Text Classification of Drug Abuse Patients and Embedding Analysis of Clinical Notes

Ben Fox
Introduction to Natural Language Processing
University of California, Santa Barbara

December 11, 2018

Abstract

The following aims to build multi-label classifiers for drug abuse patients' and reveal text correlations through clinical notes data. Random forest, logistic regression classifier chains, logistic regression label power-set, multi-label k-nearest neighbors, and long short term memory recurrent neural networks were used as multi-label classifiers to predict drug abuse. Further, word embeddings were learned from the corpus to examine correlations between drug abuse patients and other textual features. The dataset used was the MIMIC-III (Medical Information Mart for Intensive Care III) database, which is comprised of over forty thousand patients who stayed at the ICU in Beth Israel Deaconess Medical Center between 2001 and 2012. Clinical notes related to ICD-9 codes matching opioid abuse, alcohol abuse, cannabis abuse, amphetamine abuse, hallucinogen abuse, barbiturate abuse, and other drug abuse were used along with an equally random portion of the notes data not related to drug abuse to train the machine learning models. Logistic regression classifier chains produced the best results with an accuracy of 77.9% and f-measure of 69.5%. Word2Vec was used to train the word embeddings. Embeddings revealed interesting correlations between drug abuse patients and their careers, specifically bankers, chefs, bartenders, waitresses, clerks, and bookkeepers.

1 Introduction

Recently, electronic health records have become more commonplace in the medical community. With this consistent recording of patient information, data scientists have become

more and more interested in the large amounts of data, especially considering its value in predicting disease, examining relationships between diseases, survival analysis, and many other factors. Most of these data consist of numerical entries from lab work, vital sign recordings, or binary or multi-class representations of patient demographics, sex, etc. These data are very structured and were not used in the classification models in this report.

However, there exists a largely unstructured set of data within these records; specifically, the clinical notes written by doctors upon patient visits. These notes include a variety of information about the patient diagnosis, family history, previous visits, and much more. This paper aims to provide a model for classifying drug abuse and learning word embeddings using these clinical notes.

2 Related Work

2.1 Classification Tasks

Using clinical notes as a text classification task has not seen much attention over recent years. This is mainly due to the fact that other methods, outside of text classification can be used to classify disease. For example, using laboratory values to classify disease would reveal a much more accurate and reliable method as opposed to text classification.

Though, with diseases that are harder to diagnose such as HIV or critical limb ischemia and harder to diagnose mental health disorders, text classification can be extremely useful. In HIV patients, NLP and text classification helped reveal text indicative of high risk behavior and classifiers were trained with accuracies and F-measures around sixty percent [4]. Also, in critical limb ischemia, researchers were able to build a an NLP algorithm to identify critical limb ischemia patients via their clinical notes. The clinical notes classifier ended up having a higher positive predictive value than the original method of diagnosis [1].

Other NLP text classification tasks have used clinical notes to predict homelessness via clinical notes associated with medical records of veterans at the Veterans Affairs facilities[5]. This model used a human-reviewed corpus to train a classifier to predict homelessness. It resulted in high accuracy and f-measure greater than ninety percent. Lastly, NLP tasks have been used to predict suicidal ideation for psychiatric patients. This NLP model, though less predictive than structured models, could potentially help identify those at risk for suicide based off of a few simple, general mood questions [2].

2.2 Embeddings

Word embeddings in clinical notes, contrary to text classification of disease, has seen more and more interest over recent years. Many publications have attempted to learn classic methods such as GloVe, Word2Vec, and latent dirichlit allocation given their corpus of clinical notes; these are detailed by Dubois [3]. One method examines representing patients' diagnoses state in a recurrent neural network to predict future occurrences of disease and to extract relevant features from the data. Another, treats clincial notes as documents and learns document level embeddings, but has been found to be hard to reproduce [3].

Additionally in many of these papers, clinical notes are anonymized (rendering the words un-ordered) [3], requiring the words to be regenerated in a random sequence for embedding training. This is contrary to what is done in this paper.

3 Data

3.1 Format

The dataset used was the MIMIC-III (Medical Information Mart for Intensive Care III) database [6], which is comprised of over forty thousand patients who stayed at the ICU in Beth Israel Deaconess Medical Center between 2001 and 2012. Clinical notes related to ICD-9 codes matching opioid abuse, alcohol abuse, cannabis abuse, amphetamine abuse, hallucinogen abuse, barbiturate abuse, and other drug abuse were used along with an equally random portion of the notes data not related to drug abuse to train the machine learning models. There were a total of 9,269 hospital admissions in the dataset. 4,733 were related to drug abuse. Twenty percent of these data were used for a test set and ten percent for a validation set. These notes contain all notes for patients and are not out of sequence as seen in previous papers [3]. The notes are also in a structured format, which made it easier to remove unnecessary text.

3.2 Preprocessing

Clinical notes were tokenized, normalized, and stemmed using the Porter2 stemmer. Notes were analyzed for structures to remove unnecessary text such as dates, laboratory values, and other irrelevant information. Further, there was a class imbalance, since alcohol abuse is by far the most common drug abuse. Class weights were thus calculated for each class and applied to the machine learning algorithms.

