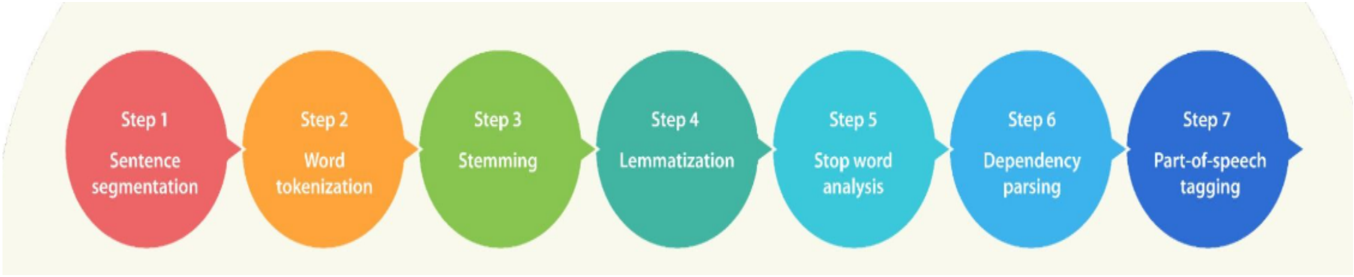
Research Paper Simplification Pipeline Design

Sentence Simplification (SS) can be done using fine-tuned Transformer models, specifically BERT and GPT-2. Transformer models have revolutionized Natural Language Processing by using self-attention mechanisms to understand and generate text. These models can be pre-trained on extensive text data retrieved from the research paper and then fine-tuned for specific tasks. BERT is a bidirectional model that considers the context from both sides of a word, while GPT-2 is unidirectional, making it faster.

1. Pre-processing Tasks:

Our initial task of pre-processing module is responsible for handling raw research papers which includes the following steps:
a. **PDF to Text Conversion**: To extract text content from PDF files.
b. **Tokenization**: To break down the text into words and sentences.
c. **Text Cleaning**: To remove unnecessary characters, symbols, and formatting.
d. **Sentence Segmentation**: Split the text into sentence



1.1 Training a tokenizer :

We choose to train a byte-level Byte-pair encoding tokenizer (same as GPT-2), with the same special tokens as RoBERTa. Hereby, we recommend training a byte-level BPE because it will start building its vocabulary from an alphabet of single bytes, so all words will be decomposable into tokens.

2. Definition of Simplification (Feature-wise):

Simplification in research papers involves various aspects:

a. Language Simplification: Simplify complex and jargon-rich language.
b. Sentence Shortening: Reduce sentence length while maintaining clarity.
c. Paraphrasing: Reword sentences while preserving meaning.
d. Content Summarization: Condense the paper's main points

Also simplification refers to specific changes achieved through transformer models in sentence restructuring. These alterations include:

2.1 Reducing Complexity: Making complex sentences easier to understand by rephrasing or restructuring them, particularly for younger audiences or non-native speakers.

2.2 Shortening Sentences: Transforming long or complicated sentences into shorter, more straightforward structures while maintaining the original meaning.

2.3 Enhancing Clarity and Readability: Improving the overall clarity and readability of sentences through the use of simpler vocabulary, clearer syntax, and grammatical structures.

2.4 Preserving Meaning: Ensuring that, despite simplification, the fundamental message and vital information of the original sentence remain unchanged.

2.5 Improving Accessibility: Adapting language to be more understandable for a broader range of readers, including those with language difficulties, learning challenges, or cognitive impairments.

The main objective of employing transformer models in sentence simplification, is to make the content more easily comprehensible without compromising its core meaning, thus catering to a more diverse audience.

3. Scope of simplification

The scope of research paper simplification is to improve the accessibility and comprehension of scholarly papers for a broader audience, including students, non-experts, and those with limited domain knowledge. It entails simplifying complex sentences while preserving the research's core concepts and reducing complexity by removing intricate language and employing paraphrasing. Future directions include exploring the application of these simplification techniques in controlled sentence simplification, widening the potential applications beyond research papers. The ultimate aim is to bridge the comprehension gap between intricate research content and a more diverse readership through efficient text simplification.

