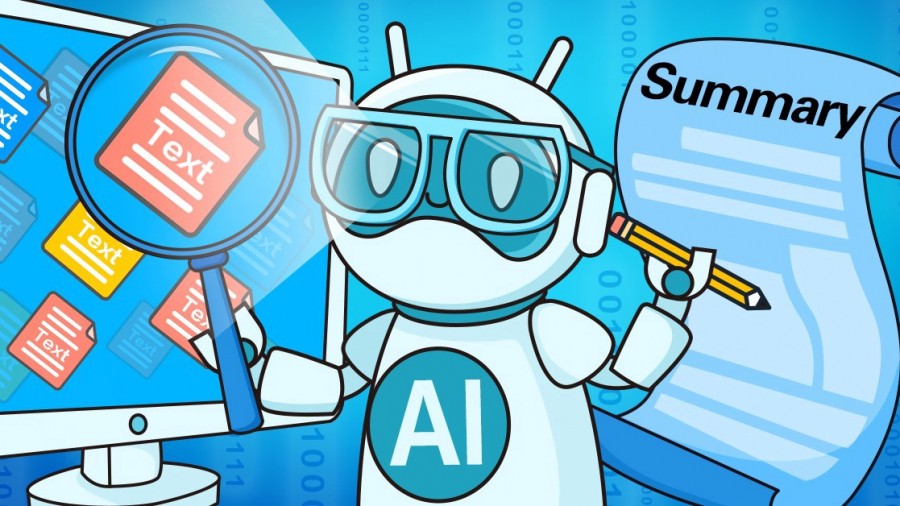
TEXT SUMMARIZATION USING TRANSFORMERS (BART)

Advanced Machine Learning (64061)

By Prof. Murali Shanker



By Sindhu Vegiraju

**Introduction:**

The aim of summarization is to create a concise summary from a long document or articles such that no information is lost. In recent times, computing text summaries using Deep Learning has gained popularity.

The amount of text data available online is increasing at a very fast pace hence text summarization has become essential. Most of the modern recommender and text classification systems require going through a huge amount of data. Manually generating precise and fluent summaries of lengthy articles is a very tiresome and time-consuming task. Hence generating automated summaries for the data and using it to train machine learning models will make these models space and time-efficient.

Deep learning technologies have proven to be particularly promising for this endeavour since they attempt to replicate the way the human brain functions by managing multiple levels of abstraction and nonlinearly translating a given input into a given output.

**Summarization Techniques:**

Nowadays, the AI community has two ways to approach automatic text summarization, Extractive Summarization and Abstractive Summarization. However, in this system I focus on Abstractive Summarization because it is more advanced and closer to human-like interpretation. It has more potential (and is generally more interesting for researchers and developers).

**Extractive Summarization**: The Extractive approach selects the most important phrases and lines from the documents. It then combines all the important lines to create the summary. So, in this case, every line and word of the summary actually belongs to the original document which is summarized.

**Abstractive Summarization**: The abstractive approach uses new phrases and terms that are different from the original document, keeping the meaning the same, just like how humans do in summarization. So, it is much harder than the extractive approach.

**Dataset:**

The dataset used is BBC News dataset which was generated from a dataset used for text classification. It consists of 2224 BBC news website documents relating to stories. The BBC News Dataset was used for generating summaries that consisted of text and human generated summaries which are summaries that have been written by humans.

**Preprocessing:**

This dataset consists of long news articles along with short summaries for comparison. The raw dataset was then cleaned using various pre-processing techniques such as:

**Lower casing** - To convert the input text into the same casing format so that all capital, lower case and mixed case are treated similarly.

**Eliminate Punctuation** - HTML tags and links- Removal of punctuations, links and tags that do not add meaning to the text such as “!"#$%&\'() \*+, -. /: ;<=>? @ [\\] ^\_ {|} ~`” to standardise the text.

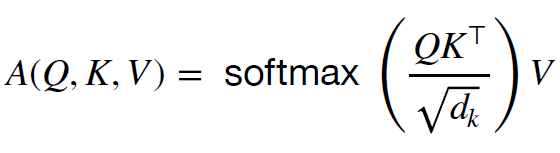
**Eliminate Stopwords and frequently occurring words** - Removal of common words such as ‘the’, ‘a’, etc that are frequently used in a text but do not provide valuable information for downstream analysis.

