Exploring the Relationship between Complaint Narratives and Categories

The Enchanted Badgers
Alexander Einarsson, Sergio Servantez, Marko Sterbentz
December 1, 2020

1. Introduction

Monitoring complaints filed against the police offers a rare glimpse into the inner workings of the police force who have been tasked with protecting and serving the public. It is clear that complaints and their categorization play a meaningful role in being able to predict the likelihood of future misconduct on the part of police officers [1]. However, predictive capabilities are only as good as the data with which the models are built. Making sure that the categorization of complaints filed against the Chicago Police Department is both complete and meaningful is of critical importance for ensuring the analysis of the data provided by the Invisible Institute is as correct as possible. We posit that confirming that the existing categorizations are meaningful and accurate, as well as assigning meaningful categories to unknown and uncategorized complaints would be a valuable asset for those trying to analyze the data. As a result, the theme of our project was to identify meaningful categories for uncategorized complaints by analyzing the relationship that exists between the complaint category and the complaint report narrative, and to use this new information to explore various aspects of complaint investigations and outcomes. In particular, we examine the capabilities of language models in predicting the categories and outcomes of complaint narratives and determine alternate categorizations derived directly from the text of complaint report narratives.

2. Data Exploration

One of our earliest findings, which served to focus our project, was the result of our querying the database about what percentage of complaints was categorized as unknown or not categorized (year over year). The query results can be seen in Figure 1, and we also generated a line graph to visualize the results as seen in Figure 2.

	year	percentage
1	2018	1.1029411764705883
2	2017	1.085883514313919
3	2016	12.940275650842267
4	2015	8.124280782508631
5	2014	4.208654416123296
6	2013	1.7169811320754718
7	2012	0.897021887334051
8	2011	0.8166533226581264
9	2010	0.5057408419901586
10	2009	0.35285815102328866
11	2008	0.24088093599449414
12	2007	0.297303036738161
13	2006	0.5994405221792993
14	2005	0.9466708740927736
15	2004	0.3922034704064654
16	2003	0.33438038301752965
17	2002	0.24187975120939878
18	2001	0.18585858585858586
19	2000	0.3485928128364177

Figure 1 — Query results showing the percentage of complaints that are uncategorized for years between 2000 and 2018.

