Machine Learning Text Mining Methods for Clinical Studies Classification by Clinical Domain

By

Connie Sosa

A Capstone Project Paper Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

In

Data Science

University of Wisconsin – La Crosse

La Crosse, Wisconsin

(December 2017 of Degree Award)

ACKNOWLEDGEMENTS

I would like to thank Sheri Tibbs at the Duke Clinical Research Institute (DCRI) for the opportunity to develop this machine learning text classifier using the Aggregate Analysis of Clinical Trials database. Also, I would like to thank Dr. Karen Chiswell at the DCRI for providing the dataset with clinician labeled clinical studies.

ABSTRACT

Machine Learning Text Mining Methods for Clinical Studies Classification by Clinical Domain

Connie Sosa

This project explored automated machine learning methods to classify clinical studies into clinical domains. This includes unsupervised clustering methods and supervised classification algorithms for text analysis on the 'Brief Summary' description of clinical studies. Cross Industry Standard Process for Data Mining methodology was employed for the data mining process.

Latent Dirichlet Allocation, a probabilistic generative model, was applied to help view the frequent words used in different groups of clinical studies. Supervised machine learning methods were applied to a dataset that contains clinician labeled studies. Classification algorithms SVMs, LDA, Maximum Entropy, Bagging, Boosting, and Random Forests were used for the model fitting process. Results of various performance metrics from each model were evaluated.

Random forests algorithm, a decision tree-based algorithm that pools decisions from multiple classifiers, has shown to be the best resulting model. Additional strategies were presented to further improve classification.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS 3

ABSTRACT 4

LIST OF TABLES 6

LIST OF FIGURES 7

1. INTRODUCTION 8
1.1 Project Background 8
1.2 Objectives 8

2. LITERATURE REVIEW 9

3. METHODOLOGY 10
3.1 Sample Population 11
3.2 Data Collection 12
3.3 Statistical and Machine Learning Techniques 13
3.3.1 Model Selection 13
3.3.1.1 Algorithms 14
3.3.1.2 Performance Metrics 14
3.3.1.3 Resulting Best Model 18

4. RESULTS AND PROJECT OUTCOMES 20
4.1 Results 20
4.2 Project Outcomes 21

5. SUMMARY AND CONCLUSION 23
5.1 Summary 23
5.2 Conclusion 23

REFERENCES 25

LIST OF TABLES

Table 3.2-1 AACT Data Dictionary (Clinical Trials Transformation Initiative, 2017)..... 13

Table 3.3.1.2-1 Confusion Matrix SVMs 15

Table 3.3.1.2-2 Confusion Matrix LDA 15

Table 3.3.1.2-3 Confusion Matrix Maximum Entropy 15

Table 3.3.1.2-4 Confusion Matrix Boosting 15

Table 3.3.1.2-5 Confusion Matrix Bagging 16

Table 3.3.1.2-6 Confusion Matrix Random Forests 16

Table 3.3.1.2-7 Overall Statistics for Bootstrap Aggregation (Bagging) 17

Table 3.3.1.2-8 Overall Statistics for Random Forests 18

Table 3.3.1.3-1 Overall Algorithm Performance 18

Table 4.1-1 Random Forests Model Assessment by Class 21

Table 4.1-2 Variable Importance - Top 20 21

LIST OF FIGURES

Figure 3-0-1 CRISP-DM Process Diagram (Bukralia, 2015) 11

Figure 3.1-1 Sample Population Clinical Study Domain Distribution 11

Figure 3.2-1 AACT Data Schema (Clinical Trials Transformation Initiative, 2017) 12

Figure 3.3.1.1-1 Latent Dirichlet Allocation (LDA) Topic Modeling (Phase 4 Clinical Studies) 14

Figure 4.1-1 Random Forests, Estimation of the Prediction Error by Class 20

Figure 4.2-1 An Example of the Classifier Labeled Oncology Clinical Study (Clinwiki, 2017). 22

1. INTRODUCTION

1.1 Project Background

National Institutes of Health (NIH) defines clinical trials as

> "*research studies that explore whether a medical strategy, treatment, or device is safe and effective for humans. These studies also may show which medical approaches work best for certain illnesses or groups of people. Clinical trials produce the best data available for health care decision making.*" (National Institutes of Health)

ClinicalTrials.gov provides the description of the system for tracking clinical trials. ClinicalTrials.gov is a registry and results database of publicly and privately supported clinical studies, this registry currently contains over 100,000 research studies conducted in more than 170 countries and is widely used both by medical professionals and the public. New research studies are being submitted to the registry by their respective sponsors (or sponsors' designees) at a rate of approximately 350 per week. (National Institutes of Health, 2017).

Clinical Trials Transformation Initiative, a public-private partnership by FDA and Duke University, developed the Aggregate Analysis of Clinical Trials (AACT) dataset that aggregates the content of ClinicalTrials.gov and makes access of clinical studies information available to the public. It provides health care professional, patients, and their family members easy access to information about clinical studies. (Tasneem et al., 2012).

1.2 Objectives

Traditionally, clinical trial studies are classified manually into Medical Subject Headings - MeSH. (Tasneem et al., 2012). This project aimed to utilize machine learning methods to help facilitate the classification of studies in the clinical research space. To increase the efficiency of categorization, the use of automated methods and text mining algorithms was explored to analyze the brief summary text description of clinical studies from the Aggregate Analysis of ClinicalTrials.gov (AACT) dataset. The objective of this project was to explore text mining algorithms to classify clinical trials into clinical domains, which can potentially be used by clinical researchers, clinicians, healthcare partners with the goal of improving the navigation of clinical domain information.