4. Problem Formulation:

Input:

- Complex sentences in the raw text format extracted from a research paper.
- Fine-tuned transformer models, specifically BERT and GPT-2.

Output:

- Simplified versions of the input sentences that preserve their original meaning.

Models:

- Utilize fine-tuned transformer models, specifically BERT and GPT-2, to learn the transformation process from complex sentences to simplified versions during the fine-tuning phase.

Loss Function:

- The loss function for sentence simplification encompasses the following components:

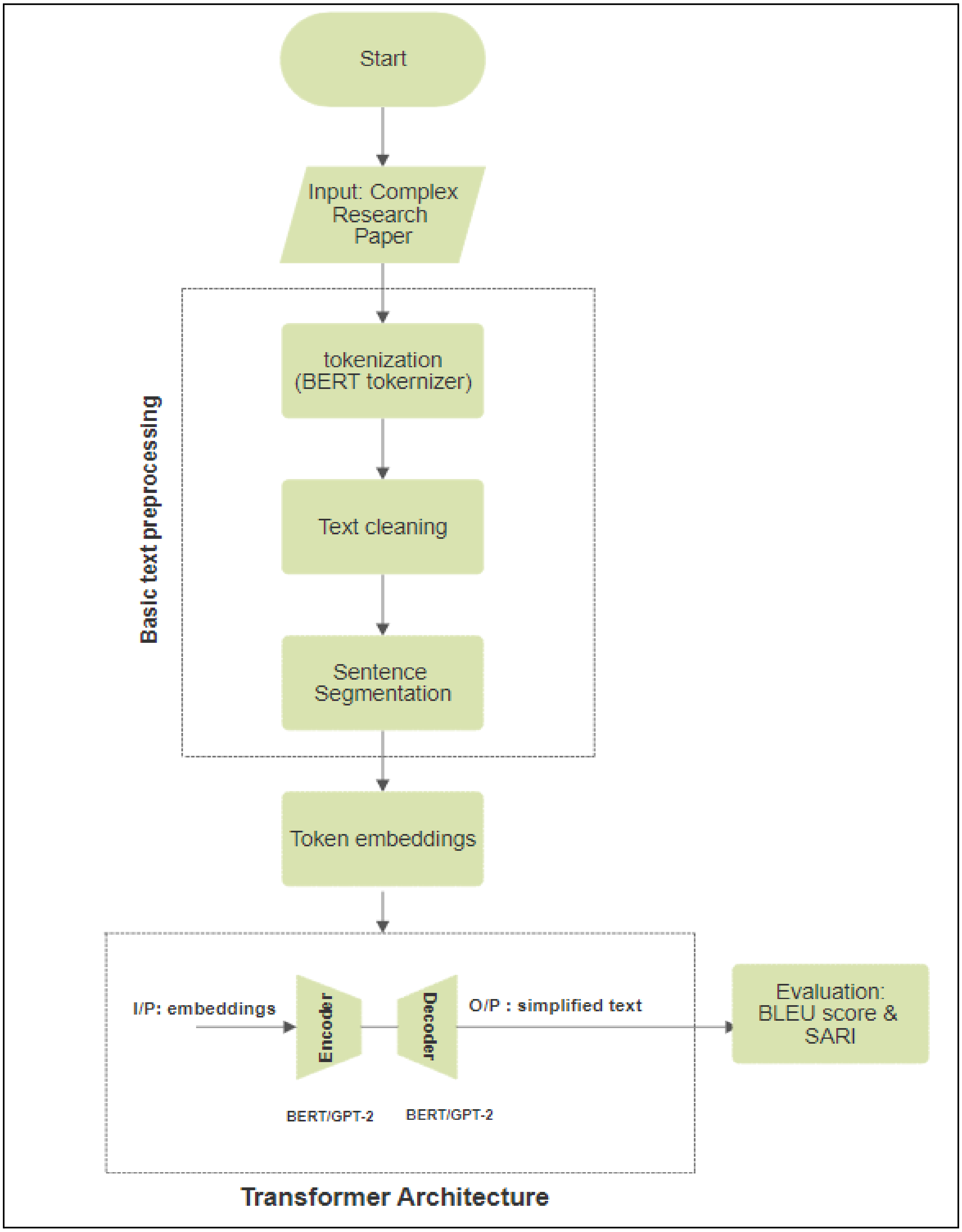
1. Linguistic Quality: Measures fluency and grammatical correctness using metrics such as BLEU or language model perplexity.
2. Meaning Preservation: Assesses how well the simplified sentences maintain the meaning and semantics of the original sentences, quantified by metrics like SARI (System output against references and against the input sentence).

