Data Preprocessing

MIMIC - III Clinical Database

url: https://physionet.org/content/mimiciii/1.4

```
import pandas as pd
import nltk
import re
import numpy as np
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
```

D_ICD_DIAGNOSES has text describing each ICD9 code.

The long and short titles are combined to get the distribution of words using document term matrix

```
# Load datasets
notes = pd.read_csv("NOTEEVENTS.csv", low_memory=False)  # Load clinical notes
diag = pd.read_csv("DIAGNOSES_ICD.csv", low_memory=False)  # Load diagnosis codes
iDiag = pd.read_csv("D_ICD_DIAGNOSES.csv", low_memory=False)  # Load ICD descriptions

# Combine the SHORT_TITLE and LONG_TITLE columns
iDiag['CombinedTitles'] = iDiag[['SHORT_TITLE', 'LONG_TITLE']].fillna('').agg(' '.join, axis=1)  # Merge title columns
```

The keyword kinetic accident regular expression developed by Dr. Jim (Co-supervisor)

```
vehicleRegex = (
    r'\b(?:mva|mba|vehicle|bus|pedestrian|passenger|ute|ped|bike|dirtbike|motorbike|pushbike|scooter|truck|'
    r'bicycle|motorcycle|driver|driving|rtc|rta|\d*km[a-zA-Z/]*|skateboard|surfing|surf|horse|collision|'
    r'crossing|buggy|ebike|jetski|vs car|car vs|car accident|moving car|traffic light|traffic lights|'
    r'hit by car|hit by a car|car hit|airbag|airbags|T boned)\b'
)
```

```
iDiagFiltered = iDiag[iDiag['CombinedTitles'].str.contains(vehicleRegex, case=False, na=False)]  # Filter titles
```

```
# Text preprocessing function
def preprocessText(text):
    text = text.lower()  # Convert to lowercase
    text = re.sub(r'[^a-z\s]', '', text)  # Remove non-letters
    tokens = word_tokenize(text)  # Tokenize text
    stopWords = set(stopwords.words('english'))  # Load stopwords
    tokens = [word for word in tokens if word not in stopWords]  # Remove stopwords
    lemmatizer = WordNetLemmatizer()  # Initialize lemmatizer
    tokens = [lemmatizer.lemmatize(word) for word in tokens]  # Lemmatize tokens
    return ' '.join(tokens)  # Return processed text

iDiagFiltered = iDiagFiltered.copy()  # Make a copy to avoid modifying the original DataFrame
iDiagFiltered['ProcessedText'] = iDiagFiltered['CombinedTitles'].apply(preprocessText)  # Preprocess titles
```

LDA modelling is used to identify the distribution of the keywords in the ICD9 descriptions in the long and short title words.

The keyword distribution in the ICD9 title words are analysed to identify the relevant topics. The number of topics or value of k is chosen as 5 because of possibility of different distribution of words relating to different types of kinetic accidents.

```
# Create a document-term matrix
vectorizer = CountVectorizer(max_df=0.95, min_df=2, max_features=1000)  # Initialize vectorizer
dtm = vectorizer.fit_transform(iDiagFiltered['ProcessedText'])  # Create document-term matrix

# Fit LDA model
lda = LatentDirichletAllocation(n_components=5, random_state=42)  # Initialize LDA
lda.fit(dtm)  # Fit LDA model

# Display the top words per topic
def displayTopics(model, featureNames, noTopWords):
```

```
    topicWords = {}
    for topicIdx, topic in enumerate(model.components_):  # Iterate topics
        topWords = " ".join([featureNames[i] for i in topic.argsort()[:-noTopWords - 1:-1]])  # Get top words
        topicWords[topicIdx] = topWords  # Store topic words
        print(f"\nTopic {topicIdx}:")
        print(topWords)  # Print topic words
        print("\n" + "-"*80 + "\n")
    return topicWords

noTopWords = 1000
topics = displayTopics(lda, vectorizer.get_feature_names_out(), noTopWords)  # Display topic words
```

```
Topic 0:
vehicle rider animal accident injuring animaldrawn occupant veh cycle pedal mvoth ped drawn anim person motor cyc accani

--------------------------------------------------------------------------------


Topic 1:
vehicle motor accident injuring traffic collision mv involving motorcycle unspecified acc pedestrian passenger nature pe

--------------------------------------------------------------------------------


Topic 2:
gas motor involving dental activity exhaust poisoning vehicle surface surf caries poisonexhaust whether undetermined acc

--------------------------------------------------------------------------------


Topic 3:
vehicle accident injuring nontraffic motor snow offroad involving veh oth ntraf motordriven boarding alighting person mo

--------------------------------------------------------------------------------


Topic 4:
injuring accident railway involving collision oth coll rr object motor nontraffic vehicle fall passenger person cyclist

--------------------------------------------------------------------------------
```

Topic 2 is removed because the words in the topic 2 distribution are less relevant to the kinetic accident keywords.

```
# Get the topic distribution for each document
topicDistribution = lda.transform(dtm)  # Get topic distribution

# Identify documents NOT associated with Topic 2 (less relevant keywords in the distribution of topic 2)
nonTopic2Docs = iDiagFiltered[topicDistribution.argmax(axis=1) != 2]  # Filter out irrelevant topics

# Re-create the document-term matrix for the remaining documents
dtmRelevant = vectorizer.fit_transform(nonTopic2Docs['ProcessedText'])  # Recreate document-term matrix

# Re-run the LDA model on the filtered documents
ldaRelevant = LatentDirichletAllocation(n_components=4, random_state=42)  # Initialize new LDA
ldaRelevant.fit(dtmRelevant)  # Fit new LDA model

# Display the top words per topic in the filtered dataset
topicsRelevant = displayTopics(ldaRelevant, vectorizer.get_feature_names_out(), noTopWords)  # Display topics
```

```
Topic 0:
vehicle motor traffic accident injuring mv collision involving motorcycle unspecified passenger nature acc reentrant non

--------------------------------------------------------------------------------


Topic 1:
loss offroad highway control fall motor due collision vehicle injuring without accident traffic person mvpers passenger

--------------------------------------------------------------------------------


Topic 2:
vehicle accident injuring involving nontraffic motor oth rider animal occupant collision coll animaldrawn person snow ob

--------------------------------------------------------------------------------


Topic 3:
injuring accident railway involving pedal cyclist collision rr cycl acc coll train person cycle pedestrian specified mvt
```

The matched topics are used to get the unique ICD9 codes which are used to extract the relevant notes.