2. LITERATURE REVIEW

Various papers regarding text mining in the medical related field were reviewed along with general literature in supervised and unsupervised methods. Key papers considered includes:

Spasic et al. described in the International Journal of Medical Informatics paper how text mining uses techniques from natural language processing, data mining, and machine learning to process large corpus of text for mapping documents to categories based on their content. The SVMs, random forests, boosting algorithms were explored in their paper. (Spasic et al., 2014).

Blei, in his 2012 ACM article, detailed several probabilistic topic models that provide statistical solutions to the problem of managing large corpora of documents. Latent Dirichlet Allocation (LDA), a probabilistic generative model, was one of his algorithms explored. LDA is a popular topic modeling technique for document clustering. It is a probabilistic analysis of text data that guides the keyword selection and content optimization strategies. (Blei, 2012).

Additionally, decision tree-based methods were considered. Decision tree-based methods, discriminative models, involve segmenting or grouping the predictor space into number of simple regions. Multiple decision trees, such as found in Random Forests, Bagging, and Boosting are meta-algorithms that pool decisions from multiple classifiers. They use ensemble methods that combine and average the estimates from multiple machine learning algorithms to make better predictions (Hastie, Tibshirani, James, and Witten, 2015).

DeVille explained in his *Text Mining with 'Holographic' Decision Tree Ensembles* paper that multiple decision-trees classifiers can dramatically increase classification and prediction performance over single-methods classifiers. (DeVille).

These articles contributed to the basis for analysis and applying predictive methods for this capstone project.

3. METHODOLOGY

There are multiple phases of the data mining process. This project employed the "Cross Industry Standard Process for Data Mining" CRISP-DM methodology. It is a widely used analytics framework in the industry. CRISP-DM has six major process steps. Listed below is the brief description of these phases (IBM SPSS Modeler CRISP-DM Guide, 2011).

- Business Understanding - This phase involves understanding the project objective from a business perspective. This knowledge is used for defining the data mining problem and a plan to solve such problem.

- Data Understanding – This step involves data collection and seeks to understand the data question including its quality and its characteristics.

- Data Preparation – This phase involves making the initial data into a final data form ready for use by the modeling tools. This can be an iterative process and might be performed multiple times. This may include handling missing values, data cleaning or transforming data values into usable types (e.g. binary) for data mining.

- Modeling – This phase involves the application and the selection of models, or the training of models where parameters are determined to derive at the optimal model. This step can be an iterative process, which might necessitate going back to the previous phase of data preparation.

- Evaluation – This phase involves the comparisons of different algorithms using various performance metrics. During this phase, the resulting models are evaluated for suitability of the business objectives, and the decision on the use of the data mining classifier is reached.

- Deployment – This phase involves the preparation of the model for use by the client. As the requirement dictates, this can be simple generation of reports, produce data in client defined file formats, or deploy a complex data mining process system.

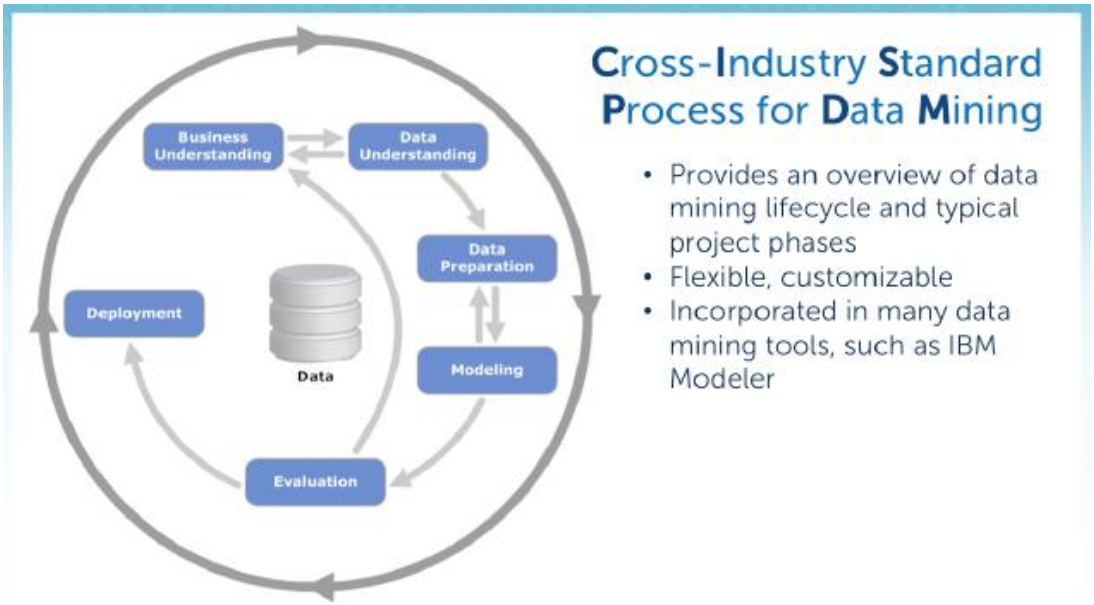

Figure 3-0-1 CRISP-DM Process Diagram (Bukralia, 2015)

3.1 Sample Population

The 1,000 clinical studies were obtained as follow. All the clinical studies that were registered after September 2007 and were restricted to the 'Interventional' study type. The content was downloaded from ClinicalTrials.gov on September 27, 2010. This subset contains 40,970 studies. Subsequently a random sample of 1,000 of these 40,970 studies were selected. (Tasneem et al., 2012).