**Transformer Model:**

The Transformer network is solely based upon multiple attention layers. It does not make use of RNN and is reliant on attention layers and positional encoding for remembering the sequence of words in the input sequence. The global dependencies created with the help of multiple attention layers help in creating parallelization in processing the input.

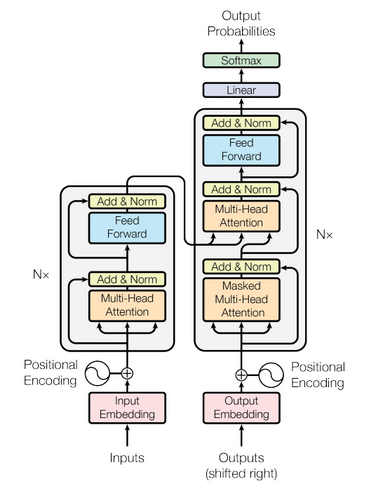
The transformer model contains encoder and decoder layers, where each is connected to a multi-head attention layer and feed forward network layers. The model remembers the position and sequence of words with the help of cosine and sine functions that creates positional encoding. The multi-head attention layer in the encoder and decoder layer applies a mechanism called self-attention. The input is fed into three connected layers to create query (Q), key (K), and value (V) vectors.

These vectors are split into n vectors.



Self-attention is applied on n separate vectors to create multi-head attention.

Below Figure depicts the architecture of a transformer model. It contains an encoder and decoder layer and the various normalization and multi-head attention layers are also depicted in the figure.



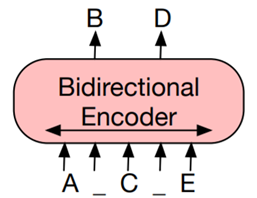
**Pretrained Models based on Transformers** - Hugging face works as an opensource for providing many useful NLP libraries and datasets. Its most famous library is the Transformer library. The transformer library consists of various pre-trained models like BERT, GPT, BART, T5, PEGASUS to predict summaries of texts that can be fine-tuned for any dataset. Here let us discuss about a pre-trained model called BART that is tuned and implemented for the BBC news dataset to give fairly good summaries.

**BART:**

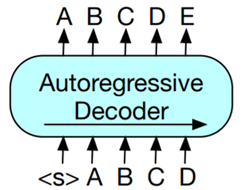
At the end of 2019, researchers of Facebook AI Language have published a new model for Natural Language Processing (NLP) called BART (Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension).

BART stands for Bidirectional and Auto Regressive Transformers. It is built with a seq2seq model trained with denoising as a pre-training purpose. It uses a standard seq2seq model architecture combining an encoder similar to BERT and a GPT-like decoder.

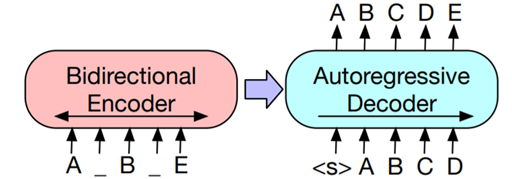
**BERT: Random tokens**are**replaced**with the token**[MASK],** and the document is**encoded bidirectionally. Missing tokens**are**predicted independently**, so BERT cannot easily be**used for generation.**



**GPT: Tokens**are**predicted auto-regressively,** meaning**GPT can be used for generation.** However,**words**can only**condition on leftward context**, so it**cannot learn bidirectional interactions.**



**BART:** The pre-training task involves changing the order of the original phrases randomly and a new scheme where text ranges are switched with a single mask token. The large model of BART consists of twice as many layers as are present in the base model. It is quite similar to the BERT model but BART contains about 10% more features than the BERT model of comparable size. BART's decoder is autoregressive, and it is regulated for generating sequential NLP tasks such as text summarization. The data is taken from the input but changed, which is closely related to the denoising pre-training objective. Hence, the input sequence embedding is the input of the encoder, and the decoder autoregressively produces output. I have used the "facebook/bart-large-cnn" pre-trained model and then the Bart tokenizer, which is constructed from the GPT-2 tokenizer. Hence words are encoded differently depending on their position in the sentence.



