PROJECT REPORT

DATA SCIENCE SEMINAR - FALL 2021

The Majestic Roadrunners

Saumya Singh

Geethanjali Vasudevan

Shubham Shahi

THEME:

‘Who shot first?’

A *closer look at police officers who have fired their weapons on duty*

In the light of a continued effort to analyze the data provided by the Chicago Police and create insight into the increasing crime rate in Chicago, our class partnered with Invisible Institute. Our team started analyzing the notion that there were only a few bad apples in the Chicago Police which were bringing a bad name to the entire police department. But later when we started examining the Tactical Response Report (TRR), there were shocking revelations on the number of misconducts happening around the city. What started out as a question to find the pattern in bad apples totally changed to a belief that there is a serious problem with the way these misconducts are being reported at an alarming rate. However, time constraints and data limitations pushed us to narrow down our research to have a closer look at the police officers who used firearms. Going through the TRR there was a clear pattern on the races involved in most of these firearm cases which made us question the community bias, which can be a major reason for these firearm cases. There is also a huge possibility that these reported firearm cases can be unlawful at times. We also tried to take into account different factors like the different types of force types, subject, and officer injuries. In our analytics, we aim to explore the potential of systemic racism, the distribution of crime using firearms among different units, and the proportion of subjects injured to that of officers injured in the use of firearms.

CHECKPOINT 1: Relational Analytics

In Relational Analytics, we aim to identify the factors in different crime scenes that led to the use of firearms. We tried to answer the following questions- What were the most common types of force (Physical force, verbal command, etc) exhibited by the officer that ended up in the use of the firearm? What is the proportion of subjects injured to that of officers injured in the use of firearms? Based on Demographics (Age, Gender, and Race), which group of subjects are most likely to be involved in firearms? What proportion of the subjects carried weapons in the events that involved the use of firearms?

10 rows

	firearm_used	force_type
1	1022	Firearm
2	879	Member Presence
3	856	Verbal Commands
4	270	Other Force
5	119	Physical Force - Stunning
6	49	Physical Force - Direct Mechanical
7	48	Physical Force - Holding
8	12	Chemical
9	3	Taser
10	3	Impact Weapon

Table1: Firearm is set to True, with the total number of cases grouped by the ForceType

From our analysis, we concluded that the most common force type was Firearm(1022), Member Presence(879), Verbal Commands(856), and other Physical Forces. A lot of methods were tried but still, the maximum cases stated in Tactical Response Report (TRR) involves the use of Firearm.

2 rows

	officer_injured	firearm_used
1	false	841
2	true	185

Firearm set to true, Officers Injured

2 rows

	subject_injured	firearm_used
1	true	666
2	false	358

Firearm set to true, Subject Injured

From the data obtained from the trr_trr table, we analyzed the ***78% of the times Subject is getting injured to that of 22% of Officers.*** This indicated that among the firearm cases, the weapon was used by the officer in almost 90% of the scenarios.
Based on the demographics, we concluded that the young Black males were involved in maximum misconducts. Even a significant number of firearm cases were also from the age group of less than 25, which compels us to think that the use of illegal weapons should be monitored closely in the Country.