The unique ICD-9 codes related to specific topics, merges these codes with patient and admission data to identify relevant patient records, and then merges these records with clinical notes to display matched observations.

```
# Extract ICD9 codes for the remaining relevant topics
icd9Relevant = nonTopic2Docs['ICD9_CODE'].unique()  # Extract relevant ICD-9 codes

# Print the extracted ICD-9 codes
print("\nUnique ICD-9 Codes for Relevant Topics:")
print(icd9Relevant)

# Merge DIAGNOSES_ICD with icd9Relevant to get relevant patient and admission IDs
relevantPatients = diag[diag['ICD9_CODE'].isin(icd9Relevant)][['SUBJECT_ID', 'HADM_ID']]  # Filter patients

# Drop duplicates to ensure we have unique patient and admission IDs
relevantPatients = relevantPatients.drop_duplicates()  # Remove duplicates

# Merge NOTEEVENTS with relevantPatients to extract matching observations
relevantNotes = pd.merge(notes, relevantPatients, on=['SUBJECT_ID', 'HADM_ID'], how='inner')  # Merge notes

# Display the relevant observations
print(f"Number of matching observations: {len(relevantNotes)}")
print(relevantNotes.head())  # Show relevant notes
```

```
Unique ICD-9 Codes for Relevant Topics:
['E8213' 'E8214' 'E8215' 'E8216' 'E8217' 'E8218' 'E8219' 'E8220' 'E8221'
 'E8222' 'E8223' 'E8224' 'E8225' 'E8226' 'E8227' 'E8228' 'E8229' 'E8230'
 'E8231' 'E8232' 'E8233' 'E8234' 'E8235' 'E8236' 'E8237' 'E8238' 'E8239'
 'E8240' 'E8241' 'E8242' 'E8243' 'E8244' 'E8245' 'E8259' 'E8260' 'E8261'
 'E8262' 'E8263' 'E8264' 'E8268' 'E8269' 'E8270' 'E8272' 'E8273' 'E8274'
 'E8278' 'E8279' 'E8280' 'E8284' 'E8290' 'E8294' 'E8298' 'E8299' 'E8252'
 'E8253' 'E8254' 'E8255' 'E8256' 'E8257' 'E8258' 'E8850' 'E8852' 'E8860'
 'E8869' 'E8000' 'E8001' 'E8002' 'E8003' 'E8008' 'E8009' 'E8010' 'E8011'
 'E8117' 'E8118' 'E8119' 'E8120' 'E8121' 'E8122' 'E8123' 'E9290' 'E9585'
 'E8041' 'E8042' 'E8043' 'E8051' 'E8052' 'E8053' 'V6511' 'E8323' 'E8124'
 'E8012' 'E8013' 'E8018' 'E8019' 'E8020' 'E8021' 'E8022' 'E8023' 'E8028'
 'E8029' 'E8031' 'E8032' 'E8033' 'E8125' 'E8126' 'E8127' 'E8128' 'E8129'
 'E8130' 'E8131' 'E8132' 'E8133' 'E8134' 'E8135' 'E8136' 'E8137' 'E8138'
 'E8139' 'E8140' 'E8141' 'E8142' 'E8143' 'E8144' 'E8145' 'E8146' 'E8147'
 'E8148' 'E8149' 'E8150' 'E8151' 'E8152' 'E8153' 'E8154' 'E8155' 'E8156'
 'E8157' 'E8158' 'E8159' 'E8160' 'E8161' 'E8162' 'E9685' 'E8061' 'E8062'
 'E8063' 'E8071' 'E8072' 'E8073' 'E8100' 'E8101' 'E8102' 'E8103' 'E8104'
 'E8105' 'E8106' 'E8107' 'E8108' 'E8109' 'E8110' 'E8111' 'E8112' 'E8113'
 'E8114' 'E8115' 'E8116' 'E8246' 'E8247' 'E8248' 'E8249' 'E8250' 'E8251'
 'E848' 'E8163' 'E8164' 'E8165' 'E8166' 'E8167' 'E8168' 'E8169' 'E8170'
 'E8171' 'E8172' 'E8173' 'E8174' 'E8175' 'E8176' 'E8177' 'E8178' 'E8179'
 'E8180' 'E8181' 'E8182' 'E8183' 'E8184' 'E8185' 'E8186' 'E8187' 'E8188'
 'E8189' 'E8190' 'E8191' 'E8192' 'E8193' 'E8194' 'E8195' 'E8196' 'E8197'
 'E8198' 'E8199' 'E8200' 'E8201' 'E8202' 'E8203' 'E8204' 'E8205' 'E8206'
 'E8207' 'E8208' 'E8209' 'E8210' 'E8211' 'E8212' 'E9916' 'E9934' 'E9942']
Number of matching observations: 58808
   ROW_ID  SUBJECT_ID   HADM_ID   CHARTDATE CHARTTIME STORETIME  \
0     220       66479  134640.0  2148-02-07       NaN       NaN
1      34        7429  127159.0  2126-08-20       NaN       NaN
2     528        2650  188655.0  2132-10-01       NaN       NaN
3     518        2650  188655.0  2132-09-26       NaN       NaN
4     524        2650  188655.0  2132-10-01       NaN       NaN

            CATEGORY DESCRIPTION  CGID  ISERROR  \
0  Discharge summary      Report   NaN      NaN
1  Discharge summary      Report   NaN      NaN
2  Discharge summary      Report   NaN      NaN
3  Discharge summary      Report   NaN      NaN
4  Discharge summary      Report   NaN      NaN

                                          TEXT
0  Admission Date:  [**2148-2-3**]             D...
1  Admission Date: [**2126-7-29**]       Dischar...
2  Admission Date:  [**2132-12-3**]      Dischar...
3  Admission Date:  [**2132-7-9**]      Discharg...
4  Admission Date:  [**2132-7-9**]      Discharg...
```

Text relevant to the kinetic accident keyword matching ICD9 codes are extracted from Dischargy Summary in NOTEEVENTS.csv

The extracted relevant text are labelled true (1) and is used to finetune the classification head for preliminary modelling.

```
# Step 1: Extract ICD-9 Codes for Relevant Topics
icd9Relevant = nonTopic2Docs['ICD9_CODE'].unique()

# Step 2: Filter Relevant Patients and Notes
relevantPatients = diag[diag['ICD9_CODE'].isin(icd9Relevant)][['SUBJECT_ID', 'HADM_ID']]
relevantPatients = relevantPatients.drop_duplicates()

relevantNotes = pd.merge(notes, relevantPatients, on=['SUBJECT_ID', 'HADM_ID'], how='inner')

# Step 3: Define the regex pattern for the relevant topic (kinetic accidents)
vehicle_regex = r'\b(?:mva|mba|vehicle|bus|pedestrian|passenger|ute|ped|bike|dirtbike|motorbike|pushbike|scooter|truck|bicyc

# Step 4: Extract Relevant Lines or Return None
def extract_relevant_lines(text, regex_pattern):
    lines = text.split('\n')
    matching_lines = [line for line in lines if re.search(regex_pattern, line, re.IGNORECASE)]
    return matching_lines if matching_lines else None

matched_texts_relevant_lines = relevantNotes['TEXT'].apply(lambda text: extract_relevant_lines(text, vehicle_regex))

# Step 5: Combine Relevant Lines with HADM_ID
matched_texts_combined = pd.DataFrame({
    'HADM_ID': relevantNotes['HADM_ID'],
    'Relevant_Lines': matched_texts_relevant_lines
})

# Display the extracted information
print(matched_texts_combined.head())
```

```
     HADM_ID                                      Relevant_Lines
0  134640.0  [Pedestrian struck by motor vehicle, 60 year o...
1  127159.0  [involved in a motor vehicle accident. She was...
2  188655.0  [post motor vehicle accident on [**2132-7-9**]...
3  188655.0  [male who was a restrained passenger in a 110 ...
4  188655.0                                               None
```

```
# Filter for rows where Relevant_Lines is not null
nnl = matched_texts_combined[matched_texts_combined['Relevant_Lines'].notnull()]

# Extract the HADM_ID and the non-null Relevant_Lines
nnl = nnl[['HADM_ID', 'Relevant_Lines']]

# Display the extracted information
print(nnl.head())

len(nnl)

nnl['label'] = np.ones(len(nnl))
nnl.head()
```

```
     HADM_ID                                      Relevant_Lines
0  134640.0  [Pedestrian struck by motor vehicle, 60 year o...
1  127159.0  [involved in a motor vehicle accident. She was...
2  188655.0  [post motor vehicle accident on [**2132-7-9**]...
3  188655.0  [male who was a restrained passenger in a 110 ...
5  191263.0  [This is a 47 year old female who was in a mot...
```

