful representations of the data. This makes it easier to train subsequent layers for specific task.

## The background of the Transformer

The Transformer model was invented by Google Research and has toppled decades of Natural Language Processing research, development and implementations.

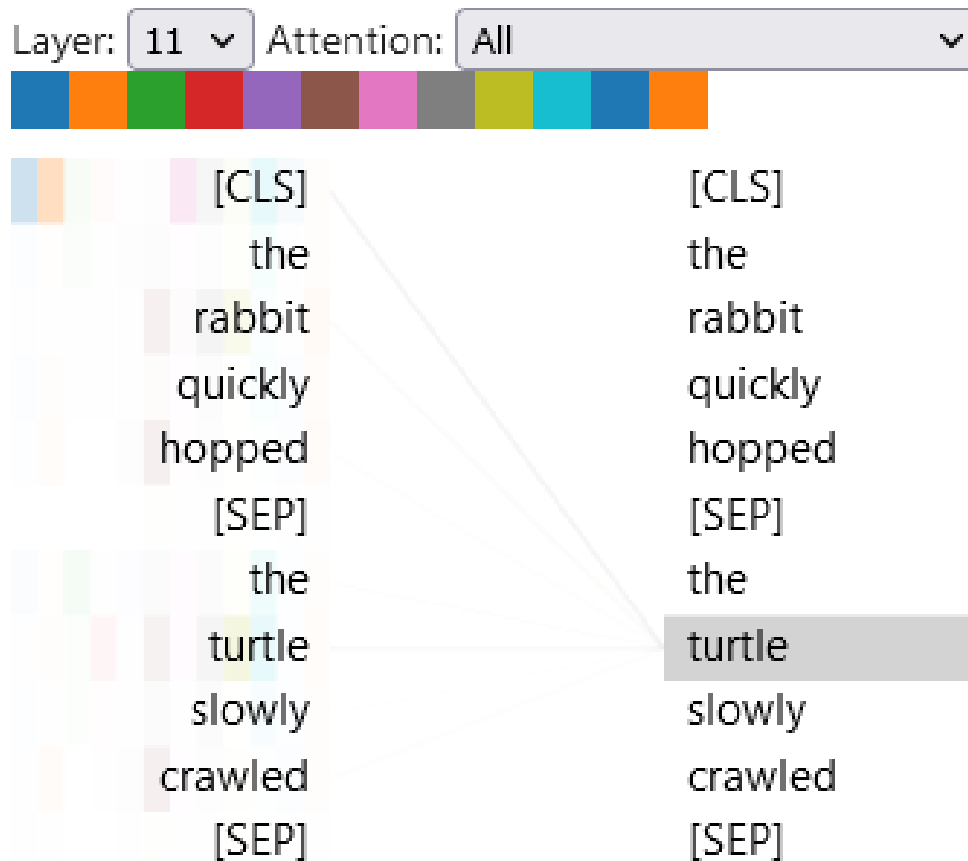


Figure 7 Multi-head Attention

The model uses an encoder-decoder architecture, where the encoder converts the input text into some intermediate numerical vector representation. The decoder converts this vector into output text. Attention vector are generated during the encoder process. The self-attention mechanism helps the encoder weigh the relevance of the words in the input text.

## The T5 Model

Text-to-Text Transfer Transformer, **T5** was introduced in the paper [Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer](#)<sup>6</sup>. **T5** is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format.

**T5** is an attempt to cast a broad range of NLP problems into a unifying sequence-to-sequence framework. It allows for the application of the same model, objective, training procedure, and decoding process for every task. Handled problem

<sup>6</sup> [Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer](#)



classes vary from summarization to sentiment analysis and question answering, among many others.

## Benchmark

Extractive summarization using **TF-IDF**<sup>7</sup> will be used as benchmark model. This is the simplest algorithm for automatic text summarization and serve as baseline performance using ROUGE as the evaluation metric. Based on the training and validation dataset, benchmarking the TF-IDF return ROUGE return **23.9**. This will be used as the baseline for further improvement.

