Checkpoint 5: Natural Language Processing

The Silver Imps

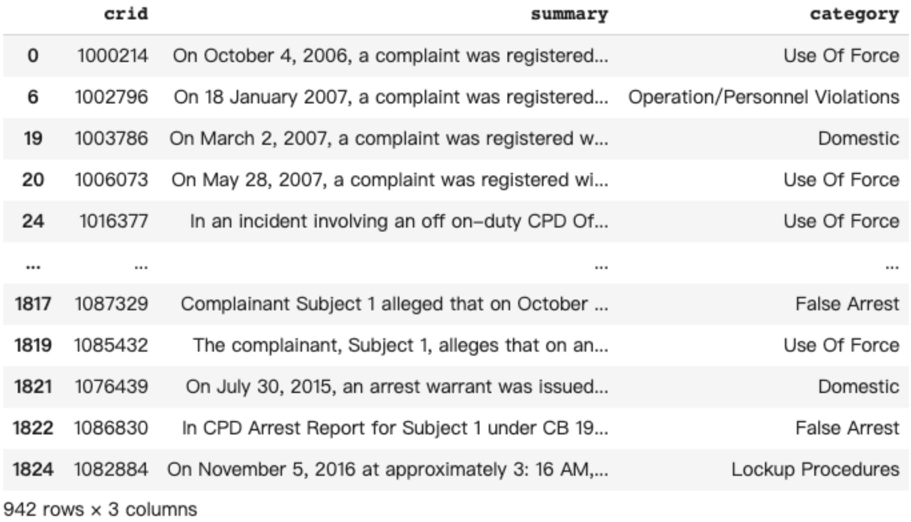
Song Luo, Jing Jiang, Yuxin Chen

Task Description

In our proposal we proposed that the for the Natural Language Processing checkpoint, we want to explore the text classification models to help classify a summary of a police misconduct to a corresponding category. We choose to use Transformer model as our architecture. Also, because our focus is on the behaviors of the “repeaters”, we decide to separate the data into two groups: summaries of allegations towards repeaters, and summaries of allegations towards non-repeaters. We also want to evaluate the trained model on the other group of data (repeater model to non-repeater test set, non-repeater model to repeater test set).

Data Preparation

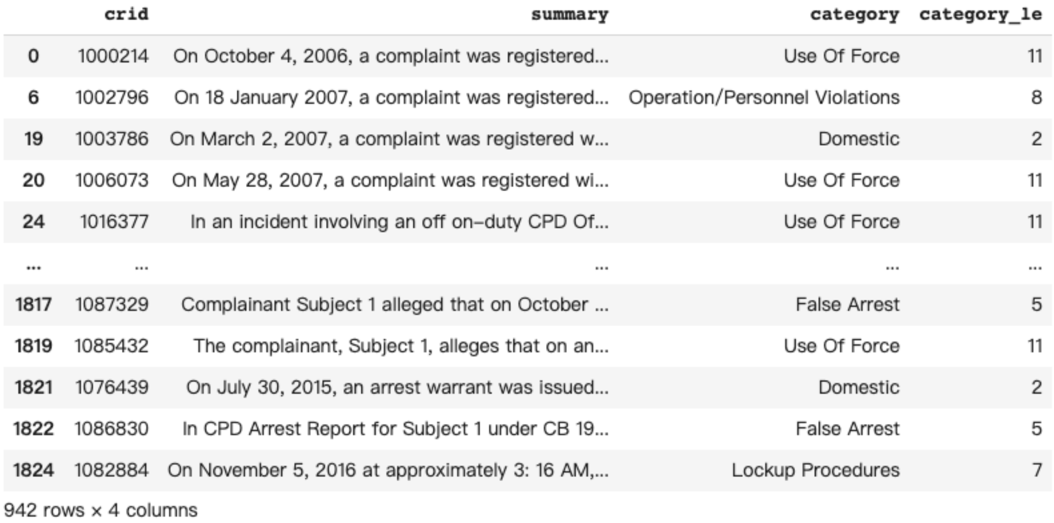
First, we use psycopg2 and pandas to query the allegations that have a non-empty summary field, and categories of the misconduct for each allegation. Then by mapping their crid field, we create the following table containing the crid, summary text, and category information.