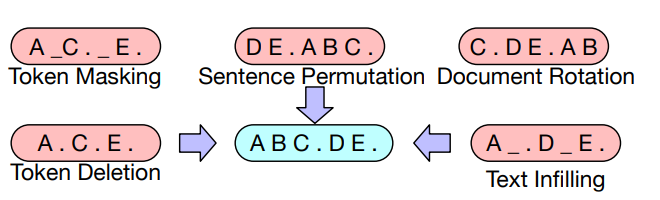
Inputs to the encoder need not be aligned with decoder outputs, allowing arbitrary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and using representations from the final hidden state of the decoder.

**BART Architecture:**

BART uses the standard sequence-to-sequence Transformer architecture. The architecture is closely related to that used in BERT, with the following differences: (1) each layer of the decoder additionally performs cross-attention over the final hidden layer of the encoder (as in the transformer sequence-to-sequence model); and (2) BERT uses an additional feed-forward network before word prediction, which BART does not.

**Pre-training BART:**

BART is trained by corrupting documents and then optimizing a reconstruction loss—the cross-entropy between the decoder’s output and the original document. Unlike existing denoising autoencoders, which are tailored to specific noising schemes, BART allows us to apply any type of document corruption. In the extreme case, where all information about the source is lost, BART is equivalent to a language model. It is experimented with several previously proposed and novel transformations, but it is believed that there is a significant potential for development of other new alternatives. The transformations used are summarized below, and examples are shown in the figure.

****

**Token Masking:** Following BERT, random tokens are sampled and replaced with [MASK] elements.

**Token Deletion:** Random tokens are deleted from the input. In contrast to token masking, the model must decide which positions are missing inputs.

**Text Infilling:** A number of text spans are sampled, with span lengths drawn from a Poisson distribution (λ = 3). Each span is replaced with a single [MASK] token. 0-length spans correspond to the insertion of [MASK] tokens. Text infilling teaches the model to predict how many tokens are missing from a span.

**Sentence Permutation:** A document is divided into sentences based on full stops, and these sentences are shuffled in a random order.

**Document Rotation:** A token is chosen uniformly at random, and the document is rotated so that it begins with that token. This task trains the model to identify the start of the document.

For generating summaries, I am fine-tuning the BART transformer model from the Hugging face library.

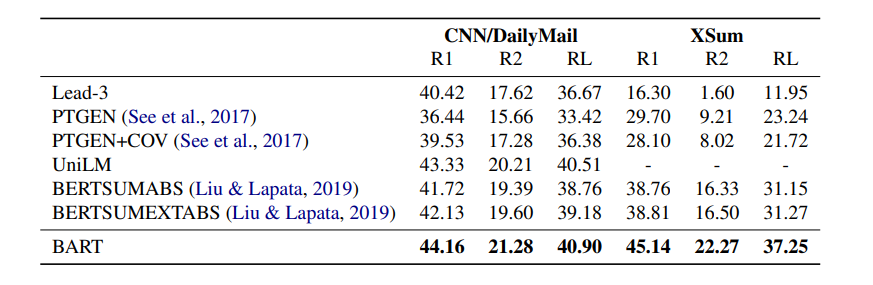
**Fine-tuning BART:** Because BART has an autoregressive decoder, it can be directly fine-tuned for sequence generation tasks such as abstractive question answering and summarization. In both of these tasks, information is copied from the input but manipulated, which is closely related to the denoising pre-training objective. Here, the encoder input is the input sequence, and the decoder generates outputs autoregressively.

**Performance on two standard summarization datasets:**

To provide a comparison with the state-of-the-art in summarization, presenting the results on two summarization datasets, CNN/DailyMail and XSum, which have distinct properties.

