# ELEC-E5550 - Statistical Natural Language Processing D
# News Article Summarization

Pranav Arora (pranav.arora@aalto.fi) (Student ID:1012372),
Yuvrajsinh Chudasama (yuvrajsinh.chudasama@aalto.fi) (Student ID: 1014244)

April 2021

# 1 Introduction

As we are moving towards big data, a lot of information is currently out there on web and different text resources. Therefore, finding relevant information is becoming a non-trivial task. Hence, we set the goal of this project is to implement text summarization of news articles to generate a clear and concise summary. This summary generation tool may find application in different areas such as academic articles, new articles or general web blogs.

Primarily, there exists two types of summarization techniques: extractive and abstractive summarization. Extractive summarization, as the name suggests, is a technique employed to extract important key words and sentences. This can be done using unsupervised strategies such as k-means, k-medoids and spectral clustering. While, the abstractive summarization focuses more on learning the language structure where the extractive summarization is a pre-processing step. This is done in a supervised manner. Key problems targeted will be topic identification and abstractive summary generation. We are employing three models i.e. BERT fine-tuned for extractive summarization, BART fine-tuned for abstractive summarization and a pretrained BART model. These three models are chosen as such because of their prior performance in the abstractive and extractive text generation tasks. Finally a a comparison is performed with the ground truth human summary using the commonly used method - Recall-Oriented Understudy for Gisting Evaluation(ROUGE) metric.

# 2 Literature Study

Text summarization is a non-trivial problem which needs both language understanding and text generation. Two kinds of summarizations exists: extractive and abstractive. It is approached by researchers primarily using three methods - cue, title and location. [1].Iqbal.et.al provide deep insight into the various approaches - probabilistic modelling, frequency based modelling and machine learning, for different kinds of texts. Each of the approach has some drawbacks although, the probabilistic models tend to capture the context better than others. Machine learning based method treat the summarization problem as a classification problem, which can be challenging in the sense that it requires reference summaries - constructing and bench-marking such a data-set is cumbersome. Wong et.al. introduced a semi-supervised approach which for the most part trains on unlabeled data with some label data [2].

For scientific paper summarization, the key method is to emphasize on the sentences where papers are cited as they contain the most relevant and concise description of the topic. [1] Hugging-face language models almost finds application in most of the linguistics tasks. Lu et.al. applied BERT to produce concise summary. [3] BERT is trained on the notion of masked language modeling where the model tries to predict the masked words. Lu.et.al showed that while co-reference resolution improves the the process of extractive summarization. Using BERT did not improve the performance justified by the ROUGE score. [3] [4]

Yasunaga et al created a hand annotated summary corpora for scientific papers. [5] They further describe methods for summary generation which provides concise description of the author's contribution and related works. They used novel hybrid methods focused on summarizing the abstract and cited text and then use adding text to the abstract from the cited text to the abstract. This methodology is based on the graph based structure to identify relation(edges) between sentences. Further a authority score is used highlight the importance(weight )of a particular reference paper. Then a graph convolution network(GCN) [6] is applied to the input and their relation graph. [5]. Based on the hybrid methods described above, the sentences are selected to be part of the generated summary. The evaluation metric on which gradient descent is then performed to minimize the cross-entropy loss using the ROUGE score between the generated summary and the gold summary - given experts. [5]

Tan et.al. used BERT and GPT2 pretrained models from hugging-face to summarize Covid-19 research text articles. [7] As expected a general practice for evaluation is to use the ROUGE score. The key problems addressed by authors is low resource challenge and problems with generating abstractive summarization using pre-trained language models such as BERTSUM. The model is divided into two parts first is the unsupervised algorithm such as k-medoid applied to the token embeddings generated by BERT. Then input of keywords and gold summary is given to the GPT-2. As GPT-2 is an autoregressive generative model it learns from the previous sequences using the masking mechanism. Therefore, training is done using a special token summarize to signify the contents to be included in the summary

Ghademi et. al. employ an interesting hybrid methodology combining extractive summarization with abstractive summarization using multiple documents. [8] Multiple documents are used to generate the extractive summary, redundancy is removed with the determinantal point process. [8] The abstractive summaries are then generated by giving the exctractive summary as input to the generative model such as T5 or GPT. Again expert evaluations and ROUGE score is used to estimate the validity of the generative summary. The model achieved a ROUGE 1 - 41.85 and ROUGE 2 - 12.17 which is on-par with the state-of-the-art(SOTA) summarizers.