The clinician-assigned classifications for the dataset are cardiovascular (CV), mental health (MH), Oncology (Onc), and Other. The sample population summary statistics showed that the dataset exhibits a modest imbalanced class distribution. Figure 3.1-1 below displays the class distribution of clinician-assigned clinical domain.

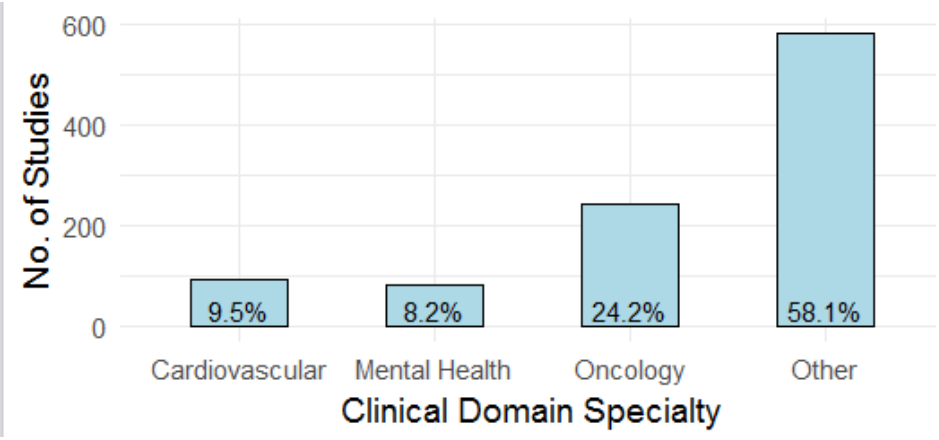

Figure 3.1-1 Sample Population Clinical Study Domain Distribution

3.2 Data Collection

- Training and Validation Dataset

The data used for this project was the validation clinical study dataset used in the publication, "The Database for Aggregate Analysis of ClinicalTrials.gov (AACT) and Subsequent Regrouping by Clinical Specialty." PloS One 7.3 (2012): e33677 authored by Tasneem, et al. The dataset contains the clinical review of 1,000 clinical studies. These are the final adjudicated results from the manual review by clinicians. This data set was used for model training and validation.

- Production Dataset

CCTI AACT database provided the data for the unlabeled clinical trial studies. The clinical trial studies dataset is publicly accessible. The AACT database contains 40+ tables, 300+ fields, with data types of Boolean, integer, character, text, etc. (Clinical Trials Transformation Initiative, 2017). This data was used as the production data set for which the classifier was applied to. Figure 3.2-1 shows the partial diagram of the AACT Data Schema. (Clinical Trials Transformation Initiative, 2017).

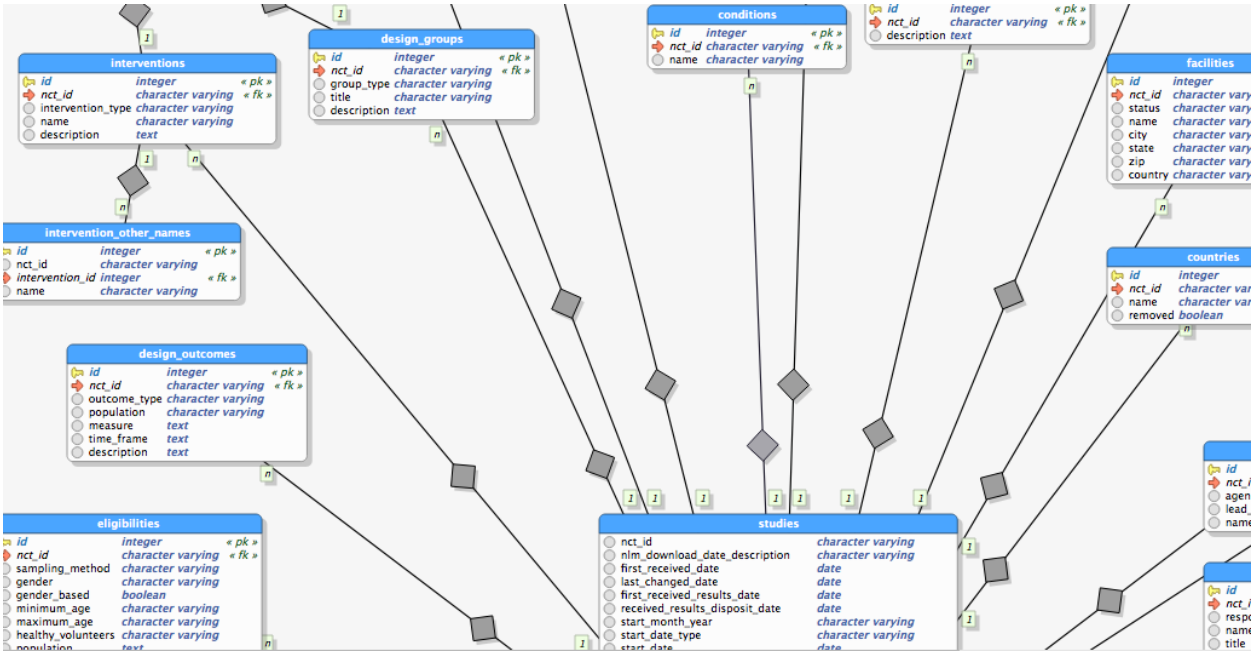

Figure 3.2-1 AACT Data Schema (Clinical Trials Transformation Initiative, 2017)