---

<sup>7</sup> <https://en.wikipedia.org/wiki/Tf%E2%80%93idf>

## III. Methodology

### Data Preprocessing

The preprocessing done after “Load BBC News summary dataset” consists of the following steps:

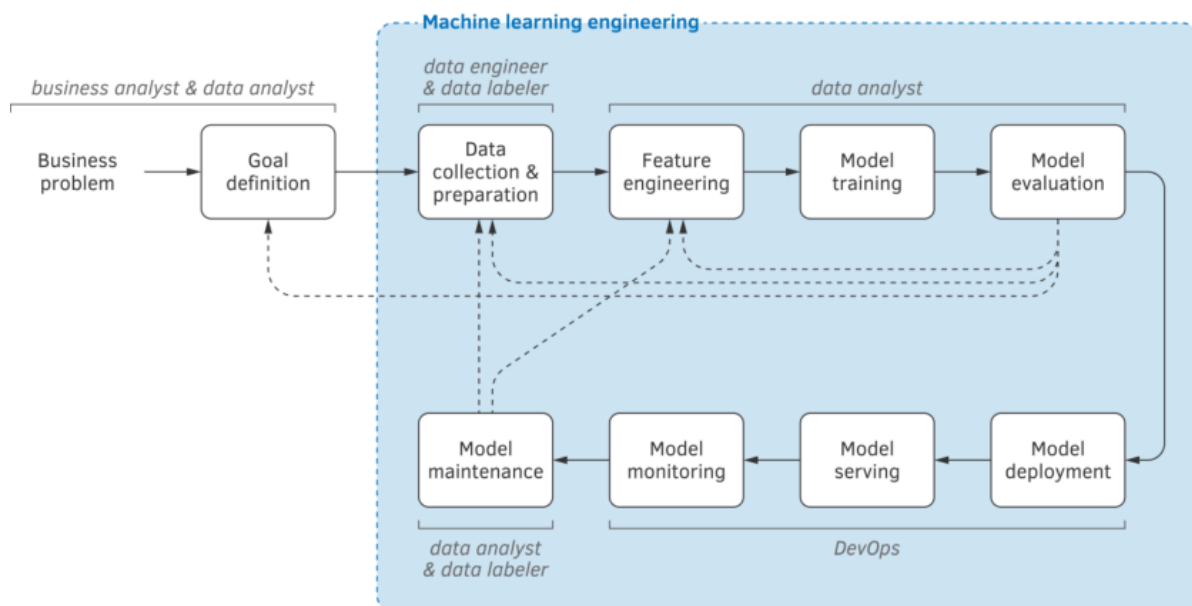
1. Removal of newline from text elements.
2. Partition the first line of the articles as title column.
3. Aligning the respective summary(label) with the news article(input).
4. Prefix inputs with the task “summarize: “.
5. Tokenize both the news article and summary.
6. Truncation of news article and summary.
7. Converting the tokens to their corresponding IDs in the pretrained vocabulary.

The preprocessing step 4 to step 6 are known as **Encoding**. To do all of this, instantiate the tokenizer with the “*AutoTokenizer.from\_pretrained*” method, which will ensure:

- A tokenizer that corresponds to the model architecture used,
- Download the vocabulary used when pretraining this specific checkpoint.

## Implementation

The workflow for approaching the solution is as in below figure:



8

The solid arrows show a typical flow of the project stages. The dashed arrows indicate that at some stages, a decision can be made to go back in the process for revision etc.

**Goal definition** and **Data collection and preparation** were in Problem statement and Data preprocessing as discussed above.

**Feature engineering**, this project requires 2 features: News articles and summaries. These 2 features are encoded and fed into the model in the next step.

**Model training** consist of fine-tuning the pretrained model using BBC news article dataset.

**Model evaluation**, **ROUGE** is defined and used as the evaluation metric to validate the model fine-tuning process.

**Model training** and **Model evaluation** are implemented using the following steps:

1. Define the pretrained model from Hugging Face transformer library. **T5** model is chosen and used in this project.
2. Define training parameters like batch size, evaluation strategy and learning rate etc.
3. Instantiate Seq2SeqTrainer with training parameters.