For evaluation of the generated summary, the most common metric is ROUGE score and expert evaluation. ROUGE scores measures the similarity(overlap) of the n-grams between a reference summary and generated summary. [4] Different modifications of ROUGE metric such as ROUGE-1, ROUGE-L, ROUGE-W are available. Mostly commonly used metrics are ROUGE 1, ROUGE 2 and ROUGE L. [4]

All the literature points to the problems in abstractive summary generation and the limited methods of the evaluation score. Some interesting architecture were observed coming limitations but the summary generation still seems to be a non-trivial task.

# 3 Methods

## 3.1 Extractive Summarization

### 3.1.1 BERT

The Transformers architecture has become a state of the art model since its inception in 2017. The Transformers use a bidirectional Encoder and Decoder based model with Positional Encoding, Multi-headed Attention and Feed Forward Network to learn contexts quickly from both the sides of a sentence. This kind of a model has been a significant improvement compared to the traditional RNN and LSTM based models in terms of fast and efficient training as they can process sentences as a whole and there is no need of recursion.

One of the most famous and used models nowadays called BERT which stands for Bidirectional Encoder Representations from Transformers is basically a set of these Encoder blocks from Transformer model stacked together. The basic BERT model called BERT base consists of 12 of these Encoder layers which results in around 110 Million parameters whereas an even large model called BERT large consists of 24 Encoder Layers which results in around 340 Million parameters. Due to the large number of parameters, it can take a significant time to get trained for a task. To solve this training problem, these BERT models are usually pre-trained on a large amount of data and then for a specific task it can be fine-tuned using the task-specific data.

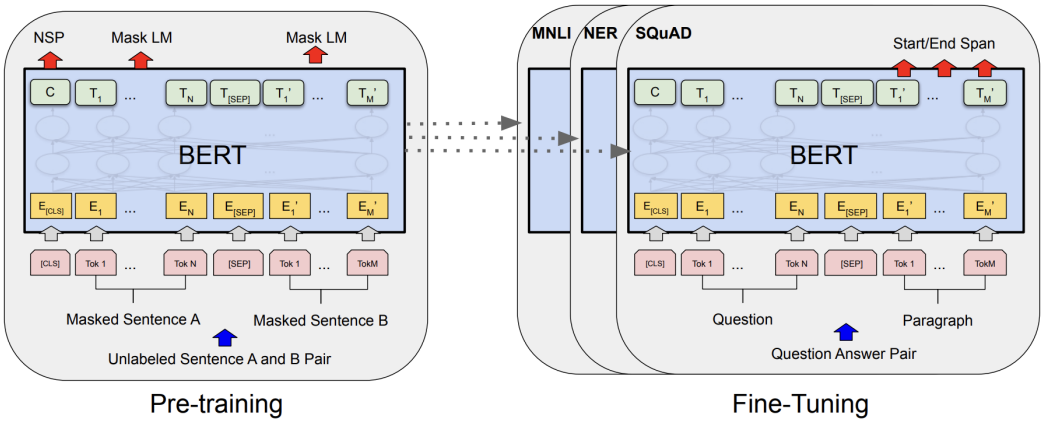


Figure 1: Training procedure for BERT [9]

Two different kind of methods are used for training these BERT models. The first is a masking based technique in which we supply sentences as input to the model with some of the words in those sentences are masked. The model then tries to predict these masked words and it can be trained using the loss between the actual masked word and the predicted words by the model. This training procedure helps the model learn context of the sentence from both left and right side of the sentence and thus the model get contextual language understanding. However, this kind of training process is not useful in making the model learn

contexts between two different sentences. To do that, a binary classification based approach is used where two sentences are fed to the model and the model tries to learn if the second sentence follows the first one or not. In this way, the model learns contexts between different sentences. In practice, for pre-training both of this procedures can be done simultaneously where two masked sentences are fed to the model and the model tries to predict the masked words as well as the fact whether the second sentence follows the first one. This training procedure as explained in [9] is described in Figure 1.

