2018

Using Natural Language Processing and Machine Learning for Analyzing Clinical Notes in Sickle Cell Disease Patients

Shufa Khizra
Wright State University

Using Natural Language Processing and Machine Learning for Analyzing Clinical Notes in Sickle Cell Disease Patients

A Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

by

Shufa Khizra
B.S.C.S., Andhra University, 2014

2018
Wright State University

Wright State University
GRADUATE SCHOOL

December 5, 2018

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Shufa Khizra ENTITLED Using Natural Language Processing and Machine Learning for Analyzing Clinical Notes in Sickle Cell Disease Patients BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Tanvi Banerjee, Ph.D.
Thesis Director

Mateen Rizki, Ph.D.
Chair, Department of
Computer Science and Engineering

Committee on
Final Examination

Tanvi Banerjee, Ph.D.

Michelle Cheatham, Ph.D.

Mateen Rizki, Ph.D.

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

ABSTRACT

Khizra, Shufa. M.S., Department of Computer Science and Engineering, Wright State University, 2018. *Using Natural Language Processing and Machine Learning for Analyzing Clinical Notes in Sickle Cell Disease Patients.*

Sickle Cell Disease (SCD) is a hereditary disorder in red blood cells that can lead to excruciating pain episodes. SCD causes the normal red blood cells to distort its shape and turn into sickle shape. The distorted shape makes the hemoglobin inflexible and stick to the walls of the vessels thereby obstructing the free flow of blood and eventually making the tissues suffer from lack of oxygen. The lack of oxygen causes serious problems including Acute Chest Syndrome (ACS), stroke, infection, organ damage, and over the lifetime an SCD can harm a persons spleen, brain, kidneys, eyes, bones. Sickling of RBC can be triggered by a number of conditions such as dehydration, acidity, low levels of oxygen, stress, and change in temperature. There is no specific medication for pain crisis and the signs and symptoms varies from person to person, making it difficult to provide a common treatment for SCD and understanding the disease. It is believed that 90,000 to 100,000 American are affected by SCD. Myriad number of studies have been working on gaining better understanding of the disease and predict pain crisis and pain level. These studies help people to mitigate or prevent pain crisis by taking precautions. However, no study has used clinical notes to predict pain score and pain sentiment. Clinical notes provide patient specific information including procedures and medication; and can therefore help in predicting accurate scores.

Our study focuses on four research problems namely patient informative, pain informative, pain sentiment and pain scores using SCD data. Notes are taken for a patient during hospitalization but only few provide beneficial information, therefore patient informative and pain informative helps healthcare professionals to scan through the notes that can provide valuable information from all the clinical notes maintained. Pain sentiment and pain score predict the change in pain and pain level for a particular note. Our study experi-

mented with two feature sets, firstly features obtained from cTAKES, a Natural Language Processing (NLP) and secondly features obtained from text using NLP techniques. Four supervised machine learning models namely Logistic Regression, Random Forest, Support Vector Machines, and Multinomial Naive Bayes are built on these different sets of features. From the results, it can be noted that cTAKES features are performing well for SCD problem for all the four research problems with F1 score ranging from 0.40 to 0.86. This indicates that there is promise for using NLP techniques in clinical notes as a means to better understand pain in SCD patients.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview

Sickle Cell Disease (SCD) is a group of hereditary blood disorders in red blood cells. Cells in tissue need oxygen, and the hemoglobin in red blood cells carries oxygen to different parts of the body. Hemoglobin is disc shaped, and it is the shape of hemoglobin that allows it to freely flow through the veins and deliver oxygen to organs; however in SCD patients, the shape of the hemoglobin is distorted into a sickle shaped hence the name sickle cell disease (demonstrated in Figure 1.1 [1]). This distorted shape makes the hemoglobin inflexible and sticks to the vessel walls leading to blockage and eventually preventing the tissues from receiving oxygen. Poor oxygen delivery to the tissues can cause a sudden and severe pain crisis. Moreover, these attacks come without warning, and the patient needs to go a to hospital for effective treatment. Lack of oxygen can also cause serious problems, including Acute Chest Syndrome (ACS), stroke, infection, organ damage, and over a lifetime SCD can harm a persons spleen, brain, kidneys, eyes, and bones [2].

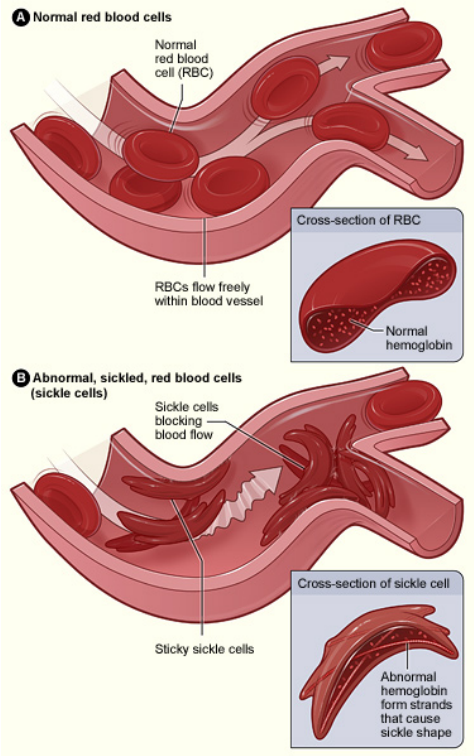


Figure 1.1: Sickled blood cells against normal blood cells [1]

While normal blood cells have a lifespan of 90 to 120 days, sickle cells can live only up to 10 to 20 days. Since the sickled hemoglobin cant change its shape and bursts eventually, people from SCD have a shorter life span and die early due to shortage of red blood cells [3].

SCD is inherited and not contagious and does not spread from person to person. SCD cannot be acquired over the lifetime if it is not identified at birth [4].

1.2 Statistics

SCD affects millions of people throughout the world, and it is believed that 80% of the SCD cases occur in Sub-Saharan African. It also occurs in people from Spanish speaking regions such as South America, the Caribbean, Central America, India, the Arabian Peninsula and in people of African origin living in other parts of the world [5].

It is believed that SCD affects 90,000 to 100,000 Americans approximately, mainly African Americans and it occurs in 1 out of every 500 African American origin people, and 1 out of every 36,000 Hispanic-American births [5].

1.3 Causes

SCD is inherited, and people with SCD have a problem with genes involved in the development of hemoglobin, a mutation that changes types of hemoglobin chains in red blood cells. It is inherited when a child receives sickle cell genes from both the parents [6].

Sickling of the red blood cells can be triggered by a number of conditions such as low levels of oxygen, acidity, dehydration of blood, stress, changes in temperature, or being at high altitude [7]. Certain organs having high metabolic rate, such as brain muscles, extract more oxygen leading to sickling of red blood cells [8]. Some organs are vulnerable to lower levels of oxygen and acidity, such as when the blood moves slowly through the liver, kidney or spleen. These conditions make the organ vulnerable to damage.

1.4 Signs and Symptoms

Some of the above mentioned causes results in sickling of the red blood cells and blocking the blood flow to the tissues as demonstrated in figure 1.2 [9], causing it to suffer from damage due to lack of oxygen. The damage causes disability in patients, leading to excruciating pain episodes of variable severity and frequency, depending on the degree of organ involvement.

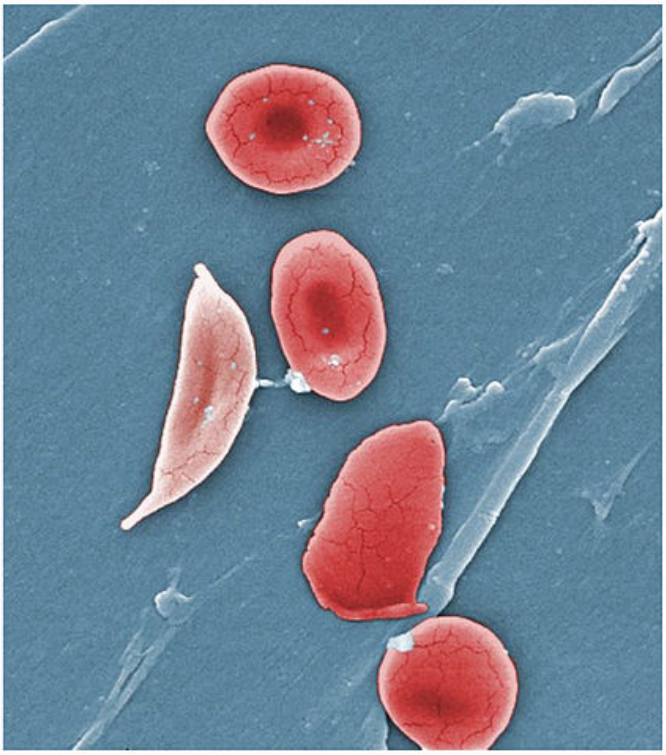


Figure 1.2: Blood clot due to sickling of red blood cells [9]

Major symptoms of sickle cell anemia include pain crisis, Acute Chest Syndrome, heart failure, bacterial infection, fatigue, Anemia, breathing problems, gallbladder disease, leg ulcers, headache, lung and heart injury, painful erections in men, eye damage, stroke etc. Individuals may feel pain in any part of the body, but often in the legs, arms, chest, belly, hand-foot syndrome, and lower back. They may also feel weakness, extreme tiredness, jaundice, and have trouble moving some parts of their body. Acute chest syndrome accounts for the largest number of hospitalizations and death [10]. SCD affects the blood, and its symptoms can occur anywhere. The signs and symptoms can vary from person to person and can change over time.

1.5 Diagnosis

SCD can be diagnosed with a simple blood test [11]. Most often it is detected at birth during the routine newborn screening tests, and it can also be diagnosed before birth.

1.6 Treatment

Changes in red blood cells can eventually lead to sickle cell crisis. Some of the medications taken during the sickle cell crisis include pain medications such as narcotics, IV fluids, blood transfusions for anemia, and curing infection.

People with SCD can often manage their pain at home by taking pain medicine and drinking plenty of liquids. However, if the pain is severe then hospitalization is a better option. There is no best treatment for SCD; treatment options vary from person to person depending on the symptoms [12].

Bone marrow or stem cell transplant is the only cure for SCD, however it carries the risk of death and other serious side effects, making it a less viable option as compared to just managing the pain symptoms of SCD patients [12].

1.7 Progress Notes for SCD

So far we had a brief discussion of SCD, its causes, diagnosis, treatment and prevention. Identification of SCD crisis before its occurrence is crucial as it can help the patient mitigate the painful episodes of the sickle cell crisis. Numerous studies are undergoing using different techniques to prevent the sickle cell disease crisis [13] or predict pain level. One of them uses vital signs and physical signs tracked by mobile technology to study and predict the pain levels [13].

The adoption of Electronic Health Record (EHR) systems has facilitated reuse of clin-

ical records for research, taking initiatives, and making informed decisions. Integration of clinical notes with electronic health records has led researchers to use this valuable textual information (such as medication, and procedures) to gain a better understanding of the problem. SCD symptoms vary from person to person, and progress notes from each patient can add medical and health status information specific to that patient to improve prediction of the sentiment and pain scores.

Nursing Notes are medical notes made by the nurse over the course of hospitalization of the patient. THEY record the assessments, current status, and care delivered to the patient. The documentation provides patient assessments, plan of care, and real time progress notes. Complete and accurate nursing notes are critical to make informed and good decisions for patient care.

Progress notes are essential to patient care, and document patients status. Progress notes represent a vast wealth of knowledge and insights. They are part of medical records and keep track of nursing assessments, care provided, patient condition, patient progress, current status, discharge summaries, and relevant information to deliver excellent care. They provide accurate condition while the patient is under the care of healthcare team [14].

Progress notes are crucial as they record the events during patients care, allowing clinicians to compare the current and past status, communicate findings, plans and opinions between clinicians and other members of the healthcare team, and make informed decisions. They are the repository of facts and clinical thinking, and are intended to communicate patients health condition and progress to those who access the health record.

During hospitalization clinical notes and vital signs are registered. When a SCD patient is hospitalized, nursing notes are documented to monitor the ongoing progress of the patient. They provide the patient's assessment of their condition; plan of care notes outlining the plan of care for the patient and goals and its outcomes; and progress notes tracking the patient's progress, medication, the level of pain on the pain scale, the patient's improvement, and child development activities.

Nursing notes are documented several times during hospitalization, and the date and the time are recorded during the patients assessment; hence, there are multiple notes maintained for each patient to track the progress over the course of hospitalization. However, there are pitfalls in some nursing notes as some of them are copied, duplicated, sometimes incomplete and non-informative, and dont provide meaningful information regarding the patient as observed from the data. Hence, reusing and revisiting non-informative progress notes for evaluation by healthcare professionals can be redundant and time consuming, which brings up the need to identify informative and non-informative notes.

Clinical notes, a storehouse of potential breakthroughs, have been used in a number of studies for analysis and prediction. For this study, progress notes from nursing notes are used.

In this thesis, we are trying to solve 4 different problems. These are identifying:

1. Patient informative notes
2. Pain informative notes.
3. Pain sentiment
4. Pain score

We explain this in detail below:

1. Informative notes with respect to a patient are notes emphasizing the patient's medical history, current state, progress, medication information and discharge information. Notes with the above information are useful for research and can lead to meaningful results, for instance identifying popular medications for a patient and effective treatment during analysis, and also save a tremendous amount of time by omitting the non-informative notes.
2. Informative notes with respect to pain are notes demonstrating information regarding a patients pain and outlines pain status when the notes were taken. Progress notes

have information regarding the patients family, progress, and playroom activities. Not all notes are informative in terms of pain as they might emphasize other aspects. SCD patients report pain level at different times during hospitalization, and these notes document the patients pain, hence it is useful to identify notes demonstrating pain versus notes emphasizing other aspects of hospitalization.

3. Pain Sentiment is broadly classified into 4 categories, namely increase, decrease, neutral and unsure. From the notes and vital signs, a pain score for each progress note is collected and classified into one of 4 classes based on the current and previous score; an increase indicates a increase in pain level, decrease indicates a decrease in pain, neutral is no change in pain and not-determined is unsure regarding pain.

4. Pain Score is classified into 3 classes namely moderate, severe, and not-determined. Pain scores are accrued from the vital signs and they are classified as moderate if it falls in the range (0-6), severe (7-10) and not-determined if there was no pain score in the notes.

Natural language processing, machine language techniques, and cTAKES are very popular with unstructured data, and for our study we are using these techniques for the prediction of patient informative, pain informative and pain sentiment and pain score.

This information is valuable as it saves a huge amount of time to track the patients progress, as there are millions of patients and tracking each progress notes is tedious and time consumptive.

1.8 Block Diagram

As discussed above, the four problems addressed in our study are identifying patient informative, pain informative, pain sentiment, and pain score from clinical notes. Figure A

demonstrates the hierarchy of the datasets and Figure B demonstrates the pipeline followed in building the models.

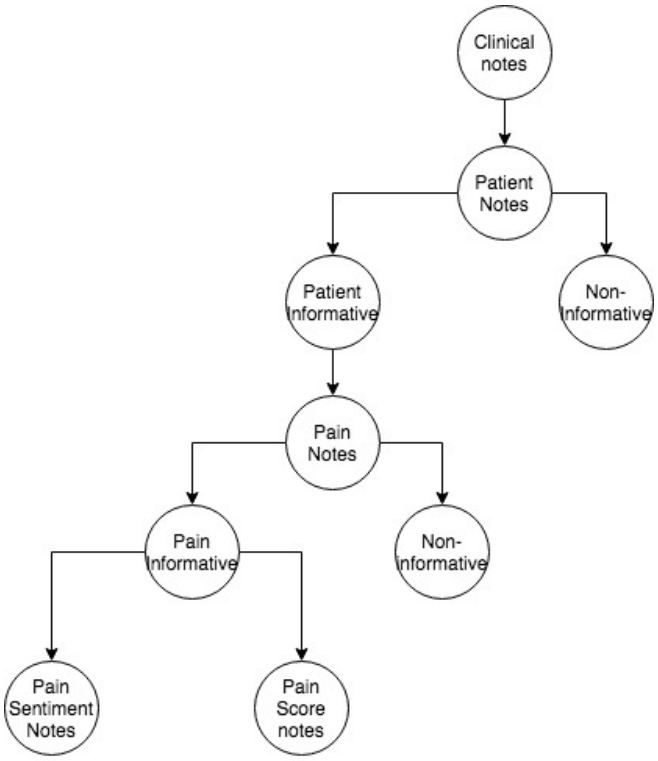


Figure 1.3: Block diagram A demonstrating hierarchy of datasets

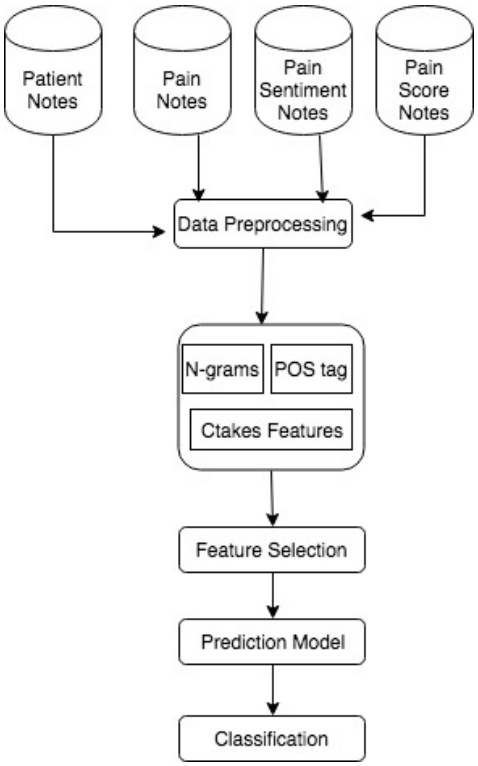


Figure 1.4: Block Diagram B demonstrating the workflow of the system.

In figure 1.3, the dataset follows a hierarchy of 5 levels. Clinical notes are on top of the tree, patient informative and patient non-informative are at level 1, pain notes are at level 2, pain informative and non-pain informative are at level 3, pain sentiment notes and pain score notes are at level 4. The levels demonstrate how each child node is derived from the parent note. Pain notes are derived from the patient informative notes alone, patient non-informative notes are discarded in the pain notes dataset, similarly sentiment and pain score notes are derived from pain informative notes, and non-pain informative notes are discarded for building the pain sentiment and pain score notes datasets.

Figure 1.4 demonstrates the pipeline for building the machine learning models. All the datasets in the study follow a similar pipeline. At stage 1, data is preprocessed; at stage 2 features are extracted, extracted features include n-grams, part of speech tags and cTAKES features; at stage 3 features are selected using a feature selection algorithm; at stage 4, cleaned and informative features are fed to machine learning models for prediction; stage 5

is classifying the data to various class labels. The steps outlined in the pipeline are further discussed in chapter 3 and 4.

1.9 Outline of Thesis

The rest of the thesis is organised as follows: chapter 2 outlines related work dealing with work on clinical notes, NLP, and SCD. In chapter 3, data collection, annotation, methods and various NLP and machine learning techniques used to compute the results and metrics to quantify our results are demonstrated. In chapter 4, results and discussion are outlined to help gain insights on SCD. In chapter 5, conclusions and future work are presented.

2 Related Work

This chapter discusses studies related to Natural Language Processing and text classification. Section 2.1 explores studies using technology with wearable sensors for healthcare, followed by section 2.2 discussing studies using clinical notes and Natural Language Processing for healthcare, and section 2.3 explores studies using technology and Sickle Cell Disease (SCD) specifically.

2.1 Technology and Wearable sensors

Data collected from wearable sensors are fed to machine learning models to make real time predictions such as fall risk and human activity recognition. Healthcare is adopting wearable devices and utilizing the real time information to build predictive models and make real time predictions as discussed in the section below.

Giansanti et al. [15] developed a neural network classifier to predict fall risk using mahalanobis distance and kinetic parameters assessed by wearable devices. Falls are risky and possibly life threatening. Their model uses a new neural network (NN) to predict falls. The NN uses kinetic parameters collected using accelerometers and rate gyroscopes during a posturography protocol. Kinetic parameters were assessed by standing sway test in three conditionS: eyes open (EO), eyes open on foam (EOF) (patients standing on foam instead of floor), and eyes closed on foam (ECF), and values were noted for ten 60s trials and two parameters $R_1 = \frac{mean(EOF)}{mean(EO)}$ and $R_2 = \frac{mean(ECF)}{mean(EO)}$ are calculated centroids from inputs and

fall risk predicted by statistical classifier based on Mahalanobis distance are fed to the NN. They trained the model on mahalanobis distance and on two groups of 30 elderly subjects with different Tinetti fall-risk scores (determines elder's risk of fall within the next year) and the model was validated on two groups of 100 individuals with different Tinetti fall-risk scores. The NN was the optimal model compared to Nave Bayes, multilayer perceptron, Support Vector Machine, Bayes Net, and statistical classifiers implemented for the study. The model had a 97% specificity [16] (measures proportion of negatives correctly identified) and 98% sensitivity [16] (measures the proportion of positives correctly identified) and 0.965 AUC [17] (Area under the curve of Receiving Operating Characteristics (ROC) is a measure to see if the model is useful, an AUC of 0.5 indicates it is no better than a random model).

Lester et al. [18] proposed a hybrid model using Generative and Discriminative models [19] (A generative model is based on joint probability, P(x,y), with x as inputs and y as output, and makes predictions based on Bayes theorem to calculate P(y—x) and selects the most likely y. A Discriminative model learns to map the inputs x directly to output y) approach to identify 10 different human activities. They recorded 12 hours of data consisting of different activities and used 80% of the data for training. Two approaches are considered for the study: for the first approach, features extracted from the data were inputted into a static classifier for prediction, features selection was performed using boosting and classified using ML techniques, A decision stump (a one level decision tree) came out to be the optimal model compared to Nave Bayes (explained in Chapter 3) with precision and recall (these metrics are explained in Chapter 3) of 98% and 84%. The second approach uses the posterior probability of the classifier as inputs to a Hidden Markov Model [20] (a model to recover the sequence of states to generate the emission from observed data) and the precision and recall (explained in Chapter 3) was 99% and 91%, with an overall accuracy of 95%. They found that the addition of the HMM to the model helped to smooth the classification errors caused by the first approach.