---

<sup>8</sup> Andriy Burkov: Machine Learning Engineering Figure 4 (2021)

4. Define the metrics computation function for model evaluation purpose.
5. Finetune the model by calling the train method.
6. Validate the training results against validation set.
7. Run evaluation metric and print results.
8. Upload the trained model to Hugging Face Hub.

## Model deployment and serving

The following models are deployed:

Model	Description
<b>TF-IDF</b>	Extractive summarization model
<b>Default Pipeline</b>	Hugging face summarization pipeline default model
<b>T5-base</b>	T5 base model
<b>T5-base finetuned</b>	T5-base finetuned with BBC News Summaries dataset
<b>T5-base Headline</b>	T5-base finetuned with BBC News Title dataset

Figure 8 Deployed models

**FASTAPI**<sup>9</sup> will serve as the backend to deploy and serve the solution model. This will respond to POST request by:

1. Load the fined tuned model from Hugging Face Hub.
2. Encode request body.
3. Run the model predict method.
4. Decode model respond.
5. Respond to the POST request with summarized text.

**REACT**<sup>10</sup> will serve as the frontend for this project web application. This will POST to the backend with user request and display summarized text and metrics. React Hook and Material UI are used for state management and UI respectively.

---

<sup>9</sup> <https://fastapi.tiangolo.com/>

<sup>10</sup> <https://reactjs.org/>

Model

Headline

Select model to use

Raw Text \*

Risks from a resurgence in Covid-19 cases and a possible tightening of measures both locally and abroad has not derailed RHB Group Research's optimism on Singapore's equity market.

Reference Text

Risks from a resurgence in Covid-19 cases and a possible tightening of measures both locally and abroad has not derailed RHB Group Research's optimism on Singapore's equity market.

SUMMARIZE

Summarized List

Headline summarized: a resurgence in Covid-19 cases and a possible tightening of measures both locally and abroad has not derailed RHB Group Research's optimism on Singapore's equity market.
rouge1 Precision:1.00 Recall:0.35 fmeasure:0.52 rouge2 Precision:1.00 Recall:0.33 fmeasure:0.50 rougeL Precision:1.00 Recall:0.35 fmeasure:0.52 rougeLsum Precision:1.00 Recall:0.35 fmeasure:0.52

Figure 9 REACT Frontend

Finally, **model monitoring** and **model maintenance** will be outside the scope of this project.

## Refinement

To improve the summarization model with **TF-IDF** f-measure 23.9 as the baseline. The final model was derived by fine-tuning the **T5** model with BBC News summary dataset, achieving f-measure 30.7.

Table 1 Benchmarking various models

Model	TF-IDF	Default Pipeline	T5-base	T5-base finetuned
<b>ROUGE</b>	23.9	24.3	20.6	30.7
<b>f-measure</b>				

## IV. Results

### Model Evaluation and Validation

Validation set was used to evaluate the model.

The final hyperparameters were chosen based on trial and error.

The follow list the hyperparameters chosen:

- evaluation\_strategy: "epoch"
- learning\_rate: 2e-5
- weight\_decay: 0.01
- num\_train\_epochs: 1

To test the robustness of the final model, test was done using local news article from various domains. The following observations are listed:

- T5 headline model has the highest ROUGE f-measure score consistently.
- Sports domain produce the least consistent results.
- Abrupt end in sentence.
- ROUGE does not take into account the sentence structure or grammatical correctness.

### Justification

After deploying the frontend web application, the following observation were made:


- The first run of any model will take the longest duration to respond. This is because the models are downloading to the backend.
- The classification delay is about 12 seconds after the first run.

The summarization model is able to generalize across domains.

With reference to table 1, the summarization model with f-measure score **30.7** perform better than the benchmark model f-measure score **23.9**.

## V. Conclusion

### Free-Form Visualization

 Text Summarization

Model

Headline ▾

Select model to use

Raw Text \*

Singapore's equity market looks optimistic, despite risks: RHB  
Amala Balakrishner Published on Wed, Oct 20, 2021 / 8:06 PM GMT+8 / Updated 12 hours ago  
Singapore's equity market looks optimistic, despite risks: RHB - THE EDGE SINGAPORE

Follow us on Facebook and join our Telegram channel for the latest updates.  
Risks from a resurgence in Covid-19 cases and a possible tightening of measures both locally and abroad has not derailed RHB Group Research's optimism on Singapore's equity market.

