Natural language processing

Introduction
In this checkpoint, we extend our analysis further by trying to uncover hidden insights relevant to firearm cases in the text of the CPDB. By skilfully exploiting the enormous amount of free-form data, we try to uncover hidden characteristics of firearm cases and exhibit a more nuanced model of the text.

Tasks
In this checkpoint, we tackle these questions-

1. Are we able to find unseen patterns/trends related to firearm cases that we could not extract quantitatively ?
2. Do we see a clear differentiation between themes of firearm cases vs non-violent ones?
3. Do the sentiments of firearm cases vary substantially from the rest of the cases?

For example, data_allegation table consists of 12,392 summaries of allegation complaint. Each record consists of a detailed summary of the relevant allegations.

Data preparation
On analysing the text present in the CPDB, we found that summaries in the data_allegation table made much sense to our central theme. It consisted of summaries involving firearms. On preliminary analysis of the summary column in the data_allegation table, we found that most of the rows had missing values. Then, on further analysis of the same table we found that there was a column “cr_text” which also contained complaint reports. Using sql operation, we combined the 2 columns(summary, cr_text) into a single column for natural language analysis

EDA and data pre-processing

1. **Wordclouds** - In the exploratory data analysis(EDA) part, we visualized the entire raw text using wordcloud. Word clouds are visual representations that give greater prominence to words that appear more in the entire corpus. We did this step to understand the overall theme of the summaries.

Figure 1: Wordcloud of raw text

From figure 1, we can see that words like allegation, complaint, report, finding, victim are the prominent words in the corpus.

2. **Missing values -** Next, we identify the records with missing summary reports. The corpus contains 1130 missing records. These are omitted since we cannot apply natural language processing on them.

3. **Filtering firearm records** - Since our motive is to exploit firearm related reports, we segregate them using the keywords like firearm, gun, shot, etc present in the summary reports. This resulted in extraction of 575 records

```
train[train['summary'].str.contains('firearm|gun|fire-arm|shot|shooting|shoot|injured|fire arm|
```

4. **Text preprocessing -** In the preprocessing part, we tokenized the sentences and removed the noise in the raw data. First, we converted the sentences into words, built

bi-grams and trigrams, removed stop words and lemmetized the words into meaningful stem words.

Model building

LDA classifies text in the corpus into particular topics. It builds a topic per document model and words per topic model. It also assumes that each document is produced from a mixture of topics. Those topics then generate words based on their probability distribution.

The LDA topic modeling requires 2 important parameters

1. **Term document frequency** - It is a statistical measure that evaluates how relevant a word is to a document in a corpus. For example, the word ‘vehicle chase’ appears many times in a document, but fewer times across other documents, it probably means that it's an important word.
2. **Vectorization** - Since words cannot be fed into the model directly, we vectorize them using a bag of words approach.
3. **Hyperparameters** - At each iteration, we analysed 200 documents. We initially clustered the 575 summaries into 4 topics. We ran the model for 200 epochs.
4. **Initial results** - Figure 2, shows the top keywords in each of the topics along with their relative importance. On analysis of the result, we found that LDA extracted some of the common words like allege, accuse, officer, gun, intake, initial from the corpus. These words did not make any sense to our central theme. We removed these words from the corpus by adding them ('report', 'accuse', 'allege', 'party', 'allegation','officer','accuse','initial','intake',"alleg", "accus", "offic", "chicago", "parti", "report","complain", "polic")) to the stop words list and reran the model.

```
[(0,
  '0.029*"allegation" + 0.026*"accuse" + 0.025*"officer" + 0.021*"weapon" + 0.020*"intake" + 0.018*"initial" + 0.018*"report" + 0.015*"fail" + 0
0.013*"police"'),
 (1,
  '0.065*"report" + 0.062*"party" + 0.049*"allege" + 0.046*"officer" + 0.022*"allegation" + 0.022*"intake" + 0.021*"initial" + 0.021*"state" + 0
0.018*"male"'),
 (2,
  '0.070*"allege" + 0.064*"accuse" + 0.063*"officer" + 0.050*"complainant" + 0.030*"gun" + 0.030*"allegation" + 0.028*"intake" + 0.024*"initial"
y" + 0.019*"report"'),
 (3,
  '0.044*"allege" + 0.037*"accuse" + 0.027*"party" + 0.023*"report" + 0.018*"arrest" + 0.017*"allegation" + 0.014*"warrant" + 0.013*"intake" + 0
+ 0.010*"possession"')]
```