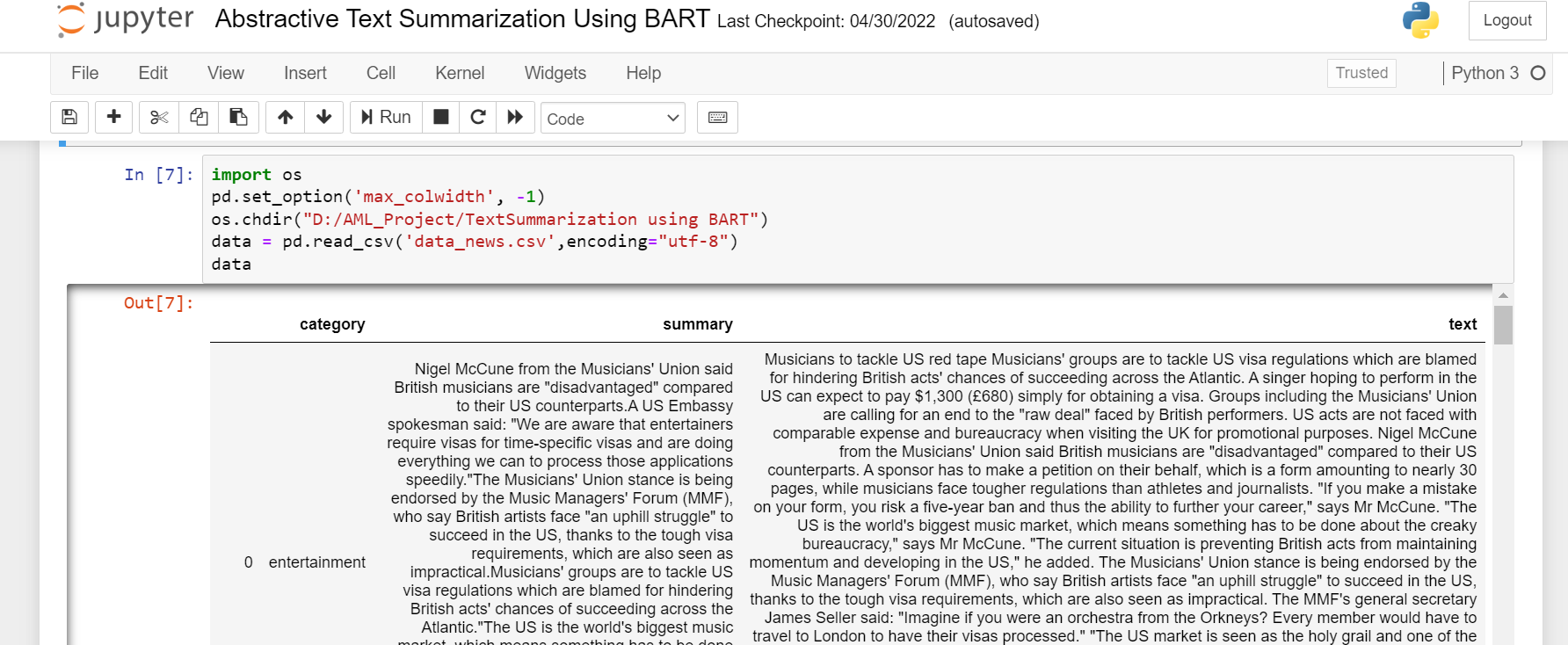
Summaries in the CNN/DailyMail tend to resemble source sentences. Extractive models do well here, and even the baseline of the first-three source sentences is highly competitive. Nevertheless, BART outperforms all existing work.

