Supplemental Digital Content 1 – Natural Language Processing Text Transformation

Natural language processing is a subfield of computer science and linguistics focused on transforming human language into structured data analyzable via machine learners. The overall process of text processing, all of which was used in this study, requires the following: careful medical and site-specific abbreviation expansion, spelling correction (spelling correction was achieved using edit distance and frequency analysis to create custom dictionaries in this study, though this can also be achieved using natural language processing / machine learning standards such as the Unified Medical Language System Knowledge Sources and related tools(1)), careful removal of less-meaningful words, and finally text normalization (lower case, removal of numbers and punctuations, truncation). The order of these processes is also important. In this study, to develop custom spelling correction, we used edit distance and frequency analyses. When analyzing text from sites outside of the Train/Test and Holdout datasets, there was small additional processing required for text processing. Although developing national standards on how to standardize procedure text would be nice, this is not the clinical reality, and thus a machine learning algorithm robust to variation in procedure text / leveraging site-specific nuances is important for the foreseeable future.

As a final step to transform text into numerical values, we adopted two methods based on classification algorithms. In the first method, used in support vector machine and random forest, we created frequency-inverse document frequency matrixes from the text. Each word in the frequency-inverse document frequency matrix is a numerical value representing the importance of the word to the text. Included were unigrams that occurred more than four times in the case pool and the top 500 bigrams (based on likelihood ratio). Terms with document frequency >0.9 were removed as these terms likely do not contain information to aid in classification. We joined the categorical and numerical features with frequency-inverse document frequency matrix and formed a large sparse matrix for use in the classification algorithm.

For our second method of text transformation we used word2vec representation(2) to maintain more context from the procedural text. Due to its sparsity, several machine learning algorithms do not work well with the frequency-inverse document frequency method and the word2vec model incorporates more of the text into its translated form. Word2vec represents each word in low dimensional continuous vector space where similar words are mapped to nearby points(3). These word vectors, trained on large corpus, can be used in classification. For more relevant word2vec training and potentially better results, we used pretrained word2vec embeddings, from biomedical text(4). Each word in this model is represented as 200 dimensional vector. Since most descriptions contain multiple words the result is a word matrix for each case. The word2vec method of text transformation was used for long short-term memory, extreme gradient boosting, and label-embedding attentive model classification algorithms. For the extreme gradient boosting model we performed an additional step: we averaged the input matrix for each case to reduce dimensions and have a sense of the average overall context of the text.

1. (US) NLoM. UMLS® Reference Manual. SPECIALIST Lexicon and Lexical Tools. Bethesda (MD)2009 Sep- [Available from: https://www.ncbi.nlm.nih.gov/books/NBK9676/ and https://www.nlm.nih.gov/research/umls/about_umls.html.

2. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. ArXiv e-prints [Internet]. 2013 January 01, 2013. Available from: https://ui.adsabs.harvard.edu/#abs/2013arXiv1301.3781M.
3. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. 2013:3111--9.
4. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S, editors. Distributional Semantics Resources for Biomedical Text Processing2013.