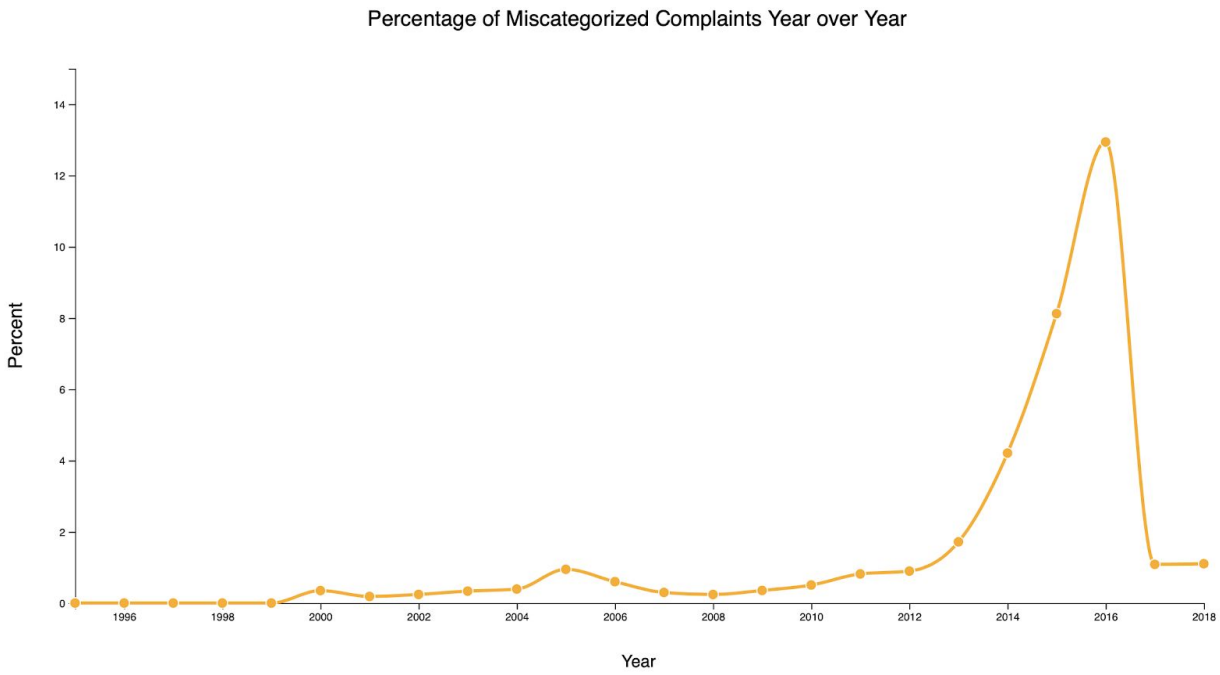


Figure 2 — Line graph showing the percentage of uncategorized/miscategorized complaints each year between 1995 and 2018.

Years occurring before 2000 all had a percentage of zero and were truncated here for brevity. While records prior to 2000 are less frequent, it still seems interesting that no complaints before this period were unclassified or classified as unknown. It's possible that new categories were introduced that complicated categorization, or maybe the intake process for complaints changed (possibly to digital records based on the time period) to a method which was more error prone.

Starting with the year 2000, we note that the Unknown/uncategorized percentage remained relatively low (sub two percent) for just over a decade. Starting in 2013, however, we see that the percentage of Unknown/uncategorized increased rapidly, to reach a peak in 2016 at almost 14 percent, over an order of magnitude more than at any point between 2000 and 2012. After this peak, the Unknown/uncategorized complaint percentage plummets, back to almost the same level as before 2013. There seems to exist a clear explanation for why that percentage would lower drastically after 2016: the Invisible Institute gained access to the CPDP data in late 2015 and started publishing articles about the police department using this data. It can be surmised that the CPD put more pressure on their officers to correctly categorize complaints as a way to avoid bad publicity after this point. Given the value in correctly classified complaints, this increased transparency into the department via the CPDP data can only be viewed as a positive.

However, the question remains why the Unknown/uncategorized complaints so drastically increased for a few years starting in 2013. We have been unable to find any evidence that there was a drastic change in the department around this time, and we lack the insight in the goings-on in the department to come up with any testable hypotheses. We recommend that the Invisible Institute do a qualitative exploration of the reports around that time to see if there was an explosion in complaints load, if the reporting of the complaints changed, or if there are other possible explanations for the sudden increase in Unknown/uncategorized complaints. Future research projects could and should test hypotheses that the Invisible Institute can form based on such a qualitative exploration project.

3. Classifying Uncategorized Complaints using a Transformer Model

In search of a quantitative way to gain insight about the uncategorized complaints, we decided to use a state-of-the-art transformer language model to predict the complaint categories for the set of uncategorized complaints. We wanted to do this to determine what complaint types were most frequently miscategorized by the Chicago Police Department, which would give some insight into whether the lack of categorization was random or not.

In order for a language model classifier to be properly trained, there needs to be a sufficient amount of training data. Within the CPDP database, there was roughly 1100 complaints that have narrative summaries associated with them. Unfortunately, this was not enough raw data with which to train these kinds of models. Fortunately, in addition to the narrative summaries that are associated with allegations in the CPDP database, there was a set of roughly 45,400

narratives external to the database that were available and provided by the Invisible Institute. After cleaning these narratives and integrating the two datasets together, we ultimately had 16,010 distinct narrative summaries and associated metadata with which to build and train our language models.

Using this richer set of narratives, we built a BERT language model [2] that was capable of predicting the complaint categories.The trained BERT model achieved an accuracy of 83.9%, indicating that we reasonably could expect that the narrative summaries could be used to predict complaint categories using this model. Given this, we used the model to predict what the uncategorized complaints should have been categorized, with the results shown below in Figure 3.