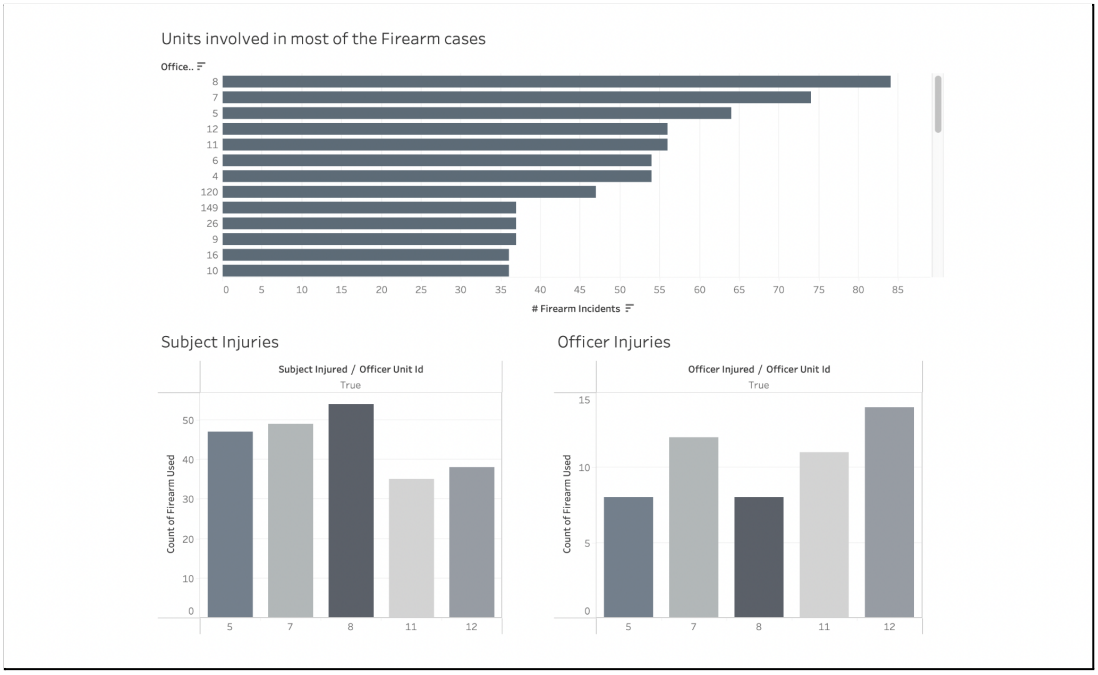
5 rows

	subject_race	firearm_used
1	BLACK	780
2	HISPANIC	159
3	WHITE	64
4	<null>	13
5	ASIAN/PACIFIC ISLANDER	10

We also concluded in our analysis for Relational analytics that in almost 80% of the cases the subject was armed. At the end of Checkpoint 1, we were left with a few open-ended questions like are there really few bad apples or we are not able to see the reality of both sides.