	crid	summary	category
0	1000214	On October 4, 2006, a complaint was registered...	Use Of Force
6	1002796	On 18 January 2007, a complaint was registered...	Operation/Personnel Violations
19	1003786	On March 2, 2007, a complaint was registered w...	Domestic
20	1006073	On May 28, 2007, a complaint was registered wi...	Use Of Force
24	1016377	In an incident involving an off on–duty CPD Of...	Use Of Force
...	...	...	...
1817	1087329	Complainant Subject 1 alleged that on October ...	False Arrest
1819	1085432	The complainant, Subject 1, alleges that on an...	Use Of Force
1821	1076439	On July 30, 2015, an arrest warrant was issued...	Domestic
1822	1086830	In CPD Arrest Report for Subject 1 under CB 19...	False Arrest
1824	1082884	On November 5, 2016 at approximately 3: 16 AM,...	Lockup Procedures

942 rows × 3 columns

Because the package we use for training (simpletransformers) requires the labels being integer values, we add a label encoding to map the category value from string to integer, as shown below. In total we have 13 classes.



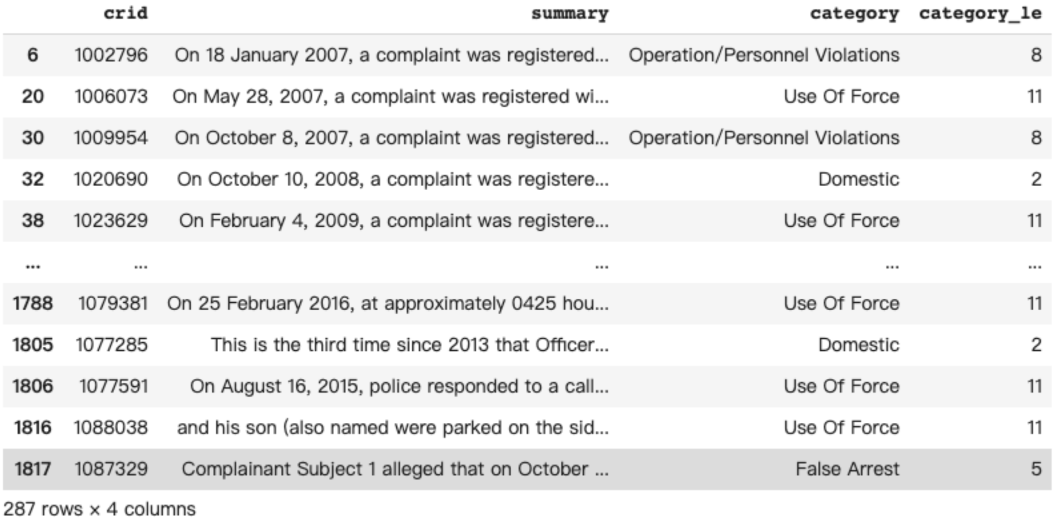
	crid	summary	category	category_le
0	1000214	On October 4, 2006, a complaint was registered...	Use Of Force	11
6	1002796	On 18 January 2007, a complaint was registered...	Operation/Personnel Violations	8
19	1003786	On March 2, 2007, a complaint was registered w...	Domestic	2
20	1006073	On May 28, 2007, a complaint was registered wi...	Use Of Force	11
24	1016377	In an incident involving an off on–duty CPD Of...	Use Of Force	11
...	...	...	...	...
1817	1087329	Complainant Subject 1 alleged that on October ...	False Arrest	5
1819	1085432	The complainant, Subject 1, alleges that on an...	Use Of Force	11
1821	1076439	On July 30, 2015, an arrest warrant was issued...	Domestic	2
1822	1086830	In CPD Arrest Report for Subject 1 under CB 19...	False Arrest	5
1824	1082884	On November 5, 2016 at approximately 3: 16 AM,...	Lockup Procedures	7

942 rows × 4 columns

Different from previous 4 checkpoints, for this checkpoint we did not divide the allegations across different timespans when querying the data. The first reason is that the summary texts are similarly structured over time. The second reason is that the data entries containing non-empty summaries are not numerous, which will lead to badly performed models after training.

However, our definition for "repeater" is consistent: the top 10% officers with most allegations. So for this checkpoint, among the allegations with non-empty summary field, we count the distinct number of officer IDs, and result is **1687**. So the top 10% definition will lead to the number of repeaters being **169**, and the number of non-repeaters being **1518**.