Figure 2: Preliminary results of the LDA model

topic_keywords	document_num
allege, accuse, officer, complainant, gun, allegation, intake, initial, party, report	0.0
report, party, allege, officer, allegation, intake, initial, state, accuse, male	1.0
allege, accuse, officer, complainant, gun, allegation, intake, initial, party, report	2.0
report, party, allege, officer, allegation, intake, initial, state, accuse, male	3.0
report, party, allege, officer, allegation, intake, initial, state, accuse, male	4.0

Figure 3: Keywords extracted by LDA in each document

Results

Result 1: LDA outputs the words that make up each hidden topic along with its probability distribution.

```
[(0,
  '0.043*"search" + 0.039*"complainant" + 0.032*"enter" + 0.031*"justification" + 0.031*"arrest" + 0.026*"weapon" + 0.024*"gun" + 0.023*"find" + 0.021*"residence" + 0.021*"victim"'),
 (1,
  '0.054*"weapon" + 0.022*"fail" + 0.020*"duty" + 0.016*"find" + 0.013*"itis" + 0.012*"police" + 0.011*"secure" + 0.011*"vehicle" + 0.011*"handcuff" + 0.011*"hour"'),
 (2,
  '0.075*"reporting" + 0.033*"state" + 0.027*"victim" + 0.019*"fail" + 0.017*"arrest" + 0.017*"male" + 0.016*"police" + 0.013*"white" + 0.013*"go" + 0.012*"offender"'),
 (3,
  '0.036*"complainant" + 0.034*"vehicle" + 0.024*"police" + 0.021*"state" + 0.019*"find" + 0.019*"gun" + 0.017*"enter" + 0.017*"none" + 0.015*"reporting" + 0.015*"call"')]
```

Figure 4 : Words in a particular topic and its distribution

Result 2: Since it is important to know why a certain document is clustered into a particular topic, we displayed the keywords generated by LDA in each of these topics.

topic_keywords	document_num
weapon, fail, duty, find, itis, police, secure...	0.0
reporting, state, victim, fail, arrest, male, ...	1.0
reporting, state, victim, fail, arrest, male, ...	2.0
search, complainant, enter, justification, arr...	3.0
complainant, vehicle, police, state, find, gun...	4.0

Figure 5: Keywords generated by LDA for each document

Also, by LDA assumption, we know that each document is a mixture of topics. We categorized the documents by the dominant topic.

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	1.0	0.8782	weapon, fail, duty, find, itis, police, secure...	[inattentive, duty, fail, weapon, go, fitting,...
1	1	2.0	0.6943	reporting, state, victim, fail, arrest, male, ...	[reporting, several, uniformed, plainclothe, r...
2	2	2.0	0.9938	reporting, state, victim, fail, arrest, male, ...	[reporting, male, black, uniformed, respond, r...
3	3	0.0	0.6852	search, complainant, enter, justification, arr...	[reporting, unknown, male, plant, drug, try, c...
4	4	3.0	0.5597	complainant, vehicle, police, state, find, gun...	[reporting, entered, residence, justification,...
5	5	1.0	0.6531	weapon, fail, duty, find, itis, police, secure...	[duty, discharge, firearm, victim, justificati...
6	6	3.0	0.9661	complainant, vehicle, police, state, find, gun...	[reporting, approach, passenger, side, vehicle...
7	7	2.0	0.8042	reporting, state, victim, fail, arrest, male, ...	[victim, enter, ciresiit, case, reporting, off...
8	8	1.0	0.9710	weapon, fail, duty, find, itis, police, secure...	[fail, properly_secure, weapon, discover, weap...
9	9	3.0	0.9928	complainant, vehicle, police, state, find, gun...	[reporting, partner, hispanic, uniform, possib...