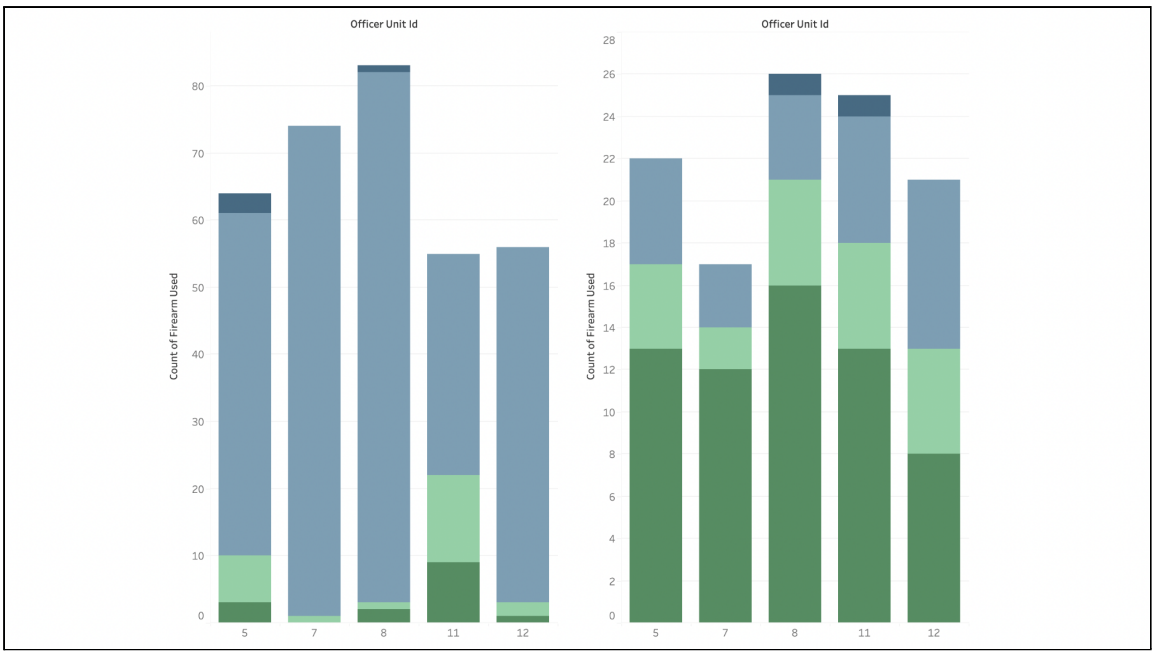
CHECKPOINT 2: Data Exploration

For this checkpoint, our main aim was to consider the distribution of misconduct among the several units. This checkpoint is also to identify if there is a particular demographic where these incidents are common.
We tried to answer the following questions in this checkpoint- Identify the Units that are most involved in all of the firearm cases, along with a comparison of Subject injuries vs officers injuries. Compare the results with other force types to narrow down on the analysis. Vertical bars to visualize the distribution of races across the firearms incidents. Treemap to compare the locations for each of the five Units where the firearm incidents have happened frequently.



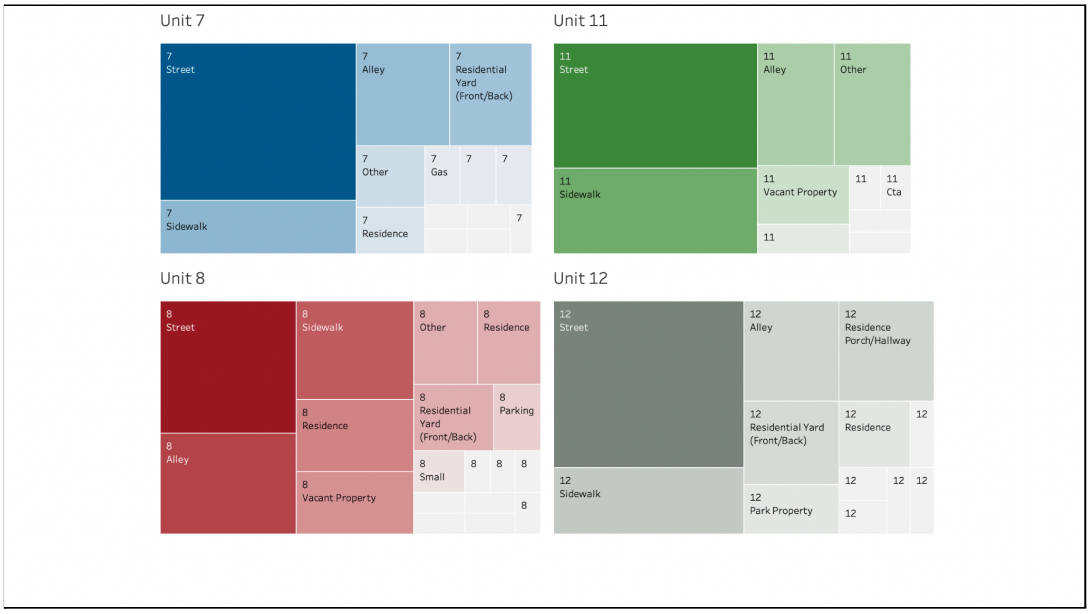
Subject injuries vs Officers Injuries

We analyzed from this checkpoint that the units that were involved were *Unit-8 (84 cases), Unit-7 (74 cases), Unit-5 (64 cases), Unit-11, and Unit-12 (56 cases each).* When then tried to filter the subject and officer injuries for these units. The subject injuries in Unit 8 were much more than the officer injuries.



Race distribution of the officers and the subject involved in the firearm cases

From further analysis, we also observed observe that the subjects mostly belong to the black community whereas the officers that are on duty are mostly from the white community. This clearly indicates a community bias.
Next, we also worked on a treemap for the demographics of the location in which the misconduct happened. From the treemap, we observed that most of the time misconduct happened in the streets in good lighting conditions.