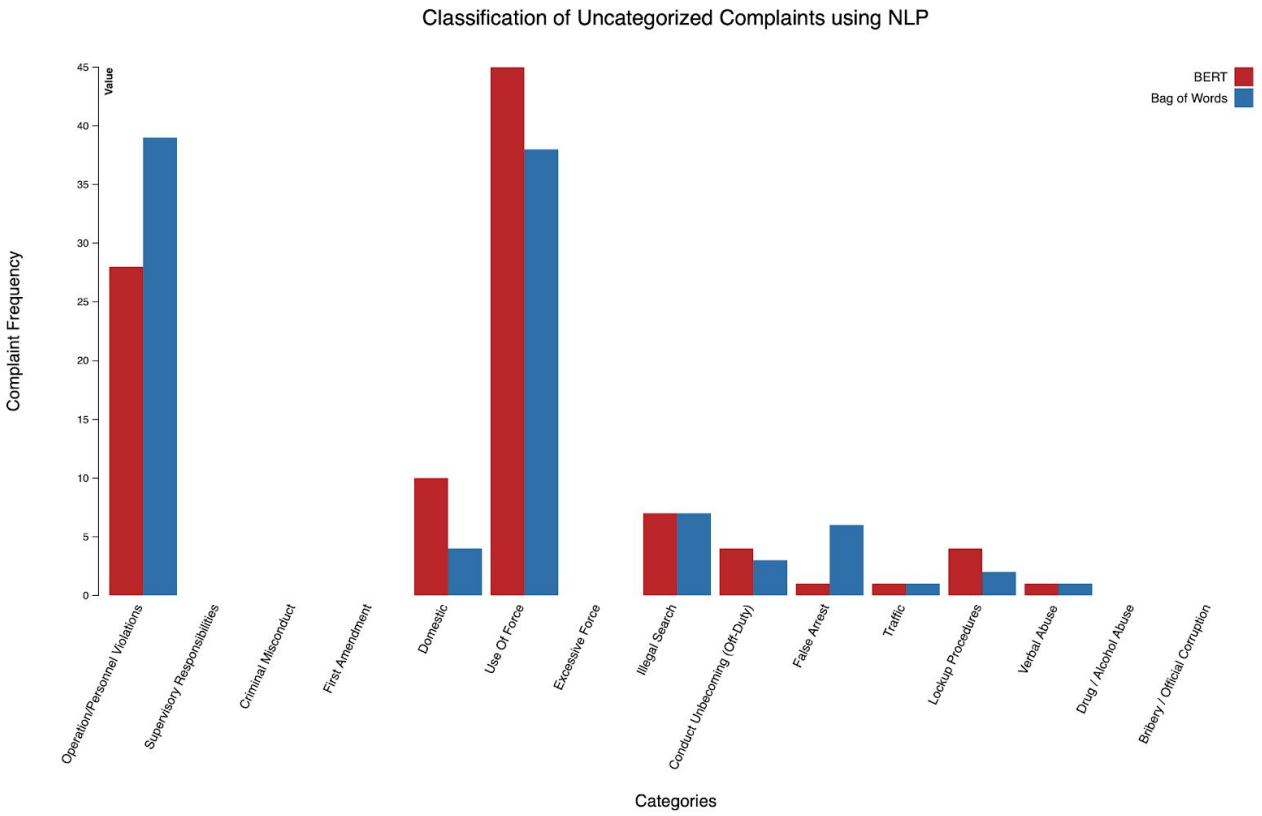


Figure 3 — A grouped bar chart showing the distribution of class categorizations predicted by our language models.