Sentiment analysis

We also went a step further and analysed the sentiment of firearm and non-firearm cases. Using sentiment analysis, we wanted to see if the sentiment of non-violent cases varied substantially in comparison to violent-cases. Using sentiment intensity analyzer, we classified the summaries into 5 categories- strongly negative, negative, neutral, positive and strongly positive. Then we visualized the results using a bar graph.

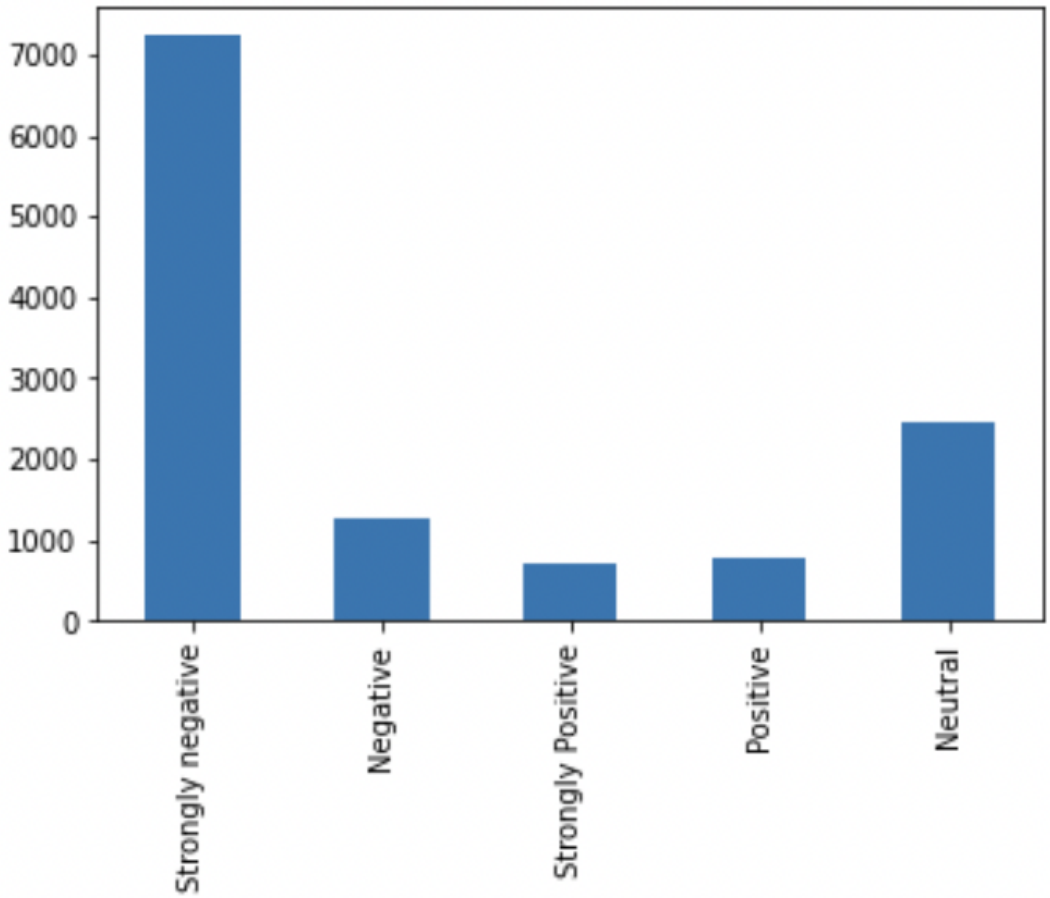

Figure 6: Sentiment of non-firearm cases

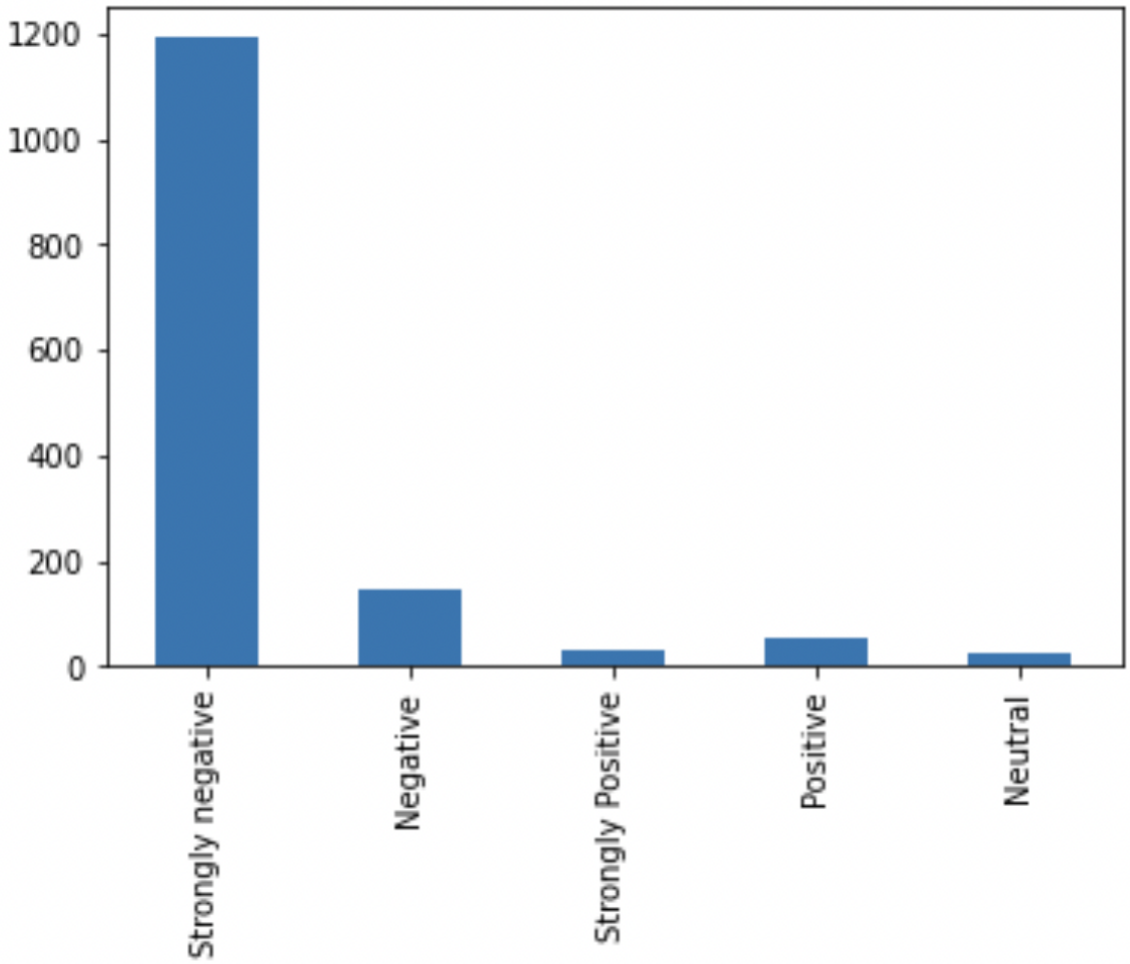

Figure 7: Sentiment of Firearm cases

Interpretation of the results

1. From interpretation of the results, some of the topics clearly pointed out a few patterns such as in topic 0, we have terms like search,warrant, enter, threaten, house, apartment, drug, etc. These indicate that many firearm cases occured during home search by officers.

