

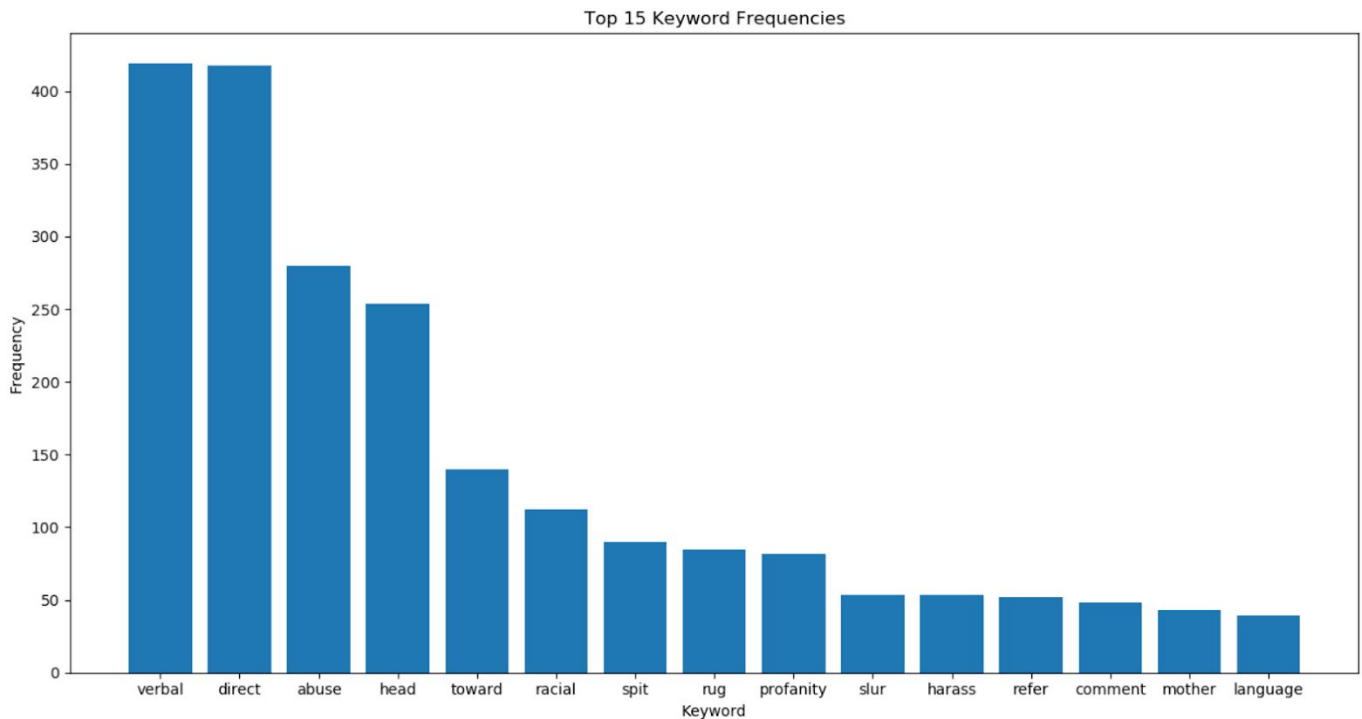
Checkpoint 5 Findings: Natural Language Processing

We had two goals for checkpoint 5. First, could we classify verbal abuse, profanity, and offensive remarks in complaint report narratives? Second, to connect it back to our theme, how can we relate these verbal abuse classifications to other officer characteristics? Initially, we predicted new officers in training to be taught to try to de-escalate situations, meaning no use of offensive language. But what if once officers join the force, they get a sense of the culture and mimic others who tend to use verbal abuse? Thus, we figured this could help us explore how police culture is spread from those of more experience/higher seniority to those of less. It could also help us explore how much verbal behavior is weighted in disciplinary decisions. However, our actual findings did not exactly align with our initial predictions.