Looking at the classification results, there are a number of interesting observations that jump out. The first is the high number of uncategorized complaints that were classified as belonging to the “Operations/Personnel Violations” category. This category seems to be a sort of catch all category, and it stands to reason that a high number of uncategorized complaints would fit into this category. This could simply be the result of the person in charge of categorizing the complaint being unsure what to report as the category, but it could also be that these typically intra-department complaints are being intentionally obfuscated in order to obscure the problems occurring within the department. In order to test this, more qualitative analysis of these

complaints will be required. It is also worrisome that so many uncategorized complaints were classified as “Use of Force” by both models. While this could just be a coincidence, the disproportionate number of “Use of Force” classifications suggests that these complaints might have been intentionally miscategorized.

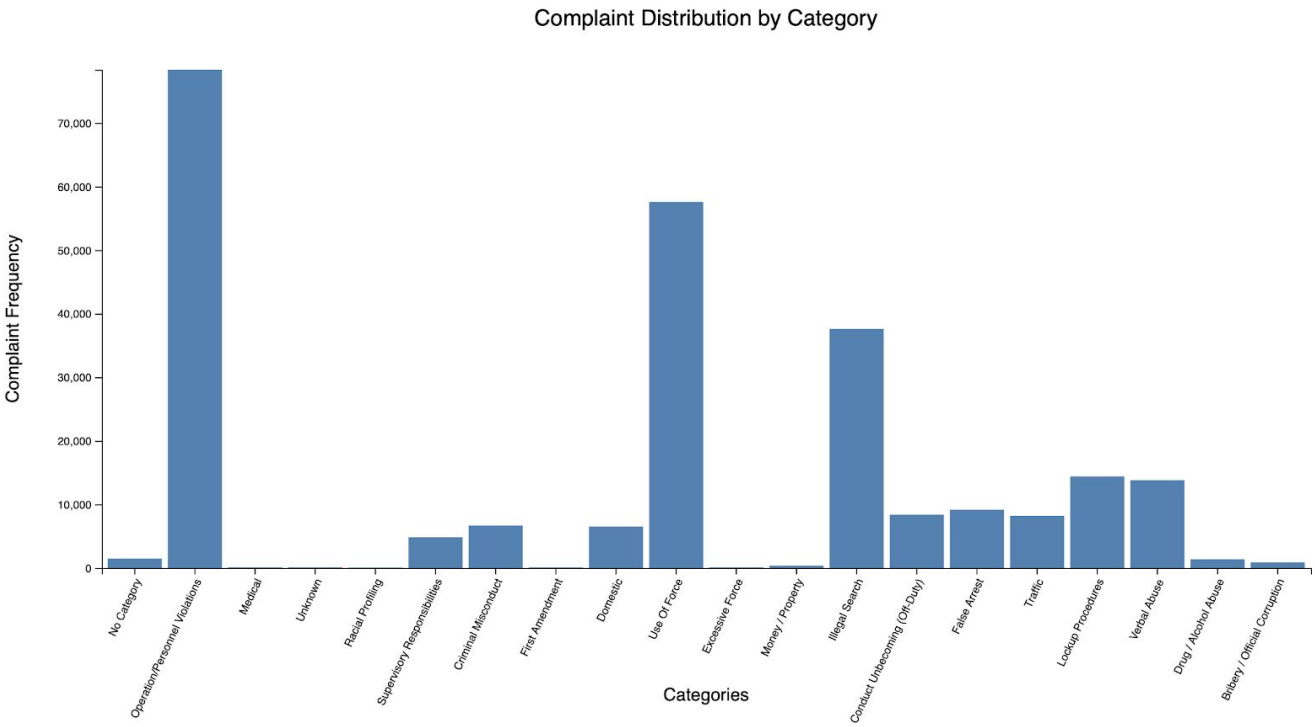


Figure 4 — Distribution of complaint categories for all complaints provided by the Invisible Institute.

Figure 4 shows the distribution of all complaints according to their categories. We can see that overall, the distribution of categories assigned to the uncategorized complaints follows roughly the same distribution as the original dataset. However, there is clearly a higher proportion of the uncategorized complaints belonging to the “Use of Force” category, which seems a little odd. Further qualitative analysis of the complaints that were classified into this category would be needed to determine whether these were intentionally miscategorized, if it was an honest mistake (i.e. multiple categories make sense), or simply a mistake by the language model classifiers. However, if it is indeed the case that these complaints were obviously miscategorized by the Chicago Police Department, this would provide compelling supporting evidence for a need to improve the reporting system citizens use to hold officers accountable. A definitive answer to this requires more work.

