Research Question

For our Natural Language Processing task, we propose to employ a neural model to obtain embeddings of allegation summaries, and then use these latent representations to improve predictions of officer complaint rate or whether a complaint will be sustained. Additionally, we can use these embeddings to do clustering or topic modelling analysis to identify groups of complaints and their common characteristics (e.g. if complaints within similar clusters emerge from similar precincts, similar officers, etc.)

In analyzing the above question, we began with a baseline machine learning classifier that attempts to predict whether a complaint is likely to be sustained, given the number of officers involved in the complaint, their demographic information, and aggregated statistics (average number of awards, average number of allegations, average number of use-of-force instances, etc.) (18 total features). Here, a complaint was considered sustained if any associated allegations were sustained. While we would ideally be predicting from information only up to the date of the complaint instead of current statistics, we don't expect that this will make too significant of a difference, under the assumption that an officer's trends are somewhat consistent on average.

Rather than making these predictions using a single decision tree, as was used in checkpoint 4 for a similar classification task, we expand upon this model by implementing an ensemble random forest classifier (maximum depth of 4). In this baseline test, where the complaint summaries are intentionally omitted, the predictor achieves the following performance:

Score Type	**Training Dataset**	**Validation Dataset**
Accuracy	0.851	0.830
F1 Score	0.887	0.766

For this evaluation and all others discussed in this document, the test set size was 1000 instances and the validation set size was 147 instances (of the 1147 complaints in the database that have an included summary).

In our second evaluation - for comparison with the above baseline - we instead made predictions using a spaCy-calculated 300-dimensional embedded representation of each summary, and no other features. Evaluating this new dataset on the random forest model described above (but with a maximum depth of 4), the model achieves the following performance:

Score Type	**Training Dataset**	**Validation Dataset**
Accuracy	0.861	0.837
F1 Score	0.892	0.760

As can be seen, the performance of the model as trained on each dataset individually are remarkably similar, with the baseline model performing ever so slightly better. This is especially interesting, given the immense variability of these summaries due to their freeform entry nature. However, we next examine the effects of using feature vectors that combine both the original baseline features and the summary embeddings (318 total dimensions per instance). With a maximum tree depth of 5, this composite model yielded the following results:

Score Type	**Training Dataset**	**Validation Dataset**
Accuracy	0.904	0.878
F1 Score	0.924	0.835

Thus, by combining both datasets, a non-negligible performance increase is attained. That is, the information embedded in a complaint summary is useful in improving predictions of whether that complaint will be sustained. However, it is difficult to draw conclusions about what aspects of the summaries have a significant impact in making predictions in their current latent representations.

Therefore, to visualize such information, we perform topic modeling on the collection of summary embeddings, to attempt to extract the information in the summaries that is most influential in predicting complaint sustainment. The following images use t-SNE dimensionality reduction to reduce the large dimensionality of the data. Topic modeling is done using non-negative matrix factorization (NMF), as we found that this method produces better results than LDA, due to its ability to take advantage of tf-idf preprocessing.

Figure 1 shows a scatter plot of the 300-dimensional word embeddings reduced to 2 dimensions, where each complaint point is colored according to one of the three topics that NMF grouped the complaints into (k=3 groups was chosen since visual inspection of the plots seems to reveal roughly three clusters). Here each topic is represented by its 10 most prominent words. The first (red) topic we will call “more serious officer complaints”, the second (green) we call “minor, typically traffic-related complaints”, and the third we will call “officer carelessness/negligence”. Obviously, some of these categories overlap, as can be seen in the figure. However, it is particularly interesting that the separation of topics corresponds so well to the rough clusters in the figure, especially considering the topics and the vectors were found using two completely unrelated methods. The figure indicates that the information from the given three topics, extracted from the summary embeddings, plays a major role in determining whether a complaint will be sustained.