The main data of interest for fitting the text classifier model are in the 'Studies' and 'Brief Summaries' tables. 'Studies' table contains the basic information about clinical studies, and 'Brief Summaries' table contains text description of clinical studies. (Clinical Trials

Transformation Initiative, 2017). Other tables and fields were used for querying criteria for the purpose data retrieval.

table	description	rows per study	db section	nlm doc
...				
Brief_Summaries	A single text column that provides a brief description of the study.	one	Protocol	BriefSummary
...				
Detailed_Descriptions	A single text column that provides a detailed description of the study protocol.	one	Protocol	DetailedDescription
...				
Studies	Basic info about study, including study title, date study registered with ClinicalTrials.gov, date results first posted to ClinicalTrials.gov, dates for study start and completion, phase of study, enrollment status, planned or actual enrollment, number of study arms/groups, etc.	one	Protocol & Results	

Table 3.2-1 AACT Data Dictionary (Clinical Trials Transformation Initiative, 2017)

3.3 Statistical and Machine Learning Techniques

Supervised Learning is a method used to learn the relationship between independent attributes and a dependent attribute. The independent attribute is the summary text description for the clinical study, and the dependent attribute is the clinical domain specialty label for the study.

In supervised learning, an algorithm uses a pre-defined inputs or predictors to predict the values of the output or response, it produces an inferred function, which is called a classifier. Classifier is a system that performs classification. A classification problem is the process of analyzing input data and predicting it an output label. The predicted label here is a discrete category in the clinical domain. Based on the training dataset, the algorithm generalizes it to unseen data, and makes a prediction of the clinical domain specialty when a new clinical study is given. (Hastie, Tibshirani, and Friedman, 2009).

3.3.1 Model Selection

This project explored various machine learning algorithms and developed a text classifier to automatically label the clinical studies into clinical domain specialty. The estimation of the performance of these different models is done in the model selection process. It is in the model selection process, the best model is chosen. (Hastie, Tibshirani, James, and Witten 2015).

These clinical studies were split into 75% and 25%, training and test dataset. Three quarters of the studies were used as training set (750 studies), and the remaining studies were used as the test dataset.

3.3.1.1 Algorithms

Several of the supervised classification algorithms and unsupervised clustering methods were applied for this problem. This includes unsupervised clustering method -- Latent Dirichlet Allocation (LDA), commonly used supervised classification algorithms -- Support Vector Machines (SVMs), Scaled Linear Discriminant Analysis (SLDA), and Maximum Entropy, as well as decision tree-based methods -- Bootstrap Aggregation (Bagging), Boosting, and Random Forests.

Exploratory data analysis was performed using unsupervised clustering method of Latent Dirichlet Allocation for topic modeling phase 4 clinical studies. LDAvis R package was used for the topic model visualization. (Sievert C. and Shirley K., 2015). It revealed many overlapping groups, see Figure 3.3.1.1-1 below.

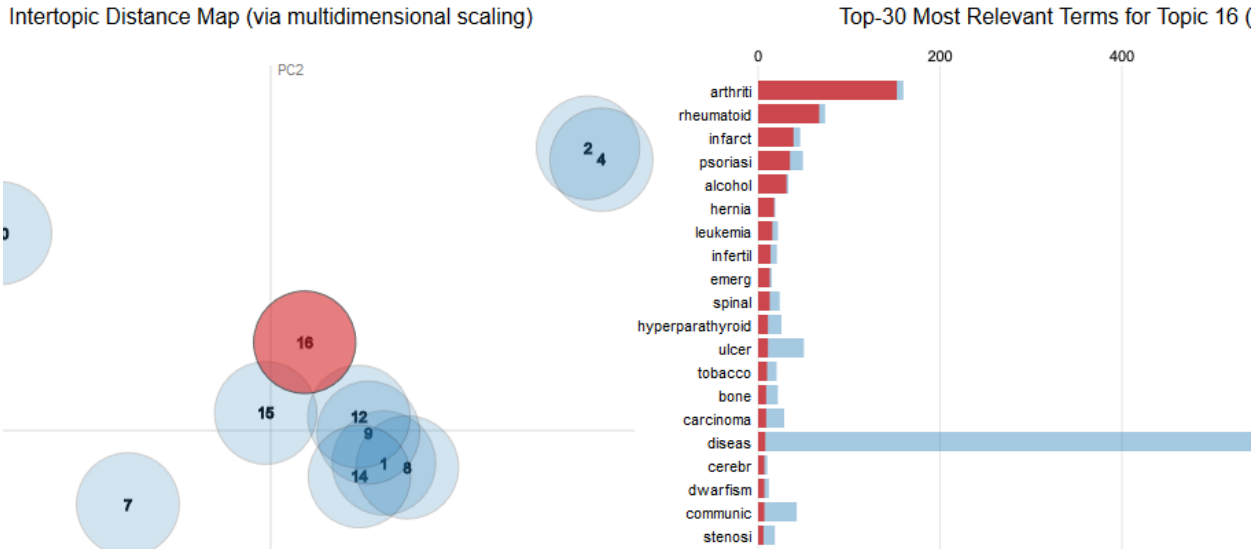

Figure 3.3.1.1-1 Latent Dirichlet Allocation (LDA) Topic Modeling (Phase 4 Clinical Studies)

3.3.1.2 Performance Metrics

- Confusion Matrix

This classification system has been trained to distinguish clinical studies among ‘CV’, ‘MH’, ‘Onc’, and ‘Other’ clinical domains. Confusion matrix summarizes the counts of the correct and incorrect predictions broken down by the four classes. Hastie, Tibshirani, James, and Witten explained that the diagonal elements of the confusion matrix indicate correct predictions, and off-diagonals indicate incorrect predictions. It provides information on the types of errors, and can be used to calculate precision and recall rates. (Hastie, Tibshirani, James, and Witten, 2015). Confusion matrix for each algorithm is shown below.

