Automated pharmaceutical study selection in systematic review using machine learning algorithms

Shubham Uniyal, Master of Science in Computer Science

University of Dublin, Trinity College, 2022

Supervisor: Professor Arthur White

In this thesis, an automated systematic review was conducted on two datasets using text mining and machine learning strategies on the abstracts in the pharmaceutical domain. Manual Systematic Literature Review had already been performed on the two datasets. The datasets were preprocessed and tokenised to ensure the availability of the best set of features for the process of feature generations. Both the datasets were split into training and test data proportionately to their structure, ensuring an appropriate amount of training and testing information and a similar split between inclusions and exclusions. Term Frequency - Inverse Document Frequency(TFIDF) is used to generate textual features from the particular corpus of documents. Features having maximum weights are used for one dataset, and features having weights closer to the variance were used for the other. 1:2 downsampling was also performed for one dataset to balance the extremely unbalanced dataset. Four Classification models namely: Logistic Regression, Support Vector Machine (SVM), Naive Bayes (Gaussian) and Bagged Classification and Regression Trees (Bagged CART) are used to make predictions on test segments of the datasets. Naive Bayes outperforms all models with a Sensitivity of 1, predicting 4/4 and 5/5 relevant documents in both the datasets. The results are also compared with similar works in the automated Systematic Literature Review domain. The analysis shows that the strategy to use machine learning algotihms for Systematic Literature Review looks promising and should be further explored.