Objective:

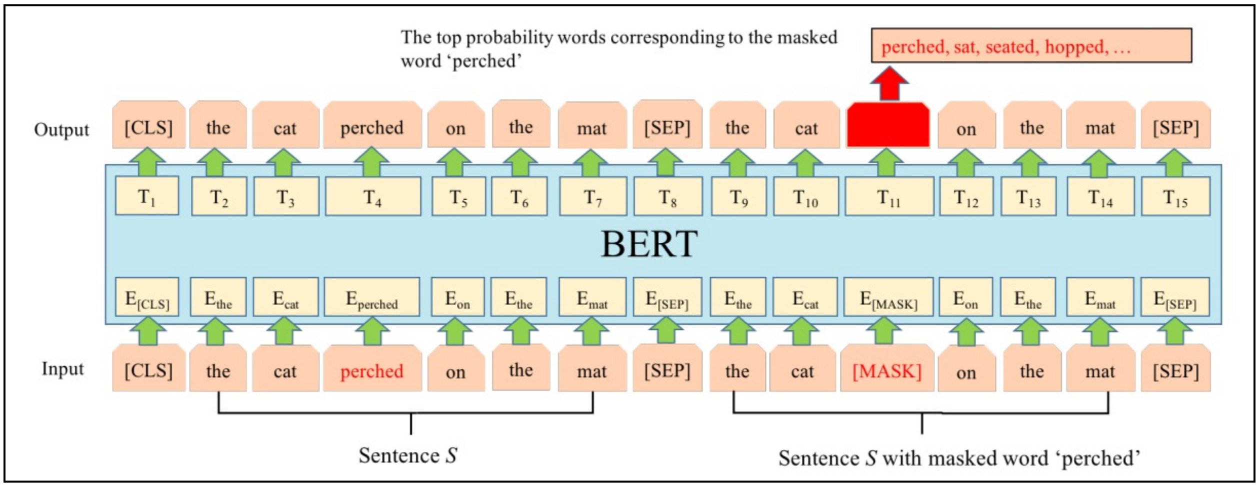
- Minimize the defined loss function to generate simplified sentences that are both linguistically accurate and retain the original meaning. Additionally, aim to improve accessibility and readability.

5. Methodology

5.1 Workflow



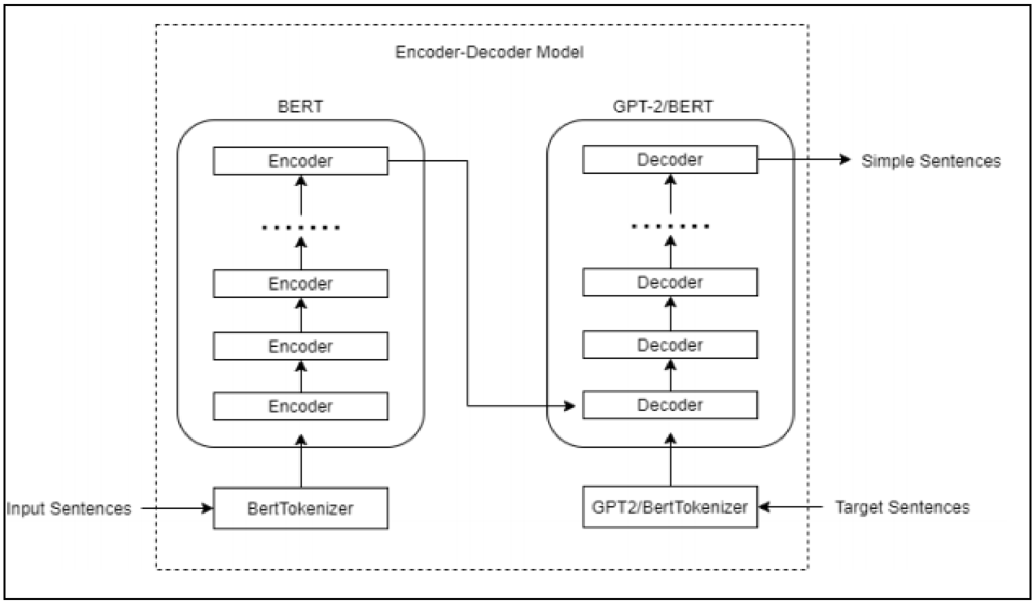
We utilize two well-known transformer models, BERT and GPT-2. BERT, which stands for Bidirectional Encoder Representations from Transformers, is a natural language processing (NLP) model developed by Google. It comprehensively understands sentence context from both left and right sides, making it proficient in grasping word meanings. On the other hand, GPT-2, another popular transformer model by OpenAI, is unidirectional, focusing only on the left context. While BERT excels at understanding sentence semantics, GPT-2 is faster and more efficient in generating text.