In the case of Extractive Summarization, we need to convert sentences into some numerical vector representations in a vector space so that we can apply clustering based approaches on it. Therefore, a non-generative model like BERT which can convert text sentences into embedding is sufficient for the extractive summarization process. There are other basic techniques available to convert text sentences into numerical vector representations such as bag-of-words or TF-IDF representation. However, these basic approaches usually return significantly sparse vectors which are not efficient for further steps and they also do not contain the context information precisely. On the other hand, BERT like model trained on a large data can produce dense embeddings for sentences with contextual information also encoded within [10].

After computing the embedding we have sentences represented in an n-dimensional Vector Space. Now we can do all kind of mathematical calculations with them like computing pairwise distances, clustering etc. For the purpose of extractive summarization, first clustering algorithm is applied to the generated embeddings. Here, the number of clusters formed is dependent on how long summary you want to generate. After clustering, we have some K cluster and their centroids which represents basically the whole text. Now, to find the actual embeddings, we can use K-Nearest Neighbor algorithm. In short, we can find 1 nearest neighbor to each K cluster centroids and converting these neighbor embedding again into the text, we get K sentences which summarizes the given text. The whole process of extractive summarization is shown in Figure 2.

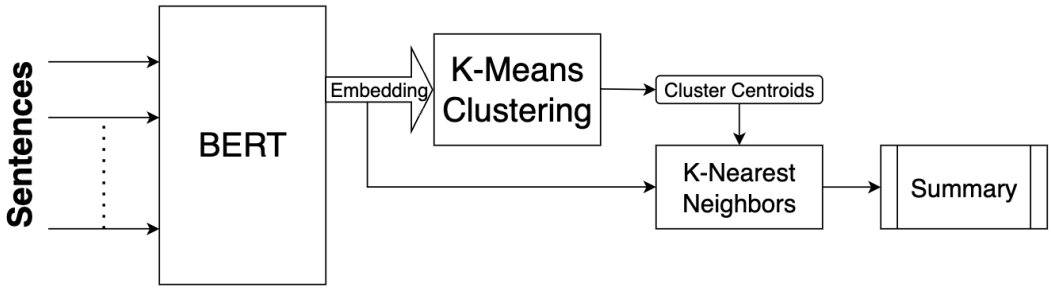


Figure 2: Process of Extractive Summarization

## 3.2 Abstractive Summarization

### 3.2.1 BART

As with any abstractive summarization task we need at least one generative model where we manually extract the text and feed the generative model with keywords, and important sentences. BART is a denoising autoencoder for pretraining sequence-to-sequence models. As sequence to sequence model the BART model comprises of two modules i.e. BERT(Bidirectional Encoder Representations from Transformers) as an encoder and GPT(Generative Pre-trained Transformer) as a decoder. The GPT model as the decoder is responsible for the summary generation task. On high level BART is trained by introducing noise in the given input text and the model is tasked with learning to rebuild the original text from the corrupted text. Although each sub-unit is performing a specific task, BERT masks some percentage of the input and then predicts the missing token independently. While the GPT learns to predict tokens auto-regressively which simply means based on the previous tokens, it outputs the next token. Now the interesting thing here that the inputs to the encoder can be arbitrary without keeping the the track of the decoder outputs, which helps in introducing noise to the input information and then the GPT learns to extract original text from the noisy text.

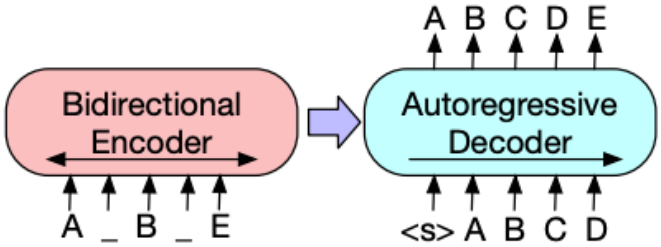


Figure 3: modules in sequence to sequence BART [11]

For pre-training BART, the input passes through a series of steps starting from tokenization. As the encoder part of BART is built using BERT, it follows a standard BERT architecture with minor modifications such as 6 layers in the encoder, decoder uses cross-attention on the final hidden layer. Also, BART's encoder differs with BERT by 10% of more parameters.