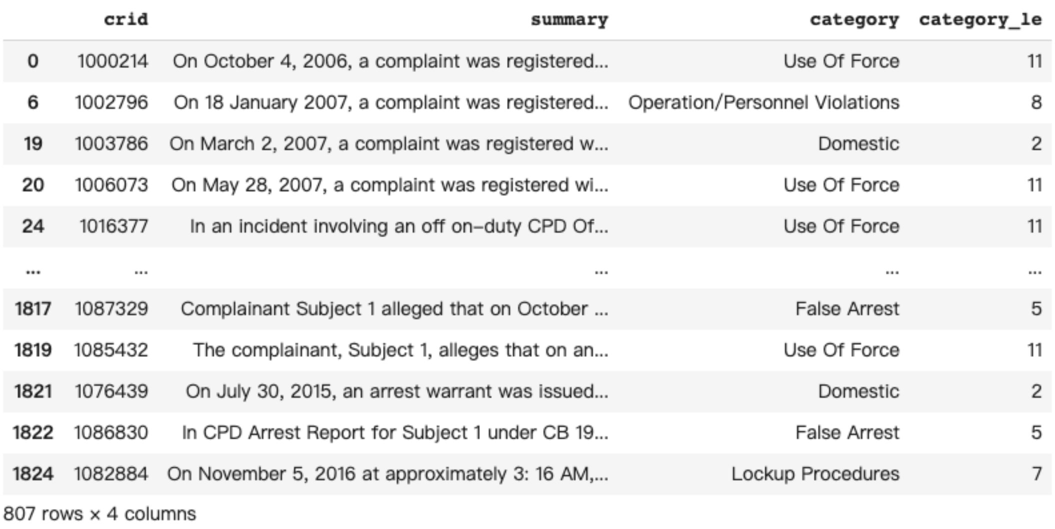
Then, we map the officer IDs back to allegation IDs, performed data cleaning, and constructed our data frames for repeaters:



	crid	summary	category	category_le
6	1002796	On 18 January 2007, a complaint was registered...	Operation/Personnel Violations	8
20	1006073	On May 28, 2007, a complaint was registered wi...	Use Of Force	11
30	1009954	On October 8, 2007, a complaint was registered...	Operation/Personnel Violations	8
32	1020690	On October 10, 2008, a complaint was registere...	Domestic	2
38	1023629	On February 4, 2009, a complaint was registere...	Use Of Force	11
...	...	...	...	...
1788	1079381	On 25 February 2016, at approximately 0425 hou...	Use Of Force	11
1805	1077285	This is the third time since 2013 that Officer...	Domestic	2
1806	1077591	On August 16, 2015, police responded to a call...	Use Of Force	11
1816	1088038	and his son (also named were parked on the sid...	Use Of Force	11
1817	1087329	Complainant Subject 1 alleged that on October ...	False Arrest	5

287 rows × 4 columns

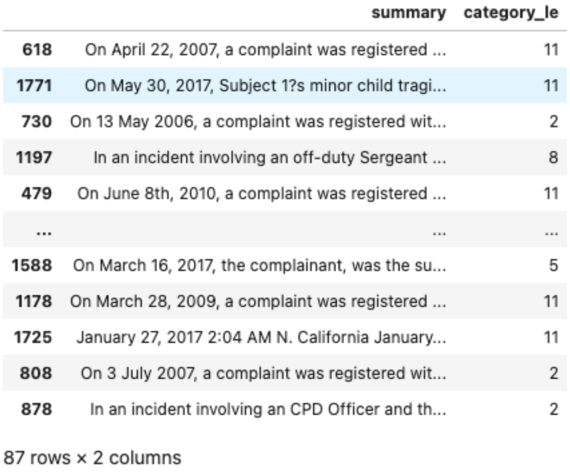
And for non-repeaters as well:



	crid	summary	category	category_le
0	1000214	On October 4, 2006, a complaint was registered...	Use Of Force	11
6	1002796	On 18 January 2007, a complaint was registered...	Operation/Personnel Violations	8
19	1003786	On March 2, 2007, a complaint was registered w...	Domestic	2
20	1006073	On May 28, 2007, a complaint was registered wi...	Use Of Force	11
24	1016377	In an incident involving an off on-duty CPD Of...	Use Of Force	11
...	...	...	...	...
1817	1087329	Complainant Subject 1 alleged that on October ...	False Arrest	5
1819	1085432	The complainant, Subject 1, alleges that on an...	Use Of Force	11
1821	1076439	On July 30, 2015, an arrest warrant was issued...	Domestic	2
1822	1086830	In CPD Arrest Report for Subject 1 under CB 19...	False Arrest	5
1824	1082884	On November 5, 2016 at approximately 3: 16 AM,...	Lockup Procedures	7

