Automatic extraction and translation of the patients smoking status from free text using natural language processing

Authors: TBD – Andrea Caccamisi\textsuperscript{1,3}, Leif Jørgensen\textsuperscript{2}, Hercules Dalianis\textsuperscript{3}, Mats Rosenlund\textsuperscript{2,3}

1) Department of Computer and Systems Sciences, (DSV), Stockholm University
2) IMS Health, Stockholm
3) Karolinska Institutet, Solna, Stockholm

Background: Despite that smoking prevalence has declined in many western societies, the total number of smokers worldwide has increased steadily since 1980, still making it a leading cause of death. Nevertheless, smoking status is difficult to capture in real world data (RWD) such as healthcare registers and databases. Although smoking habits are often mentioned in Electronic Medical Records (EMR), there are few examples of methods to assess smoking status in EMRs.

Objective: To develop and test a text-mining method using machine learning algorithms that enables an automatic classification of smoker status (smoker, ex-smoker, non-smoker and unknown status) using data extracted from Swedish EMRs. The model performance in terms of presence of misclassification and sensitivity/specificity was compared with a semi-automatic rule-based model.

Methods: 32 predictive models were created with a machine learning software (Weka version 3.9), tweaking sentence frequency, classifier type, tokenization and attribute selection using a database of 80,000 classified terms. The models were evaluated using F-Score, ROC Area and Accuracy. The rate of misclassification of these models was calculated and the best performing model was compared with the rule-based model. The error weight matrix was used to select the best model.

Results: The top two performance measures were the Support Vector Machine (SVM) Sequential Minimal Optimization (SMO) and Nearest Neighbor (NN) algorithms. The best performing model was created adopting the SMO classifier and combinations of unigrams and bigrams as tokens. Sentence frequency and attributes selection did not improve the model. With data not used during the algorithm development, SMO achieved 98.12% accuracy and 0.981 F-Score versus 79.32% and 0.756 for the rule-based model.

Conclusions: A model using machine learning algorithms to automatically classify patients smoking status was successfully developed, enabling extraction of smoking variables directly from EMRs without extracting complete case notes. This opens for using smoking status in retrospective research without manual reading and classifying of free-text fields in EMRs.

Keywords: Data Mining, Text Mining, Smoking, Electronic Medical Records, Health Information Technology, Clinical Informatics.\textsuperscript{1}

\textsuperscript{1} This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden I Stockholm), permission number 2014/54-31/3 and 2013/267-31/3