UNIVERSITÀ CATTOLICA DEL SACRO CUORE

INTERFACULTY ECONOMICS - BANKING, FINANCE AND INSURANCE SCIENCES MASTER OF SCIENCE IN STATISTICAL AND ACTUARIAL SCIENCES DATA ANALYTICS FOR BUSINESS AND ECONOMICS

1921 UN SECOLO DI STORIA DAVANTI A NOI —2021

UNIVERSITÀ CATTOLICA del Sacro Cuore

Abstractive Text Summarization

Student: Ogulcan ERTUNC
4811415

Supervisor: Prof. Andrea BELLI

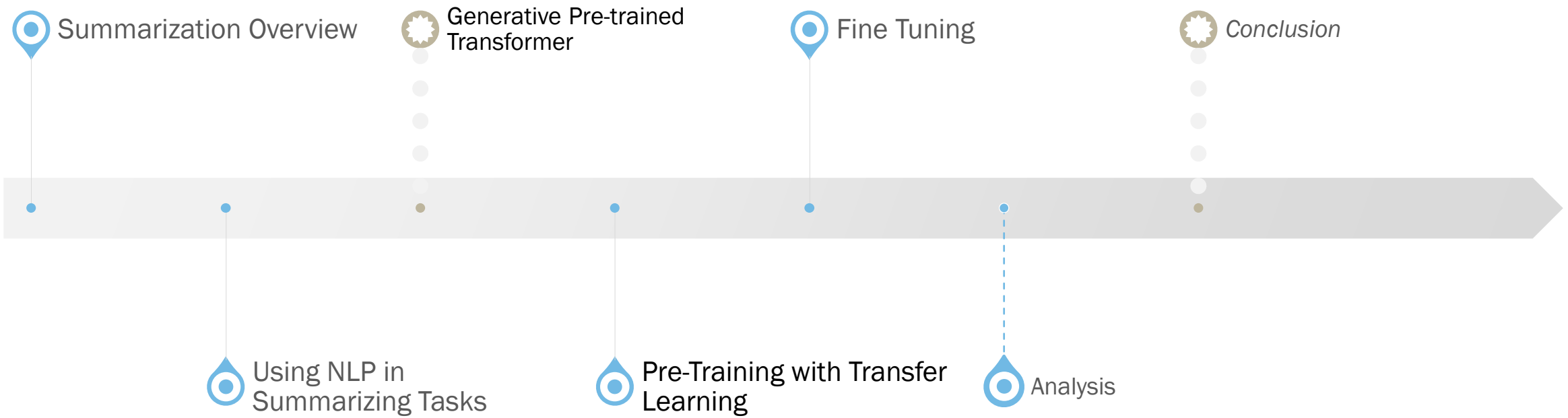
Research Object

Using GPT2 and Transformers to create an **abstractive summary**.

Research Question

Learning that **meaningful** and **unique human-made summaries** can be fine-tuned by training the GPT-2's text-rendering capabilities with a dimensionally small dataset that contains almost any new text, **using the transformer** originally designed for translate work.

METHODOLOGY/PROJECT STEPS



EXTRACTIVE VS ABSTRACTIVE SUMMARIZATION

- ## Extractive Summarization:

  Extractive summarization essentially involves the extraction of certain text fragments based on predefined weights assigned to important words, where the choice of the text depends on the weight of the words in it. Usually the default weights are assigned according to the frequency of occurrence of a word. Here, the length of the summary can be changed by defining the maximum and minimum number of sentences to be included in the summary.

- ## Abstractive Summarization:

  Whereas abstract summarization involves intuitive approaches to training the system to try to understand the whole context and to create a summary based on that understanding. This is a more human-like way of generating abstracts, and these summaries are more effective than inferential approaches. However, creating this summary is very difficult and complex.

MODEL ARCHITECTURE

Our project has 3 sub-tasks,

- Pre-training through data set
- Fine tuning on the same data set
- Obtaining the summary.