807 rows × 4 columns

Then, we dropped the ‘crid’ and ‘category’ columns because these two won’t be needed during training. After that, we only have ‘summary’ field as input, and ‘category_le’ field as labels in each data frame. And we divide each of our data frame into train set and test set, with a ratio of **7:3**. For example, the repeater data test set looks like this:



	summary	category_le
618	On April 22, 2007, a complaint was registered ...	11
1771	On May 30, 2017, Subject 1?s minor child tragi...	11
730	On 13 May 2006, a complaint was registered wit...	2
1197	In an incident involving an off-duty Sergeant ...	8
479	On June 8th, 2010, a complaint was registered ...	11
...	...	...
1588	On March 16, 2017, the complainant, was the su...	5
1178	On March 28, 2009, a complaint was registered ...	11
1725	January 27, 2017 2:04 AM N. California January...	11
808	On 3 July 2007, a complaint was registered wit...	2
878	In an incident involving an CPD Officer and th...	2

87 rows × 2 columns

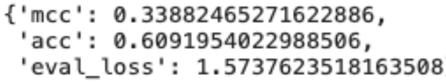
Model Training and Evaluation

We use Transformer as our text classification model. Transformer model consists of an encoder-decoder architecture. The encoder includes a self-attention mechanism and a feed-forward neural network. And the decoder of Transformer includes a self-attention mechanism, an attention mechanism over the encodings, and a feed-forward neural network.

We use simpletransformers package to train and evaluate our model. It is a package built upon the popular PyTorch framework and designed specifically for transformer models. We choose RoBERTa model type. RoBERTa builds on BERT and modifies key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates.

During training, we mostly used the default hyperparameters (learning rate: 4e-5, adam epsilon: 1e-8, batch size: 8, weight decay: 0, etc.) except that we set the number of epochs to be 10. First, we observe the evaluation metrics if applying to the test set of its own data frame group (repeater model with repeater test set, non-repeater model with non-repeater test set).

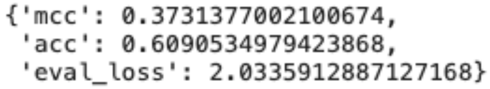
- The result for repeater data frame is:



```
{'mcc': 0.33882465271622886,
 'acc': 0.6091954022988506,
 'eval_loss': 1.5737623518163508}
```

The ‘mcc’ is the Matthews correlation coefficient, ‘acc’ is the accuracy, ‘eval_loss’ is the evaluation loss value.

- The result for non-repeater data frame is:

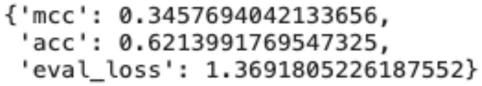


```
{'mcc': 0.3731377002100674,
 'acc': 0.6090534979423868,
 'eval_loss': 2.0335912887127168}
```

We can see that the accuracy of the model is around **60 percent**, which shows that the model can fairly perform this 13-class text classification task for the complaint summary. However, the results indicate overfitting. It is worth noting that for both datasets, overfitting happens early (as early as around epoch 5 for repeater data). As we analyze the results, we identified the limitations of this data and discussed them in Limitations section below.

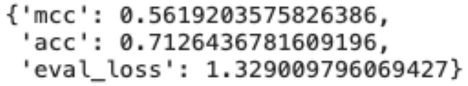
We also cross-apply the models to the other group of data frame.

- The result for repeater model to non-repeater test set is:



```
{'mcc': 0.3457694042133656,
 'acc': 0.6213991769547325,
 'eval_loss': 1.3691805226187552}
```

- The result for non-repeater model to repeater test set is:



```
{'mcc': 0.5619203575826386,
 'acc': 0.7126436781609196,
 'eval_loss': 1.329009796069427}
```

We found that the accuracy if applying non-repeater model to repeater test set can go up 10%, but the accuracy if applying repeater model to non-repeater data remains nearly unchanged. We think this is because there are more data entries for non-repeater allegations than the repeater data. And maybe because the repeaters are more likely to do certain types of misconduct, leading to easier overfitting of the model. But again, this result can be inconsistent because of the limitations we discussed below.

Limitations

For this checkpoint, we observe that there are several limitations that hurts the performance of the model:

1. The data size of both repeater allegations and non-repeater allegations are too small.
2. The structure of the summary is similar.