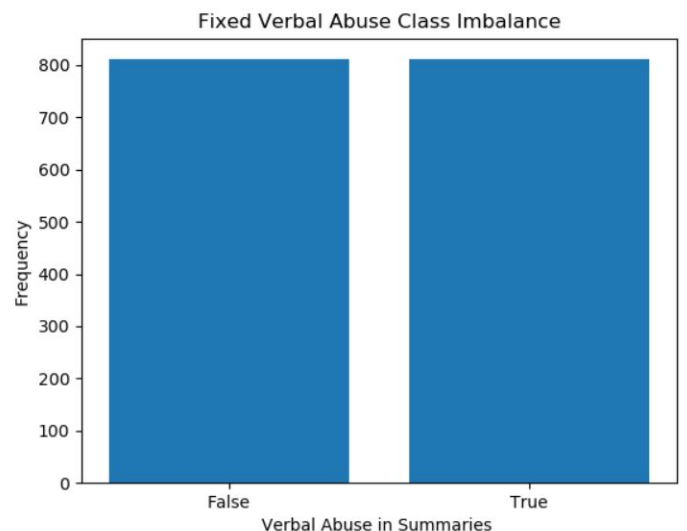
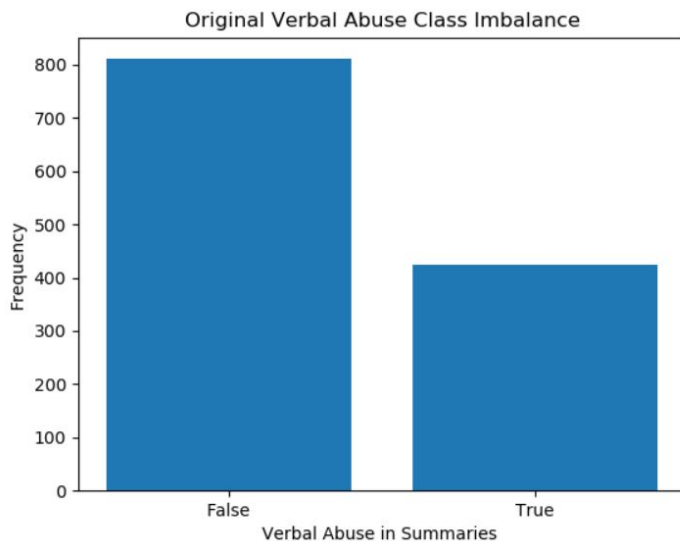
Creating a Verbal Abuse Classifier:

We noticed many allegations are mislabeled. The summary can contain instances of verbal abuse, yet the misconduct category does not list it as such. This is because allegations only list the primary misconduct. This makes it more difficult to identify cases of verbal abuse since the data is not properly tagged.

So we had labeled data to train a classifier, we did a divide-and-conquer on the summaries. We started compiling lists of keywords and phrases we saw in the summaries that implied verbal abuse. We then assigned weights ranging from 1 to 3 to these keywords to reflect how strongly they imply verbal abuse. For example, if we see the n- or f- word in a summary, it is pretty guaranteed there was some verbal abuse or profanity, so these terms have a weight of 3. However, if we see words like ‘race’ or ‘verbal’, they could be referring to verbal abuse or racial slurs from an officer, but they also could be referring to something else, so they have a weight of 1. For each summary, we identified keywords that appeared, totaled up the weights, and marked it as verbal abuse if the total was above our threshold of 2. We set a threshold of two to prevent false-positive tagging (tagging something as verbal abuse when it is not). We did some manual verification to make sure the tagging was generally comprehensive. The figure below shows the frequencies of the top 15 keywords among all summaries, though the actual list of keywords is much longer.



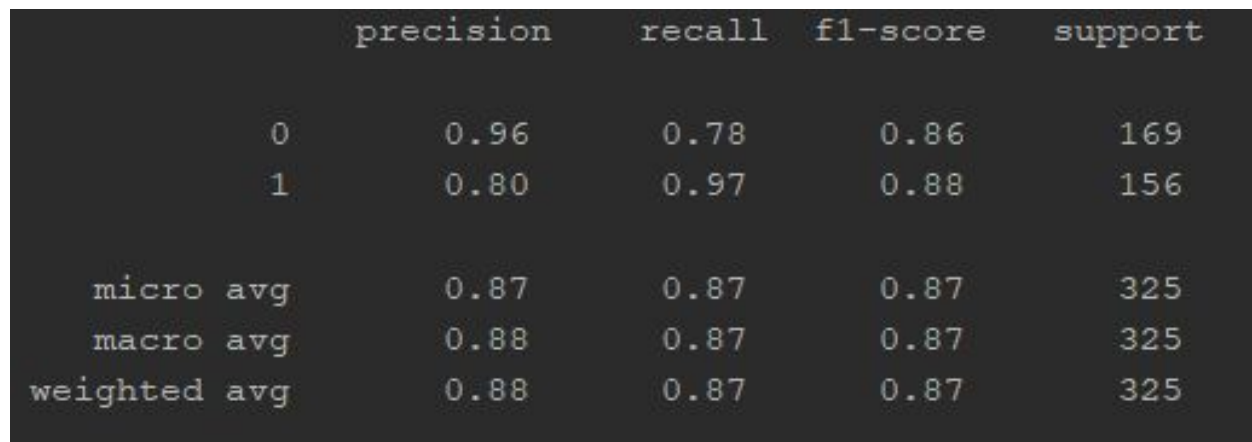
Once the data was labeled, we performed some data cleaning and preprocessing. First, we fixed class imbalances, since ‘no verbal abuse’ is twice as likely as ‘verbal abuse’ in summaries as shown in the figures below.