Dawadi et al. [21] presented a model to assess cognitive health and activity quality using smart home technologies. Study develops an automatic framework to extract features from smart home sensors data that reflect an individuals performance or ability to complete the activity. The features are fed to the machine learning model to predict cognitive health as cognitive healthy or dementia and predict the activity quality for complex set of smart home activities. To validate the model the complex activities of 263 participants are recorded, and these activities are performed in smart home testbed. Support Vector Machine (SVM are explained in Chapter 3) could predict cognitive health or dementia with 0.80 area under ROC, and significant correlation of 0.54 was obtained between observation scores and predicted activity quality.

Alshurafa et al. [22] developed a model using time-frequency decomposition in a wearable necklace using a Piezoelectric sensor to recognize Nutrition Intake. The sensor can detect skin motion and produce output voltages with varying frequencies over time. They developed an algorithm based on time-frequency distribution, and spectrogram analysis to distinguish between food types such as solid, liquid, hot, cold drink, hard and soft foods. The model works by detecting swallow movements and create spectrogram, a 3-dimensional energy plot, further statistical features are generated from the spectrogram and given to classifier for recognition. The model implemented 10 fold cross validation and was tested on data collected from 10 subjects. Random Forest (explained in Chapter 3) was the best performing model predicting liquid or solid class with an F1 score (explained in Chapter 3) of 91.2%, it was also the best performing model for the rest of the classes.

Suutala et al. [23] proposed a model to automatically classify human activity recognition including standing, sitting, running etc from on-body accelerometers. The dataset consisted of signals assessed by wearable device from 13 different subjects performing 17 daily activities and some of the general activities were combined to form 9 daily activities. Time domain features were assessed from the device and a sequential learning model was built using Support Vector Machine (explained in Chapter 3) that discriminatively learns

individual input output mappings and Hidden Markov Model [9], a generative approach to smooth temporal time dependent activity sequences. The accuracy for the model was 94% in identifying 17 daily activities and 96% in identifying 9 daily activities.

Morris et. al. [24] developed a system called RecoFit to automatically track repetitive exercises including weight training and calisthenics using wearable sensors. The model has three stages, segmentation to identify exercise periods from non-exercise periods, it is a binary state machine; followed by recognition of the exercise, the model was trained to recognize 26 exercises circuit labeled in the training data; followed by counting the number of repetitions. It is validated by collecting data from 114 participants over 146 sessions using cross validation. The results for the 4, 7, 13 exercises circuit has recognition of 99%, 98% and 96%, counting was 1 repetition 93% of the time.

Shoaib et al. [25] built a model to recognize complex human activities using smartphone and wrist-worn motion sensors and study the effect of window size on complex activities. For the study, data was recorded from three motion sensors namely accelerometer, gyroscope, and linear acceleration sensor and have studied 7 window sizes (2-30) on 13 different activities. Their study discusses that less repetitive activities including smoking, eating etc. are difficult to detect at smaller segmentation windows and increasing window sizes helps in identification of complex activities. So far we have discussed some of the papers using technology in healthcare sector, and these studies reflect the fact that different algorithms were the best performing models for different problems, and also demonstrates adoption of technology by healthcare to discover and validate new findings such as fall risk, predicting human activities. This proves using technology in healthcare can help in prediction of problems to take preventive measures.

2.2 Clinical notes and NLP

As part of new standards clinical notes are added to the EHR data, making it available to people and open for research. It is a repository of potential information, and understand the problem domain better by analysing the data. Myriad number of studies and researchers are using clinical notes and NLP in predictive modeling. The rest of this section analyses some of the past works using NLP and Clinical Notes.

A study by Joshi et al. [26] focused on prediction of acronym expansion in clinical notes. Acronym disambiguation is a well-known concern when uncovering information from EHR data. For the study they have created 7,738 disambiguated instances of 16 ambiguous acronyms. They used different features including unigram, bigrams, POS tags (explained in Chapter 3) and their combination, and implemented Nave Bayes, SVM and decision tree as the machine learning algorithms for prediction. Their baseline model reported accuracy below 50% but with unigrams as features, SVM outperformed others obtaining an accuracy of 90% when the disambiguation acronyms in the dataset were less than 50%, which is very difficult as the majority sense is very low. As the distribution of the disambiguation acronym increased from 50- 80%, all the features were performing well, with an accuracy of above 80%; as the distribution of acronym was greater than 80%, the differences in the accuracy of the features decreased to further extend and the accuracy for all the features was above 90%; and in all the cases SVM performed the best among all the machine learning algorithms. Their study also indicate that unigrams are better features compared to other features used in the study.

A study by McCart et al. [27] used clinical notes to predict falls from ambulatory care clinical documents, and statistical text modeling to serve their needs. Their aim was to predict the categories fall and not fall. They had 4 datasets from different sites, site A, B, C, D, and took 70% of the data from site A to train the model and use the rest and other sites to test the data. They employed feature selection techniques such as, gain ratio (ratio of information gained from information) [28], χ^2 (in Chapter 3 under section 3.5) and log OR,

and used Logistic Regression, SVM and SVM-cost models to build the classifier. Different feature selection techniques were used on different models, for instance, χ^2 was used for both SVM and SVM-cost, and gain ratio was used for logistic regression. SVM-cost model obtained the highest AUC scores, ranging from 0.953 to 0.978, and F1 scores ranged from 0.745 0.853, however the AUC and F1 scores for all the models were relatively the same with a difference in AUC being 0.02 and a difference in F1 score ranging from 0.02-0.05, on any test sets. The study demonstrated that statistical text mining could reliably identify falls in clinical documents.

A study by McCormick et al. [29] used NLP on clinical data to classify patient smoking status. It is a multiclass classification, classifying each patient into 3 categories namely smoker, non-smoker or unknown. For the study they have generated two feature sets using two different techniques, the first approach uses lexical features captured from the data i.e. bag of words, and other was semantic features generated by MedLee, a clinical NLP engine. Boostexter and rule based algorithms Xquery classifier were used for the study with the above-mentioned feature sets. The best performing model used semantic features and BoosTexter algorithm with a macro average F1 score of 0.75. The study demonstrated that semantic features are helpful in identifying smoking status, and therefore we tried using semantic features in our study using CTakes, a NLP system to extract information from clinical notes.

A study by Meystre et al. [30] used NLP to extract medical problems from electronic clinical notes, their aim was to provide help in maintaining the problem list updated, accurate and complete. For the study, they tried to identify 80 medical problems from clinical notes. UMLS MetaMAo Transfer (MMTx) application was used to detect the problem from notes and a negation detection algorithm to identify negation in the sentence and remove the negated medical problem from the notes, and generate a list of medical problems from clinical notes. By creating a custom subset of MMTx, containing the 80 concepts and all related sub concepts they have improved the performance of the system and achieved

a recall of 0.896 and precision of 0.691 and F1 score of 0.780. Their study compared the problem detection for 20, 40 and 80 problems and reveal that as the number of problems to detect increased recall decreased.

Khalifa et. al [31] studied on adapting existing NLP resources for cardiovascular risk factor identification in clinical notes. Eight categories of information are associated with risk of heart disease including smoking status, diabetes, hypertension etc. The model used Ctakes (discussed in Chapter 3) to identify the smoking status and rule based and pattern matching modules to detect medication and laboratory results and Textractor to detect disease and risk factor terms. The model was trained on 790 documents and tested on 514 clinical notes. The overall micro-average f1 score of the system is 87.47%. F1-score (discussed in Chapter 3) for identification of smoking status using Ctakes was 87%, demonstrating the usefulness of Ctakes in healthcare studies.

A study by Melton et al. [32], an automated detection of adverse events used NLP on discharge summaries to detect 40 adverse events. They used MedLEE, natural language processor to convert discharge summaries to coded form (entity identification of phrases to problem, certainity and status. For instance the sentence the patient may have a history of mi is coded as problem: myocardial infarction, certainty: moderate, status: past history), and input the code to a program, mapping it to a list of events that have appeared to occur during admission. Of the 57,452 discharge summaries, the system identified 1590 adverse events in 1461 cases with a sensitivity [5] (measures the proportion of positives correctly identified) of 0.28, and specificity [5] (measures proportion of negatives correctly identified) of 0.985.

NLP is applied in numerous studies using text data, and the most common machine learning algorithms associated with the studies were Logistic Regression, SVM and decision trees. Unigrams was identified as the best performing features among POS tags, bigrams and their combination, however there are few studies trying to incorporate features from the NLP system such as Ctakes for prediction. In our study we are using techniques

to analyze different features from NLP system and their combination in building predictive models.

2.3 Machine learning and SCD

With the adoption of Machine learning and Natural Language Processing in healthcare sector, multitude of papers are published each day with new discoveries and findings. As discussed in previous section many healthcare areas including SCD in addition to cancer, Down's Syndrome, etc. has adopted machine learning. Predictive analysis is used in SCD to uncover new ways to decrease the mortality rate, predict and alleviate pain crisis before its occurrence.

A study by Desai et al. [33] on discovering a novel gene signature for elevated Tricuspid Regurgitation Velocity (TRV) in SCD may help in decreasing mortality rate in SCD. An elevated TRV on transthoracic echocardiography and right heart catheterization (RHC)-defined pulmonary hypertension are independently responsible for increasing mortality in SCD. The study implemented SVM to identify a 10-gene signature to discriminate patients with and without increased TRV and validated it against the validation cohort of patients with RHC-defined PH using SVM, and it was able to identify PH in SCD with an overall accuracy of 90%. The study validated genomic signature as a potential biomarker in SCD-associated elevated TRV.

Xu et al. [34] developed a deep convolutional neural network for classification of red blood cells in SCD. The study can help in potentially assessing the severity of SCD based on the shape of the RBC. The red blood cells in SCD are found in multiple shapes including the sickle shape, and they developed a classifier to predict the heterogeneous shapes in SCD. They build their model employing a hierarchical RBC patch extraction method followed by a shape-invariant RBC patch normalization technique (to exclude background patches and save training time) and fed it to neural nets. This study used 434 raw mi-

croscopy images of 8 different SCD patients and implemented 5 fold cross validation for the results. The model performed well in identification of the 5 classes of red blood cells namely Granular and Echinocytes, Discocytes + Oval and Elongated + Sickle and Reticulocytes, and the mean accuracy was 89.28%. The study also evaluated the dataset for 8 different red blood cells resulting in the mean accuracy of 87.5%. They have also experiment on classification of deoxygenated RBCs and the model obtained a high recall of 93.8% and a low precision of 60%.

Allayous et al. [35] used machine-learning algorithms for prediction of Acute Splenic Sequestration Crisis (ASSC), a serious symptom of SCD. For the study they have used data from 42 children described by 15 characteristics, and employed feature selection to select 4 attributes namely HT (basic ht), NEU (neutrophyle number), DELTAPLT (variation of plates of blood), and HYGRO (hygrometry). They observed that ASSC is characterized by decrease in hemoglobin level, therefore excluded HB (rate of basic hemoglobin) from the feature set as they defined ASSC with HG (hemoglobin level). Their aim was to predict the severity of SCD as mild or severe, and employed four different machine learning algorithms namely Adaboost, Bagging, Boosting tree ranking, and Logistic Regression. All the models performed well with an overall AUC curve of above 0.80%, and Adaboost performed well with AUC of 0.92%.

Yang et al. [36] employed machine learning techniques to predict pain scores using physiological measures in an effort to reduce rehospitalization and predicting pain to take appropriate measures and prevent pain crisis. The dataset contains data from 40 patients with 6 vital signs as features, and predicted pain on 11, 6, 4 and 2-point pain scale. Multiple imputation by Fully Conditional Specification (FCS) was employed to solve the problem of missing values. They made analysis on intra and inter-individual level; the intra-individual level considers the samples from individual patients and trained multiple models for each patient, where as for the inter-individual model the differences between individual patients are omitted and therefore patient label was considered as one of the features along with

6 vital sign features for this analysis. They have employed different machine learning models including Multinomial Logistic Regression (MLR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF). Feature selection using χ^2 was employed and they found that all the 6 vital signs are significant and none of them are correlated. The results demonstrate that MLR performed well from the rest for intra-level for a 11 point scale with an average accuracy of 0.58, and F1 score of 0.50, and analyzed that due to class imbalance the weighted F1 was lower than accuracy. For the Inter-individual level, the best performing algorithm for 4-point scale was MLR with an accuracy of 0.681 and weighted F1 of 0.673 for dataset with patient labels, explaining the fact that accuracy increases with reduction of classes.

Milton et al. [37] presented work on prediction of Fetal hemoglobin (hbF) in Sickle Cell Anemia using an ensemble of 14 models. HbF levels are heritable and interpersonal variability is regulated in part by three quantitative trait loci (QTL). Studies identified that single nucleotide polymorphisms (SNPs) are found in QTLs and explain only 10-20% HbF variability. Combining Genetic Risk Score (GRS) and SNP can explain large amount of variance in HbF, however it is challenging to select optimal number of SNPs to combine with GRS. The dataset for the study consists of 841 sickle cell anemia patients and the results were tested in three independent cohorts. For the study they developed an ensemble of 14 models using GRS composed of variable number of SNPs, and the ensemble model explained 23.4% of variability in HbF. The ensemble model can predict HbF Levels with an accuracy ranging between 0.28 - 0.44.

The above papers emphasize on the importance of associating SCD with Machine learning. It helped in analysing SCD and identify ways to better understand the domain, which wouldn't have been possible without technology. Different areas of research are studied in SCD, however there are no studies on using text data in analysing pain in SCD patients. Progress notes are valuable source of information they provide resourceful data in text format such as pain status, medications taken, improvement in pain level, etc. which

help in building robust models, whereas other data forms such as vital signs lack more detailed patient information. Specifically, in the context of SCD, medications, procedures undertaken in the hospital play an important role to reduce pain level in patients, and text data can capture these minute details. For our study, we are using NLP techniques to analyze clinical notes to better understand pain in SCD patients.

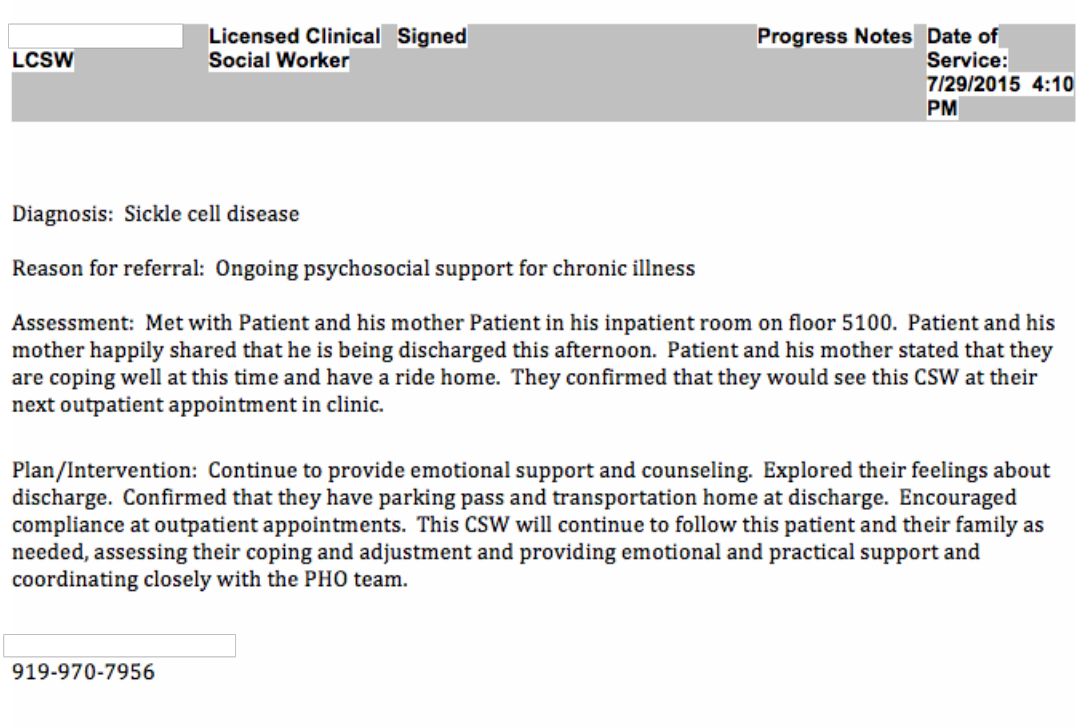
3 Methods

In this section, methods used for analyzing progress notes of SCD patients are detailed. To deal with textual data a framework is followed, it is more a linear and iterative process. The framework is a pipeline, and it is outlined as follows: (i) data collection and description, (ii) data annotation, data preprocessing, (iii) feature extraction, (iv) feature selection, (v) data modeling, and (vi) data evaluation. All the steps in the framework are discussed in detail in this section.

3.1 Data description

Our study used data collected from 112 in-patient participants with clinical notes recorded during their hospitalization by Duke University Hospital from June 2015 17. As discussed earlier clinical notes have many subsets including nursing notes, pan of care notes and progress notes; and for our study progress notes are utilized. Dataset was anonymized by replacing names of each patient with unique labels. Patients self-reported pain score was recorded along with the vitals signs in Electronic Health Records (EHR). Pain score is the pain experienced by the patient when vitals were recorded, and it ranges from 0 10, where 0 indicates no pain and 10 indicates severe pain. Data in progress notes is unstructured text data in Word format. Multiple progress notes are noted for a patient during the course of hospitalization, and by considering each progress notes for a patient as a single data point, a total of 504 data points are accrued from a dataset of 50 patients. Figure 3.1 shows a

sample of progress notes from the dataset.



	Licensed Clinical Social Worker	Signed	Progress Notes	Date of Service: 7/29/2015 4:10 PM
LCSW				

Diagnosis: Sickle cell disease

Reason for referral: Ongoing psychosocial support for chronic illness

Assessment: Met with Patient and his mother Patient in his inpatient room on floor 5100. Patient and his mother happily shared that he is being discharged this afternoon. Patient and his mother stated that they are coping well at this time and have a ride home. They confirmed that they would see this CSW at their next outpatient appointment in clinic.

Plan/Intervention: Continue to provide emotional support and counseling. Explored their feelings about discharge. Confirmed that they have parking pass and transportation home at discharge. Encouraged compliance at outpatient appointments. This CSW will continue to follow this patient and their family as needed, assessing their coping and adjustment and providing emotional and practical support and coordinating closely with the PHO team.

919-970-7956

Figure 3.1: Progress notes of a patient describing the assessment, plan and intervention during the course of hospitalization

3.2 Annotation

For our study, five scenarios namely patient informative, pain informative, pain sentiment and pain score prediction are evaluated. For each scenario different rules are followed to label the dataset. In this section, annotation process and rules followed to label the dataset in the five scenarios are addressed. The dataset is self annotated, after taking inputs from clinical collaborator from Duke University.

3.2.1 Patient informative

Nurses record clinical notes during course of hospitalization for a patient, and it contains information regarding patients health, family and other activities. Notes written are a com-

bination of informative and non-informative regarding patient, therefore identifying the informative notes is the goal in this scenario. A note is considered as informative if it satisfies any of the criterions below, and the criterions are decided based on the inputs from our clinical collaborator. Informative regarding Patient

1. Current health status
2. Current pain status
3. Health concerns
4. Medical background
5. Activities during hospitalization
6. Current therapy
7. Medical advice
8. Patients background
9. Admission information
10. Discharge information

Figure 3.2 and 3.3 below demonstrates example of informative and non-informative notes based on the criteria discussed above.

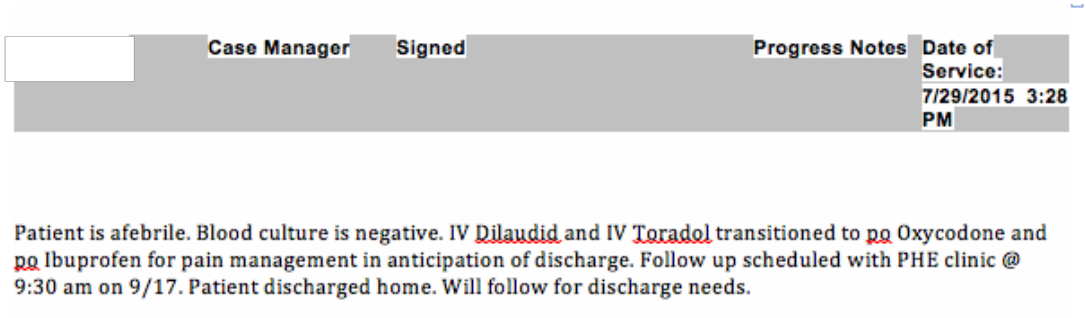
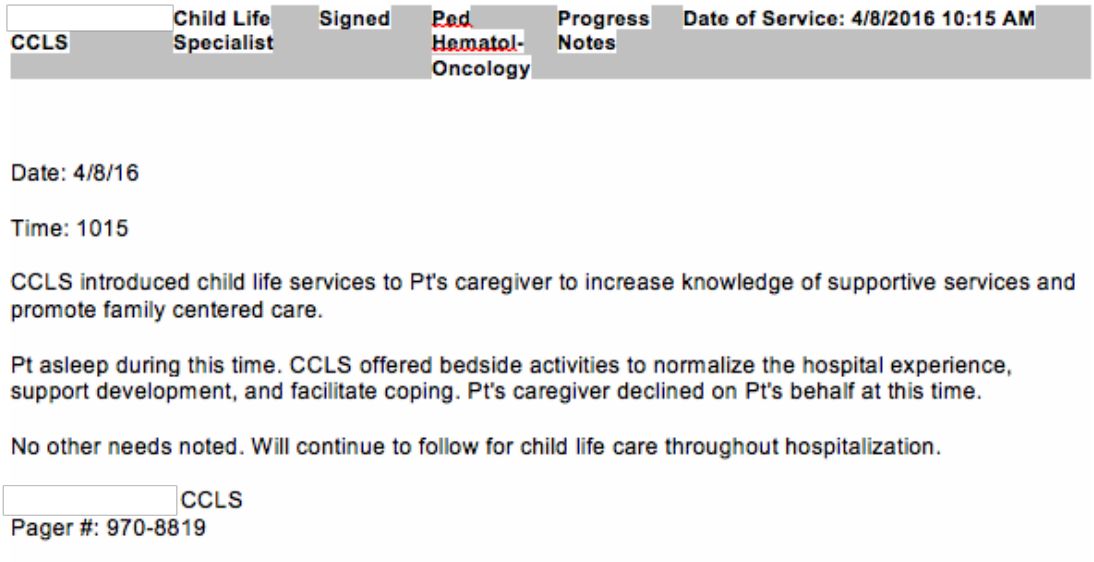


Figure 3.2: Notes taken at the time of discharge is an example of patient informative notes

Notes in figure 3.2 is considered as informative as it shares information regarding the patients discharge. Progress notes sharing information regarding other details or irrelevant information regarding patient are considered as non-informative. Figure 3.3 demonstrates an example of non-informative notes. Notes in the figure states patient is asleep and shares irrelevant information hence annotated as non-informative.