In addition, we can also perform the same dimensionality reduction on the non-summary baseline feature model and the combined model to see whether this embedded information is similarly captured by other features. For example, in Figure 2 we can see that the embedded information doesn’t appear to correlate well with the baseline features, and is therefore independent of them (with the exception of the tight cluster towards the bottom of the figure

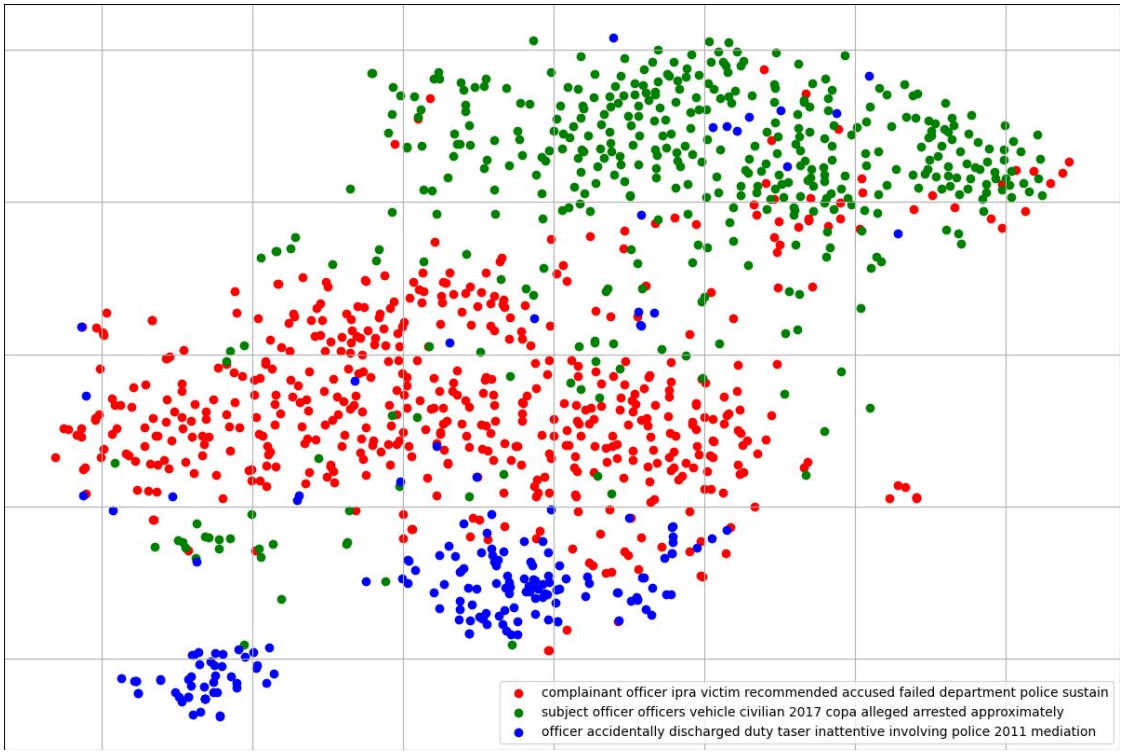

Figure 1: t-SNE dimensionality reduction of 300-dimensional complaint summary embeddings. Points are colored according to topics extracted using non-negative matrix factorization (NMF) of summary embeddings.