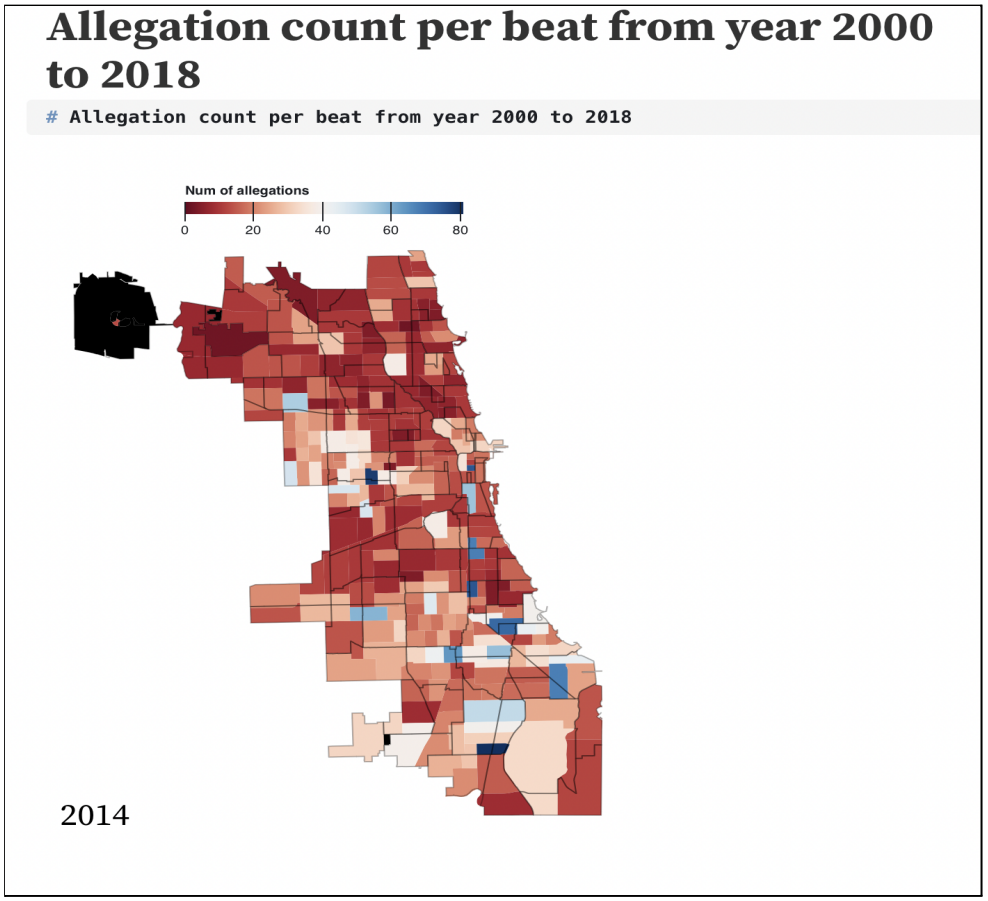


Neighborhoods where most of the firearm cases took place

This led us to the conclusion that the maximum use of firearms happened in the poor areas of Black neighborhoods.

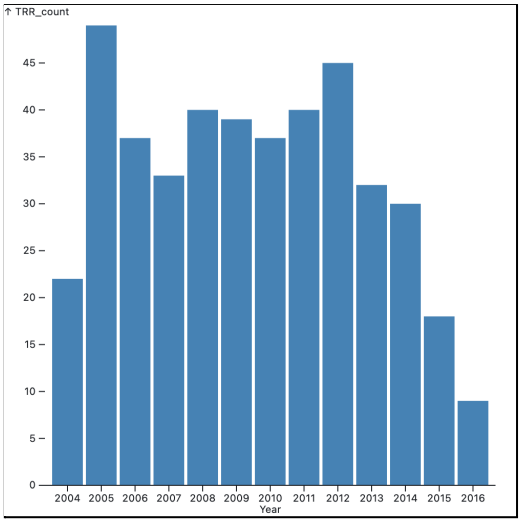
CHECKPOINT 3: InteractiveVisualization

For this checkpoint Interactive visualization, our goal is to iterate our analysis from the previous checkpoints by examining the trend in the number of allegation counts spread across different beats over the period of 10 years. We tried to answer the following questions- To use a choropleth map with a sliding time window and visualize the TRRs coming from per capita. Using horizontal bars and a search option for the race to identify districts from where the TRR percentage is higher is occupied predominantly by a particular race.