CCLS Child Life Specialist Signed Ped Hematol-Oncology Progress Notes Date of Service: 4/8/2016 10:15 AM

Date: 4/8/16

Time: 1015

CCLS introduced child life services to Pt's caregiver to increase knowledge of supportive services and promote family centered care.

Pt asleep during this time. CCLS offered bedside activities to normalize the hospital experience, support development, and facilitate coping. Pt's caregiver declined on Pt's behalf at this time.

No other needs noted. Will continue to follow for child life care throughout hospitalization.

CCLS
Pager #: 970-8819

Figure 3.3: Notes providing information of child life services and no specific information regarding patient hence non-informative

3.2.2 Pain Informative

A patient hospitalized for Sickle Cell crisis suffers excruciating pain and health professionals try to alleviate pain during the course of hospitalization. Clinical notes record details regarding patients current status, reaction to medications, recovery, and pain status. Numerous details regarding patient is recorded in clinical notes, and hence some notes provide information regarding other aspects of hospitalization. Some criterions are used to label the data point as informative or non-informative. A note is pain informative if it satisfies any of the criterion below: Informative regarding patients

1. Current pain status

2. Pain medications

3. Current health status

For this case, it is intuitive to consider only informative notes with respect to patient, because non-informative note with respect to patient indicates its non-informative with respect to pain. Therefore the dataset in this case is a subset of patient dataset.

Figure 3.4 and 3.5 below demonstrate examples of informative and non-informative notes. The notes is considered informative because it has pain mention, and it explains patients pain is 2/10 and following criteria 1. Whereas the other notes is non-informative because it shares no information regarding pain in patient, its a discharge note with no pain specific information.

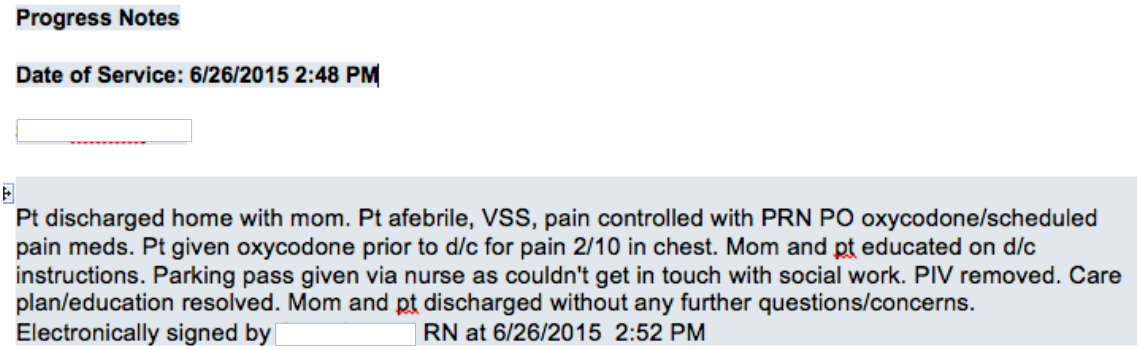


Figure 3.4: Notes describing about the pain of the patient is considered as pain informative notes

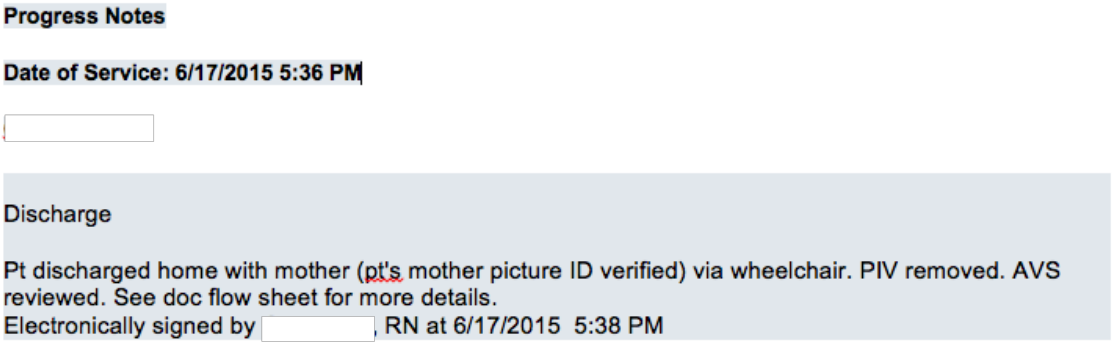


Figure 3.5: Discharge notes with no mention of pain, hence considered as pain non-informative

3.2.3 Pain Sentiment

Pain sentiment is predicting the sentiment of patients pain at a given point of time. It can be increasing, decreasing, neutral or not-determined. Pain sentiment is labeled as increase, neutral or decrease by the increase, decrease or no change in pain scores in two consecutive notes for a patient. Pain scores are recorded along with vital signs. Progress notes and vital signs are time stamped when they are recorded, however the notes and vital signs were never noted together as they were no identical timestamps identified in the two datasets (notes and vital signs), hence it was difficult to record pain score for progress notes. To address this issue, a two-hour window for the date and time in the progress notes and vital signs is considered to identify pain score for a progress note. Some of the pain scores were missing in vitals; therefore these missing scores were labeled as not-determined. It is difficult to identify if the pain is increasing or decreasing for the first progress notes recorded for individual patient, and therefore those progress notes were also labeled as not-determined. The labelling was done in collaboration with a clinician.

For this case, pain informative data points are considered for labeling of sentiment. It is intuitive to consider only Positive class because non-pain informative data points will not provide information regarding sentiment, as they were unable to identify pain.

3.2.4 Pain score

Pain scores are accrued from the vital signs (self-reported pain scores) and identified for each progress notes, they are categorized on a 0 10 pain scale, where 0 indicate no pain, and pain level increases with the increase in pain scores and 10 indicate severe pain. Pain score is classified into three categories namely:

1. Moderate (0-6 on a pain scale)

2. Severe (7-10), and

3. Not-determined

3.3 Preprocessing

Data collected is often inconsistent, incomplete and inaccurate. Preprocessing is essential to eliminate noise, extract meaningful information, knowledge from raw data and prepare for further processing. Data preprocessing techniques utilized in this study are stemming, stop word removal, removing numbers, lower casing etc.

3.3.1 Stemming

Stemming [38] is a process of reducing inflected words to its base or root form. Its a process that maps related words to the same stem. For instance, the words likes, likely, liked and liking from the dataset are mapped to the root word like. Stemming addresses the issue of adding new terms to the word lists when two words share same information.

3.3.2 Stop word and number removal

Stop words [39] are irrelevant words to the problem being solved, for a given purpose any set of words can be considered as stop words. For our study filler words in the sentences such as a, an, the, which, on, etc. are considered as stop words and eliminated from the text, as they dont provide meaningful information. Numbers are further removed.

An example notes, following the stages of preprocessing is explained as follows. Below is the sample notes.

Pt discharged home with mom. Pt afebrile, VSS, pain controlled with PRN PO oxycodone/scheduled pain meds. Pt given oxycodone prior to d/c for pain 2/10 in chest. Mom and pt educated on d/c instructions. Parking pass given via nurse as couldn't get in touch with social work. PIV removed. Care plan/education resolved. Mom and pt discharged

without any further questions/concerns.

Text follows preprocessing steps such as stemming, stop word and number removal, the final text after preprocessing is below. *Patient discharged home mom patient afebrile vss pain controlled prn po oxycodone scheduled pain meds patient given oxycodone prior discharge pain chest mom patient educated discharge instructions parking pass given via nurse couldnt get touch social work piv removed care plan education resolved mom patient discharged without questions concerns*

3.4 Feature Extraction

For our study we are using text data derived from progress notes and our goal is to build a predictive model to make predictions using four different models. For a model to perform well it must be fed with resourceful and meaningful information extracted from data. Feature extraction or feature engineering is a step towards building an accurate predictive model by extracting knowledge from raw data. In feature extraction, features are extracted from text data. There are various ways to extract features, however a simple approach is to use bag of words. Bag of words technique represents each data point as a vector of all the words in the dataset. Furthermore, for our study we have extracted features using Apache cTAKES [40], a natural language processing system to extract knowledge from clinical notes. Following are the features utilized for our study:

3.4.1 N-grams

N-grams [41] are the continuous sequence of n-words in a sentence, considering it as a single unit. An n-gram of size 1 is referred to as unigram, size 2 as bigram, size 3 as trigrams. The unigram and bigram generation for the sentence *patient is feeling well today, patient is being discharged today* are demonstrated in table 3.1. Vectors of n-grams are constructed

from bag of words and given to predictive models. Bag of words for the example after preprocessing is patient, feeling, well, today, patient, discharged, today Different weighing factors can be used to convert text to feature vectors: Following are the two weighting factors adopted in out study:

Table 3.1: Unigrams and bigrams generated from text

Features	Unigrams
Unigrams	patient, feeling, well, today, patient, discharged, today
Bigrams	Patient feeling, feeling well, well today, today patient, patient discharged, discharged today

3.4.2 Count

In this technique, the frequency of the occurrence of the n-grams for each data point is calculated and assigned to the extracted n-gram in the feature vector. For instance, a unigram feature vector for the sentence *patient is feeling well today, patient is being discharged today* is calculated by first converting the sentence into bag of words and then counting the number of times each word occurs in the sentence, in the example sentence the term patient occurs twice in the sentence hence the count for the term patient is 2. The bag of words constructed for the sentence above looks like: Patient, feeling, today, discharged

Table 3.2: Count of unigram features generated from text

Features	Count
Patient	2
Feeling	1
Today	2
Discharged	1
Today	2

3.4.3 TF-IDF

In this technique each extracted n-gram is assigned a term frequency/ inverse document frequency. TF-IDF is a statistic to evaluate how important a term is to a document in the entire corpus. The importance of the term increases proportionally with the occurrence of term in the document but is offset by the number of documents in the corpus that contains the word.

3.4.3.1 Term frequency

TF [42] is the occurrence frequency of each n-gram in a given document. Since documents vary in length, it is possible to have a term occurring many more times in longer documents than in shorter ones. Therefore term frequency is normalized by the length of the document.

$$TF = \frac{\textit{Number of times term appears in a document}}{\textit{total number of terms in the document}}$$

3.4.3.2 Inverse Document Frequency

Inverse document frequency [42] measures how rare a term is across documents in the dataset. It is a measure to scale up the rare ones and weigh down the common occurring words. The rarer the term is across the documents, more is the IDF score.

$$IDF = \frac{\textit{log of corpus size}}{\textit{occurrence frequency of the n-gram in the whole corpus}}$$

$$TFIDF = TF * IDF$$

For our study we are limiting the n-gram size to 2, as the size of the n-grams increases the frequency of the terms in the text decreases and data preprocessing steps and elimination of words makes it less viable to extract meaningful information. According to a study

using n-grams on text classification, using longer sequences (more than 2 and 3) reduces classifiers performance [43].

3.4.4 Part of Speech (POS) tag features

Part of Speech tagging (POS tag) is often called by many names including grammatical tagging and word category disambiguation. It is a process of marking the terms in the sentence to a particular part of speech. There are eight main parts of speech [44] in English namely nouns, verbs, adjectives, pronouns, interjection, preposition, adverb and conjunction. However most POS are divided into subclasses and for our study we are using 33 POS tags. For instance the POS tag [44] for the sentence *patient began complaining of worsening sharp lower chest pain.* is detailed in the table 3.3.

Table 3.3: POS tags generated for each term in the text

Term	POS tag
Patient	NN - Noun
Began	VBD Verb past tense
Complaining	VBG verb present participle
Of	IN preposition
Worsening	VBG Verb present participle
Sharp	JJ - Adjective
Lower	JJR Adjective, Comparative
Chest	NN - Noun
Pain	NN - Noun

3.4.5 cTAKES Features

cTAKES (Clinical Text Analysis and Knowledge Extraction System) [40] is a Natural Language Processing tool to extract information from EHR text notes. UIMA (Unstructured Information Management Architecture) framework [45] and OpenNLP [46] toolkit was used to built cTAKES. It processes clinical notes to identify named entities including dis-

eases and disorder, signs and symptoms, anatomical sites, procedures, and medications. For our study we are using all the five named entities as features to build predictive models. All five attributes are self-explanatory, procedure represents procedures taken, signs/symptoms are signs and symptoms identified for the patient during routine checkup, disease/disorder is the disease patient is suffering, and medication extracts information regarding the drugs prescribed for the patient. Clinical notes is the input to cTAKES system and the outputs are the five attributes. For a given clinical note, cTAKES tags each term in the notes with a named entity mentioned above, and outputs an XML file with all the terms in the clinical notes tagged to one of the five entities discussed. The final dataset is obtained by passing these XML files to a script to count the number of occurrences of each entity and create a vector of counts for each clinical notes. Table 3.4 demonstrates tagging of named entities by cTAKES for a sample text.

Patient is a 15 year old with Asthma and Sickle Cell Disease type SC who presented to Duke Children's Hospital with pain admitted with pain crisis. Hgb is 10.6 g/dL. C/O pain in right lower back and right thigh. Patient is afebrile. Blood culture is pending. Patient is receiving IV Dilaudid PCA/continuous and scheduled IV Toradol for pain management. Heat packs and Lidoderm patch to right thigh. Anticipate discharge home when afebrile, cultures negative at 48 hours and pain managed with po pain medications with MS Contin and Oxycodone.

Table 3.4: Named entities generated from cTAKES are tagged to 5 categories

cTAKES Named Entities	Terms
Procedure	Blood culture, PCA, pain management
Anatomical site	Lower back, right thigh, thigh, Sickle cell, cell
Sign / Symptoms	pain in right lower back, afebrile, po, patch, management
Disease / Disorders	Asthma, Sickle Cell Disease
Medication	Packs, Oxycodone

The feature vector is generated by counting the number of terms in each feature, ta-

ble 3.5 below is the feature vector for progress notes for a patient.

Table 3.5: Feature vector representation generated from the text tagged by cTAKES

cTAKES Named Entities	Terms
Procedure	3
Anatomical site	5
Sign / Symptoms	6
Disease / Disorders	2
Medication	2

3.5 Feature Selection

After extracting features from the data, the next step is feature selection. Feature selection is the process of selecting subset of features from the vast set of feature set for model construction. Feature selection is an essential step towards building a model for several reasons firstly, features can be redundant and secondly, some features are irrelevant to the problem being solved, therefore discarding these features would reduce complexity without loss of information. Selecting informative features will increase the performance of the model, as irrelevant features may affect the working of the model and decrease the overall accuracy. Therefore, feature selection is crucial as it avoids the curse of dimensionality problem, simplifies the classification model by eliminating redundant and irrelevant features, avoids overfitting, helps in generalization, shortens training time as less number of features would require less computation. The goal of feature selection is to identify minimum number of features to build a high performing model without much loss of information. In our study, we are using the following filter methods for feature selection.

3.5.1 Filter methods

Filter methods are feature selection techniques independent of the machine learning algorithms utilized during the clinical decision-making. The method chooses subset of features on the basics of scores in statistical measures between the independent variable and target variable, and selects the ones with highest scores. The computational time of these methods is faster than other feature selection techniques, and can still capture the meaningful information from the feature set [47]. It selects the relevant features using a univariate statistic. This measure is helpful in exploring the relationship between the features and doesnt include the assumption of the classifiers. A filter method using variance as threshold, works by computing variances of all the features and selecting the feature set based on a user defined threshold, or select the top k features with largest variance from the ranked features. A study by Yang et. al. [48] assessed various feature selection techniques such as document frequency, information gain [49] , Chi-square [50], mutual information [51] for text classification and found that Chi-square and information gain is the most effective feature selection technique reducing 98% of unique terms from the dictionary, with loss in performance of the classification using KNN (K-nearest neighbours) and LLSF (linear least square fit mapping) and therefore for our study the two feature selection methods employed are Chi-squared and mutual information.

3.5.1.1 Chi-square

Chi-square test [50] is a statistical test to determine the dependency between the feature variable and the target variable. It is a test of independence, and is calculated between each of the feature variables in the dataset and the target variable and observes the existence of a relationship between them. If the feature variable is independent of the target variable, it indicates that the feature is irrelevant to the problem being solved and hence can be discarded from the subset. If they are dependent, the feature variable is important in building an accurate predictive model. Chi-square test is statistical hypothesis test with a null hy-

pothesis that the observed frequency of the variable is equal to the expected frequency of the target variable. It assumes that variables are random and drawn from a sample of independent variables. The test result is a test statistic having a chi-square distribution and can be interpreted to either reject or accept the null hypothesis (expected and observed frequencies are same). If the observed frequency of a feature and the expected frequency of the target variable are same indicate that the two variables are independent of each other.

$$\chi^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k}$$

$$O_i = \textit{Number of observations in Class } i$$

$$E_i = \textit{Number of Expected observations in Class } i$$

3.5.1.2 Mutual Information

Mutual information [51] between two variables is a measure of dependence between the feature variable and the target variable. It determines the amount of information obtained from the feature variable about the target variable. Mutual information is a non-negative value. If the mutual information between the two random variables is zero, indicates they are independent of each other. The higher the mutual information, the higher is the dependence between the two random variables and large reduction of uncertainty. The lower is the mutual information then lower is the dependence and small reduction of uncertainty. Entropy is closely related to mutual information, and quantifies the amount of information in the random variable. Mutual Information is a useful measure because it maximizes the information between the target variables and the joint distribution in a dataset with many features. It is symmetrical and averaged. Mutual Information quantifies how much knowing of one random variable reduces the uncertainty about the other. For instance, if X and Y

random variables are independent then knowing X does not provide any information about Y, therefore the mutual information between them is 0. If X is a deterministic function of Y and vice versa then the information in X is also in Y, and knowing X will determine Y. Therefore the mutual information is the information stored in X, which is entropy of X or Y.

Mutual Information for two discrete random variables X and Y is defined as:

$$I(x,y) = \sum_{x,y} P(x,y) \ln \frac{P(x,y)}{P(x)P(y)}$$

P(x,y) is the joint probability function of X and Y

P(x) and P(y) are the marginal probabilty functions of X and Y

Features having the highest mutual information are selected for building the prediction model.

3.6 Prediction Models

For our study we implement machine-learning models to predict our outcomes. Machine learning uses statistical techniques to allow a system to learn from the data without being explicitly programmed. Supervised machine learning, a branch of machine learning is employed in our study. Machine learning is about predicting the outcome based on the inputs, supervised machine learning is a process of learning when the data provided for learning consists of input and output pairs, it is supervised because the target or response variable is given to the machine during the learning phase. In supervised machine learning a labeled training dataset is fed to the machine-learning model for training. A machine-learning model analyses the training dataset and produces a function, which can be used to

map new data. In our scenario we our using supervised machine learning models to predict for 4 scenarios including patient informative, pain informative, pain sentiment and pain score.

Our problem is a classification problem, therefore we adopted four widely used classification models in text classification namely Logistic Regression [52], [53], Multinomial Naive Bayes [54], Support Vector machine [55], and Random Forest [56], [57] to predict out outcomes. The former two are simple algorithms while the later two are complex and Support Vector Machine is known to perform well in text classification. Below is the brief discussion about the models used in the study.

3.6.1 Logistic Regression

Logistic Regression [52], [53] is a simple machine learning model implemented for binary classification. It predicts the probability of an outcome. Logistic function is also called the sigmoid function, it is an S-shaped curve that takes a real value number as input and maps it into a value falling in the range 0 to 1, but never at 0 or 1. Binary logistic regression predicts the odds of being a case for an input based on the independent variables. The odds is defined as the probability of an outcome to be a case A divided by the probability that it not a Case A.

$$p = \frac{1}{1 + exp(-(a + bx))}$$

p is the probability

a and b are the parameters of the model

Logistic regression is a linear model, but the outputs are transformed using the logistic function. The function estimates the probabilities between 0 and 1 based on one or more features in the data, and uses a threshold to transform the output to either 1 or 0. The model does not perform classification like other models, it estimated the probability of the

output for given inputs and uses a threshold; if the probability of an input is greater than the threshold it classifies it to one class, and if it is below the threshold, it classifies it to the other class. The main advantages of using Logistic Regression are its computational speed; provide insights into data as the output of logistic regression is more informative compared to other classifiers; easy to interpret the model and understand significance of features in predicting the outcome.

3.6.2 Multinomial Naive Bayes

Naive Bayes [54] is a simple machine learning algorithm commonly used in text classification especially with Natural Language Processing, it is often used as a baseline model. Naive Bayes classifier is a simple probabilistic classifier based on Bayes theorem, and assumes independence, rather than a particular distribution between the input features. It works by computing probability for the inputs and predicts probability of the features belonging to a particular class, and the class having the highest probability is selected as output. It assumes each feature from the input set contribute independently to the probability of that feature belonging to a class, regardless of any correlation between the input features.

The probability that the given set of features $(x_1, x_2, x_3.x_n)$ *belongs to class* C_k *is:*

$$P(C_k|x_1, ..., x_n) = P(C_k) \prod_{i=1}^{n} P(x_i|C_k)$$

The Naive Bayes classifier model combines the Naive Bayes probability model with a decision rule. The rule is known as the maximum a posteriori, used to select the class with the highest probability.