After the BERT procedure, from the outputs sampling is done and tokens are replaced with masks. Then random tokens are truncated from the input. Here the objective is to make the model learn where relative position of the missing inputs. After that text infilling is done to enable the model for predicting the number of tokens that are not present in the input. To capture the generality of the input, randomly the sentences are shuffled, finally document are rotated to teach the model to identify the document using the starting word in the document.

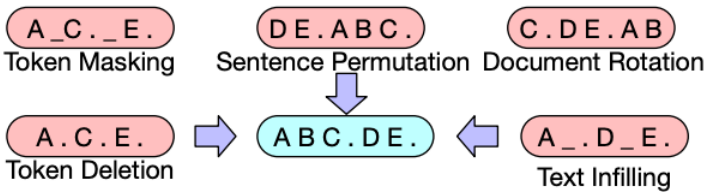


Figure 4: Noising the input for BART [11]

#### 3.2.2 Pre-trained BART-Large(pipeline methodology)

The pipelines are an effective method to use myriads of pre-trained models for inference. They are basically high-level APIs which abstract out most of the code and simplify inferencing for a variety of tasks apart from text summarization, such as text classification, question-answer generation. This follows the same setup as described above in the BART, the only difference here being that this model is pretrained on a different news article dataset i.e. CNN daily mail dataset and the number of layers in the encoder and decoder i.e. 12. Further, as fine tuning for this model is already performed, this model is used directly for inferencing from the pipeline API by huggingface.

### 3.3 Evaluation metrics

Evaluation methods for a model generated summary are limited. Recall-Oriented Understudy for Gisting Evaluation(ROUGE) is used in all the literature along with expert evaluation. Basically ROUGE compares the model generated summary with a gold summary written by human using the overlap of N-grams. Rouge-1, Rouge-2 measure the overlap of uni gram and bi gram respectively. ROUGE-L considers the longest common sub sequence which is common to most of the sentences in the gold summary. Python has a great library for rouge score calculation which is used in our implementation. Mathematically ROUGE-N is computed as follows:

$$ROUGE_N = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count(gram_n)}$$

Here n denotes the length for n-gram and $Count_match(gram_n)$ is the maximum n-grams which co-occur in a model generated summary and a set of reference gold summaries.The rouge score is based on the recall measure.

## 4 Dataset and Experiments

### 4.1 Dataset

The BBC news summary article dataset we used is from Kaggle. This particular dataset is ideally used for extractive summary as per the description on kaggle. It contains 2224

news articles from different topics along with their gold summaries which will be used for calculating ROUGE score and performance. Following topics are presented in the dataset:

1. Business
2. Entertainment
3. Politics
4. Sports
5. Technology

In the preprocessing step, the article and summary were read from their respective files and the text in them were represented as a list of sentences. In all the sentences, some of the special characters such as foreign language characters, new line characters and unnecessary white spaces were removed. The preprocessed text was stored in a dataframe which had for four columns. One each for topic of article, article number, text in the article and summary in the article.

## 4.2 Extractive Summarization

### 4.2.1 DistilBERT and K-means

In our experiments, to convert sentences into embeddings, we used a pre-trained DistilBERT model instead of BERT base model. This was done to speed up the process of generating embeddings and thus to speed up the whole extractive summarization process since we wanted to focus more on the abstractive summarization. The DistilBERT model has a size which is 40% less than the original BERT model. However, the language understanding capabilities of DistilBERT is 97% of BERT and it is also 60% faster in generating embeddings than the BERT base model [12].

DistilBERT takes as input tokenized sentences with a specific length MAX_LEN. For tokenizing our sentences from the article, DistilBERTTokenizer was used which is compatible to DistilBERT and also adds the special characters related to start and end of the sentence in the tokenization process. The MAX_LEN parameter was set to 256 which means that a sentence can only have maximum 256 words. However, in the data the sentences can have less than 256 or more than 256 words. To resolve that, the sentences are padded with zeroes if they have less than 256 words and truncated if they have more than 256 words. To supply this padding information to the model, attention masks were also created which specifies which values are words and which are just zero paddings.

The process followed for extractive summarization is the same as mentioned in Figure 2. Embeddings were generated from the DistilBERT model and then the embdeddings were clustered using K-Means clustering. We also need to supply how many clusters we want which is basically how many sentences we need in out extractive summary. In our

experiments, we used a length of 7 sentences for the summary. After clustering, these summarizing sentences were found as the nearest neighbor each cluster centroid.