	Predicted			
Actual	Cardiovascular	Mental Health	Oncology	Other
Cardiovascular	7	0	1	12
Mental Health	0	5	2	15
Oncology	0	0	43	10
Other	3	3	1	148

Table 1: SVM Confusion Matrix

Table 3.3.1.2-1 Confusion Matrix SVMs

	Predicted			
Actual	Cardiovascular	Mental Health	Oncology	Other
Cardiovascular	8	0	2	10
Mental Health	0	10	2	10
Oncology	1	0	38	14
Other	6	9	5	135

Table 2: SLDA Confusion Matrix

Table 3.3.1.2-2 Confusion Matrix LDA

	Predicted			
Actual	Cardiovascular	Mental Health	Oncology	Other
Cardiovascular	8	0	0	12
Mental Health	1	10	2	9
Oncology	0	0	44	9
Other	21	9	6	119

Table 3: Max Entropy Confusion Matrix

Table 3.3.1.2-3 Confusion Matrix Maximum Entropy

	Predicted			
Actual	Cardiovascular	Mental Health	Oncology	Other
Cardiovascular	10	0	0	10
Mental Health	1	10	1	10
Oncology	0	2	44	7
Other	22	12	2	119

Table 4: Boosting Confusion Matrix

Table 3.3.1.2-4 Confusion Matrix Boosting

	Predicted			
Actual	Cardiovascular	Mental Health	Oncology	Other
Cardiovascular	8	0	0	12
Mental Health	0	7	1	14
Oncology	0	0	43	10
Other	6	2	2	145

Table 5: Bootstrap Aggregation Confusion Matrix

Table 3.3.1.2-5 Confusion Matrix Bagging

	Predicted			
Actual	Cardiovascular	Mental Health	Oncology	Other
Cardiovascular	9	0	0	11
Mental Health	0	7	1	14
Oncology	0	0	47	6
Other	5	0	4	146

Table 6: Random Forests Confusion Matrix

Table 3.3.1.2-6 Confusion Matrix Random Forests

- Precision

Precision metric refers to how often the algorithm predicts a clinical study as belonging to a class that actually belongs to that class. In other words, precision tells us the portion of studies an algorithm deems to be about CV are actually about CV (based on the gold standard of clinician assigned labels). (Jurka et al. 2013). Precision score of 1.0 for a class CV means that every study labeled as belonging to class CV does indeed belong to class CV.

Precision = Pr (study's true label is CV | study is predicted as CV**)**

The precision rate for classifying cardiovascular studies is

$$Precision = \frac{Total\ number\ of\ correctly\ predicted\ CVs}{Total\ number\ of\ model\ predicted\ CVs}$$

- Recall

Recall is the true positive rate. Recall measures the probability of detection. (Hastie, Tibshirani, James, and Witten 2015). It gives the percentage of true CV that are identified. Recall metric refers to the portion of clinical studies in a class the algorithm correctly assigns to that class. In other words, out of all the clinicians labeled CV clinical studies, how many did the classifier classify as CV.

Recall = Pr (study is predicted as CV | study's true label is CV)

The recall rate for classifying cardiovascular studies is

$$Recall = \frac{Total\ number\ of\ correctly\ precticed\ CVs}{Total\ number\ of\ all\ the\ studies\ that\ are\ actually\ CVs}$$

- F-score

A machine learning algorithm with high recall for class CV means the classifier can identify most of the CV clinical studies out there. A machine learning algorithm with high precision for the CV class means that most of the studies labeled as belonging to class CV indeed belong to class CV. F-score produces a weighted average of both precision and recall. One being the highest level of performance, and zero being the lowest. F1-Score Wikipedia page defines F-score as the harmonic mean of precision and recall. (Wikipedia, July 2017).

$$F1 = \frac{2\ x\ Precision\ x\ Recall}{Precision\ +\ Recall}$$

Tables 3.3.1.2-7 and 8 below summarize the overall statistics for Bootstrap Aggregation (Bagging) and Random Forests algorithms.

```
Accuracy : 0.812
95% CI : (0.758, 0.8585)
No Information Rate : 0.62
P-Value [Acc > NIR] : 3.892e-11

Kappa : 0.6273
Mcnemar's Test P-Value : NA

Statistics by Class:
                     CV          MH          ONC         OTHER
                     Class: 1    Class: 2    Class: 3    Class: 4
Sensitivity          0.4000      0.3182      0.8113      0.9355
Specificity          0.9739      0.9912      0.9848      0.6211
Pos Pred Value       0.5714      0.7778      0.9348      0.8011
Neg Pred Value       0.9492      0.9378      0.9510      0.8551
Prevalence           0.0800      0.0880      0.2120      0.6200
Detection Rate       0.0320      0.0280      0.1720      0.5800
Detection Prevalence 0.0560      0.0360      0.1840      0.7240
Balanced Accuracy    0.6870      0.6547      0.8980      0.7783
```