In Naive Bayes, features are considered independent and there is no information regarding the distribution of features. The assumption on the feature distribution is called as

event model of the Naive Bayes classifier. Multinomial Naive Bayes is one of the popular distributions for text classification and it assumes multinomial distribution for each of the input features.

The main advantages of Naive Bayes classifier includes scalability - its highly scalable, requires less training data, requires less computation time, and not sensitive to irrelevant features [58].

3.6.3 Support Vector Machine

Support Vector Machine (SVM) [55] is one of the most popular machine learning models for text classification commonly known to outperform other classifiers. It is a non-probabilistic classifier. An SVM model represents data points in the space to obtain a clear line dividing them into separate categories, and ensuring the gap is as wide as possible. To predict a data points class, SVM maps it to the same space and predicts the category based on which side of the line the data point falls. The algorithm optimizes to find the line with the largest separation between two classes. SVM is also known to perform well with non-linear classification using kernel and mapping the data points to higher dimensional space.

The SVM constructs a line in two-dimensional space and a hyperplane in higher dimensional space for classification. It optimizes to construct a hyper plane that has the largest distance to the nearest training data point of any class. Though the data points can be mapped to a finite dimensional space, the classes may not be linearly separable; therefore mapping the finite dimensional space into a higher dimensional space can make the separation easier by making it more obvious. Kernels are functions used to map the low dimensional space to higher dimensional space. SVM uses the kernel trick to make complex transformations to the data and finds an optimal hyperplane between the output categories.

For our study we are using the Linear SVM, which uses the linear kernel, and finds a linear hyperplane to separate the classes using the selected features.

SVM has many advantages including higher accuracy in prediction of outcomes, works well on smaller datasets [59], helps in gaining expert knowledge about the problem [60].

3.6.4 Random Forest

Random Forest [56][57] is the ensemble of numerous decision trees (explained in the paragraph below). It is an ensemble learning method for classification. Trees are constructed during the training phase and to predict a class for an unseen data point, the model returns the mode of the classes of all the trees built during training. Random forest is popular to avoid overfitting to the training data and is commonly used when the training data size is very small to address the issue. It is build using a technique called bagging; it is based on the idea that a combination of learning models increases the overcall accuracy of the model. Random forest constructs multiple decision trees, and merges them together to build a model; to test a new example, the inputs are given to all the trees in the model, each trees predicts or votes for a particular class, the algorithm finally chooses the class having the highest votes.

A decision tree [61] is a classification tree with each internal node representing input features in the dataset, the edges representing the possible values of the input features, and the leaf representing the class labels of the target feature or a probability distribution over the classes in the model. It works on a basic principle to divide the entire dataset into subsets based on input features until it reaches a smaller set containing data points falling under a single class label. It builds the tree by selecting nodes at each level by recursive partitioning, a procedure considering all the features in the input space and trying and testing different split points using a cost function. The split with the lowest cost is selected as the node. Therefore the root node is the best predictor or feature in the feature space to predict the outcome. The cost function uses Gini Index, entropy or Information Gain to find the best split. Trees that grow very deep tend to include irrelevant features for classification,

and overfit the training data; these trees have low bias and high variance. Random forest reduces variance by training multiple models on different parts of the training dataset.

For a training set $X = x_1, ..., x_n$ with output variable $Y = y_1, ..., y_n$, bagging technique repeatedly (G times) selects random samples of data from the training set with replacement and fits tree models to these samples.

For $g = 1, ..., G$, samples n training examples from X and Y, called X_g and Y_g, and build a tree h_b with these samples X_g and Y_g.

Prediction h' for the new data point x' is made by averaging the predictions from all the decision trees on x'.

$$h' = \frac{1}{G} \sum_{g=1}^{G} h_g x'$$

Random forest differs from bagging in one way, at each candidate split the tree learning algorithm selects random subset of features. This process is called feature bagging. This is a measure to avoid selection of same features by many decision trees in the model. If the training set has one or more strong features in predicting the target variable, these features will be selected in one or more trees in the ensemble of trees, causing them to be correlated. For a training set with p features, features are used for each candidate split.

The main advantages of random forest includes handling of thousands of features, gives unexcelled accuracy, estimates the variable importance, has methods of balancing error in unbalanced datasets.

3.7 Oversampling

Imbalanced dataset is an unavoidable issue when dealing with data. In real world, it is hard to find even distribution of classes, and this problem is common both in binary and multiclass classification. Class imbalance problem causes the model to learn more about a specific class and predicts new instances as majority class. There are ways to solve this

problem including oversampling and undersampling. Since our data set is having limited data (507 data points), we are following oversampling technique in our study. Oversampling is a way to adjust the distribution of class in a dataset. In this strategy data points are generated using a sampling technique and the minority class is oversampled to make the distribution balanced. SMOTE [62], an oversampling technique is implemented for our study.

3.7.1 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) [62], an oversampling technique that creates data points for minority class. It works by taking a sample from the dataset and considers the nearest neighbors from it. SMOTE creates synthetic data point by taking a vector between the current data point and one of the k nearest neighbors, then multiplies the vector with a random number from the range (0,1). This vector is added to the current data point to create a new synthetic point. SMOTE does not create multiple instances of the same datapoint, but it creates new synthetic data points.

3.8 Evaluation Metrics

Predictive models performance improves by tweaking the model, and making modifications based on the models performance. Model performance is measured by evaluation metrics. Evaluation metrics explains the models performance, and are essential to identify the best performing model for a particular problem. They are used to assess how well the model is performing in correctly labeling the classes, and used to determine the success of the model in prediction. There are several measures to evaluate the performance of the model, and this section discusses few of the metrics used in the study to evaluate the model performance. Confusion matrix [63], precision [63], recall [63] and F1 score [63] evaluates

the performance of the model and are used to discriminate between various models in the study.

3.8.1 Confusion Matrix

Confusion matrix [63] or error matrix is used to evaluate the performance of the classifier. It is constructed on a set of test data with previously known target values. It is presented in a contingency table to evaluate the performance of the classification algorithms, especially supervised. Each row in the confusion matrix represents instances in actual class and each column represents instances in prediction class. It allows in gaining detailed analysis than accuracy (proportion of correctly classified instances). Accuracy is a not a good measure to evaluate a model when the data is unbalanced, as the model biases towards the class with more data, therefore accuracy reports inaccurate estimation of the model. It is called confusion because it makes easy to evaluate the misclassified instances, or identify instances confused by the model and mislabeled. Table 3.6 reports the following:

1. True Positive (TP): number of instances correctly predicted as Positive
2. True Negative (TN): number of instances correctly predicted as negative
3. False Positive (FP): number of instances incorrectly predicted as Positive
4. False Negative (FN): number of instances incorrectly predicted as negative

Table 3.6: Confusion matrix to evaluate the performance of a model

		Predicted class	
		Predictive Positive	Predicted Negative
Actual	Actual Positive	True Positive	False Negative
	Actual Negative	False Positive	True Negative

The other metrics namely precision, recall, and F1 score are derived from confusion matrix.

3.8.2 Precision

Precision [63] also called, as Positive predictive value is defined as the fraction of the correctly classified relevant instances among the total instances retrieved. It is a measure of quality or exactness of the classification algorithm. It is the number of correctly classified instances divided by the total number of instances retrieved.

$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)}$$

High precision indicates the classifier has returned more instances of relevant results than irrelevant ones, and low precision indicates more false Positives than true Positives. A precision of 1 means every document retrieved is relevant.

3.8.3 Recall

Recall [63] also known, as sensitivity is the fraction of the relevant instances retrieved over the total number of relevant instances in the test data. Recall is the measure of the quantity or completeness. It is the number of instances of true Positives divided by the total number of Positive instances in the dataset.

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ negatives)}$$

A high recall indicates that the machine-learning model has retrieved most of the relevant instances. A recall of 1 indicates all relevant documents in the dataset were retrieved.

3.8.4 F1 Score

F1 score [63] is another metric to test the performance of the model. F1 score considers both recall and precision to calculate the value. It is the harmonic mean of precision and recall. Harmonic mean is used instead of an average to punish extreme values. F1 score

is also known as F1 measure because both precision and recall are evenly weighted in the equation. It is a better measure compared to precision and recall alone as it finds a balance between the two measures, and when there is an uneven class distribution precision and recall are biased.

$$F1 = 2 * \frac{Precision * Recall}{(Precision + Recall)}$$

F1 score reaches its best at 1 when both precision and recall are 1, and worst at 0. For a model trying to achieve a balance between the precision and recall, F1 score is maximized.

4 Results and Discussion

This section explores the distribution of data and discusses about the performance of various classifiers in different scenarios and identifies the best performing algorithm. Below are the different scenarios studied for the research.

1. Patient informative classification,
2. Pain informative classification,
3. Pain sentiment categorization and
4. Pain score classification

Data distribution helps in analyzing the nature of data; and results are discussed and interpreted to understand the above mentioned scenarios in Clinical Notes.

4.1 Data Distribution

In this section, data distribution in the four scenarios of prediction namely patient informative, pain informative, sentiment and pain scores are detailed. For each scenario, dataset varies and therefore four distribution histograms are plotted. Understanding the data distribution is essential before building a model as it gives insights into the data and helps in yielding better results. The data distribution for four datasets is discussed as follows.

4.1.1 Patient Dataset

In this case, notes informative regarding a patient is annotated as informative and notes non-informative regarding patients health, status etc. are annotated as non-informative. The details of annotation are detailed in Chapter 3 under Annotation subsection (Chapter 3.2). Of the total 504 notes in the entire dataset, 432 (8%) notes are labeled as informative and 72 (15%) are labeled as non-informative. The distribution of the data is highly skewed, as there are more informative notes than non-informative. To solve the class imbalance problem SMOTE (as discussed in Chapter 3) is implemented. Figure 4.1 demonstrates the histogram of the distribution.

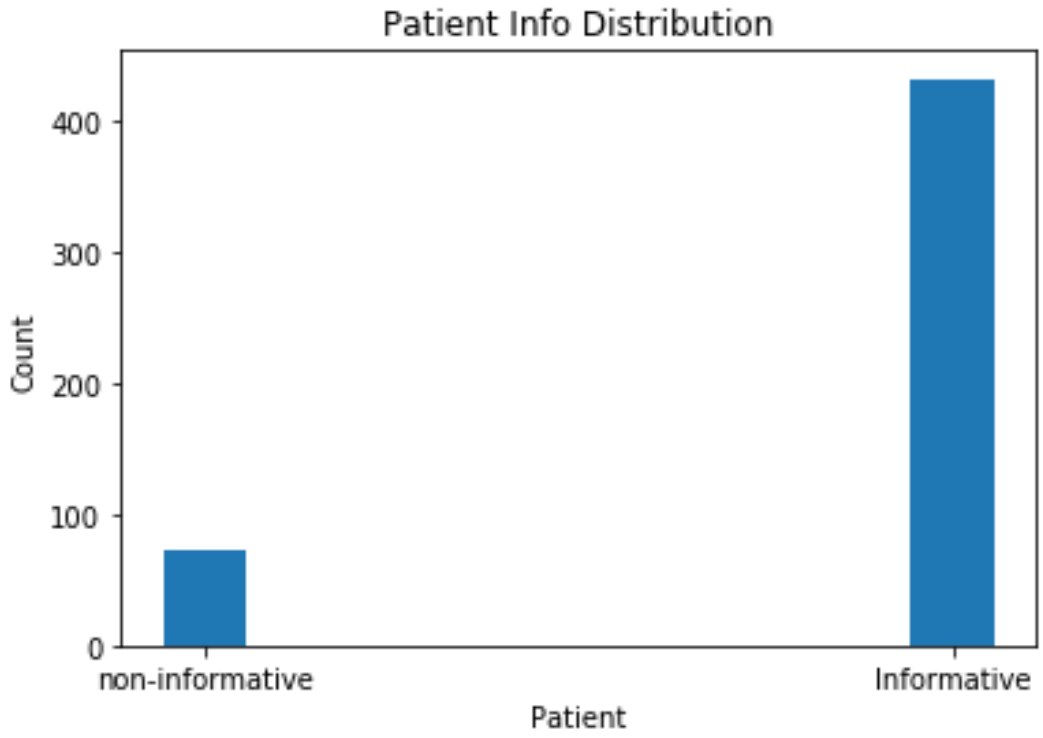


Figure 4.1: Data distribution of patient dataset with 504 data samples

4.1.2 Pain Dataset

In this case, notes informative regarding pain is annotated as informative and notes non-informative regarding patients pain such as pain status, current status etc. are annotated

as non-informative. The details of annotation are detailed in Methods section (Chapter 3) under Annotation sub-section (Chapter 3.2). Of the total 432 notes in the entire dataset, 247 notes (57%) are labeled as informative and 185 (42%) are labeled as non-informative. As described in Chapter 1, the notes informative regarding patient can be informative regarding patients pain. The distribution is balanced, and hence the machine learning models can be build on the data without further modifications to the data distribution. Figure 4.3 below demonstrates the distribution of pain data.

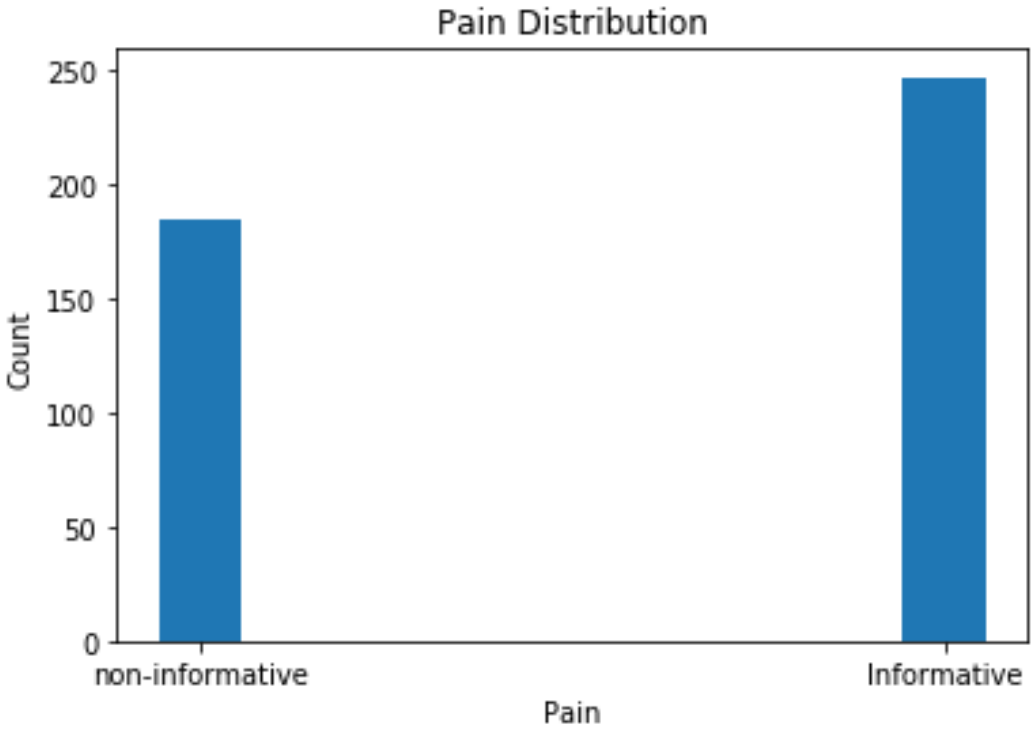


Figure 4.2: Data distribution of pain dataset with 432 data samples

4.1.3 Sentiment Dataset

In this case, sentiment is annotated based on the pain score values accrued from vital signs. Sentiment is labeled as decrease, if there is a decrease in pain level from the pain level identified in previous notes (ground truth); and increase if there is an increase in pain level,

neutral if the pain score is neutral, and not-determined, if the sentiment cannot be assessed due to unavailable pain scores. Of the 247 notes, 25 (10%) notes are labeled as increase, 43 (17%) notes are labeled as decrease, 62 (25%) notes are labeled as neutral and 117 (47%) notes are labeled as not-determined. As described in Chapter 1, notes informative regarding pain can be informative regarding patients sentiment. The distribution of the data is skewed, as there are more not-determined notes than decrease, increase, or neutral. To solve the class imbalance problem SMOTE (as discussed in Chapter 3) is implemented. Figure 4.3 below demonstrates the histogram of the distribution.

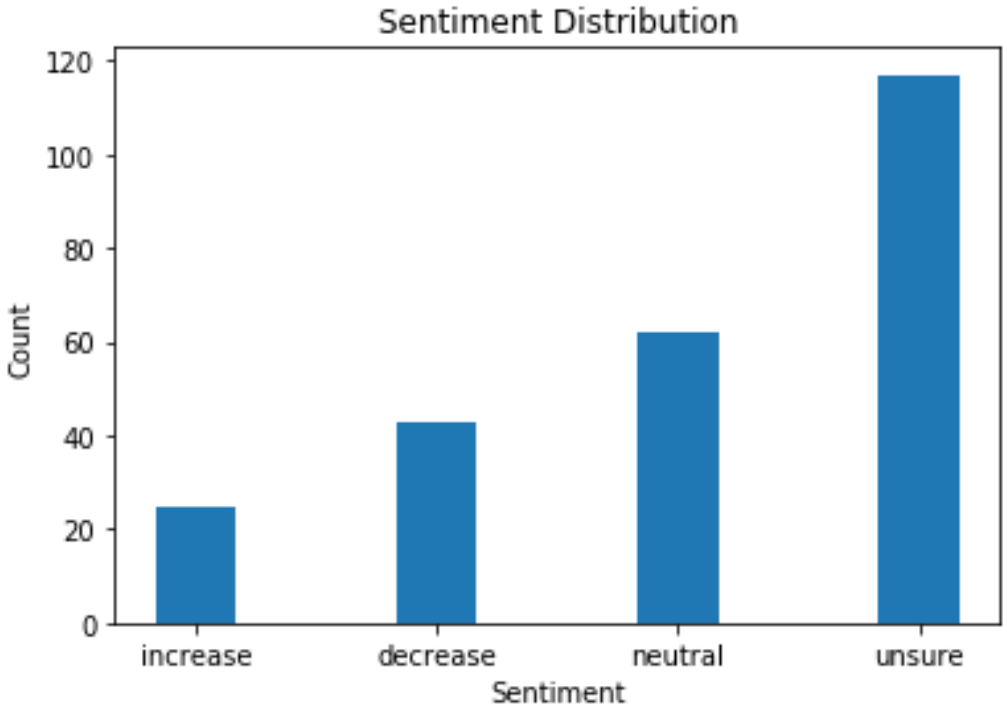


Figure 4.3: Data distribution of sentiment dataset with 247 data samples

4.1.4 Pain Score Dataset

In this case, pain scores for each notes is accrued from vitals signs. They are labeled as moderate if the pain score falls in the range (0-6), severe if it falls in the range (7-10) and not-determined if there is no pain score associated with the progress notes. Annotation

is detailed in the Methods section Chapter 3. Of the total 247 notes, 85 (34%) notes are labeled as moderate, 108 (44%) are labeled as Severe, and 54 (22%) are labeled as not determined. The distribution of the data is skewed, due to more severe notes than moderate and neutral. To solve the class imbalance problem SMOTE (as discussed in Chapter 3) is implemented. Figure 4.4 demonstrates the histogram of the distribution.

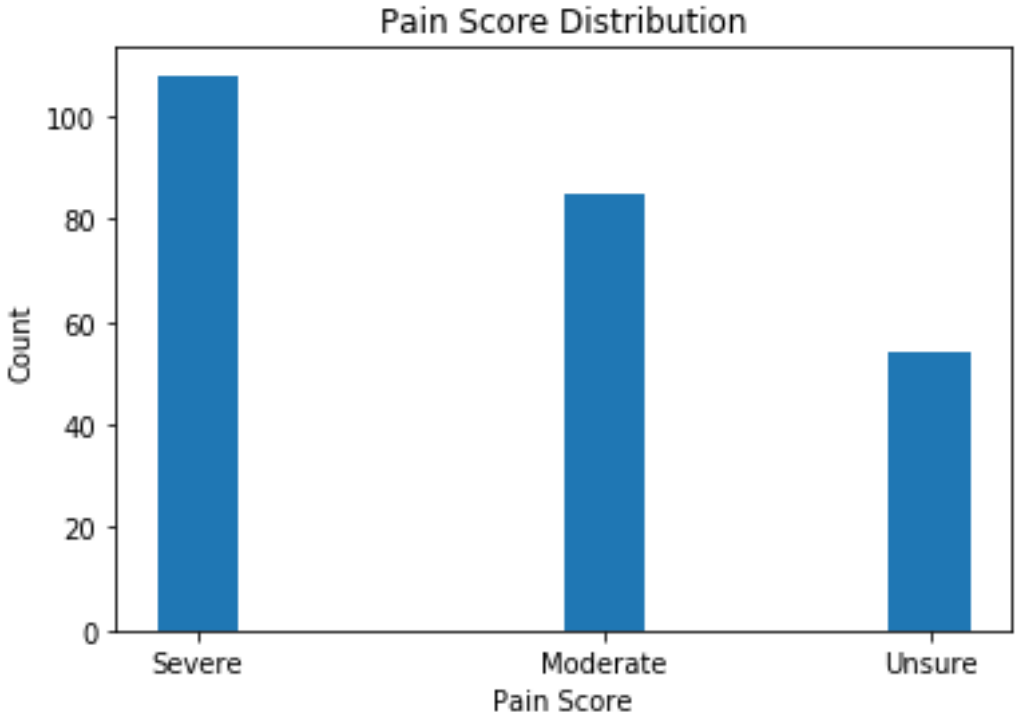


Figure 4.4: Data distribution of pain score dataset with 247 data samples

4.2 Supervised Machine Learning results

After understanding the distribution of data, next step to follow is building a machine learning model. For this study, supervised machine learning models are implemented; these models are trained using labeled data. Classifier learns from the data during training phase and predicts the outcome for the test data. Our problem deals with classes so the algorithms used for predicting the outcome are classification algorithms. In our study, we

have four scenarios to evaluate (patient informative classification, pain informative classification, sentiment classification and pain score categorization), so the section has four subsections for each scenario and results are discussed for each subsection.