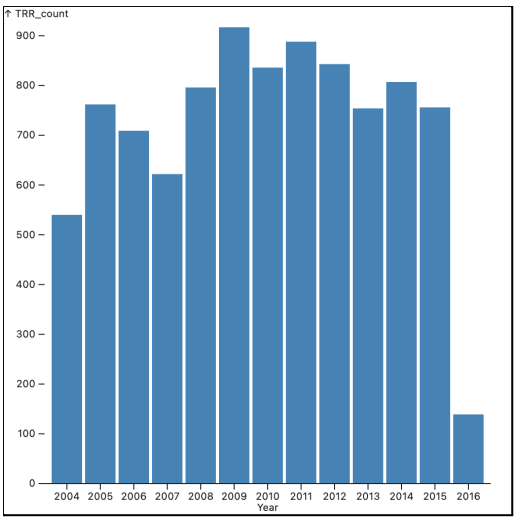


Link: *https://observablehq.com/d/284ee7694c5204e6*

We concluded from the map, that there is a consistent decline in the number of allegations from 2000 to 2018. But interestingly there is one beat(#129) in the center that is consistent and is a hotspot.



Allegation count relating to Asians



Allegation count relating to Hispanic

Link: *https://observablehq.com/@3d648a34c9857641/q2*

We also made an attempt to plot districts from where the TRR percentage is higher is occupied predominantly by a particular race. It can be concluded that the subjects for most of the time belong to the Black community, clearly indicating a racial bias.

CHECKPOINT 4: Graph Analytics

The objective of this checkpoint is

- To find a relationship between officers who are co-accused in firearm-related allegations
- To identify officers and to narrow down on the officers who are most likely to be involved in the firearm cases.

Method used - Triangle count and Page rank
Nodes - The officers
Edges - The allegations that the officers are co-accused with.
Weight of edges - The number of common allegations between the given two officers.

Challenges Faced

- Lack of common trr_id for two officers - Due to lack of common trr_id, we could not analyze officers involved in the same trr.
- From the previous checkpoints, we have observed that in all of these cases all the necessary people are being informed and being communicated to.

Solution

- We tried to extract using officers involved in the same event or present in the same datetime. But we could not find appropriate columns and event_id was related to a particular officer. So, So, we used trr_trrstatus to find officers' id and linked them to get meaningful graph analytics.

Analysis and Insights

- Triangle count - After running the different algorithms, we found that only 4 to 5 officers have a high page rank, and triangle count makes it evident that only a few officers in the department are highly connected and involved in the firearm cases.
- Page rank - We found an interesting observation, that the officers having high triangle counts are not high in the page rank table. This made us deduce that the officers though not part of a particular clique as per the triangle count, do hold prominence in the CPD. hence the allegations are not done by only a group but can be performed by a group that is being influenced by another officer.

CHECKPOINT 5:Natural Language Processing

In this checkpoint, 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. We tried to find unseen patterns/trends related to firearm cases that we could not extract quantitatively

Methods used - Topic modeling

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. 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,...

Interpretation and Insights found

1. 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 occurred during home search by officers.