This is a case of Unsupervised learning and thus, we do not need to make different training and testing sets from the data. All the available data was used to generate extractive summary using this process and the generated summary were evaluated by computing Rouge scores with the human generated gold summary.

## 4.3 Abstractive Summarization

### 4.3.1 BART

The input arguments are "articles" and "summary". The dataset is spilt training, validation and test with ratio 70:15:15. The trained model was evaluated on the test set. Using the BART fast tokenizer which takes care of all the pre-processing steps such as lemmitization, stemming etc, and the mapping function the input is prepared with the following specification:

1. Size of encoder input - 500
2. Size of decoder output - 160
3. batch size - 16
4. padding size - max encoder length
5. truncation - True

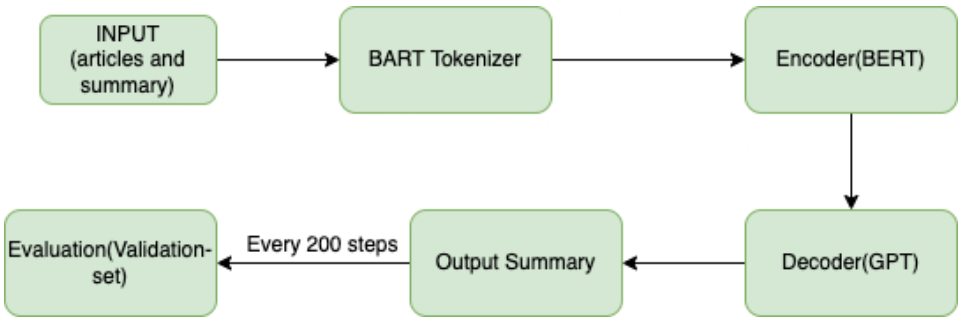


Figure 5: The sequence flow in a BART model

As BART is a very large model, a lot of parameter configurations are possible. Therefore we tried only a few configurations while keeping most of the parameters to default values. BART almost has 140 million parameters with myriads of training arguments which makes it quite a task to fine tune the model. Therefore, focusing on some key parameters such as learning rate , weight penalty etc. , we experimented with the model. The list of some model and training parameters are tabulated in the table 1 below:

Table 1: Some key parameters for BART model.

| Parameters | Values |
|---|---|
| *Bart model configuration:* | |
| length penalty | *2* |
| Number of beams | *4* |
| encoder layer dropout | *0.2* |
| No repeat ngram size | *3* |
| Activation Function | *GeLU* |
| *For training BART:* | |
| Eval steps | *200* |
| weight decay | 0.01 |
| Number of Epochs | 8 |
| Learning rate | *5e-5* |
| Loss Function | *Cross Entropy* |

#### 4.3.2 BART-Large-CNN

On BART large CNN, training is not performed as it has already been fine tuned, so there are no experiments done with this model for parameter tuning. Rather it is only used for comparison with the two other approaches. As the Hugging -Face deems the model to be great at summarization tasks. The maximum length of the output summary is set to 160 which is similar to the decoder's output in the BART base model.

### 4.4 Code

Please find all the relevant code and data files on **news article summarization**[1]

## 5 Results

In this section, we show and analyse the results from all implemented models., We are comparing the human annotated summary with the model generated summary both on qualitative and quantitative grounds. The ROUGE score as the quantitative metric.

### 5.1 Quantitative Analysis

Comparing three approaches for text summarization one for extractive summarization and two for abstractive summarization following result in Table 2 were obtained on the test-set. Here, as described earlier, DistilBERT was used in extractive summarization, and BART LARGE CNN along with BART base fine tuned model was used for abstractive summarization.

[1]Code inspired by Vincent K and dishankjani

Table 2: Quantitative results

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-Lsum |
|---|---|---|---|---|
| DistilBERT | 58.1102 | 44.5113 | 37.7012 | 37.7208 |
| BART LARGE CNN | 39.4549 | 27.2049 | 29.0128 | 29.014 |
| **BART base fine tuned** | **66.3223** | **56.9004** | **47.8824** | **47.8492** |

From the above result it is clear that the fine-tuned BART on BBC news article dataset outperforms the other two models. Here, we can see that although the BART large CNN has more parameters when compared to the BART-base (fine-tuned), but it fails to achieve good performance in terms of the quantitative aspect. Further, even extractive summarization using DistilBERT achieves better performance when fine-tuned for the summarization task. Hence, although the pipeline model may generate more fluent sentences if fails to capture the nuances and important words which reflect the important ideas in a summary.

