


Automated Research Paper Categorization



Problem statement

Develop an innovative solution for automated paper categorization in submission platforms by analyzing paper titles and abstracts.



Preprocessing Data

- Used pandas to load and manipulate the data, converting it into a pandas DataFrame.
- Added categories to all the columns as needed for analysis or model training.
- Utilized Python libraries such as **re** (regular expressions) for text pattern matching and manipulation.
- Employed the **nltk** library for natural language processing tasks.
- Formed a list of stopwords using the **nltk.corpus.stopwords.words('english')** method.
- Removed stopwords from the text data to reduce the number of tokens and enhance the quality of the data for analysis or model training.

Model Development

- Split the dataset into training and test sets using a 90-10 split, selecting 30,000 samples for training.
- Implemented a custom class **DatasetSentiment** for tokenizing the input data and preparing the attention mask using the pre-trained DistilBERT model (**distilbert-base-uncased**).
- Set parameters for model training:
 - a. Batch size: 4
 - b. Maximum token length: 512
 - c. Learning rate: 10^{-5}

Model Development

- Utilized the Transformer architecture for the model, which enables parallel computation for entire sequences.
- With Transformer, entire sentences can be processed simultaneously, rather than sequentially one word at a time from left to right.
- The major innovation of the Transformer architecture is the combination of attention-based representations and a CNN-like style of processing.
- Two main ideas in Transformer are:
 - a. Self-attention: Each word in the input sequence attends to all other words to create a representation.
 - b. Multihead attention: Allows the model to focus on different parts of the input sequence simultaneously.

Model Development

DistilBERTModel:

- It: DistilBERT base model with the following components:
- Embeddings:
- Word embeddings: Embedding layer with 30,522 tokens and 768 dimensions, including padding.
- Position embeddings: Embedding layer with 512 positions and 768 dimensions.
- Layer normalization: Layer normalization applied to the embeddings.
- Transformer Layer: Consists of 6 Transformer blocks.

Predictions

Validation Dataset

Accuracy: 0.1097

Loss: 3.35

Prediction Process:

Saved the trained model after training on the validation dataset.

Loaded the saved model for inference on the test dataset.

Ran the model on the test dataset and generated predictions.

Saved the predictions in a CSV file in the required format

Thank You !!

Team Members

- Rajeshwari Jadhav
- Sunidhi Prakash
- Arhita Kundu