Stanford | Data Science

Identifying Behavioral Health Conditions in Police Encounters

Stanford Data Science for Social Good

August 2022

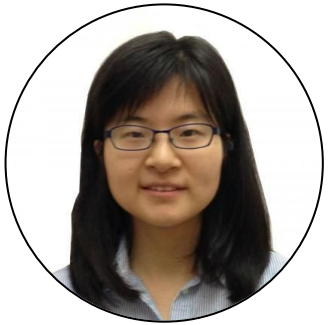
Our Team

Alex Lerner
DSSG Fellow

Amanda Xu
DSSG Fellow

Grishma Bhattarai
DSSG Fellow

Salil Goyal
DSSG Fellow



Qian Zhao
Technical Mentor

Dr. Balasubramanian Narasimhan
Faculty Advisor

Lisa Pickoff-White
Community Partner

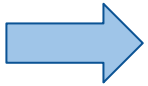
Agenda

- **Project Background**
- **Data Hierarchy**
- **Our Pipeline**
  - Page Type Classification
  - Forms Pipeline
  - Narratives Pipeline
    - DistilBERT Model Justification
    - Text Cleaning
    - Training Labels
    - Model Performance
- **Conclusion and Future Work**

Reporting shows need for law enforcement accountability

California Senate Bill 1421

*made police records relating to officer **use-of-force incidents**, sexual assault and acts of dishonesty **publicly accessible***



NEWS

Bakersfield Police Broke 31 People's Bones in Four Years. No Officer Has Been Disciplined for It

By Lisa Pickoff-White, KQED; Ross Ewald and Danielle Echeverria, Stanford University; Anne Daugherty, UC Berkeley Jun 16, 2021

On Nov. 24, 2017, Robert Cruz Jr. biked north along Baker Street, on a quiet block straddling Bakersfield's once-thriving old town and struggling new, restaurants interspersed with a rehab center and a prepaid phone store. A little before midnight, two officers noticed that the 37-year-old Cruz didn't have a front light on his bicycle. A patrol officer chased Cruz to a nearby yard. There, Cruz crouched behind a child's play tunnel, and the officer struck his arm with a baton.

In California: **3,600+ people seriously injured or killed** by law enforcement officers (2016-2020)

Source: https://www.kqed.org/news/11878013/bakersfield-police-broke-31-peoples-bones-in-four-years-no-officer-has-been-disciplined-for-it
Source: https://www.kvpr.org/local-news/2022-04-12/bakersfield-police-department-fails-to-identify-people-in-crisis-thwarting-reform

Behavioral health factors require police de-escalation

"Behavioral health conditions":

- Entails both mental health and substance abuse

Why does it matter?

- **Officers are supposed to attempt to de-escalate situations** with mentally ill or disabled or intoxicated people, who may not respond to commands in the same way

Bakersfield Police Department fails to identify people in crisis, thwarting reform

Cases of serious injury/death that involved behavioral health conditions in Bakersfield:

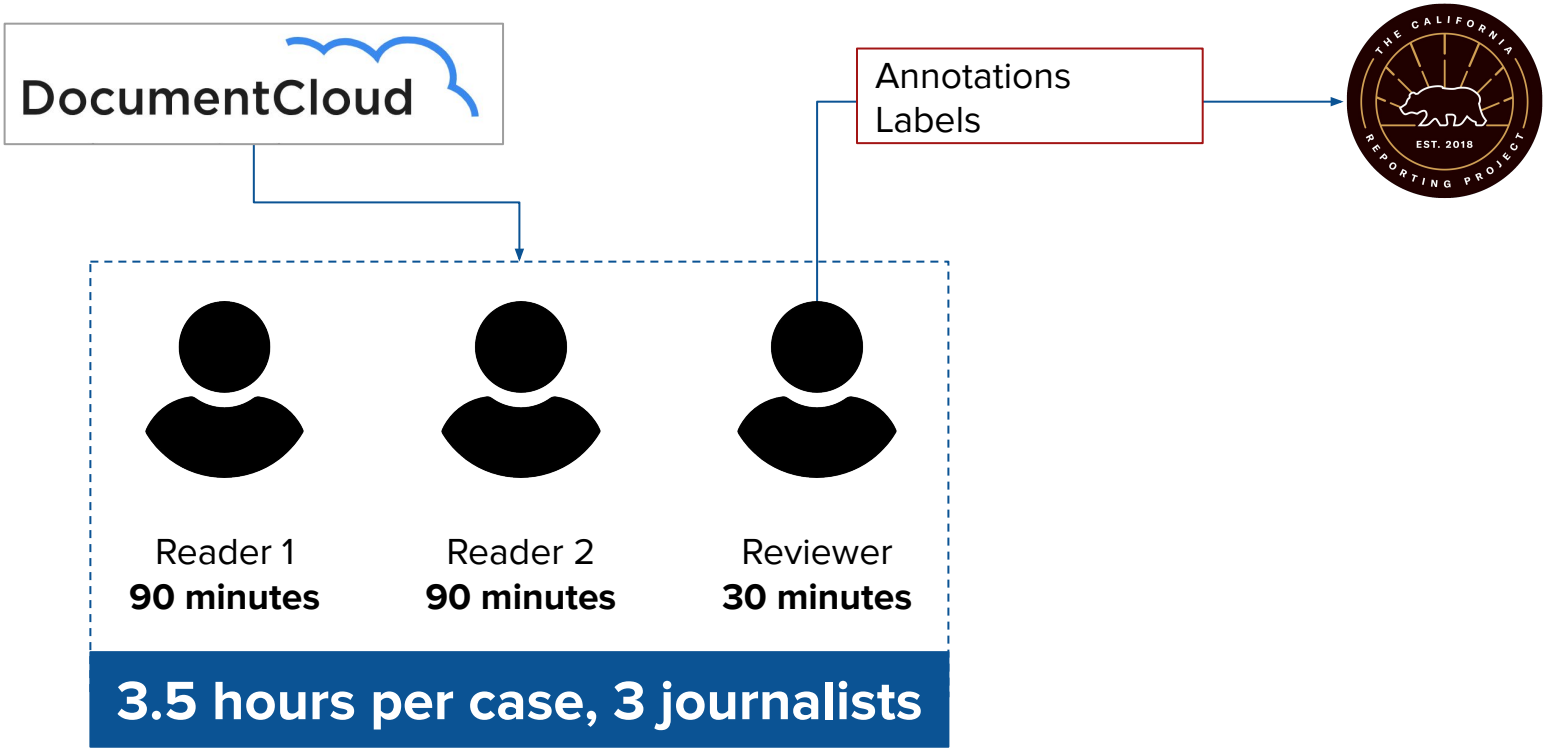
3% **reported rate**

vs.

44% **actual rate**

Source: https://www.kvpr.org/local-news/2022-04-12/bakersfield-police-department-fails-to-identify-people-in-crisis-thwarting-reform

Laborious case review process requires 3.5 man hours



Laborious case review process requires 3.5 man hours

DocumentCloud

Annotations
Labels

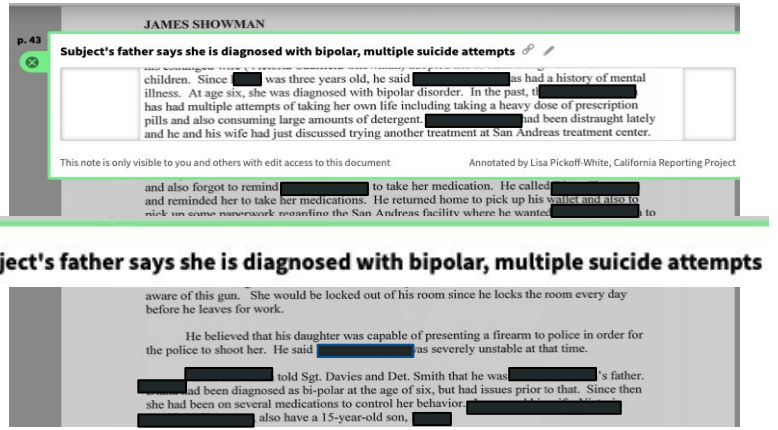
THE CALIFORNIA REPORTING PROJECT EST. 2018