	HADM_ID	Relevant_Lines	label
0	134640.0	[Pedestrian struck by motor vehicle, 60 year o...	1.0
1	127159.0	[involved in a motor vehicle accident. She was...	1.0
2	188655.0	[post motor vehicle accident on [**2132-7-9**]...	1.0
3	188655.0	[male who was a restrained passenger in a 110 ...	1.0
5	191263.0	[This is a 47 year old female who was in a mot...	1.0

Labelling non-kinetic accident cases

The complement of the ICD9 codes, which don't match the kinetic accident keywords, is used to randomly sample the same number of observations that were labeled as true kinetic accident cases for the purpose of fine-tuning BioClinicalBERT.

```
# Merge notes with diag to ensure ICD9_CODE is in notes
notes_with_icd9 = pd.merge(notes, diag[['SUBJECT_ID', 'HADM_ID', 'ICD9_CODE']], on=['SUBJECT_ID', 'HADM_ID'], how='left')

# Step 1: Filter NOTEEVENTS to exclude rows where ICD9_CODE matches icd9Relevant
non_matching_notes = notes_with_icd9[~notes_with_icd9['ICD9_CODE'].isin(icd9Relevant)]

# Step 2: Randomly sample len(nnl) rows from the non-matching notes
random_sample = non_matching_notes['TEXT'].sample(n=len(nnl), random_state=42)

# Step 3: Display the sampled rows
print(random_sample.head())

print(len(random_sample))
```

```
9299169     Pneumonia, bacterial, community acquired (CAP)...
10461640    Chief Complaint:\n   24 Hour Events:\n   Comfo...
23852980    Neonatology - NNP Physical Exam\n\nPlease see ...
3693201     Demographics\n   Day of intubation:\n   Day of...
3381758     Heart failure, right, isolated (Cor Pulmonale)...
Name: TEXT, dtype: object
22786
```

```
# Step 1: Merge 'notes' with 'diag' to ensure 'ICD9_CODE' is included
notes_with_icd9 = pd.merge(notes, diag[['SUBJECT_ID', 'HADM_ID', 'ICD9_CODE']], on=['SUBJECT_ID', 'HADM_ID'], how='left')

# Step 2: Filter 'NOTEEVENTS' to exclude rows where 'ICD9_CODE' matches 'icd9Relevant'
non_matching_notes = notes_with_icd9[~notes_with_icd9['ICD9_CODE'].isin(icd9Relevant)]

# Step 3: Randomly sample 'len(nnl)' rows from the non-matching notes
random_sample = non_matching_notes[['HADM_ID', 'TEXT']].sample(n=len(nnl), random_state=42)

# Step 4: Assign label 0 to the sampled data and rename 'TEXT' to 'Relevant_Lines'
random_sample_df = pd.DataFrame({
    'HADM_ID': random_sample['HADM_ID'],
    'Relevant_Lines': random_sample['TEXT'],
    'label': 0
})

# Step 5: Ensure 'nnl' has label 1
nnl['label'] = 1

# Step 6: Concatenate 'nnl' with the sampled data
combined_df = pd.concat([nnl, random_sample_df], ignore_index=True)

# Step 7: Display the first few rows of the combined DataFrame
print(combined_df.head())
```

```
     HADM_ID                                     Relevant_Lines  label
0  134640.0  [Pedestrian struck by motor vehicle, 60 year o...      1
1  127159.0  [involved in a motor vehicle accident. She was...      1
2  188655.0  [post motor vehicle accident on [**2132-7-9**]...      1
3  188655.0  [male who was a restrained passenger in a 110 ...      1
4  191263.0  [This is a 47 year old female who was in a mot...      1
```

```
print(combined_df['label'].value_counts())
```

```
label
1    22786
0    22786
Name: count, dtype: int64
```

Saving the data for preliminary modelling - train.csv with 45k observations

```
# Save the combined DataFrame to a CSV file
combined_df.to_csv('train.csv', index=False)

# Confirm the save by printing a message
print("combined_df has been saved to 'train.csv'")
```

```
combined_df has been saved to 'train.csv'
```

Anotation evaluation and statistical analysis

The anotation evaluation is performed to understand why the preliminary modelling resulted in a poor performance when testing on the unseen gold standard dataset, where true classes were

identified with a precision of less than 27% and an overall accuracy of less than 40%

```
data = pd.read_csv("train.csv")
```

```
data.head()
```

	HADM_ID	text	label
0	134640.0	Pedestrian struck by motor vehicle, 60 year ol...	1
1	127159.0	involved in a motor vehicle accident. She was ...	1
2	188655.0	post motor vehicle accident on **2132-7-9**, s...	1
3	188655.0	male who was a restrained passenger in a 110 m...	1
4	191263.0	This is a 47 year old female who was in a moto...	1

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45572 entries, 0 to 45571
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   HADM_ID  45360 non-null  float64
 1   text     45571 non-null  object 
 2   label    45572 non-null  int64  
dtypes: float64(1), int64(1), object(1)
memory usage: 1.0+ MB
```

Evaluation of true labels

True label cases were given more priority since its a binomial classification problem

```
ka = data[data["label"] == 1]
```

```
ka["text"]
```

```
0        Pedestrian struck by motor vehicle, 60 year ol...
1        involved in a motor vehicle accident. She was ...
2        post motor vehicle accident on **2132-7-9**, s...
3        male who was a restrained passenger in a 110 m...
4        This is a 47 year old female who was in a moto...
                              ...
22781    Specifics: 19 year old male unrestrained drive...
22782    post-MVA, c/b dens fx-T8 fx, and hemoperitoneu...
22783    50 yoF w/ a h/o MVA/T2 paraplegia since **2142...
22784    50 yoF w/ a h/o MVA/T2 paraplegia since **2142...
22785    50 yoF w/ a h/o MVA/T2 paraplegia since **2142...
Name: text, Length: 22786, dtype: object
```

Manual evaluation of 10 random samples with 100 observations each from the true case subset