"For 4Q2021, market outlook will continue to depend on how well stocks and sectors can contend with the current Covid-19 constrained growth [and] elevated inflation amidst supply chain disruptions, expectations of an early interest rate hike and corporate efforts to maximise operational efficiencies," analyst Shekhar Jaiswal explains in an Oct 20 note.

He adds that Singapore continues to offer opportunities to accumulate stocks that leverage on an economic re-opening.

Counters that offer better earnings visibility either from business restructuring or structural and inorganic growth, look promising too, reckons Jaiswal.

For now, he has taken a liking for banks as well as the consumer, healthcare, industrial, telecom and transport sectors.

This is based on "expectations of a better 2022," elaborates Jaiswal.

The analyst views Singapore's economic reopening as an eventuality that will sustain over the next 12 months.

His comments came as the Covid-19 cases here are typically seen amongst patients are either asymptomatic or have mild symptoms.

The resultant simplification of testing protocols and the expansion of Vaccinated Travel Lanes (VTLs) to 11 countries are also steps closer to Singapore's reopening.

Meanwhile, Jaiswal observes that market valuation is compelling again.

"While the Straits Times Index (STI) underperformed in 3Q2021 as investors awaited clarity on normalisation of business activities, the gradual addition of Sea (SE US, NR) into the MSCI Singapore Index has put a pause on the strong price performance for banks and the STI's forward P/E of above +1SD – a level that the STI has struggled to sustain since Jan 2008," he explains.

Jaiswal adds that the benchmark index – which is at 12.8x forward P/E – is trading below its historical average and remains the cheapest market in Asean.

His target is that the STI will end the year at 3,420 points, down from his previous 3,410 point prediction. This is based on 14.0x forward P/E.

The STI closed at 3,198.08 points on Oct 20, down 0.93 or 0.03%.