Pre-Training

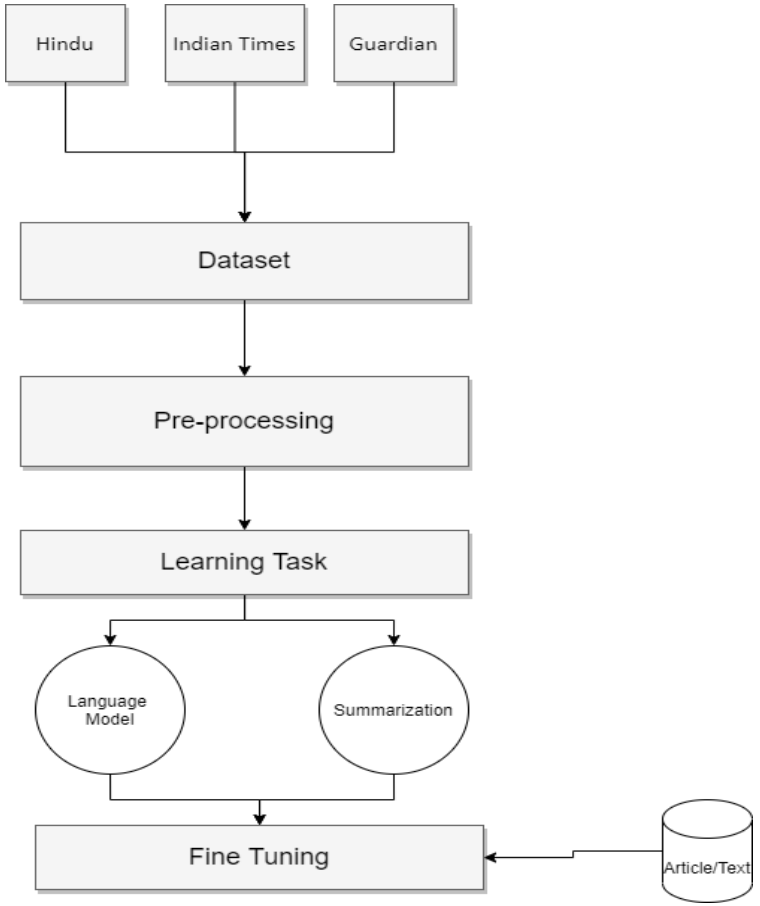
Since the dataset we will use to train the model comes from sources such as Hindu, Indian clocks, and Guardian, certain parts of the dataset will fail, or the source text will not be consistent at times. We tried to make a summary by combining the sentences that emerged by establishing relationships between the texts in our data set and the keywords we obtained from them.

Fine adjustment

Using the pre-trained model, we will fine-tune the network as per our request. We used transformer for this, mostly two types of transformer architecture are highlighted, transformer encoder and transformer decoder.