```
samples = []

for i in range(10):
    sample_df = ka.sample(100, random_state=i)
    samples.append(sample_df)
```

```
for i, sample_df in enumerate(samples):
    print(f"Sample {i}:")
    for hadm_id, text in zip(sample_df["HADM_ID"], sample_df["text"]):
        print(f"HADM_ID: {hadm_id}, Text: {text}")
    print("\n" + "-"*80 + "\n")  # Add a separator between each sample's output
```

```
Sample 0:
HADM_ID: 194437.0, Text: 34F pedestraina vs car, with pneumomediastinum
HADM_ID: 183088.0, Text: s/p Motor vehicle crash, 24F unrestrained driver of car s/p collision with tree. +ETOH;, s/p Mo
HADM_ID: 129470.0, Text: A: hypertension, hypernatremia, failure to wean s/p being hit by a car
HADM_ID: 100210.0, Text: INDICATION: Motorcycle crash with multiple facial lacerations and decreased
HADM_ID: 169193.0, Text: 22 yr old female MVA vs Tree.Found under Dashboard,extrication time 30 minutes.Placed in cervic
HADM_ID: 175238.0, Text: bike **9-2**, no LOC, unhelmeted, EtOh 117, with anterior PTX by OSH CT, Clinical impression /
HADM_ID: 190394.0, Text: Admitting Diagnosis: MVA
```

```
HADM_ID: 115234.0, Text: service s/p MVC in which she was passenger. Pt wearing hard collar,
HADM_ID: 161048.0, Text: 40 Female with known baseline anisicoria, s/p ped struck intubated at, Assessment and Plan: 50F
HADM_ID: 177663.0, Text: Reason: TRAUMA, S/P PEDESTRIAN STRUCK, HISTORY: Status post pedestrian struck.
HADM_ID: 187131.0, Text: 18 year old man with mva -unresponsive, INDICATION: 18 year old man, status post MVA, unrespons
HADM_ID: 192214.0, Text: TITLE: 45yM s/p T-Bone MVC on motorcycle, in neurogenic shock
HADM_ID: 179517.0, Text: 51F passenger in rollover MVC. Self-extricated, complained of neck pain. Taken, CLINICAL INFORM
HADM_ID: 172901.0, Text: Admitting Diagnosis: MOTOR VEHICLE ACCIDENT, Admitting Diagnosis: MOTOR VEHICLE ACCIDENT
HADM_ID: 156634.0, Text: A: 19 yo male s/p motorcycle accident with multiple injuries, hct stable s/p transfusion yester
HADM_ID: 144855.0, Text: 79F s/p MVC restrained passenger, prolonged extrication, +LOC, obvious, CVA while driving causi
HADM_ID: 137059.0, Text: Admitting Diagnosis: S/P MOTOR VEHICLE ACCIDENT, 24 year old man with C2 fracture after mva
HADM_ID: 106783.0, Text: ASMT: Pt s/p ped struck with multiple injuries, slow to wake and follow commands.
HADM_ID: 155803.0, Text: HPI: 79yo M s/p MVC, restrained passenger in rollover, +LOC at scene.
HADM_ID: 146608.0, Text: REASON FOR ADMISSION: Pt is a 39 y/o male s/p MVC car vs. tree. GCS 10
HADM_ID: 194822.0, Text: "Pt is 76 yo female who was admitted to TSICU on **9-8** s/p ped struck. Injuries: small left S
HADM_ID: 192214.0, Text: 45yM s/p T-Bone MVC on motorcycle, in neurogenic shock
HADM_ID: 168558.0, Text: CHIEF COMPLAINT: Motor vehicle collision., unrestrained driver, who veered off the road hitting
HADM_ID: 134844.0, Text: HISTORY: 78-year-old male with known subdural hematoma after collision into a, car on a bike.
HADM_ID: 173851.0, Text: Reason: HIT BY CAR, 37 year old man biciclist hit by car, HISTORY: 37-year-old male bicyclist h
HADM_ID: 196360.0, Text: Admitting Diagnosis: MOTOR VEHICLE ACCIDENT;PELVIC FRACTURE;RIB FRACTURE
HADM_ID: 151385.0, Text: Admitting Diagnosis: S/P MOTOR VEHICLE ACCIDENT
HADM_ID: 151385.0, Text: 88F tx from OSH, restrained passenger MVC, T-boned, no LOC. Admitted to
HADM_ID: 145660.0, Text: Admitting Diagnosis: MOTOR VEHICLE ACCIDENT WITH UNSPECIFIED INJURIES
HADM_ID: 197571.0, Text: HPI: 55y/o male MVC head on collision w/parked trailer. Amnestic to
HADM_ID: 153555.0, Text: Admitting Diagnosis: MOTORCYCLE CRASH, 23 year old man with mva, REASON FOR EXAM: 23-year-old m
HADM_ID: 131144.0, Text: HISTORY: A 17-year-old patient hit in a car accident. She came in through
HADM_ID: 101289.0, Text: 19 year old unrestrained driver ejected from car. Ct of abdomen
HADM_ID: 193840.0, Text: Mr. **Known lastname 5758** is a 52 year old gentleman who was an unrestrained driver who hit a
HADM_ID: 137297.0, Text: 59M restrained driver involved in low speed MVC. Per outside
HADM_ID: 117325.0, Text: HISTORY: 45-year-old female in motor vehicle accident with concern for
HADM_ID: 194201.0, Text: Admitting Diagnosis: MOTOR VEHICLE CRASH, Admitting Diagnosis: MOTOR VEHICLE CRASH
HADM_ID: 143786.0, Text: 75 year old woman pedestrian struck by motor vehicle., INDICATION: 75-year-old female pedestria
HADM_ID: 191606.0, Text: hypothyroidism, transferred f/OSH **10-27** post-MVA, c/b dens fx-T8 fx,
HADM_ID: 154154.0, Text: Admitting Diagnosis: S/P MOTOR VEHICLE CRASH, Admitting Diagnosis: S/P MOTOR VEHICLE CRASH
HADM_ID: 159511.0, Text: Admitting Diagnosis: MOTOR VEHICLE ACCIDENT\\INJURIES, Admitting Diagnosis: MOTOR VEHICLE ACCID
HADM_ID: 112580.0, Text: 42M s/p bicycle vs. auto. Tx from **Hospital3 33**.
HADM_ID: 158873.0, Text: Admitting Diagnosis: MOTOR VEHICLE ACCIDENT
HADM_ID: 135820.0, Text: vehicle accident.
HADM_ID: 153013.0, Text: INDICATION: 83-year-old woman status post high speed motor vehicle collision.
HADM_ID: 125140.0, Text: 61 year old woman s/p pedestrian vs car, INDICATION: 61-year-old female status post pedestrian
HADM_ID: 162523.0, Text: 39 year old man with s/p rollover mva
HADM_ID: 158493.0, Text: Admitting Diagnosis: STATUS POST MOTOR VEHICLE ACCIDENT WITH INJURIES
HADM_ID: 190394.0, Text: Admitting Diagnosis: MVA
HADM_ID: 176320.0, Text: Admitting Diagnosis: C-SPINE FRACTURE-S/P MOTOR VEHICLE ACCIDENT, C3 lamina appear corticated,
HADM_ID: 180768.0, Text: This is a 71 yr old female who was an unrestrained driver involved in a motor vehicle collision
HADM_ID: 129645.0, Text: 66 yo female pedestrian struck by auto. Unconscious and unresponsive at
HADM_ID: 148995.0, Text: motorcycle accident, from outside hospital after a motorcycle accident. By report, he, was helm
HADM_ID: 128321.0, Text: 40-year-old unrestrained driver, who was involved in a minor, MVA after which he was assaulted
HADM_ID: 186970.0, Text: accident at his work where he was pinned between a moving truck, and a stationary truck. He was
HADM_ID: 189304.0, Text: 27 year old man s/p motorcycle crash, s/p anterior and posterior lumbar fusion, HISTORY: 27-yea
HADM_ID: 185875.0, Text: Admitting Diagnosis: STATUS POST MOTOR VEHICLE ACCIDENT WITH INJURIES
```

