HATANBAATAR VAN Erkhembileg
MICHEL Victor

This project deals with sentence classification out of the PubMed 200k RCT dataset, with two main strategies: a baseline model using Bag of words with logistic regression and an advanced model of pre-trained biomedical word embeddings (BioWordVec). Special attention has been devoted to the management issue of memory constraints, which is a commonly met difficulty when processing large text data sets.

Few major memory problems were faced while working on the project in two aspects: at pre-training word embedding loading and at TF-IDF vectorization for the text data. Few workarounds were applied for solving those problems:

Reducing max_features in TfidfVectorizer: This way, maximum features were capped at 3000. By doing so, not only much of memory can be saved, but the dimensionality of the TF-IDF vectors that were created is restricted to a number of acceptance, while at the same time yielding acceptable performance for the classification task.

Targeted Selection of Embeddings: For a model using BioWordVec, managing memory resources takes on crucial significance. We decided to compute mean vectors per sentence to reduce the dimensionality of data, and hence the memory footprint.

Standardized text cleaning in data preprocessing, the two classification approaches based on TF-IDF vectorization and respectively the use of pre-trained embeddings. The logistic regression to categorize sentences into relevant categories:

Results of the performance for both models :

Bag of words model + logistic regression: overall accuracy of 0.77 was achieved, with macro average scores for precision, recall, and F1 at 0.72, 0.69, and 0.70, respectively.

Model with pretrained embeddings — Marginal improvement over baseline model, yielding 0.78 accuracy with marginally better macro average scores.

This consistency was further intensified by cross-validation, where an average mean accuracy of 0.7701 was registered with the bag of words model.

The slight improvement with the use of the pre-trained embeddings speaks to the specialized knowledge that these vectors capture, which is critical for biomedical text classification. Such adjustments to help alleviate memory issues, like reducing max_features, have allowed conducting these analyses without significantly sacrificing the model performance.

In the project, the strength and the challenges of the approaches of text classification applied within the biomedical context, most importantly in relation to handling the memory constraints when the data scale is larger, were brought into the fore. This is an important note since even with the adjustments that are saving on memory, one can still attain robust performance.

For later work, the said methods can further improve classification while optimizing the usage of memory, like possibly looking into more advanced dimensionality reduction strategies or further enhancing parallel processing efficiency, or maybe even experimenting with designing deep learning models in those huge data amounts.