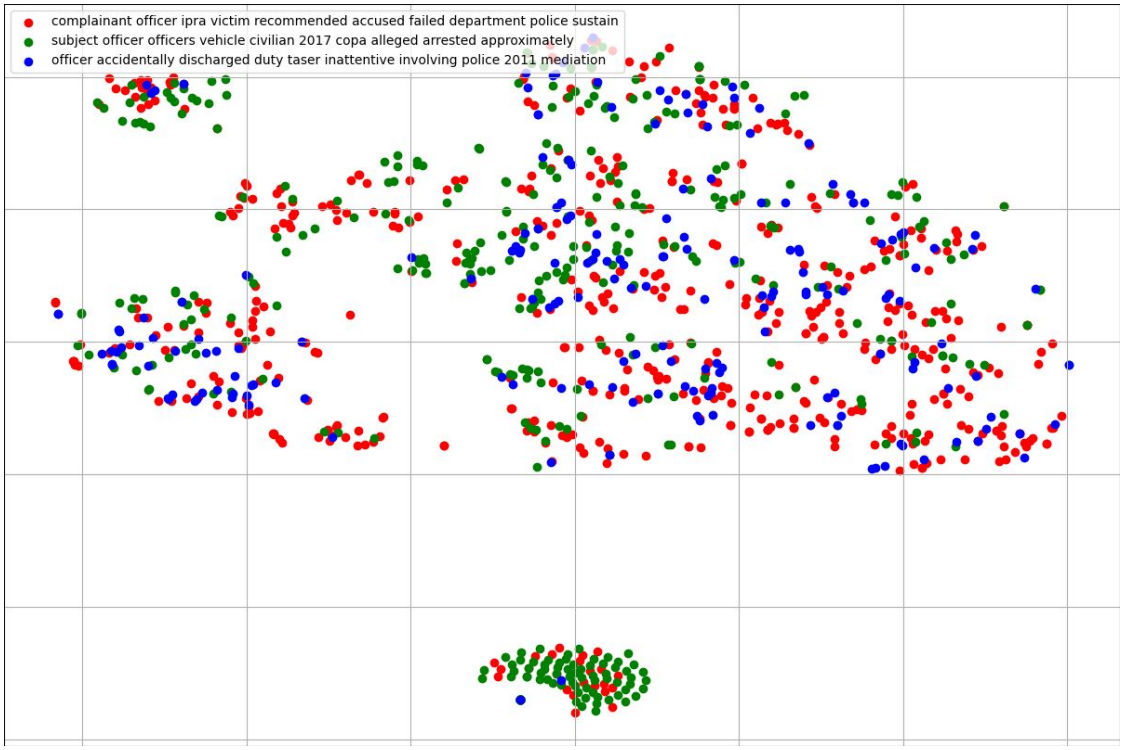

Figure 2: t-SNE dimensionality reduction of 18-dimensional complaint features. Points are colored according to topics extracted using NMF of summary embeddings.