4.2.1 Patient Dataset

The next step is building the model, and for the study we are assessing the performance of four different machine learning algorithms namely Logistic Regression (LR), Support Vector Machine (SVM) using linear kernel, Random Forest (RF), and Multinomial Nave Bayes (MNB), these algorithms are explained in Methods Chapter 3 (under section 3.6). To build our model, Scikit Learn library [65] and python [66] are used. Results and analysis are further discussed and table 7 below demonstrates the results using different features including count, Term Frequency-Inverse Document Frequency (tf-idf), Part of Speech tags (POS tags), cTAKES features, unigrams and bigrams, as discussed in Methods Chapter 3 (under section 3.4). Since the dataset is imbalanced, SMOTE as discussed in Methods Chapter 3 (under section 3.7) is implemented to balance the classes. Table 4.1 demonstrates the results; F1 score is reported for all the classifiers.

Table 4.1: Results without feature selection for different features and algorithms with F1 score.

Features		Feature Size	LR	RF	SVM	MNB
Unigram	count	2515	**0.924**	0.917	0.906	0.808
	tfidf		0.842	0.896	0.886	0.788
Unigram + POS tags	Count + POS	2550	0.920	0.921	0.911	0.808
	tfidf + POS		0.848	0.884	0.885	0.788
Bigram	count	11311	0.884	0.920	0.868	0.757
	tfidf		0.856	0.892	0.897	0.813
Bigram + POS tags	count + POS	11346	0.886	0.914	0.865	0.829
	tfidf + POS		0.853	0.886	0.901	0.807
cTAKES	cTAKES	5	0.644	**0.865**	0.503	0.622
cTAKES + Unigram	count	2520	0.858	0.853	0.828	0.748
	tfidf		0.788	0.808	0.807	0.791
cTAKES + Unigram + POS	count + POS	2555	0.838	0.867	0.802	0.802
	tfidf + POS		0.885	0.829	0.918	0.918
cTAKES + Bigram	count	11316	0.839	0.845	0.801	0.766
	tfidf		0.795	0.814	0.822	0.807
cTAKES + Bigram + POS	count + POS	11351	0.910	0.915	0.898	0.898
	tfidf + POS		0.745	0.916	0.917	0.917

From the above table, all the algorithms are performing well with F1 score ranging from 0.745 to 0.924. Though cTAKES features gives a high F1 score of 0.86, there are more false Positives (51) than true Positives (21) for non-informative class. A simple algorithm like Logistic Regression gives better results, hence this classifier is chosen for further analysis and the features selected are unigram count.

cTAKES features gives an F1 score of 0.86 with Random Forest, and uses only 5 features. Table 4.2 and 4.3 demonstrates the confusion matrix, precision and recall; these metrics are explained in methods section.

Table 4.2: Confusion Matrix of RF with cTAKES features

Prediction/actual	Non-Informative	Informative
Non-Informative	**21**	51
Informative	62	**370**

From the results it can be observed that there are more false negatives (51) than true negatives (21). They are analysed in the Discussion section (4.2.1.2).

Table 4.3: Precision and Recall of RF with cTAKES feature

Metric	Score
Precision	0.85
Recall	0.86

The LR model with unigram count as features is performing well (F1 score 0.92) without feature selection, however feature selection helps in eliminating irrelevant features, reduces training time, and it can be more generalized. Model with too many features leads to overfitting and builds a less generalized model, which can result in poor results with unseen. Also the model with cTAKES feature uses 5 features and obtain F1 score of 0.86 indicating that fewer features can result in good F1 score and hence features should be reduced using feature selection techniques.

4.2.1.1 Results with Feature Selection

Features selection techniques discussed in Methods chapter are implemented to select the most informative features uncovering the hidden structures in data and help in building an accurate predictive model. Some features are irrelevant to the research problem and are discarded to avoid redundancy. Table 4.4 demonstrates the precision, recall and F1 score for Logistic Regression with unigram and count as features for two feature selection techniques.

Table 4.4: Feature selection results for different features selection algorithms with F1 score, precision, and recall

	Feature selection	Features	Precision	Recall	F1 score
LR Unigram Count (features)	χ^2	**150**	0.950	0.863	0.903
	Mutual class info	135	0.916	0.856	0.884

It can be observed that χ^2 technique discussed in Chapter 3 section 3.5.5.1, yields better results compared to mutual information, selecting 150 features. The results and features are discussed in detail in section 4.2.1.2.4. Table 4.5 below demonstrates the confusion matrix with 150 features and Logistic regression, unigram and count as features.

Table 4.5: Confusion Matrix of LR with unigram count as features for χ^2 features selection

Prediction/actual	Non-Informative	Informative
Non-Informative	**52**	20
Informative	59	**373**

The section below discusses the results in detail.

4.2.1.2 Discussion

This section begins by discussing the best performing features. To overcome the class imbalance problem SMOTE is implemented to create artificial data points for non-informative class and balancing the distribution, as discussed in Methods chapter under section 3.7.

Furthermore the section is subdivided into algorithm analysis to analyze best performing algorithm, and feature selection, features selected, themes identification and themes clustering for text features.

4.2.1.2.1 Feature Analysis

In this section text and cTAKES features results demonstrated in table 7, 8, 9, 10 and 11 are discussed in the following subsections.

4.2.1.2.1.1 Text features: unigram, bigram, POS tags

Both unigrams and bigrams are performing well with F1 score ranging from 0.88 - 0.92, however with bigrams as the number of features increases it may add irrelevant features and decrease the performance. With unigrams the F1 was 0.92 and with bigram, F1 score

dropped to 0.88 explaining irrelevant features. From the results it can be observed that meaningful information can be captured from unigrams, such as pain, cope, better to help the algorithm differentiate between informative and non-informative nature of notes.

POS tags (33 features, discussed in Chapter 3 under section 3.4.4) does not show promising results for this dataset with F1 score of the combination unigram, bigram, POS ranging from 0.78 to 0.92, and the performance is contributed entirely by the unigram and bigram features and the POS tags may not capture the essence of the problem. The tags of adverb, noun, pronoun, verb used may not be enough to classify notes into informative and non-informative notes, as both might have both same verbs, adjectives making it difficult to obtain enough information from the tags to classify them correctly.

4.2.1.2.1.2 cTAKES Features

cTAKES features namely procedure, medication, anatomical, sign/symptom, and disease disorder are performing well with an F1 score of 0.86 with Random Forest algorithm. However the model has more false negatives than true negatives because, though the non-informative notes does not share information regarding the patient, but it has procedure, sign/symptom, anatomical and disease disorder mentions making the classifier difficult to predict the non-informative class. For instance the note below is non-informative, however it is predicted as informative because of several ***procedure***, *disease disorder*, ***anatomical*** and **sign/symptom** mentions.

Diagnosis*:* ***Sickle cell disease*** *Reason for referral: Ongoing* ***psychosocial support*** *for chronic illness. Assessment: Visited pt in her inpatient room on floor 5100. She was sound asleep at this time. No family was present. Pt lives in Apex with her mother. Mother works,* ***full time*** *at Goodwill Industries. This CSW contacted mother by phone and left a VM offering support and encouragement. This CSW also informed her that this CSW left her a parking* ***pass*** *in the room near the window.*

*Plan/Intervention: Continue to provide **emotional support** and counseling. Attempted to contact mother via phone but she did not answer so this CSW left a VM. Provided parking pass for discharge. Mother has a car to **transport** home at discharge. Consulted nurse Shirleen regarding pts progress and CSW needs. This CSW will continue to follow this patient and their family as needed, assessing their **coping** and adjustment and providing **emotional** and practical support and coordinating closely with the PHO team.*

From the example note above, though the text does not provide information regarding patients current state of health, there are mentions of procedure, sign/symptom, anatomical, and disease disorder making it difficult for the model to make accurate predictions.

Furthermore the misclassifications can be explained due to inherent issues with cTAKES such as incorrect annotations. Below is an example note with incorrect entity identification:

*Patient continues Losartan for nephropathy. Patient continues Dilaudid PCA/continuous infusion, scheduled IV Toradol every 6 hours and Lidoderm **patches** to thighs at night for pain management. Patient continues Atroven BID and prn Albuterol. Patient is receiving IV fluids. Follow up scheduled with PHE clinic @ 11 am on 7/9/2014 and PPL clinic @ 3:30 pm on 8/25/2014.Anticipate discharge home when pain can be managed with po pain medications. Dad at bedside and agrees with plan of care. Will follow for discharge needs.*

In the above text, the underlined word patches is tagged to sign/symptom, perhaps it was referring to patches in skin and therefore considered it as a sign/symptom. However, it does not provide meaningful information because it cannot be associated with neither sign nor symptom in the context of SCD. These incorrect tagging may produces incorrect results. The other example below explains incorrect entity identification of procedure.

We see another example below :

*Pain improving and transitioned to po scheduled **MS Contin**, scheduled Ibuprofen and as needed Oxycodone and. Patient continues Hydroxurea. Patient is taking adequate po fluids and IV fluids discontinued.Follow up scheduled with PHE clinic @ 1 pm on 7/1.*

Patient discharged home. Will follow for discharge needs.

For the above text, cTAKES incorrectly identifies MS (Morphine Sulphate) as procedure in the phrase MS Contin which is a drug and should be tagged to medication.

Due to incorrect tagging and providing similar information from both informative and non-informative notes, it causing misclassification and more false negatives (51) than false Positives (21).

4.2.1.2.1.3 Combination of Text and cTAKES features

F1 score with combination of features ranges from 0.74 to 0.91, however with Logistic Regression and unigram count features the model gives an F1 score of 0.92, hence the combination of features are not studied further. However the combination of cTAKES (5 features) and text features (2500 - 11000 features) results in no gain in performance, this may be because the information provided by cTAKES features is already captured by text features.

4.2.1.2.2 Algorithm Analysis

After identifying the features, the next step is identifying the best performing model. Logistic Regression is the best performing classifier with unigram and count as features with highest F1 score of 0.924. This proves that the data can be classified using a simple linear model, and does not require complex algorithms such as SVM and Random Forest.

4.2.1.2.3 Feature selection analysis

After model building, feature selection helps in avoiding irrelevant features and improve model performance. After feature selection on LR and unigram count features, F1 score obtained was 0.902. Results are further analyzed by looking at the features. Misclassification errors in informative and non-informative clases can be explained further as the term verbal is mostly associated with discharge notes, and all discharge notes are informative as they provide information regarding the patients discharge from the hospital, but few notes

also use verbal and they are not informative. However, since the majority of the notes with the feature verbal are informative, some non-informative notes containing the term are also misclassified as informative. A sample note annotated as informative explains the usage of the term verbal in discharge notes:

Discharge teaching completed with Mom who verbalized understanding, ID verified. Patient PIV removed. Nutritionist spoke with Mom before leaving. Patient left without transport in a wheelchair.

The notes with the term asleep are mostly non-informative unless the patients family share information regarding pain, health, history, and its difficult for the machine to capture such fine details from notes, and mark them as informative. Hence, there are misclassifications since the model predicts most of the notes with the term asleep as non-informative. The note below is an example of a non-informative note containing the term asleep.

CCLS (Certified Child Life Specialist) introduced child life services to Pt's caregiver to increase knowledge of supportive services and promote family centered care. Patient asleep during this time. CCLS offered bedside activities to normalize the hospital experience, support development, and facilitate coping. Pt's caregiver declined on Pt's behalf at this time. No other needs noted. Will continue to follow for child life care throughout hospitalization.

Since the patient was asleep, there is no added information regarding his or her symptom reflected in this document, we categorize this as a non-informative note. There are very few non-informative notes and certain keywords are associated with non-informative notes such as asleep thus making the prediction of non-informative notes simple compared to informative notes where every detail from current health status to past medical history are considered informative and thus making the problem complicated. Precision is higher than recall when false Positive (59) are higher than false negatives (20), since it is hard to predict informative classes due to more complexity and no clear features to distinguish informative notes, the problem has higher precision (0.95) than recall (0.86).

4.2.1.2.4 Features selected

Below are the 150 features selected by χ^2 feature selection technique, in **decreasing** order of importance, the top features are selected based on the F1 score of the model during cross validation. 150 Features in decreasing order of importance are tabulated in table 4.6. The features are further discussed in section 4.2.1.2.5.

Table 4.6: Features selected in decreasing order of importance with χ^2 feature selection

Top features	Features in decreasing order of importance
1-30	'pain', 'discharge', 'identify', 'status', 'management', 'risk , 'review', 'issue', 'crisis', 'mom', 'child', 'certified', 'initial', 'mother', 'life', 'specialist', 'resource', 'provided', 'iv' (intravenous), 'high', 'uaaf', 'po' (oral), 'screen', 'md', 'schedule', 'pet', 'pca' (patient controlled analgesia), 'prior', 'adl' (activities of daily living),
30-60	'plan', 'pm' (time), 'healthcare', 'therapy', 'system', 'history', 'post', 'license', 'type', 'nad', 'none', 'instruct', 'agreement', 'result', 'attempt', 'writer', 'past', 'medicaid', 'concurrent', 'independent', 'meet', 'network', 'asleep', 'age', 'piv' (peripheral intravenous line),'ed' (emergency department), 'patient/family', 'prefer', 'home', 'daily'
60-90	'worker', 'fluid', 'physiology', 'state', 'util', 'progress', 'well', 'planintervent', 'graduate', 'anticipate', 'comment', 'chest', 'phe' (periodic health examination), 'vss' (vital sign stable), 'medic', 'kg', 'inform', 'admit', 'remove', 'allergy', 'ss' (signs and symptoms), 'morphine', 'fund', 'food', 'promote', 'please', 'note',
90-120	'toradol', 'meal', 'full', 'pho' (primary health organisation), 'nc' (north carolina), 'session', 'phone', 'locate', 'id', 'seen', 'intervent', 'arrive', 'specific', 'orient', 'verify', 'case', 'monitor', 'letter', 'went', 'hgb', 'weight', 'infuse', 'ill', 'procedure', 'offer', 'support', 'distress', 'adjust', 'time',
120-150	'afebrile', 'knowledge', 'self', 'visit', 'remain', 'dilaudid', 'addit' (additional), 'ask', 'pass', 'stepfather', 'bilateral', 'unit', 'stable', 'rel' (relative), 'practice', 'poc' (point of care), 'ga' (gas), 'regard', 'volunteer', 'duke', 'rmd' (Room for meals at Duke), 'xray', 'asthma', 'coordination', 'month', 'understand', 'verbal', 'gdl' (g/dl gram per deciliter), 'interview', 'paperwork', 'prn' (as needed), 'ongoing', 'area', 'lb'

4.2.1.2.5 Themes identification

Based on the features selected, we have tried to identify themes in our dataset. Clustering similar topics helps in providing insights into the data, and in our problem it helps in segregating the themes identified into specific classes, thus identify notes specific to infor-

mative and non-informative. Identifying themes can help in understanding what treatments are involved in treatment of SCD, medications given at discharge or during hospitalization. Themes identified from the features selected are:

1. Discharge notes: discharge, toradol, morphine, mother, piv, fluids Discharge notes explains the condition of the patient at the time of discharge, medications given. Below is an example notes with the keywords **(morphine, discharge)** identified:

   *Patient transitioned to **Morphine** po and prn po Oxycodone. Referral to PNE clinic s/p discharge for evaluation of headaches. Follow up scheduled with PHE clinic @ 10 am on 10/28.Patient **discharged** home. No discharge needs noted. Will follow for discharge needs.*

2. Patient case management and discharge notes: allergies, toradol, morphine, resource, planning, identified, status, risk, issues, management, medicaid, concurrent, patient-family, findings, fundings, anticipate The notes explain regarding the patients case management, the state of the patient after the admit. Therefore it helps in identifying the state of patient and treatment provided, and also evaluate the effective treatment. Below is an example notes with the keywords **(morphine, toradol, anticipate)** identified:

   *Patient is a 13 year old with Sickle Cell Disease type SS who presented with chest and abdominal pain admitted with pain crisis. CXR was negative. Hgb is 9.1 g/dL and retic is 12.72%. Patient is receiving scheduled IV **Toradol** and IV **Morphine** PCA/continuous for pain management. Patient is receiving IV fluids. Follow up scheduled 7/23 @ 1 pm with Development and Behavior Clinic, 7/1 @ 9:30 am with PNE clinic and 7/1 @ 11:30 am with PHE clinic. **Anticipate** discharge home when pain managed with po pain medications. Weight: 47.5 kg Height: 145 cm Allergies: NKDA Met with mom and patient at bedside. Mom:. Patient lives at home with mom and brother in Raleigh, NC. Patient is not followed by any community*

resources. Mom will transport home at discharge. Patient has NC Medicaid to assist with medical expenses.

3. Pet therapy notes: pet, consent This theme helps in identifying what percent of patient have taken this treatment, and thereby evaluate the effectiveness of the treatment. Below is an example notes with the keywords **(pet)** identified:

   *Child Life Pet Therapy Contact Note: 1:1 Intervention(s) - Pet Therapy (Approved 11/18/15 - consent located in patient's chart) 1:1 Interventions comments: Patient seen at bedside for Pet Therapy by **Pets** at Duke Handler dog team to normalize the hospital experience and provide distraction/relaxation mechanism for anxiety and pain management issues. Patient and family reported their appreciation for visit.*

4. Child life support notes: coping, aeb, support, hospitalization This theme helps in understanding the state of different patients, and effectiveness of treatment, patients coping status. Below is an example notes with the keywords **(coping, hospitalization, support)** identified:

   *Child Life Support Note Reaction to Hospitalization/Illness: **Coping** well with overall **hospitalization**, interacted easily with CCLS Therapeutic Interventions: Education of child life services Emotional support Developmental activities provided at bedside to normalize the hospital experience, **support** development, and facilitate coping Will involve opportunities for special events/visitors Will encourage social opportunities walking incentive states he is compliant with incentive spirometer CCLS created walking incentive chart for him to increase compliance with ambulation Pain management/relaxation Will continue to facilitate expressive activities/opportunities Outcome: Will follow throughout treatment for child life care*

5. Family background: graduation, meals, food, letter, mother This theme is mainly related to family and helps in understanding the state of most of the families of the patients, understand the background. This information can help in devising preven-

tion plan for the patient based on their family situation and condition. Below is an example notes with the keywords **(mother, letter)** identified:

*Assessment: Met with Pt. and his mother in his inpatient room on floor 5100. Pt. was quiet and appeared to not be feeling well at this time. His **mother** was at his bedside and appropriately attentive to his needs. Pt. is a 15 year old boy who has sickle cell disease. He lives in Hope Mills with his mother and father. Mother works full time as a teaching assistant at a school and father is often away as a truck driver. Pt. had a 29 year old half brother and mother shared that he was tragically killed in a car accident in January 2016. Mother shared that this has been very difficult for the whole family including Pt.. Mother shared that Pt. is doing well in school and earned AB Honor Roll. She shared that his last hospitalization prior to now in May, was back in December. Mother mentioned how "grumpy" Pt. is and his "attitude" when he is hospitalized, but this CSW encouraged her patience as teens often take out their pain and frustrations on their parents while in such close quarters in the hospital. Encouraged mother to take breaks from room as needed. She stated that Pt. is attending marching band camp this summer and very excited about it. Family declined Camp Kaleidoscope. Mother shared that Pt. is currently playing the keyboard but wishes he was playing the drums. She stated that he wants to participate in Marching Band at school, but Mr. West the teacher will not let him because he has sickle cell disease. Mother was receptive to this CSW advocating and giving her a **letter** for school about this. Mother also asked for this CSW's assistance with Family Medical Leave Act (FMLA) paperwork.*

4.2.1.2.6 Themes clustering

Further these themes identified are clustered into informative and non-informative categories; table 13 below demonstrates the clusters. Certain themes are only identified in informative class, and some of them are identified in both informative and non-informative

classes, however there no themes that can only be classified as non-informative. Theme clustering helps clinicians to look only at the informative themes and avoid non-informative themes as they dont provide information regarding the patient and thus save time. It can also help in improving the writing of clinical notes, by guiding them to specific non-informative notes and identify ways to make them informative. From table 4.7 it can be also observed that these categorization are meaningful as patient case management and discharge notes are mostly informative as they provided information regarding patient and pet therapy, child life support notes, family background are a mixture as they are provide both informative and non-informative information. Informative notes:

Table 4.7: Themes categorized into informative and non-informative categories

Informative notes	Combination of informative and non-informative
Patient case management Discharge notes	Pet therapy Child life support notes Family background

After considering both text (150 features) and cTAKES features (5 features), the optimal features are cTAKES features and algorithm is RF model with an F1 score of 0.86. cTAKES model is simple with 5 features as opposed to text features (150) giving an F1 score of 0.96, therefore cTAKES model is chosen as optimal model.

4.2.2 Pain dataset

The next step is building the model by using the four algorithms as described in Methods Chapter 3 (under section 3.6). Features used for the study include count, term frequency-inverse document frequency (tf-idf), Part of Speech tags (POS tags), cTAKES features, unigrams and bigrams, as discussed in Methods Chapter 3 (under section 3.4). For this dataset, the model is trained to predict pain informative and non-informative. Pain informative notes are informative regarding patients health, pain etc. and non-informative notes

are non-informative regarding patients pain status, etc., detail annotation procedure is described in Methods Chapter 3 (under section 3.2). Table 4.8 below demonstrates the results; F1 score is reported for all the classifiers.

Table 14: Results without feature selection for different features, algorithms with F1 score

Table 4.8: Results without feature selection for different features and algorithms with F1 score.

