PROJECT REPORT

TEAM NUMBER - 22

TEAM MEMBERS:

Ch. Vamshi Krishna Babu – SE22UARI038

Ch. Vamsi Krishna – SE22UARI039

G. Rithvik Chandan – SE22UARI144

V. Raja – SE22UARI178

K. Bhuvan - SE22UARI212

Automated Book Summary Generator for Novels

INTRODUCTION :

This automated book summary generator for novels projects is generating very short and still coherent summaries for stories from a novel. Based on the given input text file that will upload a pre-trained language model, this system would process what the content it has and it will just output a coherent summary regarding the essence of the narrative.

DATASET:

Source: The sources of dataset are from variety of book summaries, research paper, the Kaggle dataset and that by Google.

Composition:

No. Of rows: 120 samples (summaries).
Number of columns.:
1. Input: Text to Summarize Original.
2. Output: Human written summary of input text.
3. Genre: the book type for whose summary is written based upon whose account; Fiction, Mystery, Sci-Fi, etc.

DATA PREPROCESSING OPERATIONS:

1. Normalization
2. Removing space
3. Stop word filtering
4. Tokenization
5. Lemmatization

This is an example of an abstractive summarization system. The summary comes in the form of reformulation and paraphrasing the input text rather than taking chunks out of it.

MODEL USED:

T5-Text-to-Text Transformer pre-trained model This has great state-of-the-art performance in abstractive summarization.

It has been trained by our self-crafted dataset by embracing the techniques of supervised learning. It is designed to hold an eye on performance while the architecture keeps the focus for the novel-based text on the parameters of coherence and fluency in narration.

SYSTEM ARCHITECTURE :

1. Input: Preprocessing text data along with tokenization, passed over the model
2. Generation: Model creates a sequence of tokens which on its own creates the summary.
3. Output: In that output is given to the token which later on decode into the summary as well, to create one summary text.

Experiment

Training Specifications: Epochs : 2 Batch Size: 8 Learning rate: 5e -5
Tools/ Libs: Frameworks Hugging Face Transformers, Tensor Flow, PyTorch
Environment: Google Colab

Test Cases: The strength and flexibility of the model were checked by feeding the model with the input length from a short passage to full-length chapters of various genres.

SCORES :

Evaluation Report:

ROUGE-1: Score(precision=0.875, recall=0.9459459459459459, fmeasure=0.9090909090909091)

ROUGE-2: Score(precision=0.8227848101265823, recall=0.8904109589041096, fmeasure=0.855263157894737)

ROUGE-L: Score(precision=0.8625, recall=0.9324324324324325, fmeasure=0.8961038961038962)

BLEU Score: 0.7190012019046286

Qualitative Evaluation:
The summaries were compared to human summaries produced. It was observed that:
Strength: Highly coherent and fluent with almost all retention of narration.
- Weakness: Actually, there were comparatively very few deviations from facts from the long passages.
Example:

Input Text: "A young wizard fights against dark forces and finds his actual destiny."
Summary Created: "A young wizard battles darkness forces and finds out what his true destiny is.

ANALYSIS & CONCLUSION :

Key Take-Aways:
- Strengths:
 - Can produce highly coherent and smooth summaries.
Extremely complex stories may be processed with near-perfect accuracy.
Weaknesses:
-It cannot accept long texts in multi-threaded format.
-Expensive computations because of the usage of transformer architecture

Improvement Possibility:
1. Improvement of the size of the database such that it may generalize
2. Language support for the multi-lingual document other than English
3. Fast model adaptation without loss of accuracy of the summarization model 4. Hybrid summary, that is, developed by both integrative and extractive techniques for fact validation purpose