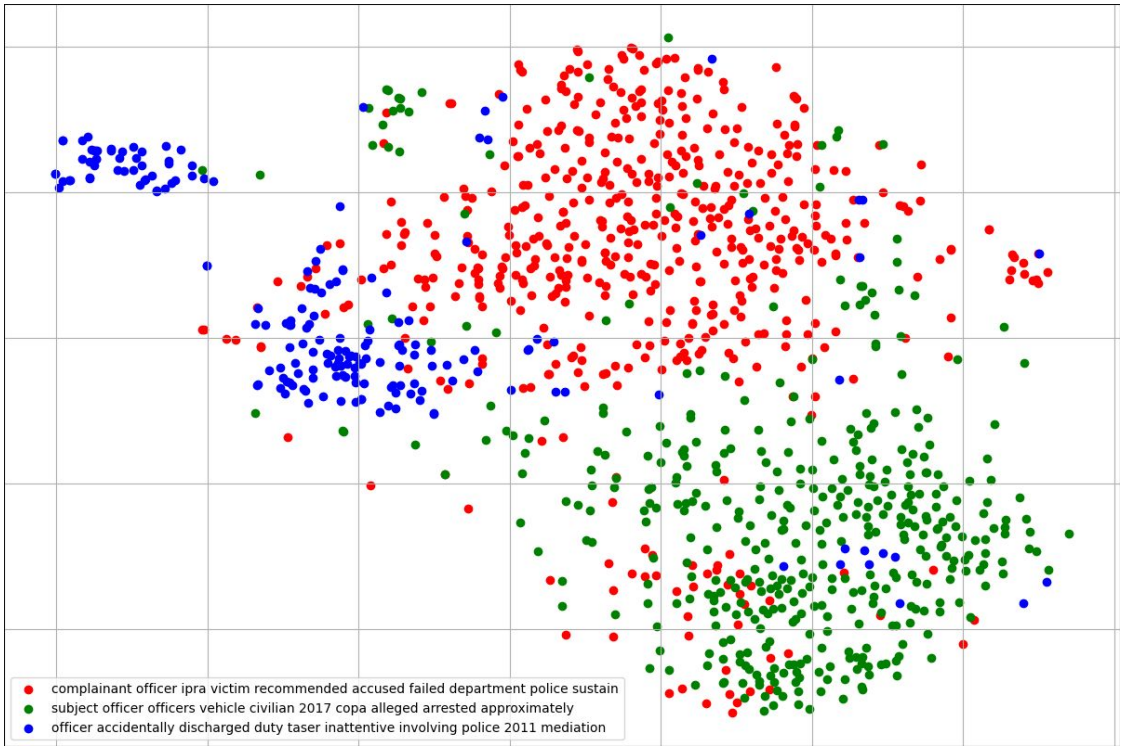

Figure 3: t-SNE dimensionality reduction of 318-dimensional complaint features + summary embeddings. Points are colored according to topics extracted using NMF of summary embeddings.

being mostly associated with “minor traffic-related complaints”. In Figure 3, however, we unsurprisingly see that there is a similar correlation as in Figure 1, since most of the features in this model are the same as those of the summary embeddings.

We can further recolor each of these plots according to the ground-truth value of “sustained” or “not sustained” for each complaint. These figures are shown below (Figures 4 - 6 respectively). For the baseline model plotted in Figure 4, we see that the spatial distribution of the summary vectors (reduced to two dimensions) represents a fairly clear division of classes. That is, the unsustained allegations appear in the upper-most cluster while the sustained allegations appear in the bottom two clusters. In comparison to Figure 1, it would appear that the complaint summaries that are more focused on minor, often traffic-related issues tend to go unsustained, while the other two summary topics tend to much more strongly correlate with sustained complaints.

A similar comparison can be made between Figures 2 and 5, where we see that, just as in Figure 2, Figure 5 doesn’t appear to show any discernible groupings, other than the bottom cluster of unsustained complaints, which again seems to coincide with the minor, traffic-related complaints. Finally, in comparing Figures 3 and 6, we see that the results are similar as in Figures 1 and 3, again likely due to the fact that the vast majority of features represented in Figures 3 and 6 (300 of 318) are complaint summary embedding features.