Features		Feature Size	LR	RF	SVM	MNB
Unigram	count	2342	0.802	0.812	0.802	0.776
	tfidf		**0.822**	0.8791	0.829	0.795
Unigram + POS tags	Count + POS	2377	0.800	0.799	0.801	0.770
	tfidf + POS		0.818	0.786	0.829	0.792
Bigram	count	9797	0.804	0.767	0.752	0.790
	tfidf		0.810	0.769	0.812	0.798
Bigram + POS tags	count + POS	9832	0.806	0.777	0.756	0.791
	tfidf + POS		0.810	0.752	0.814	0.804
cTAKES	cTAKES	5	0.686	**0.707**	0.623	0.629
cTAKES + Unigram	count	2347	0.801	0.806	0.796	0.771
	tfidf		0.808	0.805	0.821	0.786
cTAKES + Unigram + POS	count + POS	2382	0.806	0.790	0.792	0.776
	tfidf + POS		0.807	0.793	0.821	0.780
cTAKES + Bigram	count	9802	0.810	0.784	0.766	0.801
	tfidf		0.796	0.759	0.814	0.776
cTAKES + Bigram + POS	count + POS	9837	0.810	0.758	0.766	0.865
	tfidf + POS		0.794	0.762	0.812	0.776

From the above table, all the algorithms are performing well with F1 score ranging from 0.75 to 0.86. Since pain dataset is accrued from patient dataset, and the F1-score for patient was 0.92, however the maximum F1 score for pain is 0.86, it can be observed that identifying pain is a complex problem compared to identifying patient informative and non-informative as it needs to learn intricate information regarding pain from notes. A simple algorithm like Logistic Regression gives better results (F1 score 0.82), though the best F1 score was reported by MNB (F1 score of 0.86), it uses too many features (9837)

compared to LR with 2342 features, hence LR classifier is chosen for further study and the features selected are unigram and tf-idf. It can also be observed that SVM and LR gives the same F1 score of 0.82, however compared to SVM, LR is a simple model and hence considered as the optimal algorithm.

With cTAKES features, RF is the best performing algorithm with an F1 score of 0.70 and uses only 5 features. The table 4.9 and 4.10 below demonstrates the confusion matrix, precision and recall for RF with cTAKES features.

Table 4.9: Confusion Matrix of RF with cTAKES features

Prediction/actual	Non-Informative	Informative
Non-Informative	**116**	69
Informative	74	**173**

The results explain that the model is performing well, as the diagonals are higher than the non-diagonals indicating that there are more true Positives and true negatives than false Positives and false negatives, giving high recall and precision. Results are explained in discussion section (4.2.2.2.1.2). Precision and recall for the RF model are tabulated in table 4.10.

Table 4.10: Precision and Recall of RF with cTAKES feature

Metric	score
Precision	0.72
Recall	0.70

The LR model is giving good results (F1 score 0.822) without feature selection, however feature selection helps in eliminating irrelevant features, reduces training time. RF model with 5 cTAKES features gives an F1 score of 0.70 indicating that fewer features can perform well and making feature selection for text features an essential next step.

4.2.2.1 Results with Feature Selection

Feature selection is a step to improve the performance of the model and also to reduce the complexity. Features selection techniques discussed in Methods Chapter 3 (under section 3.5) are implemented to select the most informative features uncovering the hidden structures in data and help in building an accurate predictive model. Table 4.11 below demonstrates precision, recall and F1 score for Logistic Regression with unigram and tf-idf as features for two feature selection techniques.

Table 4.11: Feature selection results for different features selection algorithms with F1 score, precision, and recall

	Feature selection	Features	Precision	Recall	F1 score
Unigram tf-idf (features)	χ^2	**135**	0.749	0.951	0.837
	Mutual class info	1120	0.736	0.918	0.816

It can be observed that χ^2 technique [50] (discussed in section 3.5.5.1) yields better results compared to mutual information; 135 features are selected for the model. Table 4.12 below demonstrates the confusion matrix with 135 features using Logistic regression as classifier and unigram tf-idf as features.

Table 4.12: Confusion matrix of LR with unigram tf-idf as features for χ^2 feature selection

Prediction/actual	Non-Informative	Informative
Non-Informative	**105**	80
Informative	12	**235**

Results are discussed in the flowing section:

4.2.2.2 Discussion

Discussion section is divided into different sub sections including feature analysis to analyze different features under study, algorithm analysis to analyse best performing algo-

rithm, feature selection, features selected, themes identification and themes clustering for text features.

4.2.2.2.1 Feature Analysis

In this section results of features including text, cTAKES and their combination are discussed in different sub-sections as follows.

4.2.2.2.1.1 Text features: unigram, bigram, POS tags

Unigrams and bigrams are performing well with F1 score ranging between 0.76 - 0.82, however for SVM and RF, bigram features resulted in reduced performance, it may be because the model requires less complex features, and simple features can do the job.

The combination of unigrams, bigrams and POS tags (33 features) has an F1 score ranging from 0.76 to 0.86, however unigram and bigram alone as features has F1 score ranging from 0.76 - 0.82, indicating that addition of POS tags has so significant improvement in the performance, this may be because the information provided by POS tags is already captured by the features from unigrams or bigrams, and they dont contribute significantly in the combination.

4.2.2.2.1.2 cTAKES features

With 5 features namely procedure, anatomical, sign/symptom, disease disorder and medication, RF model is obtaining an F1 score of 0.70. However the model has high false negative (69) and false positives (74), this is because cTAKES identifies multiple instances of the same phrase and counts it multiple times, thereby increasing the feature count and learning incorrect information while model building. An example below explains cTAKES identifying multiple instances of a phrase and counting it multiple times giving more weightage to length of the phrase.

Below is an example note:

Patient continues Losartan for nephropathy. Patient continues Dilaudid PCA/contin-

uous infusion, scheduled IV Toradol every 6 hours and Lidoderm patches to ***thighs at night for pain*** *management. Patient continues Atroven BID and prn Albuterol. Patient is receiving IV fluids. Follow up scheduled with PHE clinic @ 11 am on 7/9/2014 and PPL clinic @ 3:30 pm on 8/25/2014.Anticipate discharge home when pain can be managed with po pain medications. Dad at bedside and agrees with plan of care. Will follow for discharge needs.*

In the above sample note, the bold phrase thighs at night for pain is identified as sign/symptom thrice by breaking one phrase into three different phrases containing the word pain: thighs at night for pain, night for pain, and pain. However the complete sense can be captured in one phrase, but the count for sign/symptom is 3, instead of 1. Therefore it gives more weightage to longer phrases by identifying sub phrases multiple times. And it is a common issue with all the notes.

We see another example of incorrect tagging of words to the features and multiple counting in the note below. The notes is annotated as informative but the model predicted it as non-informative.

Reviewed chart. Patient is an 18 year old with ***Sickle Cell Disease*** *type SS who remains hospitalized with* ***pain crisis****. Patient transitioned to po MS Contin and continues Morphine PCA for breakthrough pain. Patient continues scheduled Ibuprofen. Child psych consulted and making recommendations. Follow up scheduled with PHE clinic 2 10:30 am on 1/27. Will follow for discharge needs.*

In the notes above the the phrase Sickle cell disease is broken into sub-phrase Disease and counted twice for disease disorder feature. Similarly, the phrase pain crisis is broken into crisis and counted twice for sign/symptom feature. And incorrect tagging such as PCA is tagged as anatomical site, but it is a procedure.

Thus multiple counting of phrases and incorrect tagging may be the cause of misclassification errors and obtaining high false positives 74 and false negatives 69 by giving more weightage to insignificant features and giving less weightage to significant features.

4.2.2.2.1.3 Combination of text and cTAKES features

With the combination of text and cTAKES features the F1 score of the models ranges from 0.76 to 0.86. Whereas the text and cTAKES features individually with the best algorithm obtain F1 scores of 0.82 and 0.72 indicating that the combination does not significantly improve the performance. This can be explained by the fact that the features both text and cTAKES may provide same information and hence addition of them didnt unfold new aspects of the data.

4.2.2.2.2 Algorithm Analysis

After identifying informative features, the next step is to analyze the best performing algorithm for text features. Logistic Regression with unigram and tf-idf as features is the best performing algorithm among Random Forest, SVM and MNB (see Table 14), indicating that a simple linear model can distinguish between the two classes, instead of complex models such as RF and SVM. While MNB interprets all features as independent of each other, there may be some correlation between these features which could be the reason for its poor performance.

4.2.2.2.3 Feature selection analysis

Once the best performing algorithm is identified, optimal text features are selected using χ^2 and results are analyzed. From the confusion matrix of Logistic Regression with 135 unigram tf-idf features (Table 17), it can be observed that the precision is low because 79 non-informative notes are predicted as informative. This can be explained by the imbalance nature of classes, i.e. there are more informative notes (57%) than non-informative (42%), therefore non-informative notes are more often predicted as informative, leading to low precision but high recall.

On further analysis, the child life playroom notes and other notes do not provide meaningful information and are incorrectly classified as informative. The term cope (one of the features selected) in the text makes the machine incorrectly classify it as informative, be-

cause the term coping is associated with large number of informative notes (coping well etc.), therefore if the model finds the term cope in the text, it classifies it as informative. 56 instances of the total 80, misclassified as informative contains the word cope and therefore is misclassified as informative. Thus, having a low precision (0.749).

Below is an example text with the term cope and annotated as pain non-informative :

Child Life Playroom Note: Child Life Intern met with Patient and family to provide continuity of child life care. Patient is involved in daily playroom activities as provided by the Child Life staff to normalize the hospital experience, support development and facilitate coping. No other needs noted. Will continue to follow for child life care.

4.2.2.2.4 Features selected

Below are the 150 features selected by χ^2 feature selection technique, in **decreasing** order of importance, the top features are selected based on the F1 score of the model during cross validation, and the models are trained using these features. 135 features in decreasing order of importance are tabulated in table 4.13. Features are discussed in section 4.2.2.2.5.

Table 4.13: Features selected in decreasing order of importance with χ^2 feature selection

Top features	Features in decreasing order of importance
1 - 30	'pet', 'discharge', 'therapy', 'pain', 'home', 'verified', 'approve', 'dog', 'handler', 'state', 'session', 'chest', 'arrive', 'anxiety', 'distraction/relax', 'mechanism', 'monitor', 'well', 'avs' (arteriovenous), 'review', 'cope', 'id', 'understand', 'service', 'intervent',
30-60	'pca', 'verbal', 'complain', 'instruct', 'need', 'back', 'ukulele', 'condition', 'picture', 'aeb' (as evidenced by), 'belong', 'public', 'mivf' (maintenance intravenous fluid), 'stable', 'visit', 'ivf' (intravenous fluid), 'overall', 'issue', 'photo', 'ed', 'dilaudid', 'vss', 'ra' (respiration), 'music', 'appreciate', 'pass', 'acute', 'side', 'familiar', 'distress', 'outcome', 'easily',
60-90	'Vdh' (room), 'implement', 'reaction', 'stressor', 'lower', 'clinic', 'inpatient', 'leg', 'control', 'rate', 'nutrition', 'emotion', 'park', 'identify', 'ibuprofen', 'right', 'allergy', 'abdomen', 'anticipate', 'development', 'unexpected', 'license', 'afebrile', 'iv' (intravenous), 'siblings' ,
90-120	'difficulty', 'lisinopril', 'confirm', 'doc', 'flow', 'sheet', 'bolu' (bolus - dose given by iv), 'desire', 'post', 'unable', 'repeat', 'bs' (bedside), 'pharmacy', 'sign', 'normal saline', 'admit', 'assist', 'hospitalization', 'introduction', 'therapeutic', 'consent', 'cell', 'drop', 'comment',
120-150	'present', 'downstair', 'respiratory', 'support', 'annett', 'sat', 'screen', 'bed', 'chaplain', 'outside', 'kayla', 'co' (complain), 'obtain', 'skin', 'reassess', 'dose', 'later', 'staff', 'vital signs', 'appear', 'kelsey', 'stepfather', 'play', 'medications/follow', 'person', 'orient', 'commun', 'gdl' (g/dl - measure), 'nausea'

4.2.2.2.5 Themes identification

Based on the features selected, we are identifying themes in our dataset using the features. Clustering similar topics helps in providing insights into the data, and in our problem it helps in segregating the themes identified into specific classes, thus helps in identifying only pain informative and non-informative notes. Themes helps in focussing at a particular problem, for instance discharge notes themes helps in analyzing the most popular medication among patients; music therapy themes helps in analyzing the popularity of the therapy in patients; patient case management helps in analyzing the common pain score in patients. Below are the themes identified in the notes with the terms selected;

1. **Discharge notes:** discharge, avs (arteriovenous), wheelchair, verbalized, home, piv (peripheral intravenous line) This note have information regarding the patient status at discharge and medications prescribed. Below is an example notes with the keywords **(piv, wheelchair, discharge, avs)** identified:

   *Pt **discharged** home with mother (pt's mother picture ID verified) via **wheelchair.** **PIV** removed. **AVS** reviewed. See doc flow sheet for more details.*

2. **Assessment notes:** Ibuprofen, dilaudid This note assesses the state of the patient during hospitalization, medication given, identifies the state of patients and help devise new treatments. Below is an example notes with the keywords **(ibuprofen, diluadid)** identified:

   *Patient is afebrile. Blood culture is negative. IV **Dilaudid** and IV Toradol transitioned to po Oxycodone and po **Ibuprofen** for pain management in anticipation of discharge. Follow up scheduled with PHE clinic @ 9:30 am on 9/17. Will follow for discharge needs*

3. **Music therapy notes:** music, ukulele (music instrument) This note describes about the music therapy for a patient. It helps in analyzing the most popular instrument among patients, and how it helps alleviate pain. Below is an example notes with the keywords **(music, ukulele)** identified:

   *Patient received **music** therapy visit to encourage self-expression. This MT conducted a **ukulele** lesson with Pt., and gave him a ukulele as part of the Ukulele Kids project.*

4. **Child life support notes:** coping, aeb (as evidence by), support, hospitalization It describes about the patient status, and effectiveness of the treatments by analyzing the coping status of patients. Below is an example notes with the keywords **(coping, hospitalization, support)** identified:

*ccls met with pt to provide continuity of child life care. Pt continues to cope well with overall **hospitalization** at this time.Patient is involved in daily playroom activities as provided by the Child Life staff to normalize the hospital experience, **support** development and facilitate **coping**. Will continue to follow for child life care throughout hospitalization.*

5. **Pet therapy notes:** anxiety, pet, consent, appreciation Helps in analyzing the popularity of the therapy among patients and also help in introducing new therapy based on the analysis. Below is an example notes with the keywords **(pet)** identified:

   *Child Life Pet Therapy Contact Note: Intervention(s) - Pet Therapy (Approved 9/16/15 - consent located in patient's chart) Interventions comments: Patient seen at bedside for Pet Therapy by **Pets** at Duke Handler dog team to normalize the hospital experience. Patient and family reported their appreciation for visit.*

6. **Admission notes:** monitor, paperwork, ed (emergency department), mivf (Maintenance Intravenous Fluid) This note describes about the state of patient during admission, therefore it helps in assessing the time of hospitalization, whether its too early to admit a patient or its late to hospitalize and they should have admitted the patient earlier. If the patient is admitted earlier then what are the common steps to alleviate pain at home. Below is an example notes with the keywords **(ed, monitor, paperwork)** identified:

   *Patient arrived via stretcher from **ED** with transport and mother at bedside. VSS at time of arrival. Oriented to room/unit. Admission paperwork done. No further requests at this time. Will continue to monitor. Patient arrived via stretcher from ED with transport and mother at bedside. VSS at time of arrival. Oriented to room/unit. Admission **paperwork** done. No further requests at this time. Will continue to **monitor**.*

7. **Family background:** grandmother, school, parking, request, public It helps in an-

alyzing the family situation and background, and devise a treatment or prevention plan suitable to the patient. Below is an example notes with the keywords **(parking)** identified:

Diagnosis: Sickle cell disease Reason for referral: Ongoing psychosocial support for chronic illness Assessment: Met with Patient in her inpatient room on floor 5100. There was no family present again today, but Patient shared that two aunts visited with her yesterday, and her father has been coming after work. She is a sweet 11 year old girl with sickle cell disease who was admitted on 6/18. She just completed the 5th grade and has reported being a bit nervous for 6th grade next fall. Patient was painting in her room at this time and watching television. She stated that she is coping okay and was understanding that her mother had to be home with her siblings. Plan/Intervention: Provided emotional support and counseling. Continue to explore how Patient is coping with unexpected hospitalization. Provided another ***parking*** *pass for family that visits this week. This CSW will continue to follow this patient and their family as needed, assessing their coping and adjustment and providing emotional and practical support and coordinating closely with the PHO team.*

8. **Patient case management and discharge notes:** chest, allergies, abdomen, anticipate This information is regarding the patients medical background and helps in identifying a pattern in patients pain and their medical background and thus devise effective treatment plan for patients based on their pain and medical history. Below is an example note with the keywords **(abdomen)** identified:

   Case Management Assessment and Discharge Plan Admit Reason: Patient is an 11 year old with Sickle Cell Disease type SC, and recurrent episodes of neutropenia and thrombocytopenia of unclear origin who presented with ***abdominal*** *pain, constipation and splenomegaly admitted to Duke Children's Hospital with abdominal vaso occlusive pain crisis.*

4.2.2.2.6 Themes clustering

Themes identified are clustered into informative and non-informative categories; table 20 below demonstrates the clusters. Informative themes helps clinicians to analyze only pain specific themes and discard the rest, thus saving time by avoiding non-informative themes as they dont provide information regarding pain. Themes clustering into non-informative notes indicate that there is room of improvement in writing and devise ways to make all notes meaningful and save time of nurses by avoiding non-informative notes. From table 4.14 it can be observed that themes assessment and admission are mostly informative as they provide information about pain; discharge notes, music therapy and pet notes does not provide information about pain and thus non-informative; and child life support notes, family background and patient case management and discharge notes are a combination of informative and non-informative notes.

Table 4.14: Themes categorized into informative and non-informative categories

width=1

Non-Informative	Informative notes	Combination of informative and non-informative
Discharge notes	Assessment notes	Child life support notes
Music therapy notes	Admission notes	Family background
Pet therapy		Patient case management and discharge notes

Below are example theme of child life support, where one of the notes is informative and the other is non-informative regarding pain. Informative notes:

CCLS met with Pt and family to provide continuity of child life care. She here for unexpected pain crisis. She quiet during interaction, but coping well overall with unexpected admission. CCLS provided several teen items and developmental activities at bedside to normalize the hospital experience, support development, and facilitate coping. CCLS encouraged her to take walks around the unit and/or walk to the child life playroom once feeling strong and with less pain. Will continue to follow throughout hospitalization for child life care.

Non-informative notes:

CCLS met with pt to provide continuity of child life care. Patient is involved in daily playroom activities as provided by the Child Life staff to normalize the hospital experience, support development and facilitate coping. Will continue to follow for child life care throughout treatment.

After considering both text (135 features) and cTAKES features (5 features), the optimal features and the model considered as cTAKES features and LR model with an F1 score of 0.70 because a simple model with 5 features is giving a good performance. cTAKES model is simple with 5 features as opposed to text features (135) giving an F1 score of 0.82, therefore cTAKES model is chosen as optimal model.

4.2.3 Sentiment Data

After preprocessing the sentiment dataset, the next step is building the model using the four algorithms as described in Methods Chapter 3 under section 3.6. The features used for the study are including count, Term Frequency-Inverse Document Frequency (tf-idf), Part of Speech tags (POS tags), cTAKES features, unigrams and bigrams, as discussed in Methods Chapter under section 3.4. For this dataset, the model is trained to predict the sentiment increase, decrease, neutral or not-determined; these are notes informative regarding pain level. Increase indicates increase in pain, decrease as decrease in pain, neutral as no change in pain, and not-determined as not-determined regarding the pain level, detailed annotation procedure is described in Methods Chapter 3 (under section 3.2). Table 4.15 below demonstrates the results; F1 score is reported for all the classifiers.

Table 4.15: Results without feature selection for different features and algorithms with F1 score.

Features		Feature Size	LR	RF	SVM	MNB
Unigram	count	1855	0.456	0.460	0.439	0.486
	tfidf		0.501	0.484	0.508	0.503
Unigram + POS tags	Count + POS	1890	0.464	0.491	0.439	0.492
	tfidf + POS		0.511	0.493	0.513	0.517
Bigram	count	7281	0.498	0.482	0.474	0.527
	tfidf		0.514	0.447	0.513	**0.515**
Bigram + POS tags	count + POS	7316	0.506	0.469	0.474	**0.524**
	tfidf + POS		0.507	0.492	0.505	0.508
cTAKES	cTAKES	5	**0.400**	0.352	0.385	0.382
cTAKES + Unigram	count	1860	0.464	0.498	0.437	0.490
	tfidf		0.496	0.455	0.491	0.398
cTAKES + Unigram + POS	count + POS	1895	0.323	0.371	0.343	0.343
	tfidf + POS		0.296	0.339	0.318	0.318
cTAKES + Bigram	count	7286	0.521	0.472	0.500	0.518
	tfidf		0.491	0.496	0.509	0.212
cTAKES + Bigram + POS	count + POS	7321	0.358	0.312	0.355	0.355
	tfidf + POS		0.338	0.309	0.405	0.405

From Table 4.15, all the algorithms are performing with an F1 score ranging (0.29 - 0.52), and a simple algorithm like MNB gives better results. MNB with bigram and POS tag features (7316) gives an F1 score of 0.52, however MNB with bigram (7281) features gives an F1 score of 0.51, therefore MNB with bigram tf-idf is chosen for further analysis as there is no significant improvement with the addition of POS tags and a simple model is always considered optimal in comparison to a complex model when the results are not significantly different.

LR model with cTAKES (5) features gives an F1 score of 0.40. Table 4.16 and 4.17 demonstrates the confusion matrix, precision and recall; these metrics are explained in Chapter 3 under section 3.8.

Table 4.16: Confusion matrix of LR with cTAKES features

Prediction/actual	Increase	Decrease	Neutral	Non-determined
Increase	**10**	12	11	10
Decrease	2	**6**	8	9
Neutral	6	15	**21**	20
Not-determined	5	13	26	**73**

Results from Table 4.16 explain that the model is performing well, However the model is having higher misclassification in increase, decrease and neutral class. The results are discussed in section 4.2.3.2.1.2.

Table 4.17: Precision and Recall of MNB with bigram tf-idf as features

Metric	score
Precision	0.46
Recall	0.45

The model is performing with F1 score of 0.51 without feature selection, however feature selection helps in eliminating irrelevant features. Model with too many features leads to overfitting and feature selection may help in increasing the accuracy and performance of the model.

