Sustainability AI Engine
NLP Progress Report
Dibyendu Das
29-04-2023

TASK 1 (Text Preprocessing) :

We used the following preprocessing steps for text preprocessing:

- Stripping HTML tags to remove any markup or formatting in the text.
- Removing links to eliminate any URLs or hyperlinks in the text.
- Removing whitespace to ensure consistency and readability.
- Removing accented characters and lower casing the text to normalize the text data.
- Reducing incorrect character repetition and expanding contraction words to ensure consistency.
- Removing special characters and stopwords to eliminate noise and reduce computational cost.
- Performing spelling correction to correct any spelling mistakes in the text.
- Lemmatizing the text to group together inflected forms of words.

TASK 2 (Sentiment Analysis) :

When performing sentiment analysis, it is crucial to choose an appropriate model that can accurately classify text data into positive, negative, or neutral sentiments. To this end, we evaluated several pretrained models including DistilBERT, FinBERT, and Roberta from the Hugging Face library. After testing these models on a small dataset, we determined that Roberta performed the best in terms of classification accuracy. We then created a fine tuning pipeline for Roberta that allowed us to further optimize the model for sentiment analysis on our specific dataset. By choosing and fine-tuning an appropriate model, we were able to achieve more accurate sentiment analysis results, enabling us to draw more meaningful insights from the data.

TASK 3 (Keyword Extraction) :

Keyword extraction is a critical task in natural language processing that involves identifying and extracting important words or phrases from text data. To accomplish this, we tested several pretrained models including KeyBERT, DistilBERT, and the KEYword generator model. After manually checking the accuracy of keyword extraction on a small dataset, we found that KeyBERT had the highest accuracy and was the best model for our needs. We have decided to use KeyBERT for further studies on keyword extraction, allowing us to extract the most relevant and important keywords from large volumes of text data with high accuracy. This will enable us to gain deeper insights and make more informed decisions based on the extracted keywords.

TASK 4 (Text Summarization) :

Text summarization is an essential task in natural language processing that involves reducing the length of a text document while retaining its most important information. In our research, we used a correlation matrix of the sentences to summarize the text data. The correlation matrix enabled us to identify the most relevant and important sentences in the document, which we then ranked using the PageRank algorithm. This allowed us to generate a summary that contained only the most critical information from the original text. Since we used the correlation matrix and PageRank algorithm to identify the most important sentences, our approach is an example of extractive summarization. By using this method, we were able to generate accurate and informative summaries that captured the essence of the original text.

TASK 5 (Alignment Checking) :

After extracting the keywords and generating the summary, we evaluated the alignment between the extracted keywords and the headline of the original text. To accomplish this, we utilized BERTScore, a state-of-the-art metric for evaluating the quality of text generation models. BERTScore uses a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model to evaluate the similarity between two pieces of text. By applying BERTScore to our extracted keywords and the headline of the original text, we were able to determine how closely the keywords matched the main idea of the text. This enabled us to identify any mismatches and refine our approach to keyword extraction and text summarization, ensuring that our results were as accurate and relevant as possible.

TASK 6 (Keyword Ranking) :

In our study, we used a similar methodology for ranking keywords as we did for ranking sentences in the text. We employed a correlation matrix to determine the strength of the relationship between the extracted keywords and the rest of the text. This enabled us to identify the most significant and relevant keywords within the text. We then used the PageRank algorithm to assign a score to each keyword based on its relevance and importance within the text. By using this approach, we were able to identify the most critical keywords, which provided us with a deeper understanding of the main ideas and concepts within the text. This methodology ensured that our keyword rankings were accurate and informative, enabling us to gain valuable insights from the text data.