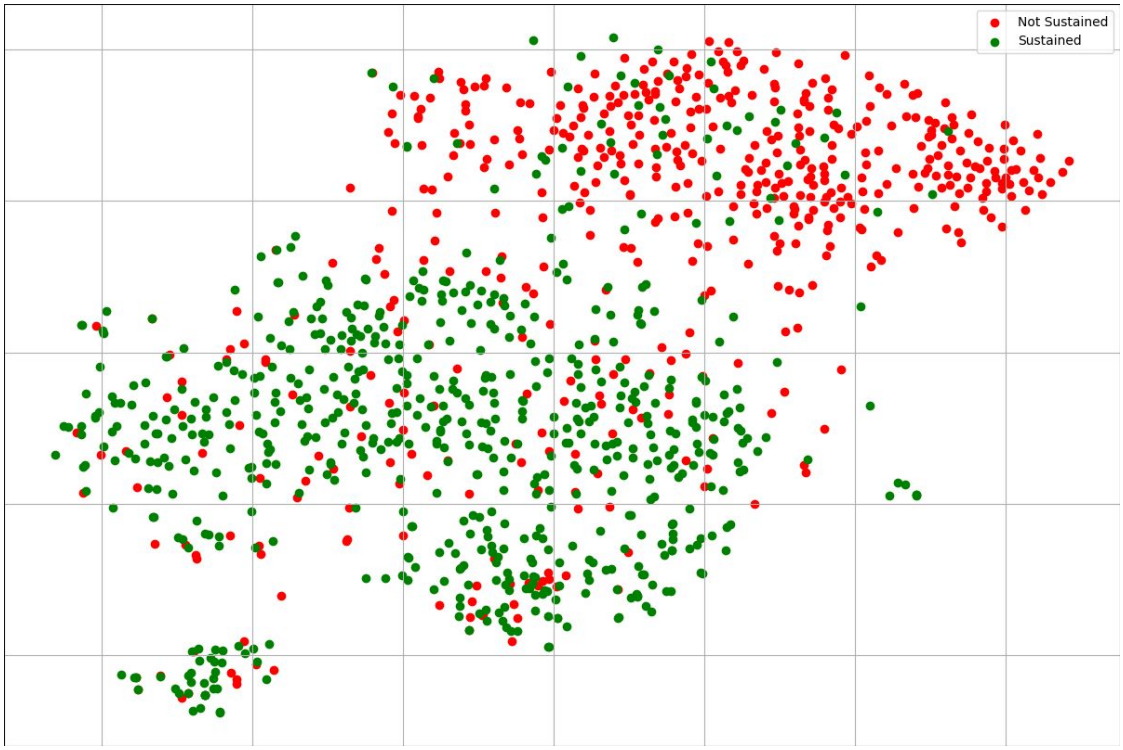

Figure 4: Same scatter plot as Figure 1, but colored by the ground-truth value for whether each complaint was sustained or not.

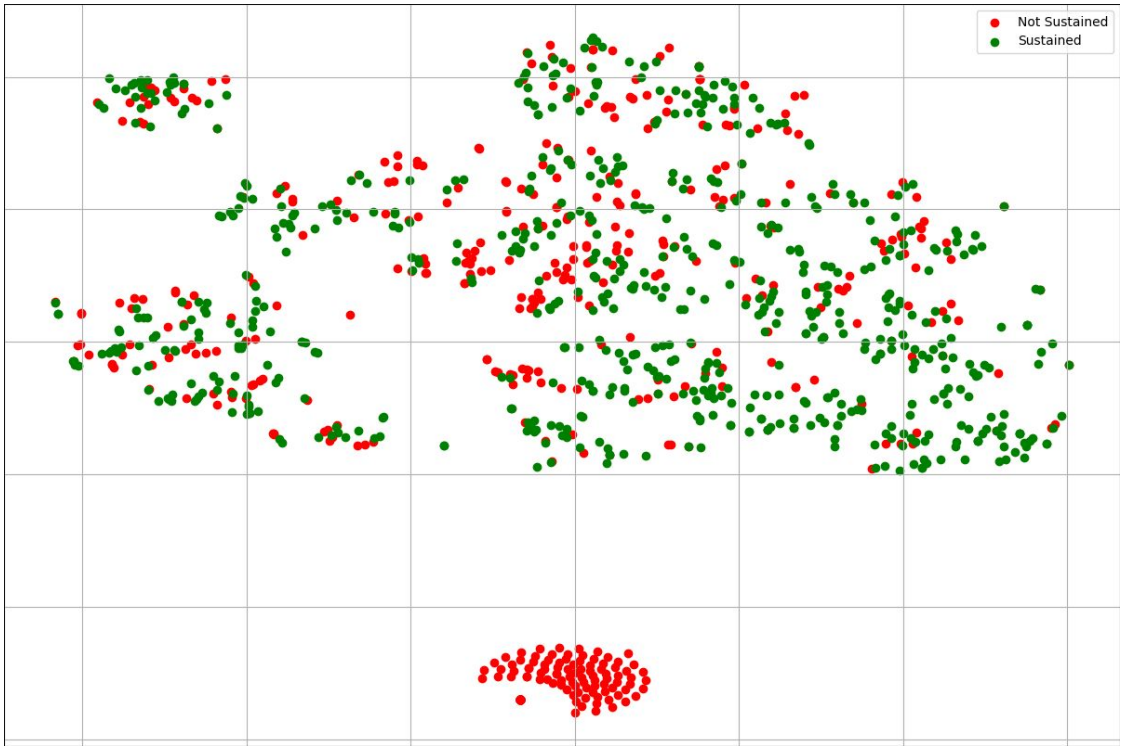

Figure 5: Same scatter plot as Figure 2, but colored by the ground-truth value for whether each complaint was sustained or not.