4.2.3.1 Results with Feature Selection

Feature selection is a step to improve the performance of the model and also to reduce the complexity. Feature selection techniques discussed in Methods Chapter 3 under section 3.5 are implemented to select the most informative features uncovering the hidden structures in data and help in building an accurate predictive model. Table 4.18 below demonstrates the precision, recall and F1 score for MNB with bigram and tf-idf as features for two feature selection techniques.

Table 4.18: Feature selection results for different features selection algorithms with F1 score, precision, and recall

	Feature selection	Features	Precision	Recall	F1 score
MNB Bigram tf-idf (features)	χ^2	**125**	0.616	0.540	0.547
	Mutual class info	250	0.498	0.402	0.414

It can be observed that χ^2 technique (discussed in section 3.5.5.1) yields better results compared to mutual information, and 125 features are finally selected for the model. The table below demonstrated the confusion matrix with 125 features and MNB, bigram and tf-idf as features.

Table 4.19: Confusion matrix of MNB with bigram tf-idf as features for χ^2 feature selection

Prediction/actual	Increase	Decrease	Neutral	Non-determined
Increase	**17**	7	10	9
Decrease	3	**9**	6	7
Neutral	10	7	**33**	12
Not-determined	22	12	9	**74**

Results are discussed in the flowing section:

4.2.3.2 Discussion

Discussion section is divided into different sub sections including feature analysis to analyze different features under study, algorithm analysis to analyse best performing algorithm, feature selection, features selected, themes identification and themes clustering for text features.

4.2.3.2.1 Feature Analysis

In this section results of features including text, cTAKES and their combination are discussed in different sub-sections as follows.

4.2.3.2.1.1 Text features: unigram, bigram, POS tags

Unigrams and bigrams yield results with F1 score ranging between 0.45 and 0.51. Bigrams are chosen as the best performing features for sentiment than unigrams, as they tend to capture modified verbs and nouns [67].

The combination of POS tags, unigrams and bigrams have F1 score ranging from 0.43 to 0.52. It can be observed that addition of POS tags (33 features) have so significant effect on the performance of the model. This can be explained by the the information captured by POS tags is obtained from unigram and bigrams and hence no significant change.

4.2.3.2.1.2 cTAKES features

The LR model is performing optimal with F1 score of 0.40 with cTAKES features (5). However it can be observed that there are more misclassifications in the class increase, decrease and neutral. This can be explained because cTAKES couldnt capture certain medication and the model couldnt attribute the decrease in pain was due to medication, and it is hard to identify the increase, decrease, neutral sentiment because it is temporal and depends on previous state of the patient (previous note). An example note demonstrates cTAKES ability to capture medication.

Below is a sample note:

Patient is a 16 year old with Sickle Cell Disease type SS and Allergies who presented with pain and oxygen desaturations admitted with vaso-occlusive pain crisis.Patient is afebrile. CXR negative. Patient is receiving IV ***Dilaudid*** *PCA/continuous and scheduled IV Toradol for pain management.Patient continues IV fluids. Patient continues HU. Follow up scheduled with PHE clinic @ 9:30 am on 4/27. Anticipate discharge home when transitioned to po pain medications and pain managed with po meds. No discharge needs noted. Will follow.*

cTAKES can capture the terms underlined in the notes, however its unable to capture Dilaudid, IV fluids as medication as it gives information regarding dehydration.

Misclassifications can also be explained due to unreliable annotations as the pain scores are determined for each notes from the vitals with a 2-3 hour window. It is difficult to identify if there is increase, decrease or constant pain from the data provided in notes. Also incorrect tagging by cTAKES, missing tags, multiple tagging for same phrase can be contributing to incorrect increase, decrease prediction, as it becomes really hard for the machine to predict. For instance the notes below is annotated as decrease but the model predicted it as not-determined.

Patient is a 16 year old with Sickle Cell Disease type SS who presented with pain admitted with sickle cell pain crisis. Hgb is 6.4 g/dL. Patient continues HU. Patient is receiving oxygen @ 31% FiO2 to maintain O2 sats ¿90%. Patient is receiving IV fluids. Patient has poor po due to nausea.Patient is constipated and receiving Lactulose. Patient is receiving ***Dilaudid*** *PCA/continuous and scheduled IV Toradol for pain management. Laurie Howlett, LCSW following for psychosocial needs. Anticipate discharge when oxygen saturations ¿90% on room air, pain managed with po pain medications and constipation resolved.Follow up scheduled with PHE clinic @ 10 am on 6/24. Will follow for discharge needs.*

Met with grandmother at bedside and discussed PRM role. NC. Patient attends school and is in the 10th grade. Grandmother will transport home at discharge.Patient is not followed by any community resources. Patient has NC Medicaid to assist with medical expenses.

Dilaudid is a medication but cTAKES could not tag it to medication, and from the text its hard to predict a decrease, as same phrases sickle cell disease, pain management are mentioned in all texts, making it hard for the model to make accurate prediction and causing misclassifications in the classes.

4.2.3.2.1.2 Text and cTAKES features

The combination of text and cTAKES features have F1 score ranging from 0.30 to 0.52, however text and cTAKES features individually obtain F1 score of 0.51 and 0.40. Therefore the combination of cTAKES to text features does not reflect significant improvement in the performance. This can be because cTAKES feature does not provide additional information to text features, hence no significant performance gain.

4.2.3.2.2 Algorithm Analysis

After identifying the best performing features, the next step is to analyze the models. Generative models (MNB) tend to perform well, over discriminative models (LR, SVM, RF) for small datasets, as proved in a study by Ng et. al. [19], therefore MNB is chosen as the best performing algorithm. The low performance can be attributed to less training data, a study by Wembo et al. [68] found that as the training data increases, performance of the model increases in identification of multiclass problem. Sentiment in healthcare is a difficult problem, and achieving an F1 score of 0.54 is state of the art in healthcare, a study by Huh et al. [69] reported an F1 score of 0.54 for text classification in healthcare.

4.2.3.2.3 Feature selection analysis

After selecting the best performing model, optimal features are selected using χ^2 and the results are analyzed from confusion matrix. There are very few examples in decrease class, therefore it is difficult to capture the meaning and predict it accurately, however the model has higher true Positive than true negative individually for each class.

One of the reasons for the F1 score to be too low, is because the sentiment (increase, decrease, neutral, not-determined) is derived from pain score in vitals, and not from notes, it is very difficult manually to identify an increase, decrease, neutral sentiment from the notes alone, however the model performed well, when manually its quite hard to annotate the labels from text alone. The other reason being, the sentiment is labeled based on previous situation, its a temporal series, and time dependent, whereas our model is inde-

pendent of time and does not take into account the previous notes and considers each note as independent of each other.

Since the sentiment could not be derived from notes alone, pain score from vitals was used to annotate the data, but there are inherent issues, the pain score was noted in a two hour window and may not be accurate. As suggested by the model, the annotation could be incorrect and it can be observed from the example notes below.

Below is an example notes annotated as not-determined:

CCLS met with Pt to provide continuity of child life care. She ***coping well at this time*** *AEB engaging easily with CCLS throughout interaction. She is excited to be D/C today. No other needs noted, will continue to follow throughout treatment for child life care.*

The phrase coping well at this time indicate the patient is doing well, and the sentiment is either decrease or neutral. However the model predicted as neutral, which is nearly accurate, as it can be annotated as either decrease or neutral based on the previous notes condition.

There are many such examples that might have caused misclassification errors, and not-determined notes can be mapped to increase, decrease or neutral, leading to high number of misclassification in not-determined class.

Some of the features in the data such as much better, complain worsen, difficulty breath, patient discharge, patient admit, began complain, worsen sharp, sharp lower, lower chest, reflect sentiment (increase, decrease, neutral).

In multiclass classification, one vs all strategy is implemented to classify the data points to the respective classes, and from the features selected (discussed in section 4.2.3.2.4) and the clusters identified (discussed in section 4.2.3.2.5 and 4.2.3.2.6) it is really hard for the classifier to distinguish between the two classes as there is overlap in them, and perhaps incorrect annotation contributes to higher precision (0.58) than recall (0.51). Higher precision and lower recall indicates that there are more false negatives than false Positives, and incorrect annotation and overlap in classes makes it difficult to club all the classes

into one class, for instance, considering the classes increase vs rest (decrease, neutral, not-determined), the model is performing better in classifying the increase class over the rest, however, it is unable to classify correctly the rest class because there is fine line between the increase and rest (decrease, neutral and not-determined) from the text which makes it harder for the classification and obtaining more false negatives. An example text below shows an example of incorrect classification and subtle differences between the class increase and decrease. The text is annotated as increase, however the model predicted it as decrease.

Patient continues to have pain *and is on a 50 mcg demand, with basal MS-Contin 30mg TID. He used 39 pushes overnight with 28 delivered doses. He indicates that* ***this dose has been managing his pain well****. Team plans to d/c the PA and start PO dilaudid and Ms Contin. If patient does well off PCA, patient will be close to discharge. Anticipate no needs at time of discharge.*

The note mentions patient continues of have pain indicating that pain is increasing/neutral and on the other side it mentions this dose has been managing his pain well indicating that the pain is decreasing and managed with medication.

4.2.3.2.4 Features selected

Below are the 125 features selected by χ^2 feature selection technique, in **decreasing** order of importance, the top features are selected based on the F1 score of the model during cross validation, and the models are trained using these features. 125 Features in decreasing order of importance is tabulated in table 4.20 below. Features are discussed in section 4.2.3.2.5.

Table 4.20: Features selected in decreasing order of importance with χ^2 feature selection

Top features	Features in decreasing order of importance
1 - 30	'child life', 'clinic social', 'license clinic', 'certified child', 'life specialist', 'social worker', 'extend hospitalization', 'hospital reaction', 'take bus', 'team page', 'worker provided', 'playroom participation', 'foster mother', 'play playroom', 'specialist provided', 'child psychology', 'admission extend', 'patient discharge', 'orient unit', 'patient admit', 'deny feel','medicaid transport', 'transport paperwork', 'room unit', 'alert came', 'answer poc' (point of care), 'bedside family', 'came bedside', 'fall implement', 'floor vss' (floor vital signs stable), 'md alert' (doctor of medicine alerted), 'unit question', 'inpatient frequent', 'age foster', 'bus raleigh', 'care system', 'foster care',
30-60	'provided patient', 'hospital need', 'state therapeutic', 'teen program', 'accompani mom', 'former foster', 'support involve', 'clinic patient', 'psych consult', 'development play', 'patient transfer', 'bad dream', 'locat child', 'active new', 'dialogue play', 'engage dialogue', 'intern development', 'movie provided', 'new movie', 'play child', 'photo id', 'pain md' (pain medicine of doctor),
60-90	provided development', 'need anticipate', 'outpatient counsel', 'parent patient', 'specialist met', 'state miss', 'visit day', 'assess introduction', 'patient arrive', 'appear uncomfortable', 'medical order', 'need vital', 'otherwise stable', 'page pain', 'page regard', 'patient sat', 'ra team' (ra- respiration), 'regard need', 'sat ra' (ra-respiration), 'stable continue', 'uncomfortable team', 'vital otherwise', 'high school', 'pain fall', 'crisis handling', 'discharge home', 'life care', 'ongoing support', 'acute distress',
90-120	'family orient', 'began complain', 'bp elevate' (blood pressure), 'breath also', 'came assess', 'complain worsen', 'difficulty breath', 'elevate patient', 'ivf addition', 'lindsay mize', 'lower chest', 'mize np', 'monitor ivf', 'morphin infuse', 'np came' (np - nurse practitioner), 'pain lindsay', 'patient began', 'ra bp', 'room well', 'sent vdh', 'sharp lower', 'state difficulty', 'sun patient', 'vdh continue', 'well dr', 'worsen sharp', 'administration patient/-family', 'initial ped', 'pack administration', 'pain ivf (intravenous fluids)', 'ped pain' (pediatric), 'patient/family orient',
120-25	'roomunit poc', 'start heat', 'unit wheelchair', 'wheelchair patient', 'orient room'

4.2.3.2.5 Themes identification

Based on the features selected, we have tried to identify themes in our dataset using the features. Clustering similar topics helps in providing insights into the data, and in our problem

it helps in segregating the themes identified into specific classes, thus helping in identification of increase, decrease, neutral and not-determined notes. Identifying themes helps in analyzing the effectiveness of a treatment, medications prescribed at different levels of pain, analyze the causes of change in pain level.

Themes identified in the notes:

1. **Child life support notes:** child life, extended hospitalization, This note helps in analyzing the patient status, effectiveness of treatment. Below is an example notes with the above keywords **(extended hospitalization)** identified:

   Medical Stressors: ***Extended hospitalization*** *Reaction to Hospitalization/Illness: Coping well with hospitalization, though pt states she is still having pain issues Therapeutic Interventions: Emotional support, Parental support, Developmental play, Playroom participation encouraged to increase ambulation and promote Positive coping during hospital admission Pt participated in afternoon playroom session to normalize the hospital experience, support development and facilitate coping Outcome: Will follow throughout hospital admission for child life care*

2. **Family background:** medicaid transport, transport paperwork, deny feel (denied any feelings of depression and sadness from text) This note helps in analyzing the family background of the patient, and identifying ways to prevent pain crisis, or alleviate pain based on their family situation and gain understanding about patient. Below is an example notes with the keywords **(medicaid transport, transport paperwork)** identified:

   Assessment: Met with Patient and his legal guardian/grandmother in his inpatient room on floor 5100. Patient stated that he was feeling much better and that he was able to walk around the floor today. Grandmother stated that they anticipate discharge tomorrow. She asked this CSW to assist with her medicaid transportation

paperwork to obtain gas vouchers. She stated that they are doing okay inpatient and Patient was using his cell phone to cope and as a distraction during his inpatient stay.

Plan/Intervention: Provided emotional support and counseling. Explored their coping with this unexpected hospitalization. Assisted with completion of ***Medicaid transportation paperwork****. Family has parking pass for discharge. This CSW will continue to follow this patient and their family as needed, assessing their coping and adjustment and providing emotional and practical support and coordinating closely with the PHO team.*

3. **Discharge notes:** patient discharge, This note helps in analyzing the medications given to patient at discharge, how well they worked if the patient is readmitted the next time. Therefore it helps in assessing pain sentiment with the medications. Below is an example notes with the keywords **(patient discharge)** identified:

   Patient transitioned to po Dilaudid every 4 hours and prn Oxycodone.Patient is afebrile.Blood and urine cultures are negative. IV antibiotics and IV fluids discontinued.Patient has infusaport central line which will not be accessed at discharge therefore will not require referral to home infusion/home health. Follow up scheduled with PHE clinic @ 11:30 am on 2/20/2014 and 10 am on 4/4/2014. ***Patient discharged*** *home. Will follow for discharge needs.*

4. **Admit notes:** patient admit, room unit, patient transfer, pain fall This note helps in analyzing the pain level of patient at admit, and the medications given to patient and their effectiveness overall. Below is an example notes with the keywords **(patient adit)** identified:

   Patient admitted *to 5119, arrived via transport accompanied by mom. Patient c/o 7/10, vs stable. PCA pump confirmed with second RN.*

Below is an example notes with the keywords **(room unit)** identified:

*Pt admitted to unit from ED, afebrile and in stable condition, with father at bedside (no photo ID available at this time). PIV infusing hydromorphone 1:1 PCA and NS @30 upon arrival. Oriented pt and father to **room** and **unit**, conducted admission interview and assessment, and reviewed POC/pain mgmt plan. All questions were answered at this time.*

4.2.3.2.6 Themes clustering

Further these themes identified are clustered into informative and non-informative categories; table 27 below demonstrates the clusters. Themes clustering helps in identifying the treatments, procedure, medications that cause decrease, increase or maintained pain during hospitalization. From the clustering in table 27, discharge notes are mostly identified with decrease, neutral or not-determined pain, because a patient is discharged if the pain is alleviated; admit notes are mostly tagged to increase or not-determined pain level, because during admission, patients pain level is increasing or not-determined (due to missing pain score reporting); child life support, family background are tagged with increase, decrease, or not-determined as they are taken during the course of hospitalization and the patients pain level is fluctuates under observation.

Table 4.21: Themes categorization into increase, decrease, neutral and non-determined categories.

Increase and Not-determined	Neutral and Not-determined	All
Admit notes	Discharge notes	Child life support notes Family background

From table 4.21 it can be observed that there are no themes that can be particularly tagged to increase, decrease, neutral or not-determined class.

After considering both text (125 features) and cTAKES features (5 features), the optimal features and the model considered as cTAKES features and LR model with an F1 score

of 0.40 because a simple model with 5 features is giving a good performance. cTAKES model is simple with 5 features as opposed to text features (125) giving an F1 score of 0.54, therefore cTAKES model is chosen as optimal model.

4.2.4 Pain Score dataset

After preprocessing the pain score dataset, predictive models are built using the four algorithms as described in Methods Chapter 3 under section 3.6. The features used for the study are including count, Term Frequency-Inverse Document Frequency (tf-idf), Part of Speech tags (POS tags), cTAKES features, unigrams and bigrams, as discussed in Methods Chapter 3 under section 3.4. For this dataset, the model is trained to predict the pain score level as moderate, severe and not-determined, which are notes informative regarding pain level, moderate indicates moderate pain level, severe indicates severe pain level and not-determined as not-determined regarding the pain level, detailed annotation procedure is described in Methods Chapter 3 under section 3.2. Table 4.22 below demonstrates the results; F1 score is reported for all the classifiers.

Table 4.22: Results without feature selection for different features and algorithms with F1 score.

Features		Feature Size	LR	RF	SVM	MNB
Unigram	count	1850	0.455	0.497	0.447	0.495
	tfidf		0.471	0.481	0.448	**0.498**
Unigram + POS tags	Count + POS	1890	0.289	0.295	0.338	0.338
	tfidf + POS		0.384	0.281	0.318	0.318
Bigram	count	7311	0.440	0.473	0.434	0.466
	tfidf		0.483	0.454	0.467	0477.
Bigram + POS tags	count + POS	11346	0.415	0.361	0.387	0.387
	tfidf + POS		0.331	0.345	0.271	0.271
cTAKES	cTAKES	5	0.377	0.416	**0.421**	0.362
cTAKES + Unigram	count	1856	0.481	0.498	0.459	0.476
	tfidf		0.493	0.491	0.483	0.441
cTAKES + Unigram + POS	count + POS	1890	0.289	0.295	0.338	0.212
	tfidf + POS		0.384	0.281	0.318	0.330
cTAKES + Bigram	count	7276	**0.522**	0.522	0.503	0.507
	tfidf		0.507	0.483	0.467	0.496
cTAKES + Bigram + POS	count + POS	7311	0.415	0.361	0.387	0.334
	tfidf + POS		0.331	0.345	0.271	0.288

From table 4.22, algorithms with different features are performing with an F1 score ranging from 0.21 to 0.52. From the results, a simple algorithm like MNB gives better results, hence this classifier is chosen for further analysis and the features selected are unigram and tf-idf. RF with 0.52 F1 score is not chosen over MNB because the feature size os huge compared to MNB.

With cTAKES features SVM performs with an F1 score of 0.42. Table 4.23 and 4.24 below demonstrates the confusion matrix, precision and recall; these metrics are explained in Chapter 3 under section 3.8.

Table 4.23: Confusion matrix for SVM with cTAKES features

Prediction/actual	Moderate	Severe	Not-determined
Moderate	**29**	27	29
Severe	23	**45**	40
Not-determined	5	9	**40**

The model with SVM and cTAKES features is performing with an F1 score of 0.42. Diagonals are higher for all the classes, further results are explained in section 4.2.4.2.1.2.

Table 4.24: Precision and Recall of RF with cTAKES feature

Metric	score
Precision	0.46
Recall	0.45

The model is performing with F1 score of 0.51 without feature selection, however feature selection helps in eliminating irrelevant features and may help in increasing the accuracy and performance of the model.

4.2.4.1 Results with Feature Selection

Feature selection is a step to improve the performance of the model and also to reduce the complexity. Feature selection techniques discussed in Methods Chapter 3 under section 3.5 are implemented to select the most informative features uncovering the hidden structures in data and help in building an accurate predictive model. Table 4.25 below demonstrates the precision, recall and F1 score for MNB with bigram and tf-idf as features for two feature selection techniques.

Table 4.25: Feature selection results for different features selection algorithms with F1 score, precision, and recall

	Feature selection	Features	Precision	Recall	F1 score
MNB Unigram tf-idf (features)	χ^2	**135**	0.609	0.608	0.582
	Mutual class info	30	0.558	0.527	0.514

It can be observed that χ^2 technique (discussed in section 3.5.5.1) yields better results compared to mutual information, selecting 135 features. Table 4.26 below demonstrates the confusion matrix with 135 bigram tf-idf features using MNB algorithm.

Table 4.26: Confusion matrix for MNB with bigram-tfidf features and χ^2 feature selection

Prediction/actual	Moderate	Severe	Not-determined
Moderate	**42**	32	11
Severe	31	**72**	5
Not-determined	12	6	**36**

Results are discussed in the flowing section.

4.2.4.2 Discussion

Discussion section is divided into different sub sections including feature analysis to analyze different features under study, algorithm analysis to analyse best performing algorithm, and feature selection, features selected, themes identification and themes clustering for text features.

4.2.4.2.1 Feature Analysis

In this section text, cTAKES features and their combination are discussed in different sub-sections.

4.2.4.2.1.1 Text features: unigram, bigram, POS tags

Unigrams and bigrams for MNB obtains F1 score ranging from 0.44 to 0.49, and unigrams with tf-idf have highest F1 score of 0.49 and chosen as the optimal features. Unigrams, bigrams with POS tags (33 features) have F1 score ranging from 0.27 to 0.49, and unigrams with the best algorithm have F1 score of 0.49, therefore the combination models are not further analysed. The results of combination is not improving perhaps because the information provided by POS tags is obtained from unigrams and bigrams and irrelevant features may decrease the performance.