Given the good results from training a BERT model to predict complaint categories, we decided to explore whether it was possible to train a model to predict outcomes from complaint summaries and categories. As before, we decided to use a state-of-the-art BERT model, but unlike the previous model our results for this model were underwhelming, even after consolidating some of the smaller outcome categories to account for the discrepancy in how the

outcomes were distributed. Our BERT model only achieved an accuracy of 69.8%, indicating that the model would be unable to consistently predict what outcome a complaint would have even after it is categorized.

4. Identifying Meaningful and Novel Groupings of Complaints using LDA Topic Modeling

In addition to classifying pre-existing complaint categories, we decided to explore whether topic modeling could be used to create new categories based exclusively on their narrative summaries. We believed the commonly used Latent Dirichlet Analysis (LDA) [3] would be a good fit for our project because we were interested in different and new ways of categorizing complaints, and we wanted to explore doing so in an unsupervised way by only taking into account what the officers wrote as a summary. By letting the LDA create the topics for us in an unbiased manner, looking only at the narrative summaries, we wanted to examine both whether the created topics were interpretable for humans, and if they provided another venue for categorization for people who want to do research on the data.

We trained 10 different models on the dataset using the hyperparameters outlined above, and decided on the final model through qualitative analysis. We ranked the models based on how human interpretable their topics were according to our group members, and then created bar charts of the word frequencies in each topic for each model. Thus, our final selection of a model was based on interpretability and distribution across topics. The full set of results can be seen below in Table 1.

Topic	Topic Words											
0	make	court	altercation	investigation	physical	lockup	involved	appear	hour	evidence	seven	10
1	handcuff	face	throw	grab	ground	push	punch	head	direct	body	incident	profanity
2	duty	fail	arrestee	medical	platform	ensure	need	meet	star	sgt	detention	aide
3	vehicle	citation	license	driver	station	car	location	issue	drive	stop	state	justification
4	member	department	hour	damage	february	property	residence	unknown	search	25	il	justification
5	return	fail	inventory	property	search	usc	personal	phone	document	remove	home	inventorie
6	order	wrong	message	9	send	hour	telephone	text	protection	itis	harass	form
7	individual	mr	rude	district	place	unprofessional	unknown	handcuff	state	hospital	threaten	make
8	false	evidence	trial	provide	release	subsequently	room	tesimony	murder	process	second	04
9	allegation	recommend	sustain	department	fail	duty	finding	base	complaint	member	incident	register
10	fail	refuse	state	arrest	provide	respond	scene	offender	regard	file	911	request
11	justification	search	stop	vehicle	detain	traffic	handcuff	impound	hour	unidentified	remove	approximately
12	male	state	white	vehicle	black	uniformed	female	traffic	accident	unknown	car	fuck
13	fail	itis	city	indebtedness	hour	unit	office	presence	2300	legal	james	location
14	search	warrant	apartmewnt	enter	residence	door	floor	justification	damage	execution	plaintiff	permission
15	child	home	harass	son	date	old	verbally	state	year	abuse	daughter	court
16	use	force	improper	weapon	point	unnecessarily	search	profanity	direct	false	excessive	warrant
17	arrest	falsely	cause	probable	plaintiff	charge	drug	possession	plant	crime	commit	believe
18	fail	department	duty	false	misconduct	recommend	provide	allegation	statement	sustain	incident	notify
19	subject	civilian	approximately	suspect	county	cook	arrest	copa	1a	attempt	search	investigation

Table 1 — The set of topics and associated keywords produced by our topic modeling.

