Data Science Seminar - MSAI 339

Checkpoint 5: Natural Language Processing

December 2nd, 2021

Professor

Jennie Rogers

Students

Aleksandr Simonyan

Dimitrios Mavrofridis

Donald Baracskay

Introduction

For the natural language processing aspect of our project, we compared the text section of the complaint reports (cr_text) against the related complaint information. In detail, we observed the relationship between the topics in cr_text, the location of the allegation, as well as the race of the complainant.

Findings & Results:

Methodology & Data Preprocessing:

In our query, we first get a collection of rows from the complainant table where the subject's race is specified. We then join this with the allegation table and filter out entries where cr_text is either blank or null. Finally, we resolve the location of the complaint report to a community. In total, this gets us the race of a complainant, the location where the complaint report was filed and the text of the complaint report.

To process the text of the complaint reports, we implemented the NLTK library in order to remove stop words and punctuation/illegal characters from the text. Additionally, we implemented the BERTopic library into our Google Colab notebook, to perform topic modeling on the clean text. For the removal of stop words, we used the reference file "Structural Topic Models.rmd" from the github repository of CPDP, as well as a set of words we observed occurred with high frequency but had little to no meaning in cr_text. As a result, we were able to convert messy and unreadable text with redundant sentences into meaningful phrases that were integrated into our BERT model. For the sake of this project, we decided to focus on two factors: race and community, which are going to be extensively analyzed in the next section.

Race:

When it comes to the relationship between the complainant race and the topics that were generated using a custom-made BERT model, we have identified a great number of interesting trends. The first is somewhat concerning: for both the black and hispanic populations, the top topic clusters were both those containing the words punched, kicked, and handcuffed. While these words are relatively common topics identified by our BERT model, they do not appear at anywhere near the same frequency for the white and asian populations. (It is important to acknowledge that Native Americans also appear in the generated data. However, they are so few in number that their complaints likely do not allow for a proper sampling.) With both the black and hispanic populations, the next most common complaint topic deals with inventory. Upon some inspection, we determined that inventory refers to the police holding evidence (and personal property) when they are supposed to return the objects to their proper owners. For white and asian complainants, however, the most common topic is rude or unprofessional tone. Complaints by Asians also have a relatively high frequency of the aforementioned punched and kicked topic.

Community:

Due to the relative lack of data and the large number of communities, the analysis by community is somewhat sparse when compared to the race. However, we still have pinpointed a few trends. For the first of said trends, we would quickly review our Checkpoint 3, in which we created a visualization of TRRs per community and filtered based on race. Notably we found that the community Austin had the greatest number of TRRs. For NLP, the topic that occurs most frequently in Austin is the aforementioned punched, kicked and handcuffed. This makes sense, and could be considered obvious, but is important to note. This trend continues for other communities with high TRR counts, and possibly illuminates a higher use of force in these communities. Another pattern is the somewhat common topic of indebtedness. Indebtedness here refers to debts to the City of Chicago, and apparently approximately 3.4% of all police officers owe money to the city. (City of Chicago, 27 Nov. 2021) This topic, however, does not appear for all communities and doesn’t seem to correlate with TRRs. We would hypothesize that

these complaints are more common in areas with lesser income. In areas without the prior two tendencies, the topic was typically rudeness or unprofessionalism.

Conclusion:

Our findings lead to a number of fairly evident conclusions. The first of which is that the black and hispanic populations both have much higher frequencies of complaints in which they described being injured or having things stolen by the police. This very possibly points to racism in the police force, but we do not have the data required to substantiate this claim. These complaints also link up with our Checkpoint 3 visualization, in which we note that the TRRs of certain communities like Austin have largely black or hispanic as the subject race. More events involving police violence occur in these districts with higher black and hispanic populations and thus there are more complaints about being injured by the police. When you compare to the frequency of topics in the white population, the injury complaints are still there, but they represent a much smaller percentage, losing out to allegations of rudeness. This indicates that whites have much different experiences with the police. Asian experiences appear to lie somewhere in between based on our topic modeling, but whites definitely seem to have the least injurious encounters as well as the fewest theft complaints.

For communities, apart from the high incidence of complaints of injuries in communities with high TRR rates, the data is a bit too sparse to make any real conclusions. It is possible to draw some inference from things like indebtedness to the city, but some communities only have a couple of complaints. These low numbers make it difficult to say whether or not the topics that come from them are truly important. For example, the community of Montclare only has two relevant complaints and their topics are detainment and falsely arrested. We could attempt to say that maybe this means that Montclare has a problem with false arrests, but with a sample size of two this is pretty much just guessing. Thus, we would say the community complaints mainly help illustrate the effects of having many TRRs and possibly could give some information on each community's socioeconomic status.

Experience with BerTopic & Google Colab:

For this project, we started with just Bertopic. It worked fairly well and it was easy to generate topics, needing only to pass a list of documents to BERT in order to get the topic of each document and the most common topics across the whole corpus. However, we had three issues with Bertopic: many topics contained useless information and many stop words, the clustering algorithm was relatively slow (taking around a minute for each run), and the visualizations of topics seemingly didn't work. We resolved the first issue by preprocessing the documents with NLTK and removing stop words. We imported our code to Google Colab to improve the speed, as Colab allows the use of GPUs, which should speed up clustering algorithms. The final problem is left unresolved. We tried importing different visualization libraries and hoped that Colab might be able to properly display the supposed chart, but the program simply passes over the lines referring to visualization as though they are not there...

Work Cited:

City of Chicago. "Employee Indebtedness to the City of Chicago: City of Chicago: Data Portal." *Chicago Data Portal*, 27 Nov. 2021, https://data.cityofchicago.org/Administration-Finance/Employee-Indebtedness-to-the-City-of-Chicago/pasx-mnuv.