Contribution ID: **16** Type: **not specified**

Transformer-based models for scientific text classification

Wednesday, 12 June 2024 09:40 (25 minutes)

The transformer model, introduced by Google in 2017, has become renowned in natural language processing (NLP). It represents a significant advancement completely departing from the mechanisms of Recurrent Neural Networks and Convolutional Neural Networks.
The features that contribute to the superior performance of transformers in NLP tasks include self-attention, multi-head attention, and positional encoding. The attention mechanism enables the model to extract dependency relationships between words, whereas the positional encoding extracts information about the position in the sequence of words.
Systematic literature reviews (SLR) help scientists to identify, select, access, and synthesize studies relevant to a specific topic. One crucial phase in conducting an SLR involves screening a vast number of articles sourced from various databases. This screening process is particularly challenging given the number of papers to be reviewed. For this reason, it is essential to perform this activity automatically with text classification.
In this context, we have employed transformer-based models to discern the primary topics covered in papers relevant to specific research areas, and group them into clusters. The primary aim is to develop an automated procedure that aids the scientific community in their search efforts by extracting pertinent information from vast amounts of documents.
During the study, we have used a range of transformer implementations, predominantly deriving from BERT, tailored to specific use cases including COVID-19, rehabilitation, and physics. Within the context of COVID-19, we have deployed models like BioBERT and PubMedELECTRA, achieving remarkable performance, particularly with PubMedBERT-large achieving a 0.9021 F1-score using 5-fold cross-validation.

Primary authors: CANAPARO, Marco (Istituto Nazionale di Fisica Nucleare); RONCHIERI, Elisabetta (Istituto Nazionale di Fisica Nucleare); TODESCHINI, Sofia Camilla (Università di Bologna); ZURLO, Giovanni (Università di Giovanni)

Presenter: ZURLO, Giovanni (Università di Giovanni)

Session Classification: Wednesday morning: Part I