The algorithm proved largely successful in finding interpretable topics purely from the narrative summaries. Several of the topics map directly onto the categories from our original dataset, such as “grab”, “push”, “punch”, “head” clearly relating to the Use of Force category, and “search”, “warrant”, “apartment”, “enter” relating to the Illegal Search category.

handcuff	face	throw	grab	ground	push	punch	head	direct	body	incident	profanity
search	warrant	apartment	enter	residence	door	floor	justification	damage	execution	plaintiff	permission

Table 2 — The key topic words are highlighted in yellow for topics 1 (top) and 14 (bottom).

In addition to some of the topics directly relating to the original category, the algorithm also proved successful in breaking categories into multiple meaningful topics. For example, the model created multiple topics including the word ‘search’, potentially related to the Illegal Search category. For ease of visualization, we generated a hierarchical bar chart showing how each original category was redistributed into the new categories. The redistribution for the illegal search category is shown below in Figure 5.

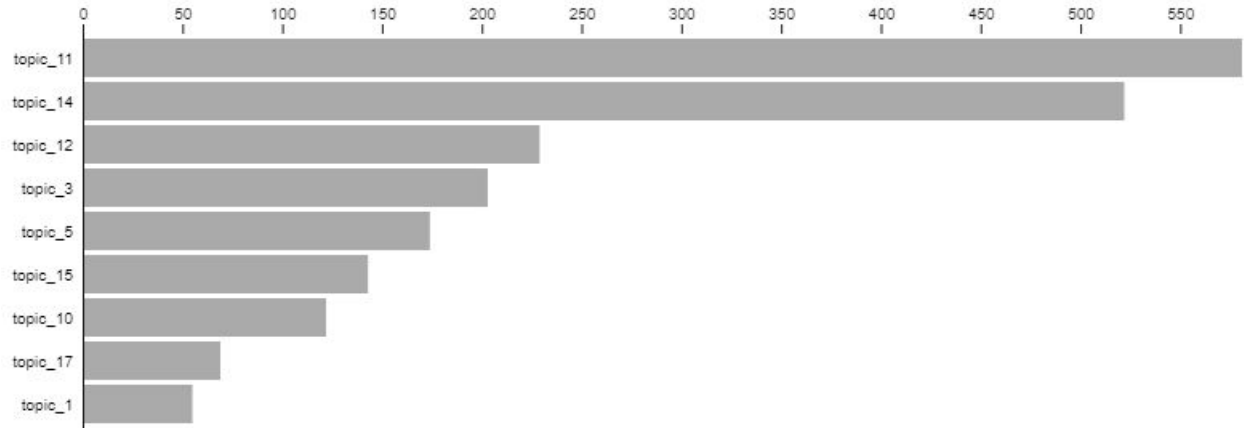


Figure 5 — The redistribution of complaints originally categorized as “Illegal Search” to various topics produced by our topic modeling.

We note that the majority of the illegal search complaints mapped onto topics #11 and #14. Looking at the words defining each topic, we saw that topic #11 appeared to relate to Illegal Search of vehicle, including words such as ‘justification’, ‘stop’, ‘vehicle’, ‘impound’, and ‘traffic’., while topic #14 seemed to indicate a relation to Illegal Search of the complainants home, containing words such as ‘apartment’, ‘enter’, ‘residence’, and ‘door’. It’s promising that the algorithm managed to break down the topics into more granular categories, as this may let researchers build categories not currently present in the larger dataset.

justification	search	stop	vehicle	detain	traffic	handcuff	impound	hour	unidentified	remove	approximately
search	warrant	apartment	enter	residence	door	floor	justification	damage	execution	plaintiff	permission

Table 3 — The key topic words are highlighted in yellow for topics 11 (top) and 14 (bottom).

We also used the final topics to examine the uncategorized category from the original dataset, and found that the 101 complaints largely mapped onto two topics, as shown below in Figure 6.