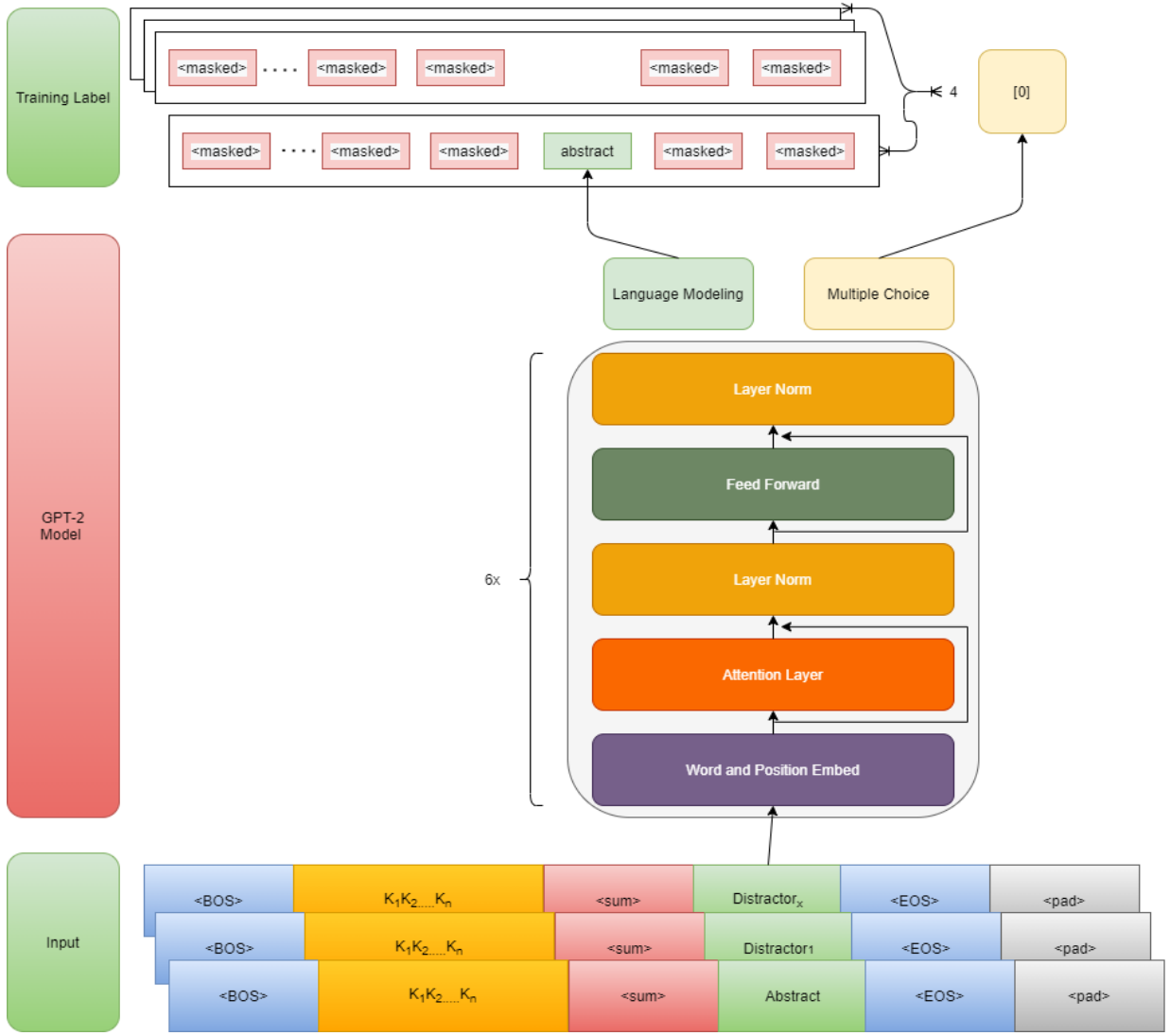
Getting the summary

We created our summaries as a result of the model we trained using the pre-trained GPT-2 architecture.

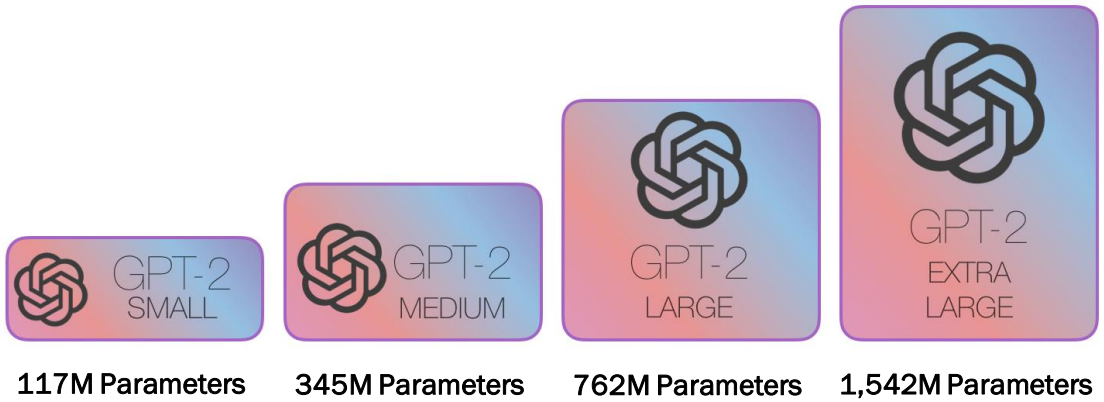


PRE-TRAINING WITH TRANSFER LEARNING

- The main challenge is the hassle of finding appropriate data to train and mask the classifier.
- We designed our GPT-2 model to train on 2 main tasks:
  - Language modeling task : it projects the hidden state onto the word embedding output layer. By applying cross-entropy loss to a text with the corresponding keyword, we get the LM loss value.
  - Multiple choice prediction task: we added the hidden state of the token in the form of "end of text" to the end of the text for the probability score we will use in the future, this was done for a simple classification task.
- Language modeling education tags are the symbol of the summary scrolled to the right by 1 token. This is because, the GPT-2 also drops automatically on its own, and the nth token output is generated from all previous n-1 token entries to the left.
- The multiple-choice education tag is the tensor of a numeric i that specifies the i'th element, the correct keyword hash pair.

