Report: Identifying smokers in unstructured data using Natural Language Processing.

Peter A Noble, July 9, 2019; Modified April 23, 2020.

Overview: The objective of the project was to extract of target words from unstructured electronic health records (EHRs) and demonstrate that an AI model can discriminate between patients who have smoked versus those who have not using words around the target and Artificial Intelligence (AI) modeling. The developed approach could be used to extract and predict any patient condition (smoking was used just as an example).

Smoking and obesity patient data was obtained from i2b2 web site https://www.i2b2.org/NLP/DataSets/Download.php. The training and testing data sets from both the Smoking and 2008 Obesity Challenges were combined. The combined file is called: EHR.txt and consists of 318,494 lines of unstructured data.

Program to determine the occurrence of all unique words in the EHR.txt file.
The occurrence of all words in the EHR.txt file was accomplished using the C++ program called 'unique_word_count.cpp'. The program outputs a text file consisting of two columns; the first being the unique word and the second being how many times it was found in the EHR.txt file. The purpose of this exercise is to identify all the words related to 'smoke'. All words relating to smoking were: "smoking", "smoke", "smoker", "smokes", "smoked", "tobacco", "cigar", "exsmoker" and "nonsmoker". These will be used as the target words in the C++ program outlined below.

Program to extract words on either side of the target word in the EHR.txt file.
The program (extract_target_words.cpp) identifies records that contain one of the possible target words (above) and extracts 10 words on either side of the target word. Here is an example:

ob_ 1	0	with	her	daughter	in	news	irv	in	she	denies
	tobacco	use	and	drinks	alcohol		rarely	allergies		
	codeine	and	benadryl		admission					

The ob_1 code identifies Obesity Patient 1. The '1' value indicates that one of the target words has been identified in the patient record. The target word was 'tobacco' and the words on either side of the target are shown.

Manual screening of words to determine if the patient actually smoked or not.
Only patient records that contained smoking related words were retained– all other data were excluded. Then the records were screened to verify if the patients had ever smoked or not (smoke.xlsx). Of the 622 patient records, 221 patients were classified as 'never smoked' and 401 patients were classified as 'smoked'. I randomized the rows to ensure no biased in the records. The top 311 records were

used for training the AI program and the bottom 311 records were used to test/validate the model.

Vectorization of the words using word2vec dataset.
The words were then converted to vectors using the word2vec dataset that consisted of 296,631 rows and 300 vector attributes: http://vectors.nlpl.eu/repository/11/3.zip. Each word in the word2vec dataset has 300 vector attributes. The program 'vector_extraction.cpp' imports the word vector dataset into RAM and then searches the 20-word extracted text for each patient. Of note, not all words are identified. For example, using ob_1 record above, the words found in the word2vec dataset were:

ob_1 0 daughter news irv tobacco use drinks alcohol codeine admission

The ob_1 identifies Obesity Patient 1. The '0' indicates that the patient was identified as a non-smoker. The remaining words were those found in the word2vec dataset. Note that the words: 'with', 'daughter', 'in', 'she', 'denies', 'and' 'rarely', 'allergies', and 'benadryl' were not found. I emphasize 'denies' because it is a crucial word in the medical record for stating that the patient does not smoke. Apparently, the word 'denys' and 'deny' were found in the vector to word dataset – but not 'denies'.

Here is an example of a patient who smoked:

ob_405 1 story complex dtr lives upstairs 2 3 pack year smoking hx quit 30 years ago no alcohol or ivdu

Here is a list of words identified by the word2vec dataset:

ob_405 1 story complex dtr pack year smoking hx quit 30 alcohol

ob_405 identifies Obesity Patient 405 who is a smoker. Note that the word2vec dataset did not find the following words: "lives", "upstairs", "23", "years ago", "no", "or" and "ivdu".

Each of the 'found' words was converted to a vector with *x* number of attributes. The maximum number of attributes for each word was 300 – but it might not necessary to use all the attributes to train the AI program. Here is an example of the word 'daughter' converted to the first 5 vector attributes:

daughter -0.105664 0.058506 0.018868 0.038212 0.024685…

The relationship between number of vector attributes and model accuracy/performance was conducted using 25, 50, 100, 150, and 300 attributes. Neuroet was used to model the data with the following parameters: min/max of 0 and 1, log sigmoid activation function for hidden layer and linear activation function for output.

The number of hidden neurons was optimized and model performance assessed based on accuracy and AUC values using test data (not used in training).

Table 1 shows that 100 vector attributes and 8 hidden neurons were optimal for predicting if a patient ever smoked based on electronic medical records. The best model yielded an AUC value of 86.3, meaning that 86% of the variability of the data was accounted for by the model.

Table 1. Relationship of number of vector attributes and model accuracy and ROC_AUC using test dataset.

Vector_attributes	Num_hidden neurons	Average Accuracy (n=3)	Average AUC (n=3)	Highest AUC
25	3	0.70	71.1	
25	5	0.73	74.3	
25	7	0.74	77.3	
25	9	0.70	72.1	
50	4	0.73	72.2	
50	6	0.75	78.4	
50	7	0.79	81.5	
50	8	0.74	78.0	
50	10	0.78	80.9	
100	5	0.78	82.2	
100	6	0.79	82.2	
100	7	0.77	81.8	
100	8	0.78	84.1	86.3
100	9	0.78	83.0	
150	6	0.78	82.1	
150	7	0.77	82.5	
150	8	0.78	83.1	
150	9	0.75	80.9	
300	5	0.73	70.7	
300	6	0.76	80.9	
300	8	0.75	79.5	