```
for i, row in sample1.iterrows():
    print(i,row["text"])
```

```
10753 Admitting Diagnosis: S/P MOTOR VEHICLE ACCIDENT, 25 year old man s/p MVA with Ptx, INDICATION: Status post MVA wit
19834 REASON FOR EXAMINATION: Evaluation of the patient after motor vehicle, collision, intubated.
19756 23 year old Ped struck now with vomiting and mental status changes, HISTORY: 23-year-old pedestrian struck by a ca
22472 Pt is a 76 yo male from Malde, MA; s/p MVA head on collision, pt hit head on windsheild and was restrained, EMS se
20234 Admitting Diagnosis: S/P MOTOR VEHICLE ACCIDENT, INDICATION: 29-year-old male post-MVA with sternotomy and repair
6474 18 M s/p MVC, was unrestrained passenger of a vehicle that rear-ended a, dump truck, reportedly pt hit head on dump
7775 49 year old man hit by a car, head, chest trauma now with increasing temps and
3991 26F pedestrian vs car about 35 mph on **3-31**
10590 Admitting Diagnosis: MVA, Admitting Diagnosis: MVA
22351 47 y.o. M driver of motorcycle no helmet arrived in our ED via Med Flight from scene. Pt agitated, noncompliat sed
4330 HPI: 27 year old male car vs tree mvc, decerebrate posturing on scene,, Trauma, s/p car vs tree mvc
12823 Admitting Diagnosis: S/P MOTOR VEHICLE ACCIDENT
15973 Reason: BIKE VS CAR THROW, INDICATION: 20-year-old male status post bicycle accident.
2476 vehicle accident with cervical spinal fracture and subsequent
20975 A- S/P MVA
7201 station but had LOC along the way suffering head on collision into, Assessment and Plan: 51M with recent left hip r
8268 Reason: s/p mva, s/p mva, HISTORY: Status post motor vehicle accident, unresponsive.
13481 Admitting Diagnosis: S/P MOTOR VEHICLE ACCIDENT
6135 HPI: 77 yo male restrained driver in a head on MVC with difficult, extraction from vehicle. Initially stable, but b
18796 Admitting Diagnosis: MOTOR VEHICLE ACCIDENT;PELVIC FRACTURE;RIB FRACTURE
11304 Admitting Diagnosis: MOTORCYCLE ACCIDENT, 36 year old woman with motorcycle accident, L leg ampuatation, T10/T11 s
15295 Admitting Diagnosis: PEDESTRIAN STRUCK
21603 "New trauma pt on T-SICU. Pt is a 73 year old married man who lives in **Hospital1 3830** with his wife, **Name (N
15097 Admitting Diagnosis: MOTOR VEHICLE CRASH;HYPOTENSION
2098 **Hospital1 19** ER s/p MVA on **2124-10-21**- bike vs car with a left open tib/fib
16314 HISTORY: **Age over 90 **-year-old woman status post motor vehicle collision. Evaluate for
11336 Admitting Diagnosis: LT HAND FX-BLUNT HEAD INJURY-S/P MVA, Status post MVA and fracture of the C5 vertebral body o
11245 Admitting Diagnosis: MOTORCYCLE ACCIDENT ,MBC, 34 year old man s/p mva, HISTORY: MVA rule out fracture.
7275 s/p pedestrian struck. R proximal humerus fracture, ORIF on
7803 49 year old man S/P MVA **8-22** w severe neurologic injury.
20617 THIS 73YO WOMAN REPORTED WAS THE RESTRAINED DRIVER OF A CAR VRS A CEMENT WALL. SHE BECAME OBTUNDED AT THE SCENE AN
2893 24 y.o.m. presents from an OSH after an MVC, car vs tree at high speed,
2770 Additional history: Parapalegic (T12 injury from car accident 30 yrs
5686 " 68F MVA, C2 body fx, bilateral vertebral foramina fx (Hangmans)., Assessment and Plan: 68F MVA, C2 body, bilater
3854 66F rollover MVC @ 60 mph +airbags +seatbelt +LOC. Prolonged
```

```
22270 A: s/p MVA. pneumonia, ?sepsis.
16782 50 year old man with mvc vs ped, HISTORY: 50-year-old male with MVC versus pedestrian, assess injury.
3926 79F s/p MVC restrained passenger, prolonged extrication, +LOC, obvious
21790 MS. **Known lastname 3005** IS A 49YO WOMAN WHO WAS OUT CELEBRATING HER BIRTHDAY AND WAS INVOLVED IN MVC ON WAY HO
18838 HISTORY: 60-year-old man status post motor vehicle collision with subdural
21932 neuro- pt a+o x3, follows commands, cooperative, epidural @ t6 w/ improved pain control, ativan rtc, pt somulent b
16357 HPI: 21M unrestrained backseat passenger found in front seat, GCS 3, intubated
11851 HISTORY: Status post motorcycle accident with trauma.
1026 s/p Motor vehicle crash, 88 yo male, restrained passenger, s/p motor vehicle crash, chest hit dashboard. + airbag d
9684 Admitting Diagnosis: S/P MOTOR VEHICLE ACCIDENT; SUBARACHNOID HEMORRHAGE; C2 FRACTURE, Admitting Diagnosis: S/P MOT
1295 s/p Motorcycle crash, 53 yo male driver s/p motorcycle crash vs truck with altered, s/p Motorcycle crash
1903 33 yo s/p high speed MCC found 100 ft from motorcycle. GCS=10 on scene,
11939 52 year old man with MVA
17858 50 year old man s/p ped struck
12102 Admitting Diagnosis: S/P MOTOR VEHICLE ACCIDENT, CLINICAL HISTORY: High speed motor vehicle accident. Multiple tub
16032 Admitting Diagnosis: S/P MOTOR VEHICLE ACCIDENT
6333 51 y/o male in motorcycle vs car collision c/ multiple injuries POD 2
2939 24 yo M presents from an OSH after an MVC, car vs tree at high speed,
6760 51yoM intoxicated s/p motorcycle accident, decellerating +flown over, +, motorcycle accident + LOC, +helmet, GCS of
16695 INDICATION: Trauma, seizure and motor vehicle collision in a 69-year-old
17458 Admitting Diagnosis: S/P MOTOR VEHICLE ACCIDENT;TELEMETRY
13597 Admitting Diagnosis: S/P MOTOR VEHICLE ACCIDENT
```

Statistical Analysis

True labels