Table 3.3.1.2-7 Overall Statistics for Bootstrap Aggregation (Bagging)

```
Accuracy : 0.836
95% CI : (0.7842, 0.8797)
No Information Rate : 0.62
P-Value [Acc > NIR] : 6.725e-14

Kappa : 0.6784
Mcnemar's Test P-Value : NA

Statistics by Class:
                     CV          MH          ONC         OTHER
                     Class: 1    Class: 2    Class: 3    Class: 4
Sensitivity          0.4500      0.3182      0.8868      0.9419
Specificity          0.9783      1.0000      0.9746      0.6737
Pos Pred Value       0.6429      1.0000      0.9038      0.8249
Neg Pred Value       0.9534      0.9383      0.9697      0.8767
Prevalence           0.0800      0.0880      0.2120      0.6200
Detection Rate       0.0360      0.0280      0.1880      0.5840
Detection Prevalence 0.0560      0.0280      0.2080      0.7080
Balanced Accuracy    0.7141      0.6591      0.9307      0.8078
```

Table 3.3.1.2-8 Overall Statistics for Random Forests

3.3.1.3 Resulting Best Model

Table 3.3.1.3-1 contains the overall precision, recall, and F-score for all the algorithms utilized in fitting the model. Three performance metrics shown are the average for all four classes. The algorithm performance Comparison shows that Bootstrap Aggregation (Bagging) and Random Forests outperformed the others in terms of precision, recall, and F-score.

	Precision	Recall	F-score
SVM	0.76	0.58	0.64
SLDA	0.67	0.61	0.64
MAXENTROPY	0.61	0.61	0.61
LOGITBOOST	0.62	0.64	0.62
BAGGING	0.77	0.62	0.66
FORESTS	0.84	0.65	0.70

Table 7: Overall Algorithm Performance

Table 3.3.1.3-1 Overall Algorithm Performance

Based on these results, random forests, with F-score of 0.7, was the best among all six algorithms. The average of four classes has recall measure of 0.65, it indicates that the random forests algorithm correctly classified 65 clinical studies into their proper domain out of 100 clinical studies. In terms of conditional probability, it can be interpreted as this. Given 100 studies with their known true clinical domain, the likelihood of this algorithm predicting the study into its true domain is 65%.

Precision measure of 0.84 indicates that out of 100 clinical studies, random forests algorithm predicted each into a clinical domain, 84 clinical studies were correctly classified. It has the following conditional probability interpretation: Given the algorithm predicted clinical domain for 100 studies, 84 studies will have the correct prediction.

4. RESULTS AND PROJECT OUTCOMES

4.1 Results

Having chosen a final model for the clinical domain specialties classification problem, performance assessment is needed to determine how well this classifier would predict the clinical domain specialty class label on new unseen clinical studies data.

The learning algorithm results were subsequently checked against the clinician labeled clinical domain to determine the accuracy of the process. RTextTools library was utilized to test the algorithm accuracy. (Jurka et al. 2013).

The application of a cross validation technique or a test set is typically used to estimate classifier's prediction error or generalization error. According to Breiman and Cutler, the developers of the random forests algorithm, there is no need for a separate test set or performing cross-validation to get an unbiased estimate of the test set error since each tree is constructed using a different bootstrap sample from the original data. (Breiman and Cutler).

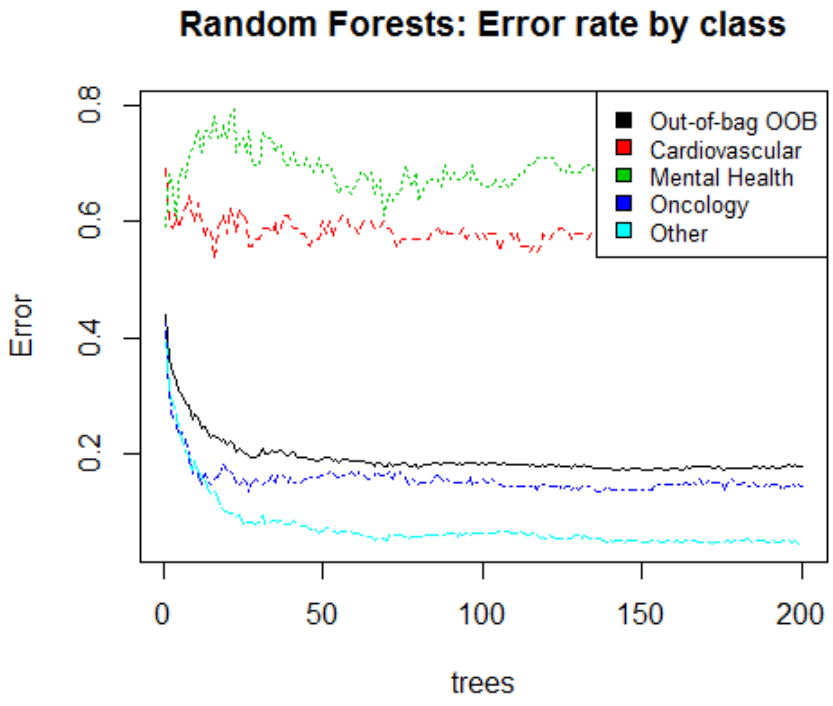

Figure 4.1-1 Random Forests, Estimation of the Prediction Error by Class

	PRECISION	RECALL	FSCORE
Cardiovascular	0.64	0.45	0.53
Mental Health	1	0.32	0.48
Oncology	0.9	0.89	0.89
Other	0.82	0.94	0.88

Table 4.1-1 Random Forests Model Assessment by Class

It was found that the classifier performs better for some clinical domain classification than for others. The high precision rate for Mental Health indicates that when the classifier predicts a clinical study to be in the Mental Health class, its accuracy is high. Low recall rate for Cardiovascular and Mental Health clinical domain is indicative of the classifier's lack of ability in finding these clinical studies.

