Prediction of Gout Diagnosis on ER admissions

by Michelle Ide

Springboard's Data Science Capstone 2

April 2021

Problem

To properly diagnose when a patient is suffering from Gout Emergency Room physicians look at the patient's Chief Complaint along with other factors. The Chief Complaint is written by medical staff, usually a nurse, recording what the patient states brought them to the emergency room. Patients often use terminology and phrases common to their region but not common to a physician's common regional dialect. These communication differences can cause barriers to proper diagnosis. When the same word means two different things to two different people, miscommunications occur.

Proposed Solution

This project creates a natural language processing tool that 'learns' the regional vocabulary and alerts physicians to the probability of the patient experiencing 'Gout'. The tool provides an alert of gout in the form of a probability based on the patient's complaint, alerting the physician to potential of Gout and follow-up on these cases.

Challenges

Medical text from written notes presents two unique challenges:

1) Abbreviations are more frequent than longform words. The preprocessing methods employed by most NLP algorithms will not only ignore abbreviations but, even when employing additional models to convert these to longform do not include the abbreviations used by medical staff in notation.

2) NER, Named Entity Recognition used by NLP processing models add weight to many of the wrong words. In medicine adverbs and adjectives are more important than nouns when performing a diagnosis. For example, I don't care who Carol is or that she is sitting (Noun, Verb) but I do care that she hurts while sitting (adverb) and her pain is extreme (adjective)

Both of these issues will greatly reduce the results by eliminating key words and placing more weight on unimportant words.

To address these issues creative preprocessing steps were tested after a baseline measurement. The baseline processed the text as normal and measure a Naive Bayes mode resulting in an average of 65%. The following processing steps were then tested prior to model tuning to address the challenges stated above.

- Tested SciSpacy model, "en_core_md", to replace medical abbreviations
- Added dictionary of known common abbreviations used in real-world medical environments.
- Eliminated stemming and lemmatization. These steps did not appear to add value and perhaps actually did more harm than good.

Method

A predictive model is built using Natural Language Processing tools from Python in a Jupyter Notebook and stored on GitHub.

Using data from MIMIC III (Medical Information Mart for Intensive Care III) database, two tables are used to train and test the model. Each table contains 3 columns of data:

1: **"Chief complaint"**: text description of patient's exact verbal complaint
2: **"Predict"**: ER physicians determination of Gout (Yes or No or Unknown)
3: **"Consensus"**: The Rhuematologist's feedback of Gout as a factor (Yes, No, or Unknown)

The model will produce a result in the form of a probability, 0% through 100%, that the patient is experiencing symptoms related to Gout where 0% means there is no indication the patient is experiencing Gout and 100% indicates the patient is experiencing Gout. This is not a diagnosis, a probability of 100% still requires follow-up by a physician. This tool is a flag to warn physicians of potential Gout circumstances that may otherwise go unnoticed due to communication barriers, allowing them to further evaluate.

DATA

The data contains 2 tables of data that are anonymized using "synthetic" formatting, replacement of identifying information with fake information such as name, address, etc. This synthetic data protects the patients personal information without influencing the model performance.

STATUS & NOTES

- Initial basic Naive Bayes model with a Count Vectorizer resulted in accuracy scores near 60%
- Removing digits improved accuracy to 66.6%
- Using a Tf-idf vectorizer improved score to 70%
- Tuning the Alpha parameter to 0.1 resulted in improvement in score to 87% with 3 false negatives and 10 false positives out of 100 results. Meaning, 89 out of 100 patient complaints are accurately labeled as having Gout, and 86 out of 100 are accurately labeled as not having Gout.
- 'Gout' is the second most common noun, and the term most associated with results
- 'Gout' was removed from the text and tested, as expected, accuracy scores were lower, but given the term was the most common and most impactful in the previous model, the score results were only reduced by 5%, from 87 to 82%. with 4 false negatives and 14 false positives. 83 out of 100 complaints accurately labeled for gout and 82 out of 100 accurately labeled when not experiencing gout.
- Additional testing of the model without stemming and lemmatization reduced the accuracy by less than 10% each and these steps were kept in the preprocessing steps.