4.2.4.2.1.2 cTAKES features

Using cTAKES features, SVM performed well with an F1 score of 0.42. There are misclassifications in the class moderate and severe and this can be explained as follows, cTAKES features are extracted under named entities including procedure, medication, sign/symptom, disease disorder, and anatomical. However there is no way to identify which medication is used for what purpose, some medications are used when the pain is severe but some are used when the pain is moderate, and there is no way to identify the difference, as a simple count of medications identified from notes are used as features.

Below is an example note, which is labeled as severe, but the model predicted it as moderate.

Patient is a 15 year old with Sickle Cell Disease type SS on Hydroxyurea, Asthma, s/p Splenectomy 1999 who presented with pain crisis. Hgb is 7.1 g/dL and continues Hydroxyurea. Patient is a Jehovah's witness. Blood Conservation team consulted and recommendations for blood conservation noted. Patient continues Pen VK for splenectomy prophylaxis. Patient continues Losartan for nephropathy. Patient continues Dilaudid PCA/continuous infusion, scheduled IV Toradol every 6 hours and Lidoderm patches to thighs at night for pain management. Patient continues Atroven BID and prn Albuterol. Patient is receiving IV fluids. Follow up scheduled with PHE clinic @ 11 am on 7/9/2014

and PPL clinic @ 3:30 pm on 8/25/2014.Anticipate discharge home when pain can be managed with po pain medications. Dad at bedside and agrees with plan of care. Will follow for discharge needs.

From the notes the medications used are Hydroxyurea, Pen VK, Dilaudid, Lidoderm, and a simple count is registered, however these medications provide insights into the level of pain moderate and severe. Lidoderm helps to reduce sharp/burning/aching pain as well as discomfort caused by skin areas that are overly sensitive to touch. This gives some information that the patient is going through severe pain, however such details are not captured with cTAKES.

Misclassifications in Moderate and Severe class can be due to subtle differences between them, and its hard to make out from the text. For instance the note below is annotated as moderate however the model predicted it as severe.

Pain improving *and transitioned to po scheduled MS Contin, scheduled Ibuprofen and as needed Oxycodone and. Patient continues Hydroxurea. Patient is taking adequate po fluids and IV fluids discontinued.Follow up scheduled with PHE clinic @ 1 pm on 7/1. Patient discharged home. Will follow for discharge needs.*

4.2.4.2.1.3 Text and cTAKES features

Unigram, bigram, POS tags (33 features) and cTAKES have F1 score ranging from 0.27 to 0.49, however unigram alone is obtaining F1 score of 0.49. Therefore addition of features is not improving the performance, because the information captured by combination of features is captured alone by unigrams.

4.2.4.2.2 Algorithm Analysis

After identifying the features, the next step is selecting accurate model for prediction. With different features algorithms perform with an F1 score ranging from 0.28 to 0.49. MNB is chosen as the optimal algorithm with F1 score of 0.49 as it is evident from the study that generative models are better for smaller dataset [19]. Also low F1 score can be attributed

to less training data, and as the training data increases performance might increase [68].

4.2.4.2.3 Feature selection analysis

After identifying the algorithm, the next step is features selection and identifying the top features for classification. The MNB model with unigram tf-idf as features obtains F1 score of 0.58. Classifying pain level is a difficult problem, and such low F1 scores are state of the art in text classification and healthcare, numerous studies reported such low F1 score [68][69]. By further analyzing the results from the confusion matrix, it can be noticed that there are more misclassification in the moderate and severe classes. The severe class instances are often misclassified as moderate and vice versa. After analysis, it was identified that certain texts with the keyword admit (one of the features) is either moderate or severe and it is really hard to determine if it is severe or moderate even for a human when the pain scores are hidden. As discussed before the texts are annotated as moderate, severe and not determined based on the pain score from the vitals and they are not labeled from the text. Therefore for some texts, it is hard even for a human to label as moderate or severe based on text. An example below demonstrates a text predicted as severe when it is labeled as moderate.

Below is an example text labeled as moderate:

Patient admitted from Duke E.D. Patient received to unit via stretcher on room air, alert and verbal, patient rates pain as 4/10 numbers pain scale to lower back, 02 sat 98% in no apparent distress, skin warm and dry to touch, able to voice needs, IVF infusing without difficulty with PCA: Morphine 5 mg/1mg. Oriented family and patient to room.

From the example notes above it is very difficult to identify the level of pain when the numbers are eliminated in preprocessing step. Therefore its obvious for a machine to misclassify when its hard to classify it manually.

The low F1 score can be explained by the same reasoning. It is hard to classify the

notes based on text, as they are not so rich in explaining the level of pain.

4.2.4.2.4 Features selected

Below are the 135 features selected by χ^2 feature selection technique, in **decreasing** order of importance, the top features are selected based on the F1 score of the model during cross validation, and the models are trained using these features. 135 Features in decreasing order of importance are tabulated in table 4.27. Features are discussed in section 4.2.4.2.5.

Table 4.27: Features selected in decreasing order of importance with χ^2 feature selection

Top features	Features in decreasing order of importance
1 - 30	discharge', 'instruct', 'id', 'oxycodone', 'photo', 'ibuprofen', 'session', 'po' (oral in medicine), 'schedule', 'home', 'arrive', 'verify', 'phe', 'por', 'resolve', 'prescript', 'adequate', 'ed' (emergency department), 'pca', 'transitioned', 'sat', 'ssx' (s/sx - signs and symptoms), 'catheter', 'excited', 'patient', 'prn' (pro re nata - when necessary), 'distress', 'deny', '
30-60	'assist', 'unit', 'infuse', 'orient', 'assess', 'remove', 'restart', 'assume', 'colescott', 'stretcher', 'ambulatory', 'destination', 'ra' (respiration), 'step', 'complain', 'dss' (department of social service), 'foster', 'hailey', 'admit', 'receive', 'liter', 'cell', 'monitor', 'distract', 'inform', 'appoint','guardian/grandmother', 'outcome', 'go',
60-90	'admission', 'oxycodone/follow', 'nausea', 'grandmother', 'usual', 'bedside', 'resource', 'alert', 'member', 'condit', 'sickle', 'awake', 'chest', 'constipation', 'need', 'sp' (status post), 'implement', 'migraine', 'topamax', 'moodi', 'yulonda', 'rib', 'abdomen', 'normalization/socialize', 'breathe', 'ac' (left ac piv),
90-120	'bed', 'dad', 'ivf (intravenous fluids)', 'take', 'culture', 'burlington', 'dr' (doctor), 'significance', 'right', 'pediatrician', 'zofran', 'oxygen', 'continuous/pca', 'phone', 'co' (complain), 'better', 'happily', 'mom', 'vdh' (room), 'hospitalizationilation', 'introduction', 'stressor', 'therapeut', 'saturation', 'mivf' (maintenance intravenous fluid), 'year', 'score', 'air', 'continue', 'opportunity', 'locate', 'playroom', 'inpatient', 'team', 'urine',
120-135	'attempt', 'high', 'transport', 'license', 'poor', 'service', 'intact', 'carolina', 'management/child', 'outreach', 'pain/patient', 'refuse', 'secondary', 'complaint', 'date'

4.2.4.2.5 Themes identification

Based on the features selected, we have tried to identify themes in our dataset using the features. Clustering similar topics helps in providing insights into the data, and in our problem it helps in segregating the themes identified into specific classes, thus helps in identifying notes of moderate, severe and not-determined class. Themes help in evaluating the effectiveness of procedures, medication, treatment of pain. Themes in Pain Score:

1. **Discharge notes:** discharge, po, oxycodone, ibuprofen, oral, mom This note helps in analyzing the medications given and the pain level at discharge. Below is an example notes with the keywords **(po, oxycodone, mom)** identified:

   *Pt discharged home with **mom**. Pt afebrile, VSS, pain controlled with PRN **PO oxycodone**/scheduled pain meds. Pt given oxycodone prior to d/c for pain 2/10 in chest. Mom and pt educated on d/c instructions. Parking pass given via nurse as couldn't get in touch with social work. PIV removed. Care plan/education resolved. Mom and pt discharged without any further questions/concerns.*

2. **Assessment notes:** oxycodone, ibuprofen This theme helps in assessing the pain of patients, effectiveness of medications and treatments in pain control. Below is an example notes with the keywords **(oxycodone, ibuprofen)** identified:

   *Pain managed with scheduled po Kadian, **Ibuprofen** as needed and **Oxycodone** as needed for pain management. Oupatient MRI scheduled@ 8:15 am and labs at CHC on 8/25.Follow up scheduled with PHE clinic @ 8 am on 8/26. Will follow for discharge needs.*

3. **Patient resource management and discharge plan:** oxycodone, chest, resource This theme provides patient's medical background and history, and thus helping in assessing the causes of pain crises from medical history. Below is an example notes with the keywords **(resource)** identified:

Patient is a 16 year old with Sickle Cell Disease type SS who presented with pain admitted with sickle cell pain crisis. Hgb is 6.4 g/dL. Patient continues HU. Patient is receiving oxygen @ 31% FiO2 to maintain O2 sats ¿90%. Patient is receiving IV fluids. Patient has poor po due to nausea.Patient is constipated and receiving Lactulose. Patient is receiving Dilaudid PCA/continuous and scheduled IV Toradol for pain management. Laurie Howlett, LCSW following for psychosocial needs. Anticipate discharge when oxygen saturations ¿90% on room air, pain managed with po pain medications and constipation resolved.Follow up scheduled with PHE clinic @ 10 am on 6/24. Will follow for discharge needs.

Weight: 57.5 kg Height: 170.5 cm Allergies: NKDA

Met with grandmother at bedside and discussed PRM role. NC. Patient attends school and is in the 10th grade. Grandmother will transport home at discharge.Patient is not followed by any community ***resources****. Patient has NC Medicaid to assist with medical expenses.*

4. **Admit notes:** abdomen, chest, arrive It helps is analyzing the level of pain of various patients and procedures taken during those times and evaluate the effectiveness of different procedures and medications. Below is an example notes with the keywords **(abdomen)** identified:

   Received pt from ER to 5110 w/ Hx Sickle Cell here w/ abdominal pain. ***Abdomen*** *distended but soft, currently rating pain at 6/10. Parents at bedside and oriented to unit and routine. Admission database completed. Placed on cardiac monitor and pulse oximetry. Continue to monitor.*

5. **Family background:** grandmother This theme helps in assessing family background of patient and devise plans based on it, and also understand causes of triggering of

pain crises. Below is an example notes with the keywords **(grandmother)** identified:

Visited with Pt in his inpatient room on floor 5100. He was out of bed today and brushing his teeth. He stated that he was still in a lot of pain today but he did appear better than yesterday. Today, we readdressed his feelings of depression and he stated that he felt better today and he denied any suicidal ideation. He confirmed that he met with Duke inpatient psych today and he has the 1800 crisis number from this CSW for outpatient. We readdressed his concern about his former foster mother Cassandra Craddock and his frustration that he has called her a bunch and she has not called back. With his consent, this CSW agreed to contact her and this CSW called her but she did not answer. She is not his foster mother anymore as he is no longer in DSS care, so she does not have to call him back, despite his desire for support from her. Discussed his support system and he named his former foster mother, his friends, his LGBT Raleigh community, his ***grandmother*** *at times, and the Hope Center in Pullen Park for foster children who have aged out of the system. Pt prayed for his health today. He shared that he is struggling with budgeting and paying rent in his apartment in Raleigh with his roommates. Pt stated that he has "many challenges in his life including having sickle cell disease, being gay, and just aging out of the foster care system."*

4.2.4.2.6 Themes clustering

Themes identified are clustered into severe, moderate and not-determined categories; table 34 below demonstrates the clusters. Themes clustering help in identifying the effectiveness of a treatment, therapy, medications as the pain level changes from severe to moderate or causes of pain change from moderate to severe. From table 4.28, admit and patient resource notes have severe or moderate pain as these notes are taken during admission and the pain level is high; discharge, assessment and family background have a combination as they are taken during hospitalization and under observation pain level fluctuates.

Table 4.28: Themes categorization into severe, moderate and not-determined categories

Severe and Moderate	Moderate, Severe, Not-determined
Admit notes	Discharge
Patient resource	Assessment
	Family background

It can be observed that there are no themes that can be specifically tagged to severe or moderate or not-determined classes.

After considering both text (135 features) and cTAKES features (5 features), the optimal features and the model considered as cTAKES features and SVM model with an F1 score of 0.42 because a simple model with 5 features is giving a good performance. cTAKES model is simple with 5 features as opposed to text features (135) giving an F1 score of 0.58, therefore cTAKES model is chosen as optimal model.

From the analysis made for all the different problems namely patient, pain, sentiment, and pain score, different models and features are considered as optimal. For patient informative and non-informative, cTAKES features and RF model with F1 score of 0.86; for the pain informative and non-informative, cTAKES features and RF model with F1 score of 0.70; for sentiment increase, decrease, neutral, and not-determined, cTAKES features and LR model with F1 score of 0.40; and pain score dataset moderate, severe and not-determined classes, cTAKES features and SVM model with F1 score of 0.42 are considered optimal. Therefore from the study cTAKES features looks promising with SCD data.

5 Conclusion

SCD is a hereditary blood cells disorder that can lead to excruciating pain crisis; number of studies are making effort to find ways to identify crisis before its occurrence to mediate it by taking precautions. Natural Language Processing and Machine Learning has been used in myriad number of studies including healthcare to predict fall risk, classify smoking status, as discussed in chapter 2 under section 2.2. There are studies on SCD using Machine Learning as discussed in chapter 2 under section 2.3, however, there are no studies on text and SCD. Our study used Natural Language Processing, Machine Learning and text to build predictive model in analyzing SCD. We are dealing with prediction for four scenarios namely patient informative, pain informative, pain sentiment and pain score. For this study, we used Natural Language Processing tool, cTAKES, Machine learning models and Natural Language Processing to build predictive models and identify themes in data. Furthermore, analyzed results and gained insights into the data. It can be noted that cTAKES features shows promising results and a scope in improving these results further. For future work, we could incorporate ways to improve cTAKES results and annotation process for sentiment and pain score; also incorporate vitals signs as features in prediction; maintaining pain scores for each notes can help with incorrect annotations; building time series models for the prediction of sentiment and pain score as they are dependent on the previous state of the patient can further improve the results.

Bibliography

[1] https://cnx.org/contents/FPtK1zmh@6.27:zMTtFGyH@4/Introduction

[2] https://www.ssa.gov/OP_Home/rulings/di/01/SSR2017-03-di-01.html

[3] https://www.webmd.com/a-to-z-guides/symptoms-of-sickle-cell-disease#1

[4] https://kidshealth.org/en/teens/sickle-cell-anemia.html

[5] https://www.cdc.gov/ncbddd/sicklecell/data.html

[6] https://www.webmd.com/a-to-z-guides/what-is-sickle-cell-disease

[7] https://www.webmd.com/a-to-z-guides/sickle-cell-crisis#1

[8] https://www.medicinenet.com/sickle_cell/article.htm#how_is_sickle_cell_anemia_diagnosed

[9] https://www.nhlbi.nih.gov/health-topics/sickle-cell-disease

[10] https://www.healthline.com/health/sickle-cell-anemia

[11] https://www.cdc.gov/ncbddd/sicklecell/facts.html

[12] https://www.cdc.gov/ncbddd/sicklecell/treatments.html

[13] Jonassaint, Charles R., et al. "Usability and feasibility of an mHealth intervention for monitoring and managing pain symptoms in sickle cell disease: the sickle cell disease mobile application to record symptoms via technology (SMART)." Hemoglobin 39.3 (2015): 162-168.

[14] https://www.practicefusion.com/nursing-notes

[15] Giansanti, Daniele, Velio Macellari, and Giovanni Maccioni. "New neural network classifier of fall-risk based on the Mahalanobis distance and kinematic parameters assessed by a wearable device." Physiological measurement 29.3 (2008): N11.

[16] https://en.wikipedia.org/wiki/Sensitivity_and_specificity

[17] Fawcett, Tom (2006); An introduction to ROC analysis, Pattern Recognition Letters, 27, 861874

[18] Lester, Jonathan, et al. "A hybrid discriminative/generative approach for modeling human activities." (2005): 766-722.

[19] Ng, Andrew Y., and Michael I. Jordan. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." Advances in neural information processing systems. 2002.

[20] https://www.mathworks.com/help/stats/hidden-markov-models-hmm.html

[21] Dawadi, Prafulla N., et al. "Automated assessment of cognitive health using smart home technologies." Technology and health care 21.4 (2013): 323-343.

[22] Alshurafa, Nabil, et al. "Recognition of nutrition intake using time-frequency decomposition in a wearable necklace using a piezoelectric sensor." IEEE Sensors Journal 15.7 (2015): 3909-3916.

[23] Suutala J., Pirttikangas S., Rning J. (2007) Discriminative Temporal Smoothing for Activity Recognition from Wearable Sensors. In: Ichikawa H., Cho WD., Satoh I., Youn H.Y. (eds) Ubiquitous Computing Systems. UCS 2007. Lecture Notes in Computer Science, vol 4836. Springer, Berlin, Heidelberg

[24] Morris, Dan, et al. ”RecoFit: using a wearable sensor to find, recognize, and count repetitive exercises.” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2014.

[25] Shoaib, Muhammad, et al. ”Complex human activity recognition using smartphone and wrist-worn motion sensors.” Sensors 16.4 (2016): 426.

[26] Joshi, Mahesh, et al. ”A comparative study of supervised learning as applied to acronym expansion in clinical reports.” AMIA Annual Symposium Proceedings. Vol. 2006. American Medical Informatics Association, 2006.

[27] McCart JA, Berndt DJ, Jarman J, Finch DK, Luther SL. Finding falls in ambulatory care clinical documents using statistical text mining. Journal of the American Medical Informatics Association: JAMIA. 2013;20(5):906-914. doi:10.1136/amiajnl-2012-001334.

[28] https://en.wikipedia.org/wiki/Information_gain_ratio

[29] McCormick PJ, Elhadad N, Stetson PD. Use of Semantic Features to Classify Patient Smoking Status. AMIA Annual Symposium Proceedings. 2008;2008:450-454.

[30] Meystre, Stphane, and Peter J. Haug. ”Natural language processing to extract medical problems from electronic clinical documents: performance evaluation.” Journal of biomedical informatics 39.6 (2006): 589-599.

[31] Khalifa, Abdulrahman, and Stphane Meystre. ”Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes.” Journal of biomedical informatics 58 (2015): S128-S132.

[32] Melton GB, Hripcsak G. Automated Detection of Adverse Events Using Natural Language Processing of Discharge Summaries. Journal of the American Medical Informatics Association: JAMIA. 2005;12(4):448-457. doi:10.1197/jamia.M1794.

[33] Desai, Ankit A., et al. ”A novel molecular signature for elevated tricuspid regurgitation velocity in sickle cell disease.” American journal of respiratory and critical care medicine 186.4 (2012): 359-368.

[34] Xu, Mengjia, et al. ”A deep convolutional neural network for classification of red blood cells in sickle cell anemia.” PLoS computational biology 13.10 (2017): e1005746.

[35] Allayous, Clara, et al. ”Machine Learning Algorithms for Predicting Severe Crises of Sickle Cell Disease.” (2008).

[36] Yang, Fan, et al. ”Improving pain management in patients with sickle cell disease from physiological measures using machine learning techniques.” Smart Health (2018).

[37] Milton, Jacqueline N., et al. ”Prediction of fetal hemoglobin in sickle cell anemia using an ensemble of genetic risk prediction models.” Circulation: Genomic and Precision Medicine 7.2 (2014): 110-115.

[38] Wiki. https://en.wikipedia.org/wiki/Stemming

[39] https://en.wikipedia.org/wiki/Stop_words

[40] 40. Savova, Guergana K., et al. ”Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.” Journal of the American Medical Informatics Association 17.5 (2010): 507-513.

[41] https://en.wikipedia.org/wiki/N-gram

[42] Jurafsky, Dan. Speech & language processing. Pearson Education India, 2000.

[43] Frnkranz, Johannes. ”A study using n-gram features for text categorization.” Austrian Research Institute for Artificial Intelligence 3.1998 (1998): 1-10.

[44] https://en.wikipedia.org/wiki/Part_of_speech

[45] https://uima.apache.org/

[46] https://opennlp.apache.org/

[47] https://en.wikipedia.org/wiki/Feature_selection

[48] Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." Icml. Vol. 97. 1997.

[49] Quinlan, J. Ross. "Induction of decision trees." Machine learning 1.1 (1986): 81-106.

[50] Pearson, Karl. "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling." The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50.302 (1900): 157-175.

[51] Cover, Thomas M., and Joy A. Thomas. Elements of information theory. John Wiley Sons, 2012.

[52] Walker, SH; Duncan, DB (1967). "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54 (1/2): 167178.

[53] Cox, DR (1958). "The regression analysis of binary sequences (with discussion)". J Roy Stat Soc B. 20 (2): 215242.

[54] Hand, D. J.; Yu, K. (2001). "Idiot's Bayes not so stupid after all?". International Statistical Review.

[55] https://www.quora.com/What-are-the-advantages-of-using-a-naive-Bayes-for-classification

[56] Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks". Machine Learning. 20 (3): 273297.

[57] Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 1416 August 1995. pp. 278282.

[58] http://www.cs.uky.edu/ jzhang/CS689/PPDM-Chapter2.pdf

[59] https://stats.stackexchange.com/questions/24437/advantages-and-disadvantages-of-svm

[60] Ho TK (1998). "The Random Subspace Method for Constructing Decision Forests"

[61] https://en.wikipedia.org/wiki/Decision_tree

[62] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

[63] https://en.wikipedia.org/wiki/Confusion_matrix

[64] Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation"

[65] http://scikit-learn.org

[66] https://www.python.org/

[67] Wang, Sida, and Christopher D. Manning. "Baselines and bigrams: Simple, good sentiment and topic classification." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics, 2012.

[68] Wang, Wenbo, et al. "Harnessing twitter" big data" for automatic emotion identification." Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom). IEEE, 2012.

[69] Huh, Jina, Meliha Yetisgen-Yildiz, and Wanda Pratt. ”Text classification for assisting moderators in online health communities.” Journal of biomedical informatics 46.6 (2013): 998-1005.