Table 4.1-2 shows the Variable Importance. It lists the top discriminating variables used in the prediction process.

	Variable Importance		Top 20
1	cancer	11	rationale
2	chemotherapy	12	combination
3	tumor	13	cardiac
4	metastatic	14	cell
5	heart	15	studying
6	tumors	16	radiation
7	schizophrenia	17	disorder
8	cells	18	phase
9	advanced	19	artery
10	coronary	20	breast

Table 4.1-2 Variable Importance - Top 20

4.2 Project Outcomes

The text classifier performed well on classifying Oncology clinical domain studies. Its ability to identify oncology clinical studies based on the 'brief summary' description of the study is 89%. Although with high performance for clinical domain 'Other', the class label for studies classified into the 'Other' domain is not very useful, since these are studies that belong to the catch all clinical domain, and need future iterations of model fitting and model training with data that contains clinical studies of finer domain categorization.

This text classifier was then used to categorize unseen clinical studies from the AWS cloud-based PostgreSQL AACT database. The query criteria were set to retrieve phases 3 and 4 clinical studies where their corresponding value for the *"were results reported"* field from the

"*calculated values"* table have value of '*true'*. The text classifier predicted the clinical trial with the title of "A Study of Long-Term Rituximab (MabThera) Maintenance Therapy in Participants With Advanced Follicular Lymphoma" as an oncology study.

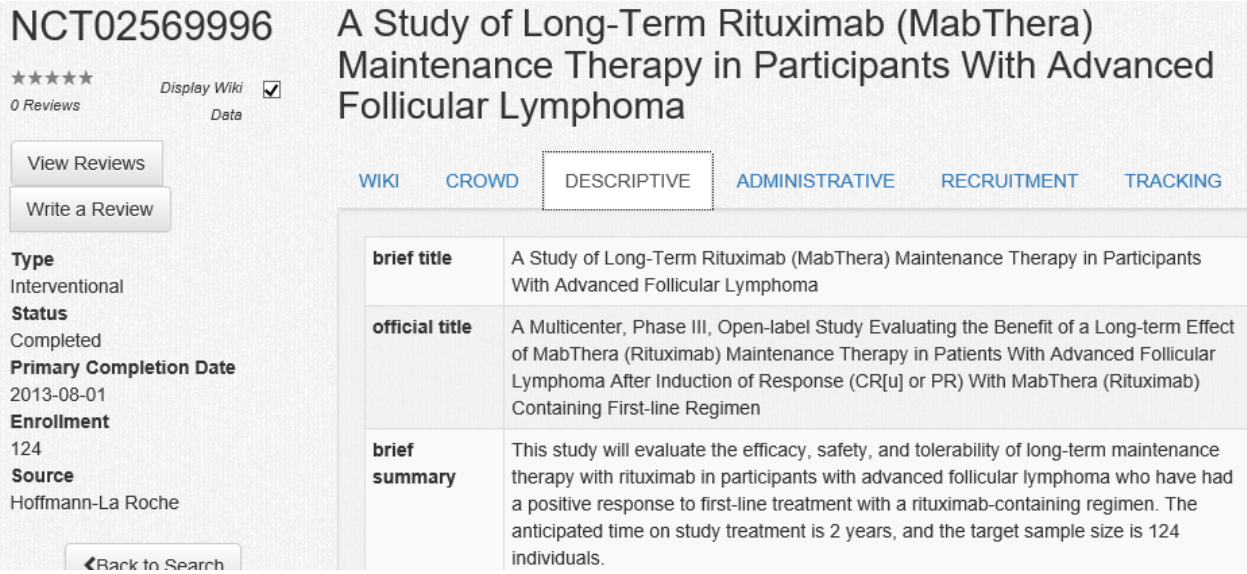

Figure 4.2-1 An Example of the Classifier Labeled Oncology Clinical Study (Clinwiki, 2017)

The resulting categorization of the clinical studies from this AACT dataset is formatted into Excel spreadsheets. The Excel spreadsheet is being provided to Sheri Tibbs at the Duke Clinical Research Institute. Methods of importing text classifier labeled oncology studies into the Clinwiki system are being explored. (Clinwiki, 2017).

5. SUMMARY AND CONCLUSION

5.1 Summary

Earlier exploratory analysis using an unsupervised clustering method, Latent Dirichlet Allocation (LDA), revealed many overlapping topics. The analysis was subjective and difficult to assess the results. After applying different classification methods, the Random Forests multi-class classifier was selected as the classifier to be utilized. The Random Forests multi-class classifier could identify 9 out of 10 oncology studies (recall), and had precision measure in the similar range. However, the classifier has low recall rate for CV and MH, it shows that the classifier may not readily detect studies that are of the cardiovascular or mental health clinical domains.

With the presence of modest class imbalance, determining the appropriate performance metrics was not as straight forward. Although accuracy for the classifier is at 83.6%, the predicted performance of some classes was not nearly as close to this number.