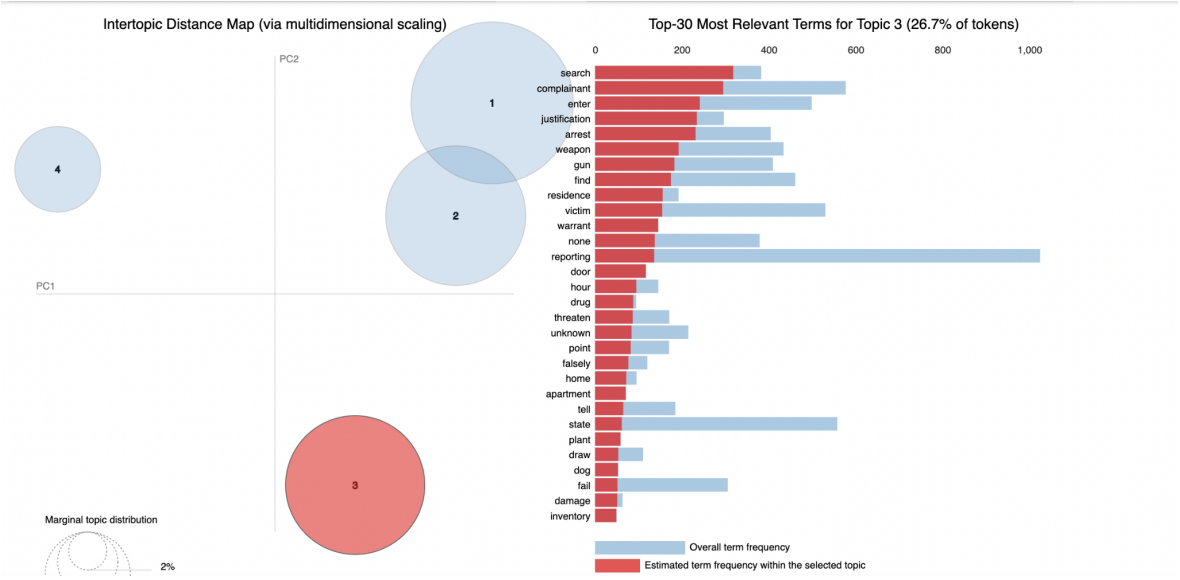


Figure 8: On the left, we can see the spatial distribution of the 4 topics. On the right, we see the top-30 relevant terms in topic 3

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 of whether 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.

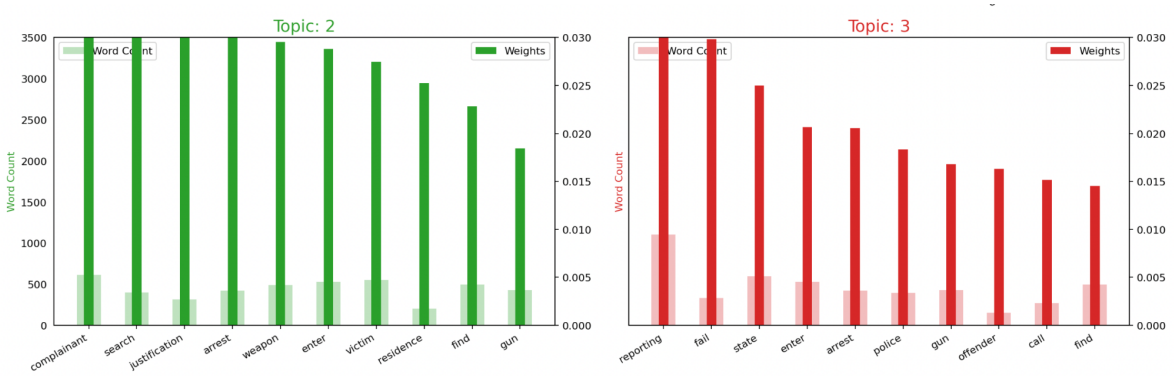


Figure 10: Topic - Word counts

Conclusion of checkpoint 5 and open questions

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. This leads us to the open question of why we could not extract proper trr_reports. Is it because the police officers are manipulating these reports that are against them or is it really a data warehousing problem?

Topic models offer a formalism for exposing a collection's themes and can be used to aid information retrieval and discover political perspectives. Using quality data, we can extract much more insights about cases that are similar to each other. This could help in a new perspective for analyzing the cases.

CONCLUSION

We started with trying to find the bad apples in the Chicago police department. Through our initial analysis, we found that there were bad apples but the data was so sparse to pinpoint the officers who were actually involved in the firearm cases. From our initial checkpoint, we found that superiors were involved, many officers were highly connected and were under the influence of other officers. On the other side of the coin, in most of the firearm cases, the subjects were armed, and incidents happened at night. This leads to the question that, were the firearm cases an act for officers to protect themselves.

We can say that not all officers are bad, but we cannot deny the saying "One bad apple spoils the bunch" meaning that one rotten apple can spoil all apples around it. The accumulation of these allegations over time is enough to signal a systematic issue. We conclude by saying that some jobs just can't have bad apples. In these jobs, everybody has to be good. Regardless of anything, policing has to change and large-scale reform must be taken.