RapidReads: TLDR Extraction from Reddit Threads with Neural Networks

Urjit Patil (up63)
Nikhil Mishra (nm116)
Atharva Bhusari (ab2414)

The problem addressed in this project is text summarization, a natural language processing (NLP) task that involves generating concise and coherent summaries Reddit posts. The challenge lies in developing models that can understand the context of the input text and produce informative and coherent abstractions. The proposed model aims to generate concise TLDR(too long; didn't read) summaries, aiding users in deciding whether to delve into the full text. The TLDR text serves as a succinct preview, enabling users to quickly assess the content and decide whether it aligns with their interests and if they want to explore it further.

The project will utilize the Summarize from Feedback dataset, available on the Hugging Face website, that has around 179k Reddit posts from multiple subreddits and corresponding summaries. In the Learning to Summarize from Human Feedback paper, a reward model was trained from human feedback. The reward model was then used to train a summarization model to align with human preferences. This is the dataset of human feedback that was released for reward modelling.

The proposed solution involves the implementation of neural network models for text summarization before which baseline models such as Frequency based and TF-IDF based models that utilize extractive methods of summarization will be implemented. The primary focus will be sequence-to-sequence models, incorporating attention mechanisms, as they have effectively captured contextual information. Pre-trained language models will also be explored to leverage their contextual understanding.

The quality of the developed models will be measured using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score. ROUGE is a standard metric for evaluating the quality of text summaries by comparing them to reference summaries. Specifically, ROUGE-1, ROUGE-2, and ROUGE-L scores will be employed to assess unigram overlap, bigram overlap, and longest common subsequence, respectively.

The successful completion of this project will result in a robust text summarization model, with the ROUGE score providing quantitative insights into its performance. The approach ensures scalability, allowing for potential future exploration of larger datasets and advanced model architectures.