The top probability words corresponding to the masked word 'perched'
perched, sat, seated, hopped, …
Output
[CLS] the cat perched on the mat [SEP] the cat on the mat [SEP]
T_1 T_2 T_3 T_4 T_5 T_6 T_7 T_8 T_9 T_{10} T_{11} T_{12} T_{13} T_{14} T_{15}
BERT
$E_{[CLS]}$ E_{the} E_{cat} $E_{perched}$ E_{on} E_{the} E_{mat} $E_{[SEP]}$ E_{the} E_{cat} $E_{[MASK]}$ E_{on} E_{the} E_{mat} $E_{[SEP]}$
Input
[CLS] the cat perched on the mat [SEP] the cat [MASK] on the mat [SEP]
Sentence S
Sentence S with masked word 'perched'

1. Input Handling: The project begins by accepting standard English sentences as input extracted from Research Paper.

2. Tokenization: These sentences are tokenized using BERTTokenizer to prepare them for further processing.

3. Encoder-Decoder Model: The token embeddings are then passed through an encoder-decoder model, aiming to rearrange or substitute words and phrases to enhance comprehensibility without losing the original information.



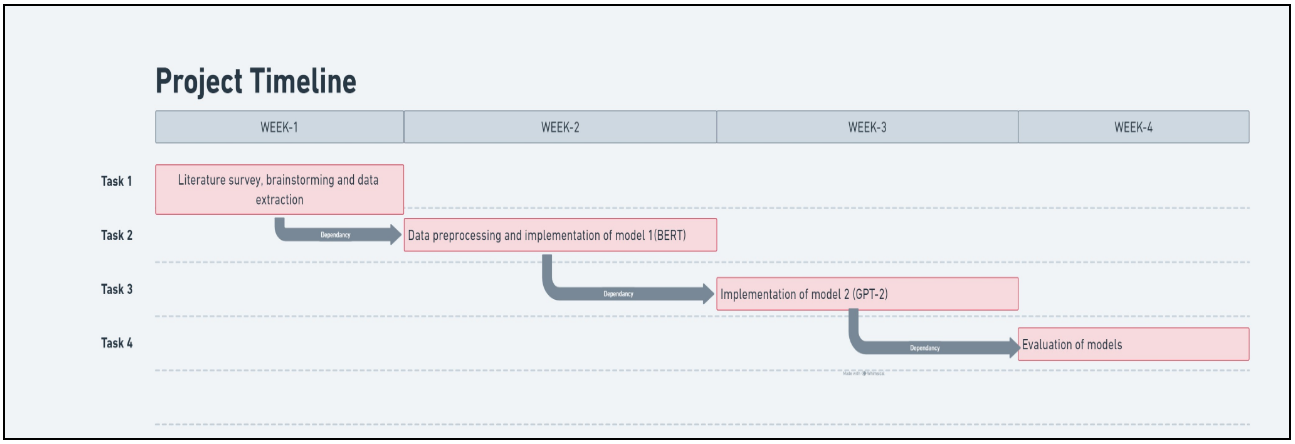
4. Output Transformation: The model generates token embeddings for simplified sentences, which are subsequently converted into readable sentences using GPT2Tokenizer.