```
# Load your dataset
data = pd.read_csv("train.csv")

# Filter the rows where label == 1
filtered_data = data[data["label"] == 1]

# Compute the length of each text (number of words) in the filtered data
text_lengths = filtered_data["text"].apply(lambda x: len(str(x).split()))

# Calculate average and maximum length
average_length = text_lengths.mean()
max_length = text_lengths.max()

# Print the results
print(f"Average Length of Text where label == 1: {average_length} words")
print(f"Maximum Length of Text where label == 1: {max_length} words")
```

```
Average Length of Text where label == 1: 18.13556569823576 words
Maximum Length of Text where label == 1: 343 words
```

False labels

```
# Filter the rows where label == 0
filtered_data = data[data["label"] == 0]

# Compute the length of each text (number of words) in the filtered data
text_lengths = filtered_data["text"].apply(lambda x: len(str(x).split()))

# Calculate average and maximum length
average_length = text_lengths.mean()
max_length = text_lengths.max()

# Print the results
print(f"Average Length of Text where label == 0: {average_length} words")
print(f"Maximum Length of Text where label == 0: {max_length} words")
```

```
Average Length of Text where label == 1: 283.88264723953307 words
Maximum Length of Text where label == 1: 6918 words
```

Further preprocess for false labels

```
# Define a function to extract the first two paragraphs and limit the word count to 512
def extract_first_two_paragraphs(text):
    if not isinstance(text, str):
        return "Not Found"

    # Split text into paragraphs using two or more newlines as separators
    paragraphs = re.split(r'\n\s*\n', text.strip())
```

```
        # Extract the first two paragraphs
        if len(paragraphs) >= 2:
            extracted_text = "\n\n".join(paragraphs[:2])  # Join the first two paragraphs
        elif len(paragraphs) == 1:
            extracted_text = paragraphs[0]  # If only one paragraph, return that one
        else:
            return "Not Found"

        # Limit the extracted text to 512 words
        words = extracted_text.split()
        if len(words) > 512:
            extracted_text = " ".join(words[:512])  # Truncate to 512 words

        return extracted_text

    # Apply the extraction function to the "text" column and replace the original text
    data["text"] = data["text"].apply(extract_first_two_paragraphs)

    # Save the updated dataset if needed
    data.to_csv("train_updated.csv", index=False)
```

Post truncation statistic of both labels

```
# Load the updated dataset
train = pd.read_csv("train_updated.csv")

# Calculate the length of each text observation in terms of words
train['text_length'] = train['text'].apply(lambda x: len(str(x).split()))

# Get the maximum, minimum, and average length
max_length = train['text_length'].max()
min_length = train['text_length'].min()
avg_length = train['text_length'].mean()

# Print the results
print(f"Maximum Text Length: {max_length} words")
print(f"Minimum Text Length: {min_length} words")
print(f"Average Text Length: {avg_length:.2f} words")
```

```
Maximum Text Length: 512 words
Minimum Text Length: 2 words
Average Text Length: 113.97 words
```

Data redundancy elimination

Eliminate number of rows with less than 3 words

```
# Load the updated dataset
train = pd.read_csv("train_updated.csv")

# Calculate the length of each text observation in terms of words
train['text_length'] = train['text'].apply(lambda x: len(str(x).split()))

# Filter out rows where the text has less than 2 words
train_filtered = train[train['text_length'] >= 2].copy()

# Save the filtered dataset to a new CSV file
train_filtered.to_csv("train_updated.csv", index=False)

# Print the number of rows eliminated
eliminated_rows = len(train) - len(train_filtered)
print(f"Number of rows eliminated: {eliminated_rows}")
```

```
Number of rows eliminated: 21
```

Remove replicates

The text is extracted from discharge summary. Different HADM_ID (Patient's each visit to hospital based Discharge Summary) can have same replicate sentence while extracting only the lines which matches the kinetic accident keywords

```
# Remove duplicate rows based on the 'text' column
train_unique = train.drop_duplicates(subset='text', keep='first')
```

```
# Save the updated dataset without duplicates
train_unique.to_csv("train_unique.csv", index=False)

# Print the number of duplicate rows removed
duplicate_rows_removed = len(train) - len(train_unique)
print(f"Number of duplicate rows removed: {duplicate_rows_removed}")
```

```
Number of duplicate rows removed: 10094
```

```
train = pd.read_csv("train_unique.csv")
```

```
train.head()
```

	HADM_ID	text	label
0	134640.0	Pedestrian struck by motor vehicle, 60 year ol...	1
1	127159.0	involved in a motor vehicle accident. She was ...	1
2	188655.0	post motor vehicle accident on **2132-7-9**, s...	1
3	188655.0	male who was a restrained passenger in a 110 m...	1
4	191263.0	This is a 47 year old female who was in a moto...	1

```
train["label"].value_counts()
```

```
0    22544
1    12934
Name: label, dtype: int64
```

Printing the true label text inputs

```
for i in range(len(train[train["label"] == 1])):
    print(i,train[train["label"] == 1]["text"].iloc[i])
```