SUMMARIZE ➤

Reference Text

Singapore's equity market looks optimistic, despite risks: RHB

Figure 10 News article using headline as reference

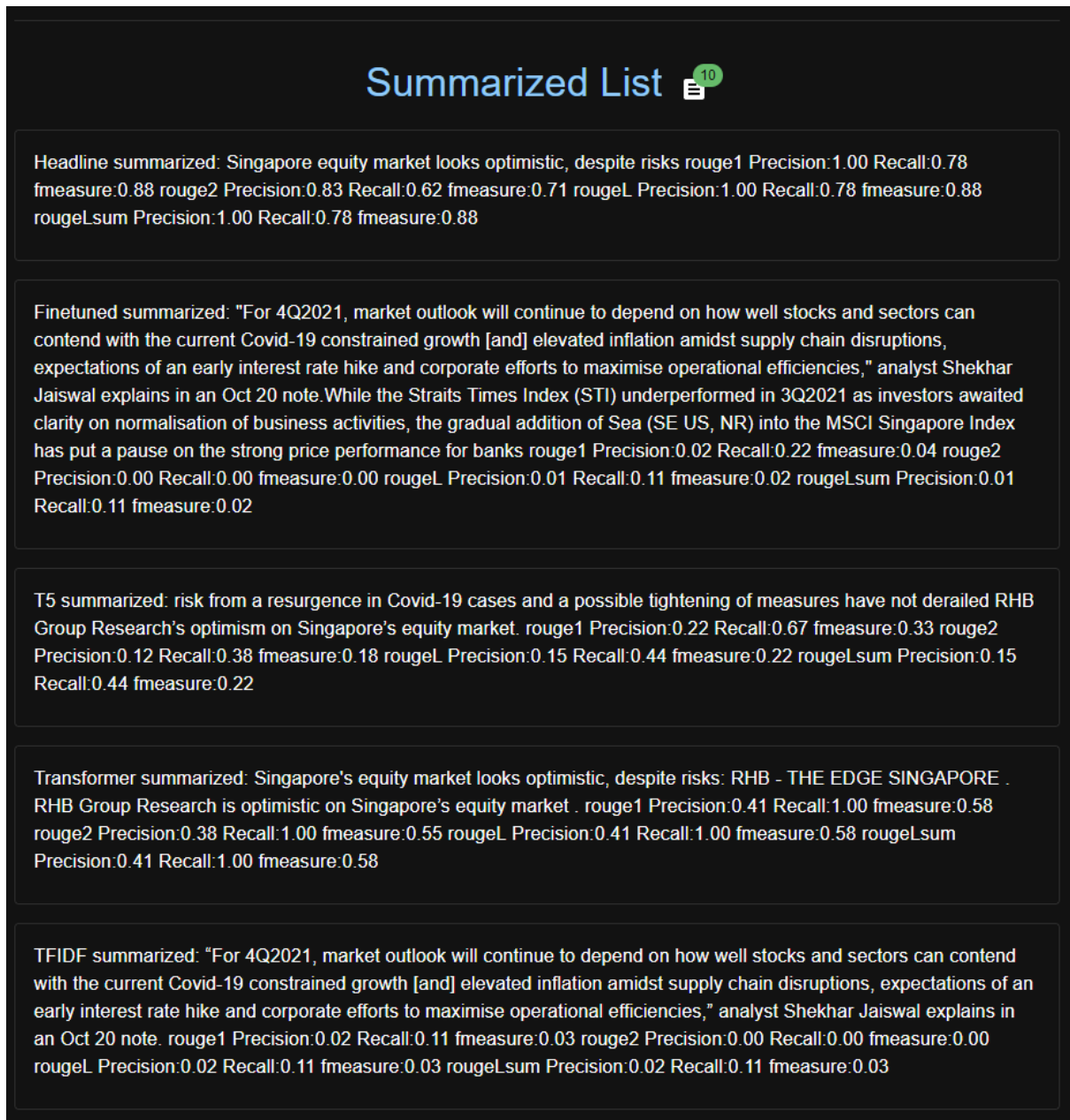


Figure 11 Results from various models deployed

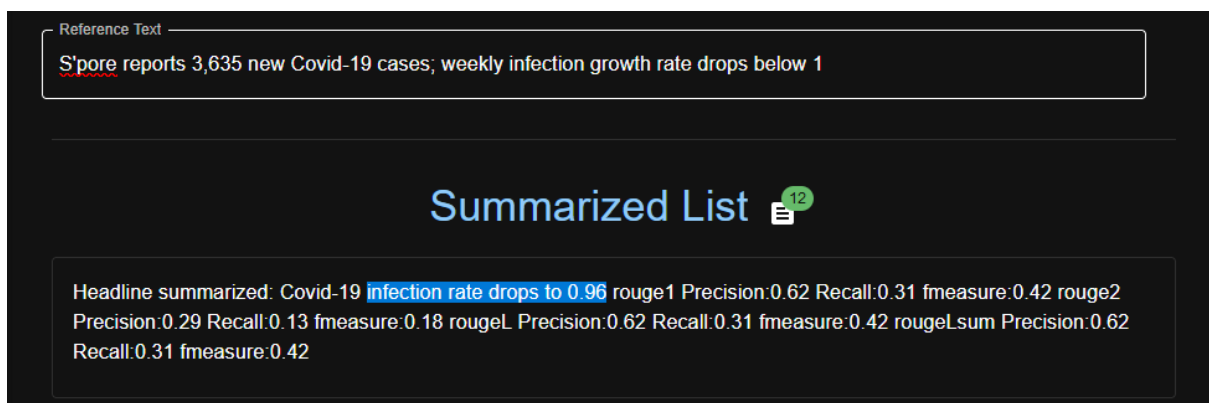


Figure 12 Interesting generalization result



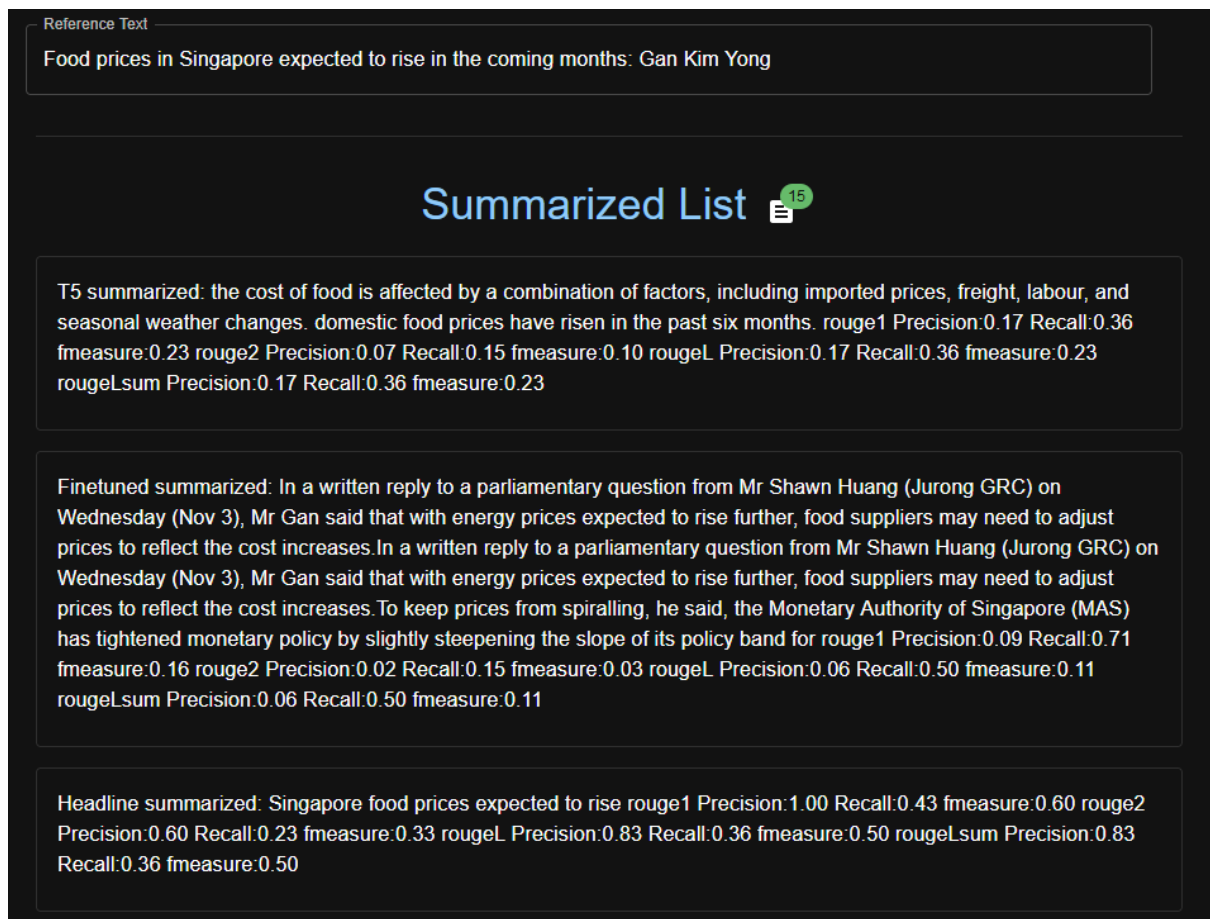


Figure 13 Headline model able to give more consistent result

## Reflection

The arrival of Transformer has created a new generation of easy-to-use artificial intelligence models. Hugging Face for example, make NLP tasks easy to implement with a few lines of code. However, understanding the algorithm and techniques behind Transformer is not easy. Transformer architecture replaces RNNs, LSTMs and CNNs from transduction problems and sequence modeling. It also produces innovations, such as positional encoding and masked multi-headed attention.

As Transformer is a black box system, I am not able to deduce how Headline model are able to produce better performance.

Transformer has allowed this project to be built swiftly and enable end-to-end build from model creation, backend and frontend implementation within short amount of time-frame. I have learned how to solution an end-to-end product using FastAPI and React.

## Improvement

Without hardware limitation, this project can be improved by using the largest T5 model – **T5-11b**. **T5-11b** model contains 11 billion parameters and require over

40GB of GPU memory. To achieve the best user experience, accuracy have to be balanced against processing speed.

This project can also be improved by enabling feedback (ratings etc.) of summarized results from the users. This will provide a valuable source of labeled feedback loop.