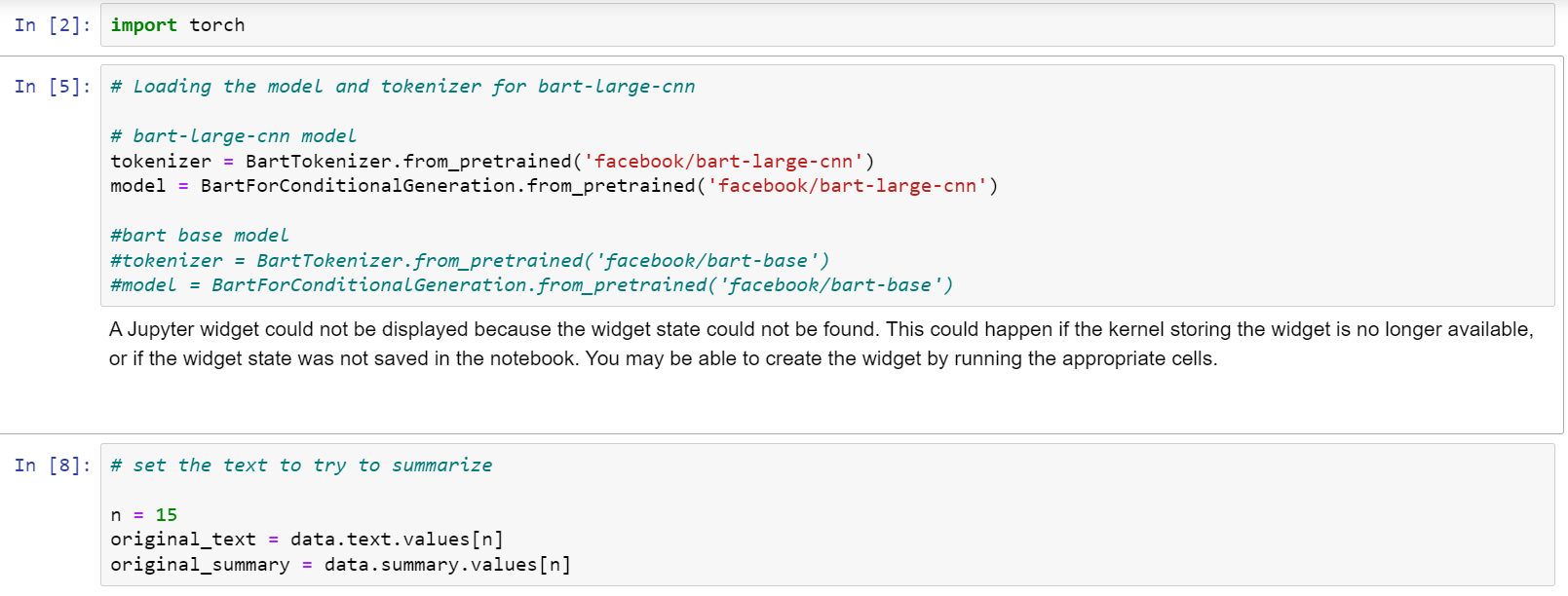
In contrast, XSum is highly abstractive, and extractive models perform poorly. BART outperforms the best previous work, which leverages BERT, by roughly 6.0 points on all ROUGE metrics—representing a significant advance in performance on this problem. Qualitatively, sample quality is high.