```
0 Pedestrian struck by motor vehicle, 60 year old female pedestrian struck at ~30 mph. Per witnesses, Polytrauma: pedest
1 involved in a motor vehicle accident. She was an unrestrained, driver with no loss of consciousness, but was hit by a
2 post motor vehicle accident on **2132-7-9**, status post
3 male who was a restrained passenger in a 110 mile per hour, motor vehicle versus tree collision in **Hospital3 4298**
4 This is a 47 year old female who was in a motor scooter accident, This is 47 yo F who presented s/p fall from scooter
5 male, status post motor vehicle accident. The patient was a, restrained driver involved in a moderate speed motor vehi
6 normal all weekend. Today while riding a bike race he had a, Mr. **Known lastname 9579** is a 47 year old male on **21
7 s/p Motor vehicle crash, 51 yo female s/p car vs. pole reportedly fleeing from police, +, s/p Motor vehicle crash
8 29 year-old male, unrestrained driver, who was brought in by EMS, for ETOH. Airbags were activated and windshield was
9 66F with MVA sustained on **2137-7-21**, restrained passenger,, hysteretcomy s/p MVA on **2137-7-21** where she was a
10 S/P pedestrian hit by snow plow
11 s/p head-on collision bicyclist vs bicyclist, 49M transfer from **Hospital1 ** **Location (un) 620** s/p bicycle acci
12 sustained in a motorcycle accident, * grand mal seizures after a head injury (also due to motorcycle
13 restrained driver high speed mvc, restrained driver in a high speed, head on motor vehicle
14 with a history of HIV and hepatitis C who was a pedestrian, hit by debris from a nearby car accident on the street. H
15 Pedestrian struck, **Age over 90 **F run over by vehicle near her home. Patient was standing in, drive way when vehic
16 41 yo female pedestrian who was struck by a moving vehicle; +LOC, 41 year old woman pedestrian struck with fever now
17 55 year old male driver; helmeted; s/p motorcycle crash on **9-5**
18 46 yo male helmeted passenger, s/p moped crash; + fatality of, driver Transferred via **Location (un) 7622** from an
19 s/p motorcycle accident, Patient is 32 yo man who presents s/p motorcycle collision vs, 1. s/p motorcycle crash
20 in MVA. He later developed headaches and presented to **Hospital1 18** ER
21 Department following a motor vehicle accident when his car, 3. Status post motor vehicle accident.
22 a pedestrian by a vehicle traveling at about 30 miles per
23 traumatic accident - bicycle vs auto, 46yo M bicyclist s/p bike vs auto, hit at 30-35mph & thrown, 46yo M bicyclist s
24 Pedestrian struck by car, 64 yo F pedestrian struck by a car at 30 mph causing her to, struck by a car traveling at 3
25 patient was involved in a bike accident with extensive bruising, of recent bike accident and concaine use. The possib
26 s/p Bicycle crash vs. auto, his bicycle was reportedly cut off by car and struck the vehicle, s/p Bicycle crash vs. a
27 pedestrian vs. car, 88-y.o. female pedestrian was struck by car, hypotensive on
28 s/p Motor vehicle crash, 69 yo female driver s/p motor vehicle crash while reportedly, s/p Motor vehicle crash
29 who suffered from a motor vehicle accident approximately one
30 right hip pain dating back to a motor vehicle collision
31 s/p motor vehicle accident, presents with diabetic ketoacidosis after motor vehicle accident, on **7-6**. She was in
32 involved in an motor vehicle accident, she was found, 1. Status post motor vehicle accident.
33 female status post a recent admission for a motor vehicle, all status post motor vehicle accident on **10-19**., moto
34 s/p Motor vehicle crash, 72 yo male s/p motor vehicle crash, he was hit on drivers side, s/p Motor vehicle crash
35 s/p Motor vehicle crash, 60 yo female s/p rollover motor vehicle collision; ?LOC She was, CLINICAL HISTORY: Motor veh
36 This is a 27 year-old male who was the driver (restrained, air, Status-post motor vehicle collision, you are sleepy w
37 fell off his bike while in a race and he was going
38 s/p Motor vehicle crash, 73 yo male driver s/p motor vehicle crash vs building, +Etoh,, s/p motor vehicle crash
39 s/p motor vehicle collision, driver in a MVC which occurred after he lost consciousness at, s/p motor vehicle collisi
40 unrestrained driver in a MVC, car versus barrier, transferred, positive airbag deployment, positive loss of conscious
41 vehicle accident and required tracheostomy placement. Subsequent, s/p MVA **4-2** w/ multiple face/pelvic/spine fx -i
42 motor vehicle accident with chest pain, driver involved in a two car MVA at 8PM on **2146-8-16**. Pt does not, recall
43 s/p Rollover motor vehicle crash, 38 yo female driver of a vehicle that was hit from the side at, the vehicle, and ap
44 s/p Motor vehicle crash, 34M rear passenger s/p motor vehicle crash with hemodynamic, s/p Motor vehicle crash
45 was "run over" by a 30-**Doctor First Name ** truck. Emergency Medical Service
46 was admitted on **2182-7-15** status post a motor vehicle crash., male who was an unrestrained passenger in a roll ov
```

```
47 vehicle accident., restrained passenger in a roll over motor vehicle accident,, 2. Status post motor vehicle accident
48 unrestrained driver involved in a motor vehicle crash with, pole. The passenger and driver had significant injury., -
49 was a restrained driver of a car hit by a bus prolonged, 1. Status post a motor vehicle accident.
50 passenger unrestrained in a head-on motor vehicle accident, a 22 year old female, unrestrained passenger in a head-on
51 revealed he was a unrestrained driver. The patient was, 4. Status post motor vehicle crash.
52 sustained a motor vehicle accident. The patient was a, passenger in the car. She had positive loss of consciousness,
53 man, unrestrained passenger of high speed motor vehicle, crash, ejected from the vehicle and thrown. The patient was
54 ADMISSION DIAGNOSIS: Motor vehicle accident., vehicle accident. She was a restrained driver. There was, airbag deploy
55 ambulance, status post a rollover motor vehicle accident, 2. Status post motor vehicle accident.
56 47yo RHF s/p bike versus car today now with L2 burst fracture, ?
57 male, status post high speed motor vehicle accident, unrestrained driver, who hit a pole, resulting in car fire
```

Balancing the dataset by undersampling false (0) cases for fine-tuning the sequential classifier model through gradient optimisation for the classification head and the 11th layer of Bio_ClinicalBERT

```
train_balanced = pd.concat([train[train["label"] == 1], train[train["label"] == 0].sample(n=train["label"].value_counts()[1]
```

```
train_balanced['label'].value_counts()
```

```
0    12934
1    12934
Name: label, dtype: int64
```

Printing the false label text inputs

```
for i in range(len(train[train["label"] == 0])):
    print(i,train[train["label"] == 0]["text"].iloc[i])
```