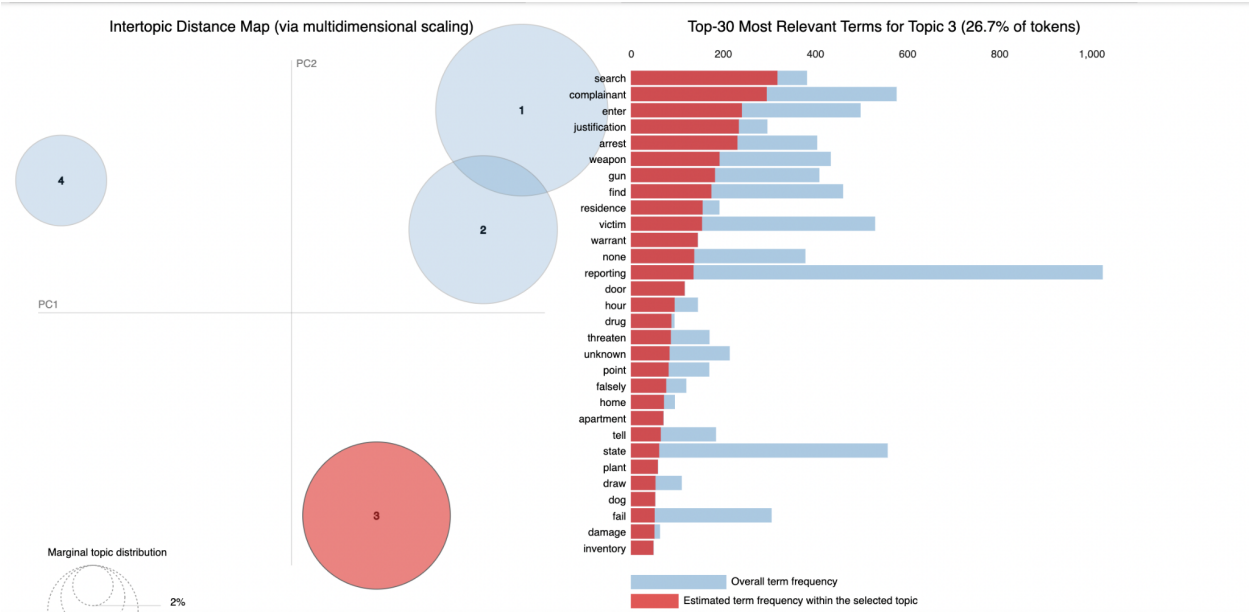

Figure 8: On the left, we can see the spatial distribution of the 4 topics. On the right, we see the top-30 relevant terms in topic 3

2. In topic 2, we found another pattern with respect to chase, vehicle, car. One of the possibilities is that the firearm cases might have happened during a vehicle chase.
3. Another key insight we gathered pointed to the term “white” which was particularly new. This led to the question if these records point to whom? On manually reading these summaries we found that it mostly represents the white officers who were involved in the firearm cases.
4. Male was another key term produced by LDA which confirms with our previous analysis that mostly men were involved in the firearm cases.

Figure 9: Word Clouds of the 4 topic clusters

5. Sentiment analysis results: Since we deal with complaint reports, it is obvious that most of them should be strongly negative/negative. The sentiment results of firearm related cases produced the same. But on analyzing non-firearm cases, we gathered many cases that were strongly positive and neutral. Around 1800 cases were positive and 2700 cases were neutral, which is uncommon.

Data Challenges and constraints

Although we analyzed the text in several tables like data_allegation, law_suit, data_attachment, the results pointed to the same themes like residence, vehicle, white officers. No prominent results with respect to demographics of the trr_reports were found. Also, the CPDB data contains a lot of duplicate and missing values, which again increases the outliers and noise in the data.

Conclusion

Understanding large corpus of unstructured text remains a persistent problem in today's data driven world. Unlike information retrieval, where users know what they are looking for, sometimes users need to understand the high-level themes of a corpus and explore documents of interest. We started topic modeling as a tool to decode the drivers and triggers of the police department. We were able to extract important qualitative insights which could not be extracted from tabular data. But on the other side, we faced several data challenges due to noise, duplicates and sparse distribution of data. Topic models offer a formalism for exposing a collection's themes and can be used to aid information retrieval and discover political perspectives.