3.3 Classification Models

The classification models used were only those that supported multi-label classification, since many patients were diagnosed with more than one drug abuse. Clinical notes text was transformed into a tf-IDF matrix with an n-gram of one to three and removing ten percent of corpus specific stop words. The random forest was trained with eighty trees and the important word features were extracted for visualization. Further, a classifier chain of logistic regression classifiers was trained. Random forest and naive bayes chains were also trained; however, logistic regression produced the best results. A Label powerset classifier with logistic regression (outperforming random forest and naive bayes) was also trained on the model. There were a total of fifty-eight power sets. Multi-label k nearest neighbors was also performed with a k of five. Lastly, a long short term memory recurrent neural network was trained with an embedding layer, dropout of 20%, and sigmoid activation function.

3.4 Word Embeddings

Word embeddings were trained on half of the entire corpus of notes (1041590 total notes) using Word2Vec with a window of 5. The vocabulary size was 72,871. This was done to visualize correlations between words relating to drug abuse. Principal component analysis (PCA) was performed to visualize this.

4 Results

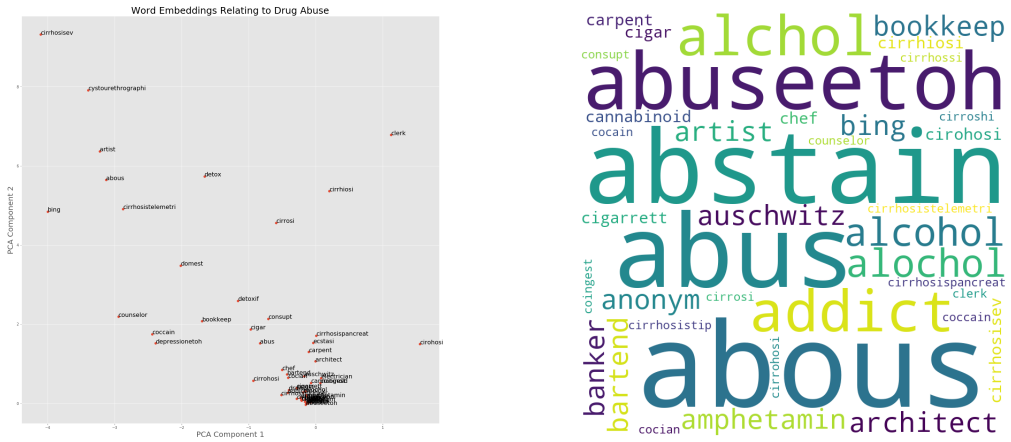
The accuracies and f-measures for the various multi-label models trained are seen below in Table 1.

Classifier	**Accuracy**	**F**
Random Forest	76.5%	63.5%
LR Classifier Chain	77.9%	69.5%
LR Label Powerset	78.0%	68.3%
ML KNN	74.0%	65.0%
LSTM	85.9%	62.7%

Table 1: Model Accuracies and F-measures

The most important splitting features from random forest did not reveal interesting results. The word embeddings most related to 'drug' and 'abus' are displayed in the PCA

plot and word cloud in figure 1.



(a) PCA Embeddings

(b) Embeddings Wordcloud

Figure 1: Visualizing Words Relating to Drug Abuse

5 Discussion and Concluding Remarks

As seen in the results section, the models trained to classify drug abuse based off of clinical notes produced good results. LSTM was the most accurate while the logistic regression classifier chain had the highest f-measure. Likely logistic regression classifier chain is the best model given the dataset because of the ability to preserve the relations between labels.

Additionally, random forest important features did not reveal interesting textual results. Word embeddings, though, revealed many interesting characteristics of drug abuse. PCA highlighted these correlations. Specifically, some careers of many of those who are drug abuse patients: bankers, chefs, bartenders, waitresses, clerks, and bookkeepers.

This study could have been improved with incorrect spelling correction, as seen in the embedding word clouds. Further, more data would have been valuable due to the class imbalance.

Overall, text classification of drug abuse was successfully modeled and word embeddings were learned from the MIMIC dataset to reveal interesting characteristics of drug abuse patients. Further studies should be done to delve deeper into the embeddings, and the text classifier models could be generalized to potentially assist in diagnosing drug abuse problems or those at risk for drug abuse without a doctor present, similar to the suicidal ideation study [2].

References

[1] Naveed Afzal, Vishnu Priya Mallipeddi, Sunghwan Sohn, Hongfang Liu, Rajeev Chaudhry, Christopher G Scott, Iftikhar J Kullo, and Adelaide M Arruda-Olson. Natural language processing of clinical notes for identification of critical limb ischemia. *International journal of medical informatics*, 111:83–89, 2018.

[2] Benjamin L Cook, Ana M Progovac, Pei Chen, Brian Mullin, Sherry Hou, and Enrique Baca-Garcia. Novel use of natural language processing (nlp) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in madrid. *Computational and mathematical methods in medicine*, 2016, 2016.

[3] Sebastien Dubois. Learning effective embeddings from medical notes. 2017.

[4] Daniel J Feller, Jason Zucker, Michael T Yin, Peter Gordon, and Noémie Elhadad. Using clinical notes and natural language processing for automated hiv risk assessment. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 77(2):160–166, 2018.

[5] Adi V Gundlapalli, Marjorie E Carter, Miland Palmer, Thomas Ginter, Andrew Redd, Steven Pickard, Shuying Shen, Brett South, Guy Divita, Scott Duvall, et al. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among us veterans. In *AMIA Annual Symposium Proceedings*, volume 2013, page 537. American Medical Informatics Association, 2013.

[6] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.