Checkpoint 5

Anam Khan (3222512 akc6310)
Madhav Khanna (3184733 mko6761)
Ashish Jeldi (3219879 aej1797)

1. To build an NLP pipeline which will read through the reports filed by the police and use it as a feature for predictive modelling. This can give the victim a good estimation of whether or not the allegation will sustain.

final_finding		
EX	219	219
NA	3301	3301
NS	685	685
SU	1030	1030
UN	428	428

race			
Asian/Pacific Islander	49	49	49
Black	3310	3310	3310
Hispanic	557	557	557
Native American/Alaskan Native	11	11	11
White	793	793	793

- **Quantitative Analysis**

  Reports from data_allegation (CPDB) were cleaned by replacing any null summaries with empty strings. Reports from the narratives.csv file were cleaned by taking rows that have a column_name of either “Initial / Intake Allegation” or “Allegation” and replacing the records which have “(None entered)” or “NO AFFIDAVIT” with an empty string to by in sync the CPDB data. Carriages returns are replaced with spaces, and records with empty strings are dropped. Duplicates were removed and rows that had the longest summary for that complaint were retained. Our target variable, ‘final_findings’, was retrieved from data_officerallegation and merged with the existing data.

  There were 6 samples where the allegation was ‘ZZ’, these were removed to avoid class imbalance issues. The data was then split into a ratio of 4:1 for training and testing respectively. Since the target variable is a string, they needed to be one hot encoded. The input features for both training and testing data were vectorized by using Count Vectorization and Tfidf.

  Naive Bayes Classifier gave us an accuracy of 0.71 on training data and 0.69 on testing data. Logistic Regression gave us an accuracy of 0.75 on training data and 0.72 on testing data. Support Vector Machines gave us an accuracy of 0.5830 on training data and 0.5825 on testing data. Support Vector Machines gave us an accuracy of 0.75 on training data and 0.70 on testing data. Decision trees gave us an accuracy of 0.70 on training data and 0.70 on testing data.

  Out of the 5663 reports that were processed, 553 had a positive sentiment, 4841 had a negative sentiment and 296 had a neutral sentiment.

- **Qualitative Analysis**

  After trial and error, we found that a decision tree of depth 3 gave us the best accuracy. And in the case of KNN, looking for 5 nearest neighbours gives the best accuracy.

  It’s not surprising that a majority of the reports expressed a negative sentiment since they are accusation reports and hence would contain language that the sentiment analyser would pick up as negative. Excluding the obviously frequent words in a report, the word cloud shows us that the most used words in the report are ‘party’, ‘justification’, ‘registered’, ‘without’, ‘authority’, ‘sustain’, ‘recommended’, ’review’, etc.

  219 Of the reports ended up in ‘EX’ category
  3301 Of the reports ended up in NA’ category
  685 Of the reports ended up in ‘NS’ category
  1030 Of the reports ended up in ‘SU’ category
  428 Of the reports ended up in ‘UN’ category

  In 1.03 percent of the reports, the victim was of race.

In 70 percent of the reports, the victim was of race.
In 14.5 percent of the reports, the victim was of race.
In 0.2 percent of the reports, the victim was of race.
In 16.8 percent of the reports, the victim was of race.