Initial task of developing a multi-class text classifier system that classifies clinical studies into different clinical domain specialties based on the summary study descriptions has shown disparity in performance among different classes. With the data available to fit the model, this classifier system indicated high fidelity for oncology domain classification, but low performance in identifying studies in the cardiovascular or mental health domains. This text classifier may show promise as a binary-classification classifier for identifying oncology clinical studies.

5.2 Conclusion

As part of reviewing results, consideration was given to how to improve the performance of the classifier. Among possible areas to consider, the nature of the data content was reviewed. According to ClinicalTrials.gov Protocol Registration Data Element Definitions, the ‘Brief Summary’ text data analyzed by the classifier is a short description of the clinical study and is limited to 5,000 characters. It is written in language intended for the public and it includes a brief statement of the clinical study's hypothesis. (ClinicalTrials.gov, Jun. 2017). Additionally, many studies contain much less than the possible 5,000 characters. As a next step to explore in improving predictability, other text fields can be considered. The National Library of Medicine documented that the data element for the ‘Detailed Description’ can contain up to 30,000 characters. It contains an extended description of the trial protocol and includes more technical information as compared to the ‘Brief Summary’ text. (ClinicalTrials.gov, Jun. 2017). Increasing the amount of text to be analyzed and utilizing the ‘Detailed Description’ of clinical study for future iteration of the text mining process could increase the fidelity of this text classifier. Though using the ‘Detailed Description’ data element looks promising in possibly improving classifier performance, special consideration needs to be considered due to the increased computational load that will result from analyzing a large corpus of text. It is recommended to

apply clustered computing or other Big Data technologies in the analysis of the larger text corpus.

REFERENCES

Blei D. (April 2012). *Probabilistic Topic Models.* Communications of the ACM Volume 55 Issue 4. Retrieved from doi:10.1145/2133806.2133826

Breiman, L. and Cutler, A. *Random Forests™.* Retrieved from https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Bukralia R. (Sept. 2015). *DS 700 Foundations of Data Science - Analytics Process.* University of Wisconsin. Retrieved Sept. 2015 from https://media.uwex.edu/content/ds/ds700/ds700_week3

Clinical Trials Transformation Initiative. (2017). *AACT Database Schema.* Retrieved 2017 from http://aact.ctti-clinicaltrials.org/schema

Clinical Trials Transformation Initiative. (2017). *AACT Data Dictionary.* Retrieved 2017 from http://aact.ctti-clinicaltrials.org/data_dictionary

Clinical Trials Transformation Initiative (Aug. 2017). *Aggregate Analysis of Clinical Trials Database.* Available from Hostname: aact-prod.cr4nrslb1lw7.us-east-1.rds.amazonaws.com

Clinwiki. (2017). *NCT02569996: A Study of Long-Term Rituximab (MabThera) Maintenance Therapy in Participants With Advanced Follicular Lymphoma.* Retrieved Aug. 2017 from https://clinwiki-dev.herokuapp.com/study/NCT02569996

DeVille, B. *Text Mining with "Holographic" Decision Tree Ensembles.* SAS Institute Inc. Retrieved from http://www2.sas.com/proceedings/sugi31/072-31.pdf

Hastie, T., Tibshirani R., and Friedman J. (2009). Springer. *The Elements of Statistical Learning-*

Data Mining, Inference, and Prediction (2nd ed.)

Hastie, T., Tibshirani R., James G., and Witten D. (2015). Springer. *An Introduction to Statistical Learning – with Applications in R*

IBM Corporation. (2011). *IBM SPSS Modeler CRISP-DM Guide.* Retrieved from ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf

Jurka, T., Collingwood, L., Boydstun, A., Grossman, E., and Atteveldt, W. (June 2013). *RTextTools: A Supervised Learning Package for Text Classification.* The R Journal Vol 5/1:6-12. Retrieved Jun. 2013 from https://www.researchgate.net/publication/283551286_RTextTools_A_Supervised_Learning_Package_for_Text_Classification

National Institutes of Health. *What are Clinical Trials?* Retrieved Aug. 2017 from https://www.nhlbi.nih.gov/studies/clinicaltrials

National Institutes of Health. (2017). *ClinicalTrials.gov Background* Retrieved 2017 from http://www.clinicaltrials.gov/ct2/info/about

National Institutes of Health. (Jun. 2017). *Protocol Registration Data Element Definitions for Interventional and Observational Studies* Retrieved Jun. 2017 from https://prsinfo.clinicaltrials.gov/definitions.html

Sievert C. and Shirley K. (2015). *LDAvis: Interactive Visualization of Topic Models.* Retrieved from https://cran.r-project.org/web/packages/LDAvis/index.html

Spasic, I., Livsey, J., Keane, J., and Nenadic, G. (Sept. 2014). *Text mining of cancer-related information: Review of current status and future directions.* International Journal of Medical Informatics, Volume 83, Issue 9. Retrieved from https://doi.org/10.1016/j.ijmedinf.2014.06.009

Tasneem, A. Aberle, L., Ananth, H., Chakraborty, S., Chiswell, K., McCourt, B., and Pietrobon, R. (Mar. 16, 2012) *The Database for Aggregate Analysis of ClinicalTrials.gov (AACT) and Subsequent Regrouping by Clinical Specialty.* Retrieved Mar. 2012 from doi.org/10.1371/journal.pone.0033677

Wikipedia. (July 2017). *F1 Score*. Retrieved July 2017 from https://en.wikipedia.org/wiki/F1_score