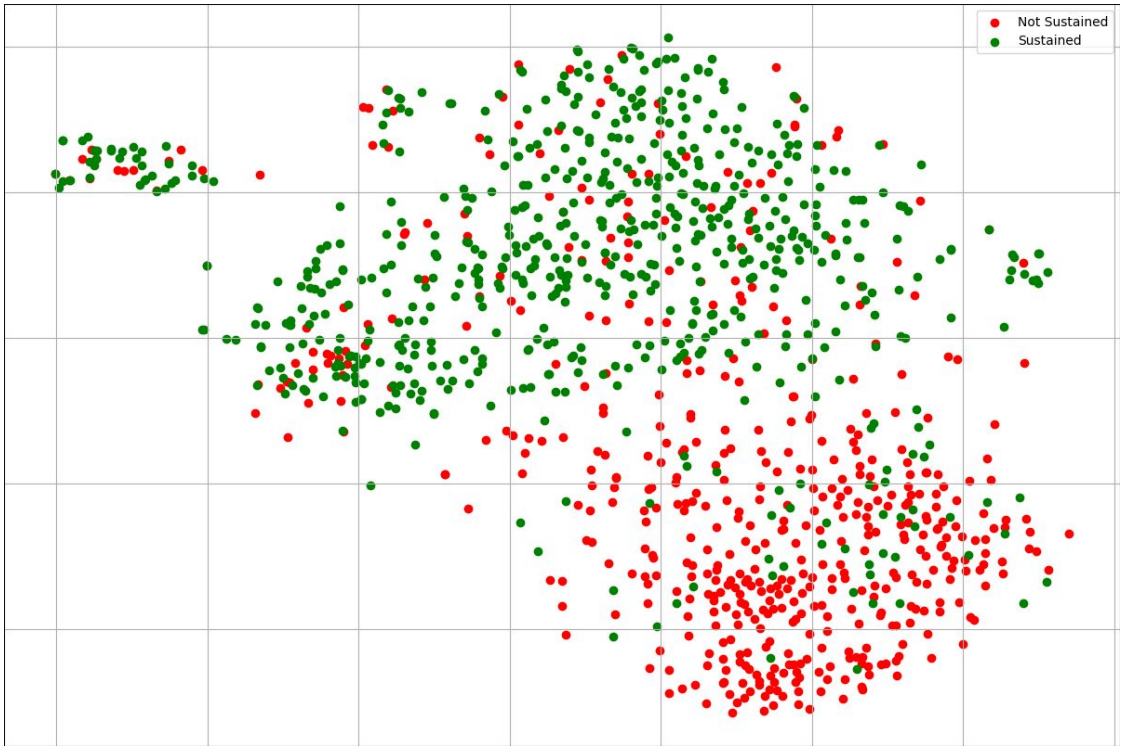

Figure 6: Same scatter plot as Figure 3, but colored by the ground-truth value for whether each complaint was sustained or not.