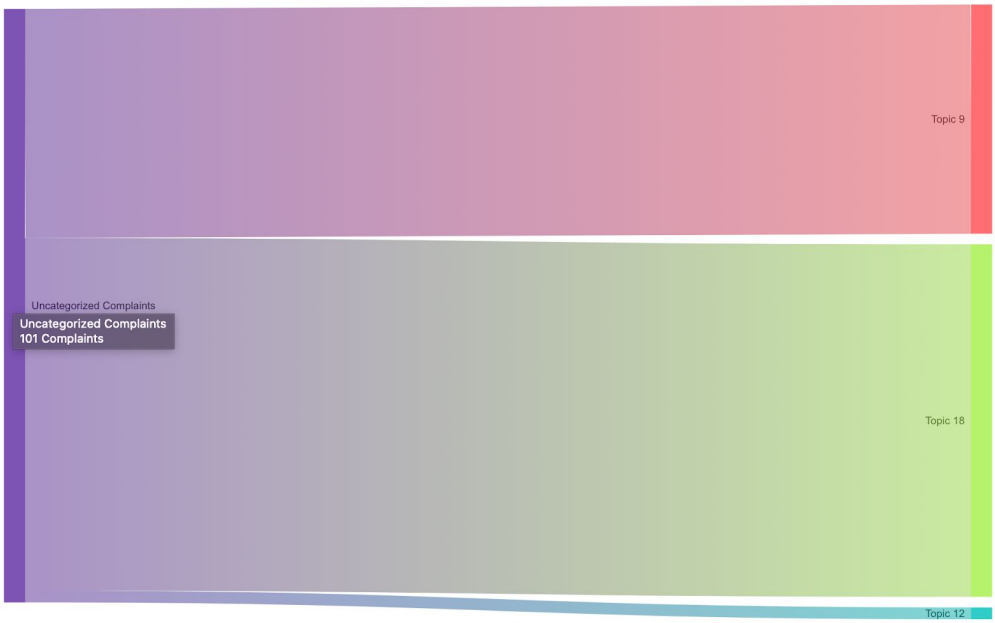
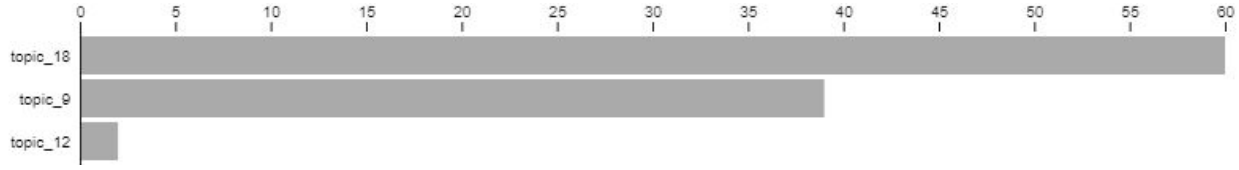
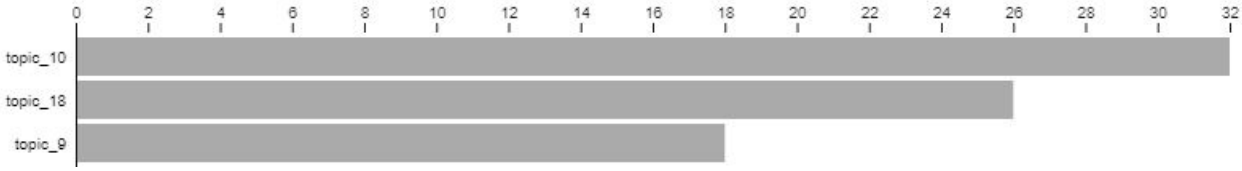


Figure 6 — Sankey diagram showing how uncategorized complaints were distributed into topics produced by our topic modeling.



Note that topics #9 and #18 also featured prominently Supervisory Responsibilities category:



Topic #10 is one of the most common topics for most original categories, but both topics #9 and #18 were less common, indicating that those topics carried more meaningful information for separating topics. The topics included words such as ‘department’, ‘misconduct’, ‘incident’ and ‘duty’, indicating that may have been department-related complaints, rather than specific officer complaints. It’s unclear whether it’s obfuscation or lack of clarity about how to file department related complaints about officers that caused this, and without an in-depth qualitative analysis about the department procedure it’s out of scope for this project to come up with any hypotheses to test. Regardless, we note it’s possible that the LDA succeeded in finding information about the unknown and uncategorized complaints where the BERT model did not.