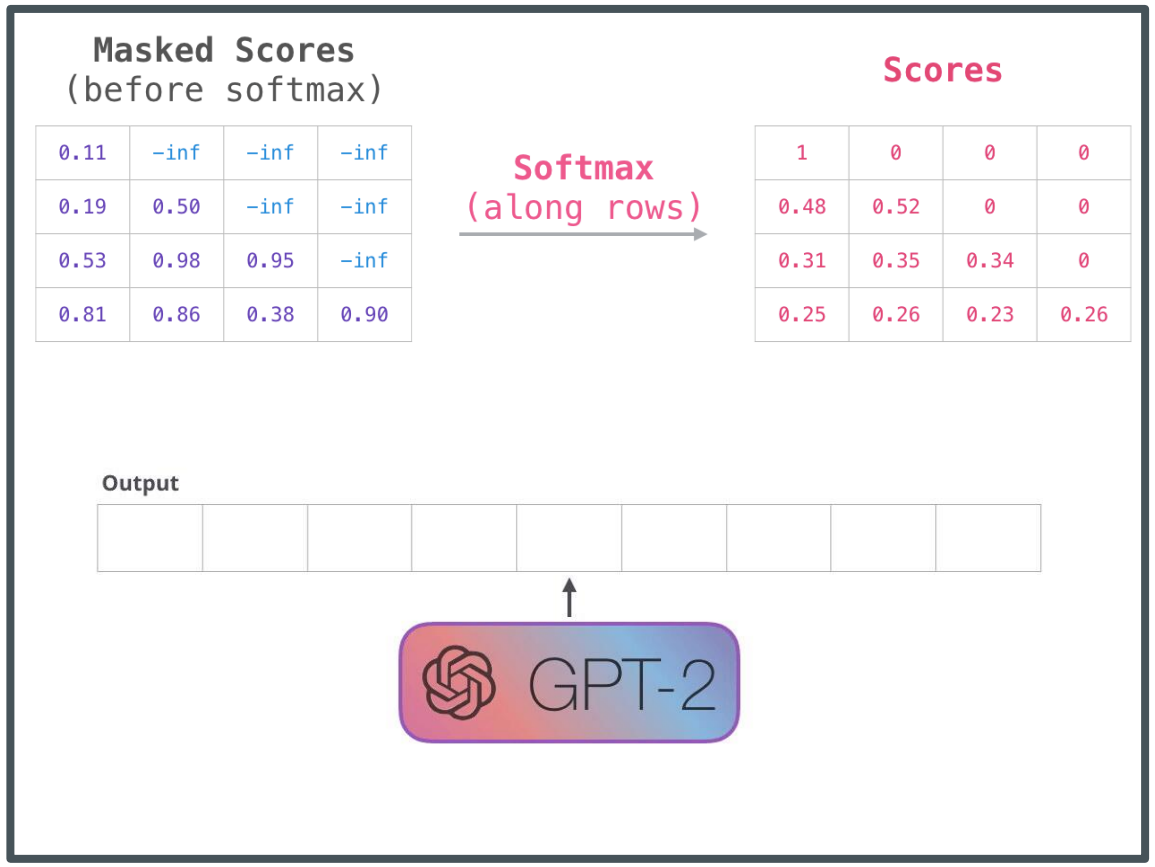


MODEL TRAINING



Due to our hardware and speed requirements, we trained our model using the DistilGPT2 version. We received training for up to 5 epochs in total. The training data set consists of 3246 training samples, each sample has 4 randomly split options. The validation dataset consists of 1572 samples, each with 4 randomly split options.

MODEL TRAINING SEQUENCE GENERATION



- The output generated by modeling GPT-2 is a tensor containing line length and word size. This is the tensor of a probability distribution that contains all the words before SoftMax.

$$P(X) = \prod_{i=1} P(X_i | x_1, x_2, x_3 \ldots . x_{i-1})$$

- First, before sampling, we can apply a scaling factor called temperature (t), which is likely to reshape the skew probability distribution before softmax(u).

$$P(X) = \frac{\exp(u/t)}{\sum l^* \exp(u_l^*/t)}$$

The high temperature dispersion caused by low probability words tends to distort; it tends to dissipate with low temperature, leaning towards high probability words.

- Afterwards, we applied the method called top-p sampling as conditional selection from those with high temperatures.

ANALYSIS

ROGUE stands for Recall Focused Understudy for Gisting Assessment. It is a measure for measuring the quality of a summary produced by the summarizer we have produced for this thesis against a human-made summary.

$$ROUGE - n = p/q$$

TEXT	Rouge 1			Rouge 2			Rouge L		
	Precision	Recall	Fmeasure	Precision	Recall	Fmeasure	Precision	Recall	Fmeasure
Epoch = 1	0.47	0.34	0.39	0.21	0.15	0.17	0.23	0.17	0.19
Epoch = 5	0.81	0.74	0.77	0.5	0.46	0.47	0.62	0.57	0.6

CONCLUSION

- Abstractive summarization is still a major challenge for natural language processing. We tried to eliminate this problem as much as we could, using a successful pre-trained model such as the GPT-2.
- In our model, we showed that the text-to-text summarization, multiple loss training strategy can be made, and our result is interpretable and logical.
- And the summaries created are to the human level performance level we want. For more successful summaries, we think keywords can be concentrated on verbs, nouns and adjectives. If we could add more information to the keyword, such as the adjective part, we could also explore how much more accuracy could be achieved. Apart from that, success can be achieved by clustering similar news and randomly adding common words to keyword sets. Or, unlike all these, we think that the ability to obtain keywords with BERT-style methods and summarize the texts more successfully can participate in keyword selection in order to increase the success of the model in the future.

THANK YOU FOR YOUR ATTENTION