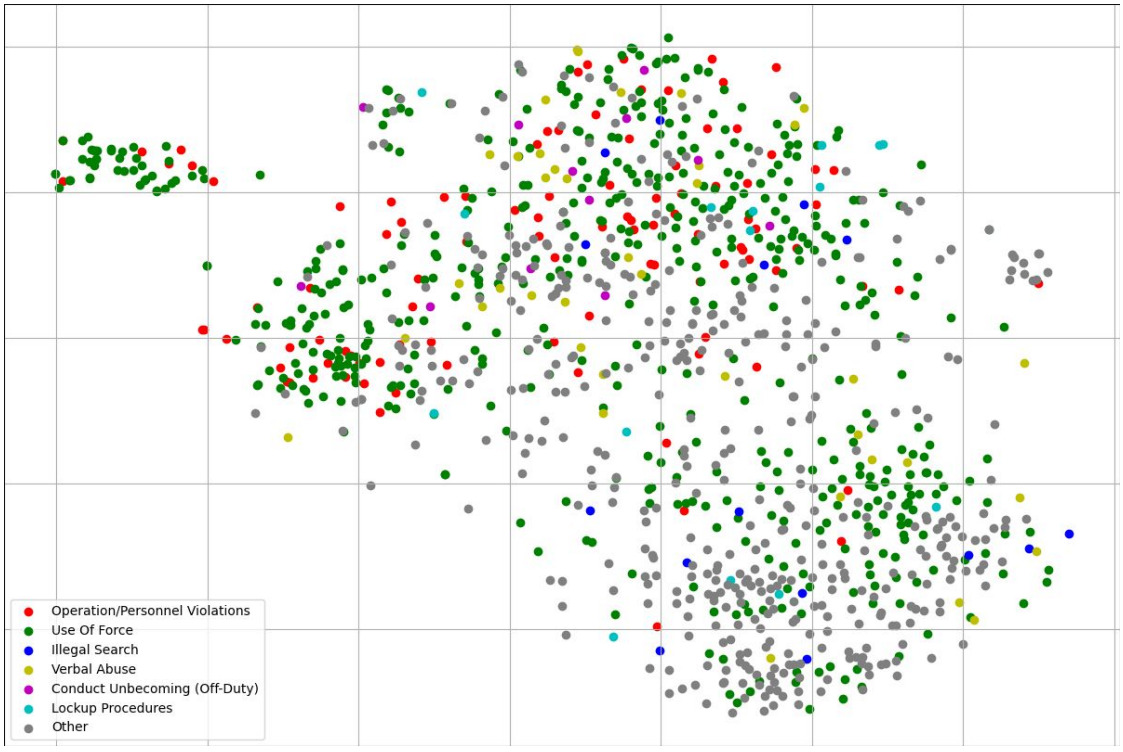

Figure 7: Same scatter plot as in Figures 3 and 6, but colored by the the most common category among the allegations for a given complaint.

One obvious question with our methods might be why it makes sense to even include topic modeling analysis on the summaries in the first place, since the allegation category is accessible in the database. However, as can be seen in Figure 7, there does not seem to be a significant correlation or grouping of the summary embedding vectors by complaint category. Therefore, it is apparent that the summaries embed information that cannot be captured by the officer-assigned allegation categories. This fact may indicate that the complaint summary as described by the victim (albeit transcribed by an officer) tells a different story than the categorization and other information included with each allegation record. However, despite having almost no grouping, the categories labeled in Figure 7 are not completely random, but instead show slight correlation with previous findings. For example more “Operation/Personal Violations” (red) and “Use of Force” (green) complaints appear in the upper clusters of this figure, indicating that these two categories often coincide with sustained allegations, as shown in Figure 6.

To sum up and to link these findings back to our original proposed project theme, we find that complaint summaries that indicate officer negligence or deeper more serious behavior are more likely contribute to the sustainment of those complaints, while summaries that indicate more minor, often traffic-related complaints are less likely to lead to sustainment. Although such a finding seems fairly intuitive, the fact that there is such a clear division with this trend (such as shown in Figure 1) whereas this is not the case for more concrete information stored in the database (such as shown in Figures 2 and 7), suggests that complaint summaries may be weighted much more heavily than other associated information when determining whether they should be sustained.

One likely area of future work with this project, specifically checkpoint 5, might be to examine the performance of multiple models in making predictions. Since decision trees and even random forests are relatively simple models, would a more complex model achieve better performance? Alternatively, since there is such a strong division between complaint summary topics and whether those complaints were sustained, perhaps a simpler model would perform just as well. Another area of future work might be to use these summaries to predict another metric, such as an officer’s number of future uses of force or number of future sustained allegations. Finally, one more potential direction for future work would be to improve the modeling of the data, such that the data that is used to make a prediction of whether a complaint will be sustained, only includes information up until when that complaint was filed.