### 5.2 Qualitative Analysis

Using the human annotated gold summary, a comparison in the below table is provided to evaluate the quality of summary. An example of gold summary which will be compared in the following table is *She made her first Carry On film in 1969 where she appeared in Carry On Again Doctor.She also appeared in Carry On at your Convenience, Carry On Matron and Carry On Loving, as well as others.Rowlands played the hard-done-by wife or the put-upon employee as a regular Carry On star.Actress Patsy Rowlands, known to millions for her roles in the Carry On films, has died at the age of 71.Agent Simon Beresford said: "She was just an absolutely favourite client She never complained about anything, particularly when she was ill, she was an old trouper.Born in January 1934, Rowlands won a scholarship to the Guildhall School of Speech and Drama scholarship when she was just 15.Rowlands starred in nine of the popular Carry On films, alongside fellow regulars Sid James, Kenneth Williams and Barbara Windsor.*

1. **Distil BERT model** is used for generating extractive summary therefore, we will only be comparing the necessary facts which are conveyed from the text article. As we can see in the table below, the summary generated by distil BERT covers most of the important facts in the human annotated gold summary but lacks the language structure,hence it receives a low ROUGE score when compared with BART base. But in terms of conveying important information it performs really well indeed.

2. **BART large CNN** is fine tuned on the CNN daily mail dataset, which is a news dataset. But, the generated summary lacks most of the facts which are present in the gold summary. Therefore, we can observe that although the BART large CNN gives a concise and short output it fails to capture the essence of text, because of that it receives the lowest ROUGE score. Further, fine tuning a model to the data is

really important even if the model is fine tuned for same task. Hence, generalization in summary generation is still a difficult task even if we have millions of trainable parameters.

3. **BART base** is fine tuned on the BBC dataset for abstractive summarization. Here , we can observe that some of the sentences match exactly. All of the information present in the gold summary is conveyed. Although the summary lacks some cohesivity as present in the gold summary but the language structure is much better than our other two models. Hence, it receives the highest ROUGE score.

| Model | Summary |
|---|---|
| DistilBERT | Rowlands starred in nine of the popular Carry On films, alongside fellow regulars Sid James, Kenneth Williams and Barbara Windsor. Rowlands died in Hove on Saturday morning, her agent said. Born in January 1934,Rowlands won a scholarship to the Guildhall School of Speech and Drama scholarship when she was just 15.After spending several years at the Players Theater in London, she made her film debut in 1963 in Tom Jones,directed by Tony Richardson. She also appeared in Carry On at your Convenience,Carry On Matron and Carry On Loving, as well as others. "She didn't mind always being recognised for the Carry On films because she thoroughly enjoyed making them. Previously married, she leaves one son, Alan. |
| BART-LARGE-CNN | Rowlands starred in nine of the popular Carry On films, alongside fellow regulars Sid James, Kenneth Williams and Barbara Windsor. She also carved out a successful television career, appearing for many years in ITV's Bless This House. Rowlands won a scholarship to the Guildhall School of Speech and Drama when she was just 15. |
| BART fine tuned | She made her first Carry On film in 1969 where she appeared in Carry On Again Doctor.Actress Patsy Rowlands, known to millions for her roles in the Carry On films,has died at the age of 71.She also appeared in carry On at your Convenience, Carry On Matron and Carry On Loving, as well as others.Rowlands played the hard-done-by wife or the put-upon employee as a regular Carry On star.She was of the old school - she had skills from musical theater and high drama, that is why she worked with the great and the good of directors.Her last appearance on stage was as Mrs Pearce in the award-winning production of My Fair Lady at the National |

Figure 6: Summary Comparison

# 6 Conclusions

We implemented one model(DistilBERT) for extractive summarization and two model BART-base and BART-large-CNN for abstractive summarization. All of these models are based on transformer(Encoder-Decoder Architecture). We showed and analysed the three models on both qualitative and quantitative front with rouge score and nuanced fact and structure checking.