```
0 Pneumonia, bacterial, community acquired (CAP) Assessment: Afebrile. Intubated via stoma 400/18/12/50%. Lungs rhonchus
1 Chief Complaint: 24 Hour Events: Comfortable o/n, denies pain or concerns Allergies: No Known Drug Allergies Last dose
2 Neonatology - NNP Physical Exam Please see attending note above for a/p today. PE: Passive alert, appropriate on exam,
3 Demographics Day of intubation: Day of mechanical ventilation: 6 Ideal body weight: 40.9 None Ideal tidal volume: 163.
4 Heart failure, right, isolated (Cor Pulmonale) Assessment: SBP >130 and UO > 100cc/hr Action: Dobutamine weaning attem
5 CCU Nursing note S-intubated, sedated. O-74y.o male c Hx CAD, CABG, AFib, CKD. Admitted c CHF and now c persistent res
6 NURSING PROGRESS NOTE RESP/CV: INFANT REMAINS ORALLY INTUBATED ON UNCHANGED VENT SETTINGS IN 36-44% FI02 TONIGHT. CBG
7 Objective Height Admit weight Daily weight Weight change BMI 163 cm 92kg (**2-20**), 94.2 kg (**2-21**) 99.3 kg (**212
8 **2130-9-6** 6:35 PM CHEST (PORTABLE AP) Clip # **Clip Number (Radiology) 67669** Reason: Please evaluate for effusion
9 NPN #2FEN: o; Infant alert active.Waking for feeds. Bottled 75cc BM 24 well this am. Changed to BM 24 with similac pow
10 Atrial paced and ventricular sensed rhythm, ventricular pacing not capture - Demand A-V sequential implanted pacemake
11 Chief Complaint: Encephalopathy, elevated lactate, hepatic failure HPI: . Ms. **Known lastname ** is a 42 year old fe
12 Alt in co:co stable, pa line d/cd.Extubated this am, and tolerating ofm at 50%, with good abgs.Lungsounds clear.Has 3
13 **2124-1-14** 1:26 PM CHEST (PORTABLE AP) Clip # **Clip Number (Radiology) 107182** Reason: s/p CVL insertion-check p
14 Chief Complaint: resp failure 24 Hour Events: PICC LINE - START **2156-1-15** 02:00 PM -Re-intubated in a.m. because
15 **2167-7-9** 6:41 PM CT C-SPINE W/O CONTRAST Clip # **Clip Number (Radiology) 39074** Reason: trauma _____
16 Mr. **Known lastname 4680** is a 69-year-old male with past medical history significant for severe COPD on home oxyge
17 67 y/o M with PMH of CAD s/p MI, DM, hyperlipidemia. admitted to OSH on **2-3** after a fall on **1-31** which he sli
18 focus hemodymics data: neuro: alert and oriented. follows commands. able to lift extremities off the bed. perla #3 an
19 SICU HPI: 70F w/ met. breast CA, recurrent bilateral malignant effusions and pericardial effusion s/p **2143-10-30**
20 **2152-11-26** 8:12 PM CHEST (PORTABLE AP) Clip # **Clip Number (Radiology) 18389** Reason: R/O Pneumothorax, PNA Adm
21 **2190-8-24** 11:52 AM CHEST (PORTABLE AP) Clip # **Clip Number (Radiology) 57910** Reason: please eval for infiltrat
22 Nursing Progress Reveiw of Systems: Neuro: Pt more awake than yesterday. Opens eyes to voice/stimulus. Occasionally m
23 **2120-12-13** 6:12 PM CHEST (PORTABLE AP); -76 BY SAME PHYSICIAN **Name Initial (PRE) 5** # **Clip Number (Radiology
24 **2154-8-18** 5:25 AM CHEST (PORTABLE AP) Clip # **Clip Number (Radiology) 104054** Reason: eval pulm process Admitti
25 Chief Complaint: vfib arrest, CVA, COPD exacrebation I saw and examined the patient, and was physically present with
26 Pt is 47 M reportedly who fell down a flight of stars at 0400 on **10-23**, per girlfriend pt was unresponsive, pt wa
27 TITLE: Chief Complaint: 24 Hour Events: TUNNELLED DIALYSIS LINE. on mica, plastics will follow got doboff and tunnele
28 **2135-10-28** 8:30 AM CHEST (PORTABLE AP) Clip # **Clip Number (Radiology) 58374** Reason: evaluate for infiltrate,
29 **Doctor Last Name 1358**,**Initials (NamePattern4) ** **Last Name (NamePattern4) 1359** MED **Hospital Unit Name 391
30 **2191-3-15** 4:20 PM CHEST (PORTABLE AP) Clip # **Clip Number (Radiology) 94525** Reason: fever, DOE, HIV +, evaluat
31 The patient is a 48-yo man with chronic ETOH abuse and depression, who was BIBA after being found by neighbors living
32 **2126-11-10** 4:15 PM BABYGRAM (CHEST ONLY); -76 BY SAME PHYSICIAN **Name Initial (PRE) 5** # **Clip Number (Radiolo
33 Sinus rhythm. Normal tracing. Compared to the previous tracing of **2132-12-31** the rate has slowed. Otherwise, no d
34 Family Information Next of **Doctor First Name **: **Name (NI) 711** **Name (NI) 9884** husband (c) **Telephone/Fax (
35 Hypotension (not Shock) Assessment: Action: Response: Plan: Pain control (acute pain, chronic pain) Assessment: Actio
36 **2103-9-4** 5:20 AM CHEST (PORTABLE AP) Clip # **Clip Number (Radiology) 102647** Reason: eval for interval change A
37 NPN #1-O: remains in air isolette at 29.8, swaddled, axillary temps 98.2-98.4. #2-O: TF increased to 100cc/k/d BM20/E
38 Nursing note: 88 year old woman admitted to SICU at 1845 from PACU s/p cerebellar hemorrhagic CVA - from OSH. Vent dr
39 **2182-8-6** 1:16 PM CT ABDOMEN W/O CONTRAST; CT PELVIS W/O CONTRAST Clip # **Clip Number (Radiology) 16832** Reason:
40 Nursing progress note #2 O: Remains in 25cc flow, 100% O2 nasal cannula. Breath sounds equal & clear with mild SC ret
41 89 yo female now with chronic ventilatory failure and re-admission from outside facility now re-admitted with signifi
42 Nursing Progress Note 1900-0700 Resp O/A: Remains in RA as of ~noon. Breathing very comfortably, mild skinny ic/sc re
43 assessment as noted res: needs multiple nts q1hour with copious yellow thick secretions. trying to cogh with weak cou
44 NPN Addendum: Agree with above note from PCA, **Female First Name (un) 4215**. During infants last feeding, **Female
45 NPN 1840 #1 F/N: TF 120cc/kg/d Neosure 22 44cc q 4 hrs. Infant bottle feeds 10-30cc q 4 hrs then tires out and requir
46 Neonatology-** 7** Progress Note **Known lastname 3075** **Last Name (NamePattern1) 1369** in her open crib, in nasal
47 NICU nursing note 1. Resp=O/Presently on bubble prong CPAP of 5. FIO2 24-26%. No spells. (Please refer to flowsheet f
48 **2126-5-17** 9:57 AM CHEST (PORTABLE AP); -59 DISTINCT PROCEDURAL SERVICE Clip # **Clip Number (Radiology) 51220** R
49 Sinus rhythm Consider left atrial abnormality Modest inferior T wave changes - are nonspecific and may be within norm
50 Neonatology Attending Progress Note: DOL #7 PMA 30 1/7 weeks CPAP 6, RA, clear, RR=30-50s caffeine increased yesterda
51 Respiratory Care Pt weaned from a/c to Cpap/ps this shift intial **9-25** resulting abg 727/45/107/22/-5/ increased s
52 Respiratory Therapy Pt remains trached w/ #7.0 Portex DIC trach. Received this AM on .50% trach mask, however pt beco
53 Chief Complaint: Respiratory failure I saw and examined the patient, and was physically present with the ICU Resident
54 59 F with CAD s/p LAD BMS x 2 on **2105-3-16**, CHF (EF 20%), pulm htn and asthma recently discharged, presented thre
```

```
55 Chief Complaint: 24 Hour Events: - CKs **Numeric Identifier 8941** (peaked @ **Numeric Identifier 8942** @1200) - mor
56 Nursing Progress Note #3-O/A- TF=150cc/kg/d of BM26 via NGT. Abd exam benign. Voiding and stooling. Heme neg stool to
57 24 Hour Events: 9:15 PM intubated for tacchypnea, respiratory distress Allergies: No Known Drug Allergies Last dose o
```

```
train_balanced.to_csv("train.csv", index=False)
```

This dataset (train.csv, overwriting the data "train.csv" used for preliminary modelling) is used in seqClassifier.py finetuning algorithm which resulted in 50% accuracy when tested on the unseen gold standard dataset