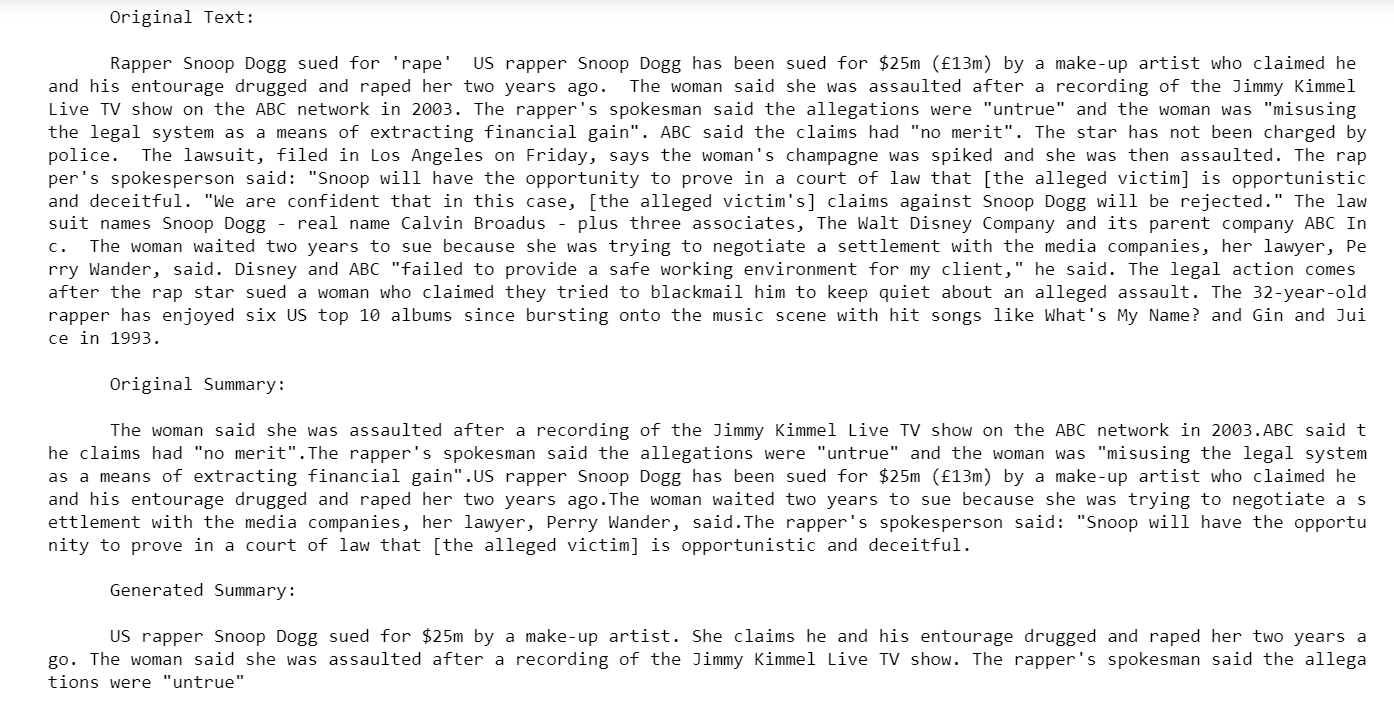
**Results on BBC News Dataset:**

Let us consider an instance of the below Original Text, Original Summary from the BBC News dataset and the Generated Summary from the code.









**Original Text:**

Rapper Snoop Dogg sued for 'rape’ US rapper Snoop Dogg has been sued for $25m (£13m) by a make-up artist who claimed he and his entourage drugged and raped her two years ago. The woman said she was assaulted after a recording of the Jimmy Kimmel Live TV show on the ABC network in 2003. The rapper's spokesman said the allegations were "untrue" and the woman was "misusing the legal system as a means of extracting financial gain". ABC said the claims had "no merit". The star has not been charged by police. The lawsuit, filed in Los Angeles on Friday, says the woman's champagne was spiked and she was then assaulted. The rapper's spokesperson said: "Snoop will have the opportunity to prove in a court of law that [the alleged victim] is opportunistic and deceitful. "We are confident that in this case, [the alleged victim's] claims against Snoop Dogg will be rejected." The lawsuit names Snoop Dogg - real name Calvin Broadus - plus three associates, The Walt Disney Company and its parent company ABC Inc. The woman waited two years to sue because she was trying to negotiate a settlement with the media companies, her lawyer, Perry Wander, said. Disney and ABC "failed to provide a safe working environment for my client," he said. The legal action comes after the rap star sued a woman who claimed they tried to blackmail him to keep quiet about an alleged assault. The 32-year-old rapper has enjoyed six US top 10 albums since bursting onto the music scene with hit songs like What's My Name? and Gin and Juice in 1993.

**Original Summary:**

The woman said she was assaulted after a recording of the Jimmy Kimmel Live TV show on the ABC network in 2003.ABC said the claims had "no merit". The rapper's spokesman said the allegations were "untrue" and the woman was "misusing the legal system as a means of extracting financial gain". US rapper Snoop Dogg has been sued for $25m (£13m) by a make-up artist who claimed he and his entourage drugged and raped her two years ago. The woman waited two years to sue because she was trying to negotiate a settlement with the media companies, her lawyer, Perry Wander, said. The rapper's spokesperson said: "Snoop will have the opportunity to prove in a court of law that [the alleged victim] is opportunistic and deceitful.

**Generated Summary:** US rapper Snoop Dogg sued for $25m by a make-up artist. She claims he and his entourage drugged and raped her two years ago. The woman said she was assaulted after a recording of the Jimmy Kimmel Live TV show. The rapper's spokesman said the allegations were "untrue"

**Conclusion:**

BART gave better results as the summaries generated were fluent, accurate, and integrated supporting evidence from the input document. Hence, the summaries generated by the BART pretrained model demonstrate that the BART model is useful for text understanding.

Bart allows many different ways of document corruption.

BART is particularly effective when fine-tuned for text generation but also works well for comprehension tasks. BART has outperformed other models in the NLP field and achieved new state- of-the-art results on a range of abstractive dialogue, question answering, and summarization tasks.

**References:**

<https://arxiv.org/pdf/1910.13461.pdf>

<https://neurondai.medium.com/automatic-text-summarization-system-using-transformers-are-you-tired-of-reading-a-long-paper-22d2cd9f5260>

<https://arxiv.org/ftp/arxiv/papers/2108/2108.01064.pdf>

<https://siddiquimubasheer.medium.com/text-summarization-using-bert-and-t5-e05dbbc757c6>

<https://www.youtube.com/watch?v=1JBMCG8rW18&t=7s>

<https://www.youtube.com/watch?v=MxNnl_gHV1Y>