In DistilBERT, the average ROUGE-2 score changing the output number of sentences is around 44, increasing the number of sentences for the output did not help to improve the performance. As there were not a lot of parameters to play around with DistilBERT, these were the theoretical limits of the model. But the model's output conveys all the facts maybe not in the coherent manner.

In BART-large-CNN, we achieved a ROUGE2 score of 27, as the model failed to capture the important facts in the news articles. The peculiar thing here, is that the BART-large-CNN model is fine tuned on CNN news dataset and the model's is supposed to do well on summarization tasks. At-least it should generalize well over different news dataset. Here we proved that it is not the case and the role of fine tuning on a particular dataset. Further, a noteworthy observation is that in BART-base we limited the input to encoder to a max length of 500 words, but here we pass the whole article, hence throwing more data and building a larger model is not the best solution.

In BART base model, we achieved an average ROUGE2 score of around 56, with varying parameters. BART-base is still a pretty large model with lot of parameters and it gave quite a good result after fine tuning. Although, not exactly similar but Narayan.et.al work in extreme classification achieved a ROUGE score of 31 for single line summaries. [13] This indicates that the model has indeed performing well on the BBC-dataset.

Further improvement to the problem may be shifting to a fully advanced auto-regressive model such as GPT 3. Due to memory constraints we did not use the full length of all the articles while training the BART-base model. Hence, including more text may improve the ROUGE score. Extreme classification can to employed to produce more concision and limit the number of sentences in the input which capture the essence of the article.

# 7 Division of labor

Both of the team members contributed well and supported each other while doing the project since the beginning of literature study, project plan, method discussion, experiment execution, presentation preparation and final report writing. During the experiment period, Yuvraj worked on the preprocessing the data, and training distil BERT model as well as evaluated and analysed the performance with multiple parameters. Meanwhile, Pranav put his efforts on fine tuning and training the BART-base model and executing the pipeline method for pretrained BART-large-CNN model.

# 8 Acknowledgment

I would like to express my deepest gratitude to Professor Mikko Kurimo for organizing such a great course. Many thanks to Ekaterina Voskoboinik for providing a informative feedback and guiding us in choosing the right approach for modelling. Also, many thanks to our peers Hans and Preetha for helping us during the project.

# References

[1] S. Iqbal, S.-U. Hassan, N. R. Aljohani, S. Alelyani, R. Nawaz, and L. Bornmann, "A decade of in-text citation analysis based on natural language processing and machine learning techniques: An overview of empirical studies," *Scientometrics*, vol. 126, no. 8, pp. 6551–6599, 2021.

[2] K.-F. Wong, M. Wu, and W. Li, "Extractive summarization using supervised and semi-supervised learning," in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK: Coling 2008 Organizing Committee, Aug. 2008, pp. 985–992. [Online]. Available: https://aclanthology.org/C08-1124

[3] X. Lu and Y. Jang, "Generate concise content: Text summarization with bert."

[4] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[5] M. Yasunaga, J. Kasai, R. Zhang, A. R. Fabbri, I. Li, D. Friedman, and D. R. Radev, "Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks," *CoRR*, vol. abs/1909.01716, 2019. [Online]. Available: http://arxiv.org/abs/1909.01716

[6] R. v. d. Berg, T. N. Kipf, and M. Welling, "Graph convolutional matrix completion," *arXiv preprint arXiv:1706.02263*, 2017.

[7] V. Kieuvongngam, B. Tan, and Y. Niu, "Automatic text summarization of covid-19 medical research articles using bert and gpt-2," *arXiv preprint arXiv:2006.01997*, 2020.

[8] A. Ghadimi and H. Beigy, "Hybrid multi-document summarization using pre-trained language models," *Expert Systems with Applications*, vol. 192, p. 116292, 2022.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] D. Miller, "Leveraging bert for extractive text summarization on lectures," *arXiv preprint arXiv:1906.04165*, 2019.

[11] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[12] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[13] S. Narayan, S. B. Cohen, and M. Lapata, "What is this article about? extreme summarization with topic-aware convolutional neural networks," *Journal of Artificial Intelligence Research*, vol. 66, pp. 243–278, 2019.