After, we tokenized, lemmatized, and removed stop words using the NLTK Python package. Then, we converted the data into TF-IDF feature vectors using Sklearn’s TfidfVectorizer.

Once we processed the input data into a more usable form, we split our data into training and test data sets using an 80/20 split. Then, we fit the training data with a Multinomial Naive Bayes classifier. Since Naive Bayes is a fast and scalable model, making training and testing much easier.

We evaluated our model in two ways. First, we used a train test split of 80/20 and used Sklearn's metrics classification report to view precision, recall, f1-score, and accuracy. The image below shows these results.



	precision	recall	f1-score	support
0	0.96	0.78	0.86	169
1	0.80	0.97	0.88	156
micro avg	0.87	0.87	0.87	325
macro avg	0.88	0.87	0.87	325
weighted avg	0.88	0.87	0.87	325

We also used 5-fold cross-validation to ensure we were not overfitting to a specific training set. With the accuracy metric, our CV results were: [0.83435583 0.78703704 0.84876543 0.77777778 0.83024691]. We do see a bit of variation here with the CV accuracies, but the values are all still pretty high, around 80%. Thus, we can confidently say our model was a success and can predict with about 80% accuracy whether a complaint report narrative depicts verbal abuse or other offensive language.

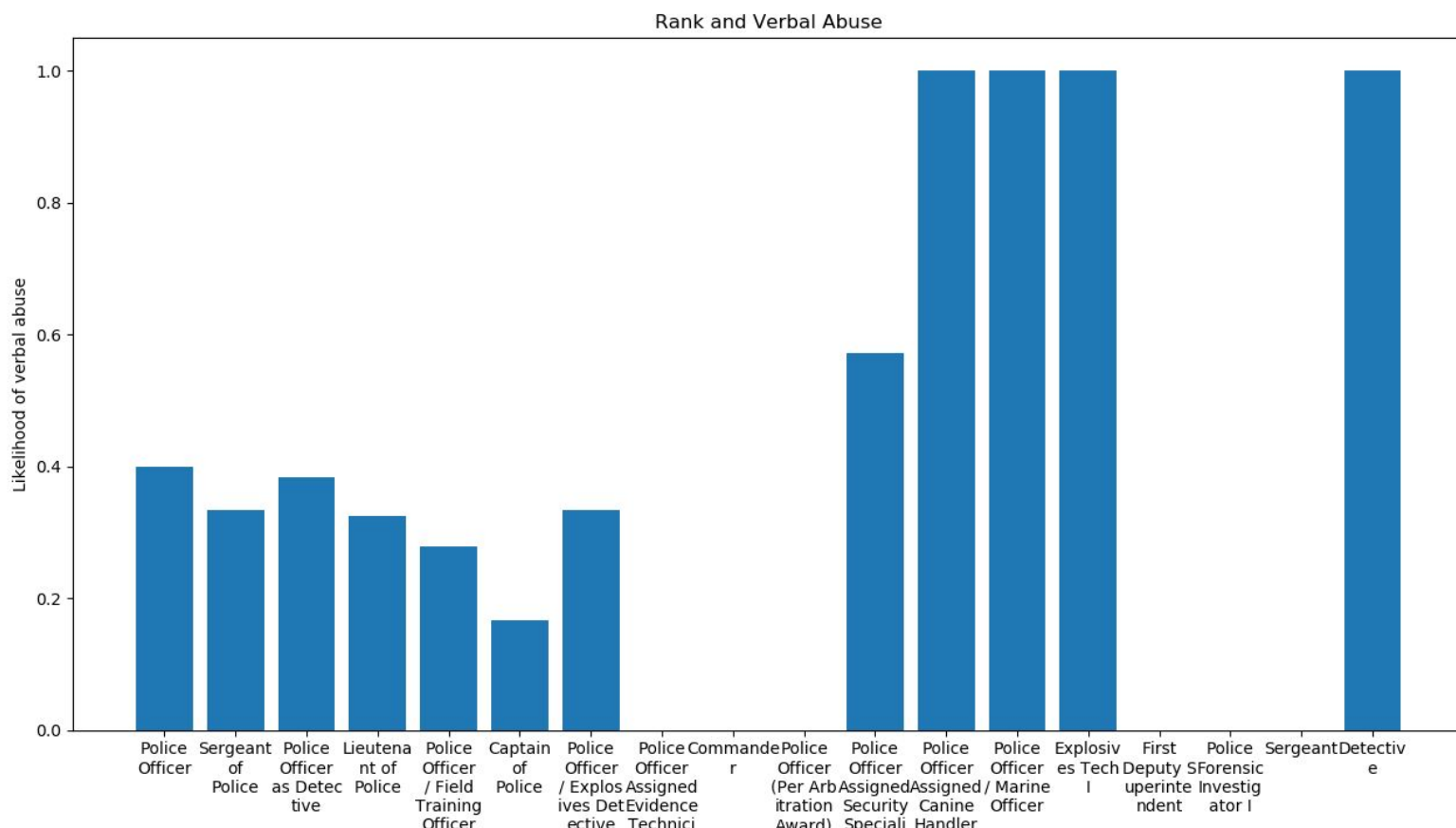
There are some limitations to our data and model. First, many allegations don't have summaries, meaning there was less data to work with here than previous checkpoints. Second, while we tried to make our keyword list and tagging comprehensive, there is likely some human error. There are probably a handful of mislabeled cases. Third, Naive Bayes does assume independence between the features, but some words are more likely given the presence of another word.

Officer Characteristics and Verbal Abuse:

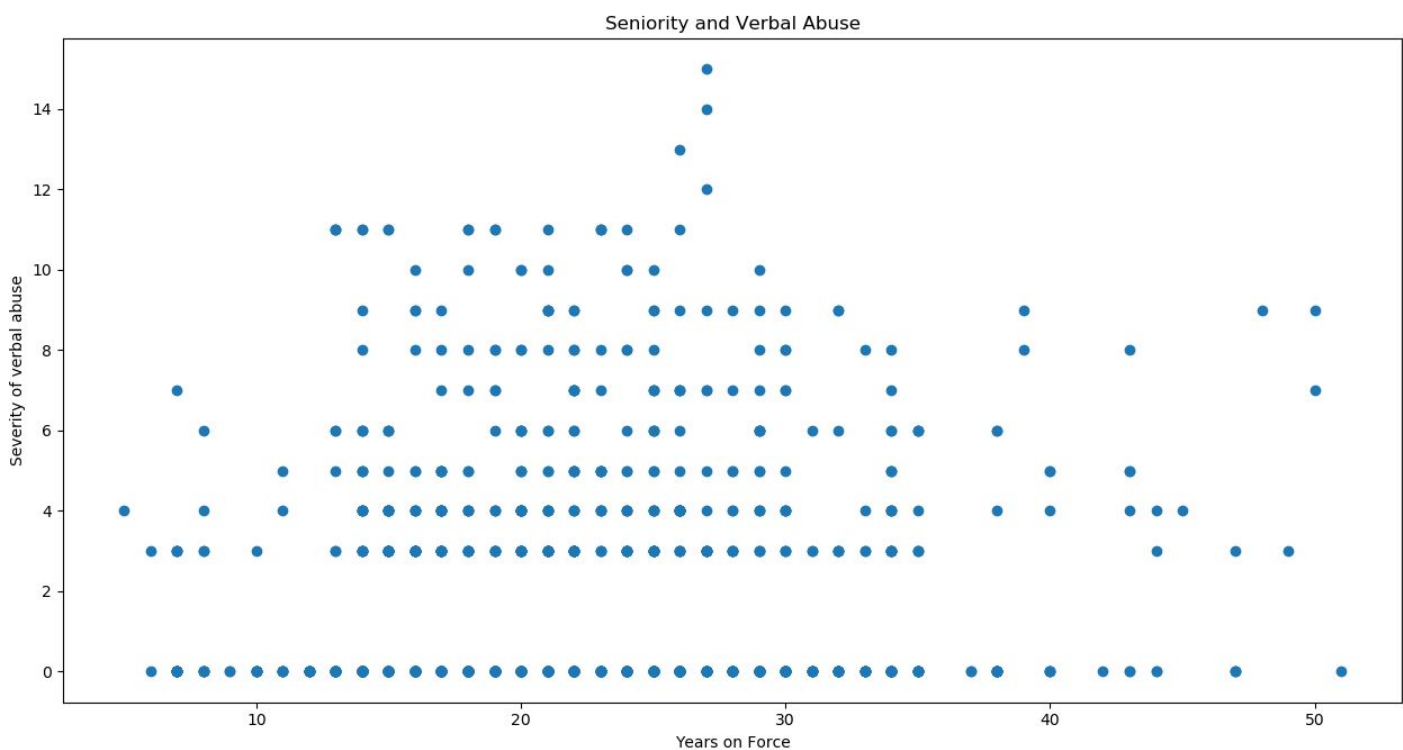
We wanted to connect our exploration of verbal abuse back to our theme. How does verbal abuse relate to certain officer characteristics? Do officers of higher or lower ranks think they can get away with verbal abuse? Do officers with more years on the force use more offensive language?