While a single unsupervised algorithm would never completely change the landscape of the handcrafted categories, we have shown that by merely using the narrative summaries of the complaints we can craft meaningful and interpretable topics for classification, and that these topics are robust enough to create consistent mapping from the unlabeled/uncategorized category onto the algorithm-generated topics. While it doesn't replace the handcrafted and very granular categorization in the original dataset, it may serve to augment people who are interested in critically reviewing the current categories. Because of the relative lack of complaints with an accompanying narrative summary, there is also a chance that the algorithm can create more meaningful topics given more data, and that more data would also allow for a greater number of topics, which would increase the granularity.

5. Discussion and Future Work

Through the project, we found that both topic modeling and language modeling can be satisfactorily applied on the CPD data to categorize uncategorized complaints, and we believe a combination of topic and language modeling should be the next step in any future research on this topic.

On that matter, we believe that, while the model topics generated by the LDA were largely interpretable, a small number of them prove difficult to map from the words to distinct topics, at least for the authors of this report. They contain mainly words that are either departmental or filled with domain-specific words. Because of the authors' relative lack of domain expertise, we opted to accept those topics as potentially worthwhile in the model. Part of our qualitative analysis was to look at the topics through visualizations, and the topics didn't appear to be "catch-all" topics. However, we believe a primary step in future research should be for domain experts to determine the value of the current generated topics, and compare them to the topics generated by other methods.

Although the LDA yielded satisfactory, and largely interpretable, results, it is at this point a rather old algorithm, and we believe it would serve the purpose of this project well to explore the potential benefits of adapting it to use word embeddings [4] instead of words in finding novel and informative topics. As the narrative summaries consist of relatively little text, this may result in the topics carrying more information, which could improve the accuracy of our language model.

We have throughout our project found some interesting and potentially nefarious results which we have been unable to find answers for through our quantitative research. Early on in our project we identified an odd peak and immediate drop-off in uncategorized complaints which occurred around the same time as the Invisible Institute started working on the CPD data. Our work found that many of these complaints could have been categorized as either Use of Force or some form of departmental complaint. However, it remains outside the scope of our project to test the hypothesis that these complaints went deliberately uncategorized, as opposed to being

a result of change in reporting, and we would need to discuss this with domain experts to gain if not understanding then at least testable hypotheses.

Similarly, while we believe that both our language model and the LDA show that the departments may have hidden complaints that would reveal serious issues within the department, such as an excessive number of Use of Force complaints or departmental complaints, we would need to discuss this with domain experts to determine whether this may be accurate. We would also need to discuss how to find a viable hypothesis to test this.

Last, we will highlight our contributions and discuss possible paths of future work we believe have been enabled by our project. Our primary contributions are twofold. First, we demonstrated that complaint narratives have predictive power over complaint categories when using a transformer model. We were able to achieve 84% accuracy using a fine-tuned BERT classifier. Second, we demonstrated that LDA topic modeling can identify novel and meaningful groupings of complaints. Using the groupings generated by the LDA model, it would be interesting to examine the predictive power of both these groupings and the groupings provided by the LDA with word embeddings model over the complaint final outcome and compare it to that of the original categories.

6. References

[1] Arthur, Rob. (2015). “How To Predict Bad Cops In Chicago.” https://fivethirtyeight.com/features/how-to-predict-which-chicago-cops-will-commit-misconduct/

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

[3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3 (Jan), 993-1022.

[4] Changzhou Li, Yao Lu, Junfeng Wu, Yongrui Zhang, Zhongzhou Xia, Tianchen Wang, Dantian Yu, Xurui Chen, Peidong Liu, and Junyu Guo. 2018. LDA Meets Word2Vec: A Novel Model for Academic Abstract Clustering. In *Companion Proceedings of the The Web Conference 2018* (*WWW '18*). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1699–1706. DOI:https://doi.org/10.1145/3184558.3191629