5. Versatile Application: The project's output has broader utility, not limited to sentence simplification, as it can benefit various NLP tasks such as Machine Translation, Summarization, and Classification.

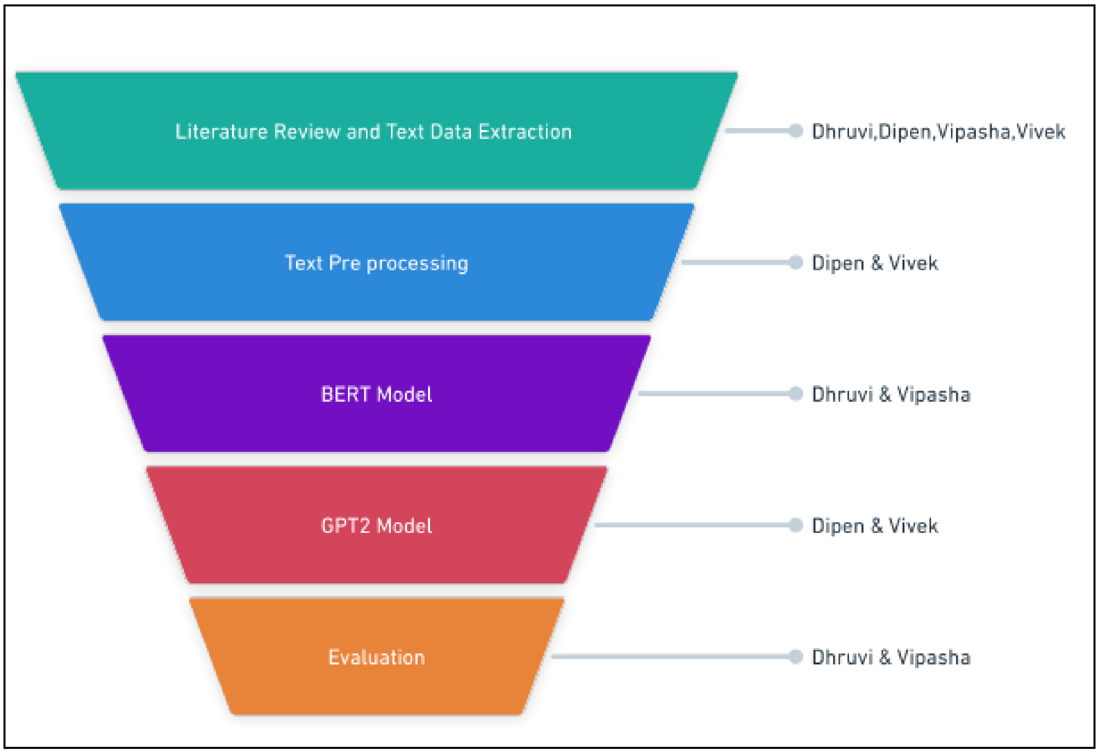
6. Model Selection: The project leverages two prominent transformer models: BERT, known for its bidirectional context understanding, and GPT-2, recognized for its speed and efficiency in text generation.

7. Quality Evaluation: The quality of sentence simplification is assessed using the widely adopted SARI metric. It compares simplified sentences to the original source and a reference sentence, utilizing n-grams to quantify additions, keeps, and deletions, with averaged scores providing a comprehensive evaluation result.

6. Timeline for completion of Project


Project Timeline
WEEK-1
WEEK-2
WEEK-3
WEEK-4
Task 1
Literature survey, brainstorming and data extraction
Task 2
Dependancy
Data preprocessing and implementation of model 1 (BERT)
Task 3
Dependancy
Implementation of model 2 (GPT-2)
Task 4
Dependancy
Evaluation of models

7. Task delegation among members



Submitted by : 202211002 - Vipasha Vaghela
202211032 - Dhruvi Shah
202211058 - Dipen Padhiyar
202211069 - Vivek Soni