If frequent verbal abuse can be tied to certain officer demographics, what does that tell us about police culture in the CPD?

First, we looked into an officer's rank. For every summary for an officer allegation, we totaled the counts of verbal abuse, organized by rank. However, since some ranks are more popular than others, we normalized these totals by the frequency of a particular rank. The results are shown in the figure below. One thing to note is that the high bars on the right on the graph are due to limited data on those ranks. But we can still see higher ranks like captain, sergeant, and lieutenant tend to have lower rates of verbal abuse than police officers. This relationship doesn't tell us anything about causality though; do individuals with more self control or a better filter tend to get promoted to higher ranks? Something that surprised us was that the pattern among ranks here differs from our checkpoint 2 results (average disciplines by rank). In our checkpoint 2 results, police officers, lieutenants, and sergeants had similar average discipline numbers. Here, we see police officers are higher than lieutenant and sergeant. However, these differences in the grand scheme of things are quite small – only a fraction of a percentage. So while there technically is a difference, the difference might not be practical enough to make any major claims.



Second, we wanted to look into a potential relationship between years on the force and verbal abuse. We were interested in this relationship because it could potentially show how behavior is influenced by CPD culture over time. Instead of just looking at a binary metric (verbal abuse or no verbal abuse), we looked at how many offensive keywords were identified in the summaries; we refer to this as the severity of verbal abuse (see the y-axis in the figure below). We predicted that verbal abuse would increase overtime on the force. However, we see a unimodal pattern when we plot the data. We see less experienced officers, less than 15 years on the force, have the lowest severity of verbal abuse. We see a sharp peak in the middle around 30 years, and then we see a drop again after the early 30s. The increase over time on the left side of the plot does support our initial hypothesis, but our hypothesis doesn't explain the right side of the plot. We have a couple of ideas that might explain this pattern. It could be a result of different police training over time. Or maybe officers with high levels of verbal abuse are removed or transferred before they reach higher years on the force.



In conclusion, we can predict verbal abuse and offensive language in complaint report narratives with over 80% accuracy using a Multinomial Naive Bayes classifier. In the future, it would be valuable to try to improve this classifier with better manual tagging of training data and better keyword lists. Also having more data would help improve the classifier since most summaries do not contain verbal abuse. We also found a relationship between rank and verbal abuse, where

The Wicked Roadrunners

Kylie Chesner, Nimesh Tripathi, Aura Ulloa-Ordonez

higher ranks tend to have lower levels of verbal abuse. Having more data would be valuable for further confirming these patterns since some ranks are much less common than others. We also see very inexperienced or very experienced cops use fewer offensive words than moderately experienced cops, and it would be interesting to further look into what is causing this behavioral pattern.