Reader 1
90 minutes

Reader 2
90 minutes

Reviewer
30 minutes

3.5 hours per case, 3 journalists



Scale of data creates need for automation

1200 hrs

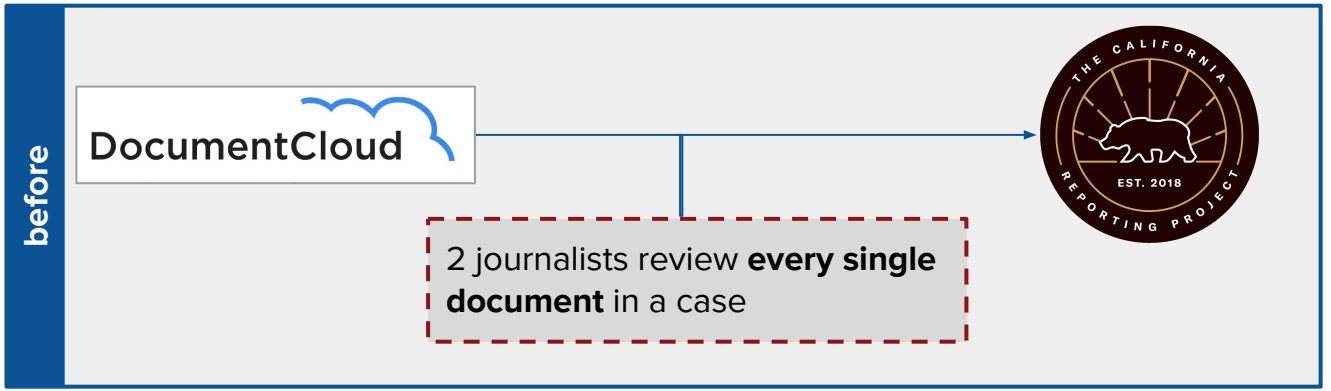
(That's 30 work weeks!)

That's how much time* journalists have spent on 6 agencies.

There are up to

464 agencies

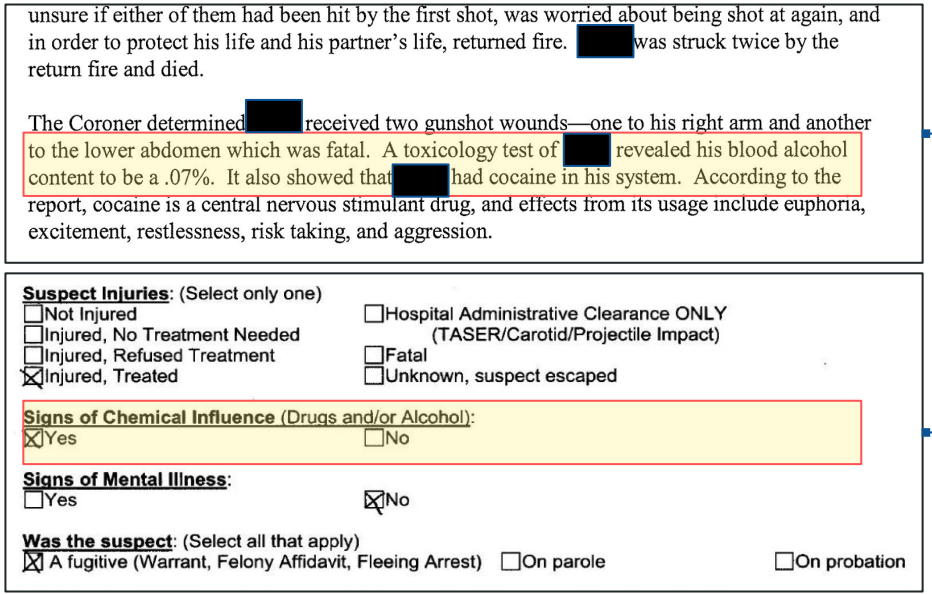
that are unreviewed.



*Note: 3.5 hours per case (based on BigLocal estimates) * 350 annotated cases = 1,225 hours

Our aim is to reduce workload via automated classification

Project Goal: *Automatically classify whether each document contains mention of behavioral health conditions.*



unsure if either of them had been hit by the first shot, was worried about being shot at again, and in order to protect his life and his partner's life, returned fire. [redacted] was struck twice by the return fire and died.

The Coroner determined [redacted] received two gunshot wounds—one to his right arm and another to the lower abdomen which was fatal. A toxicology test of [redacted] revealed his blood alcohol content to be a .07%. It also showed that [redacted] had cocaine in his system. According to the report, cocaine is a central nervous stimulant drug, and effects from its usage include euphoria, excitement, restlessness, risk taking, and aggression.

Suspect Injuries: (Select only one)
☐ Not Injured
☐ Injured, No Treatment Needed
☐ Injured, Refused Treatment
☒ Injured, Treated
☐ Hospital Administrative Clearance ONLY (TASER/Carotid/Projectile Impact)
☐ Fatal
☐ Unknown, suspect escaped

Signs of Chemical Influence (Drugs and/or Alcohol):
☒ Yes ☐ No

Signs of Mental Illness:
☐ Yes ☒ No

Was the suspect: (Select all that apply)
☒ A fugitive (Warrant, Felony Affidavit, Fleeing Arrest) ☐ On parole ☐ On probation

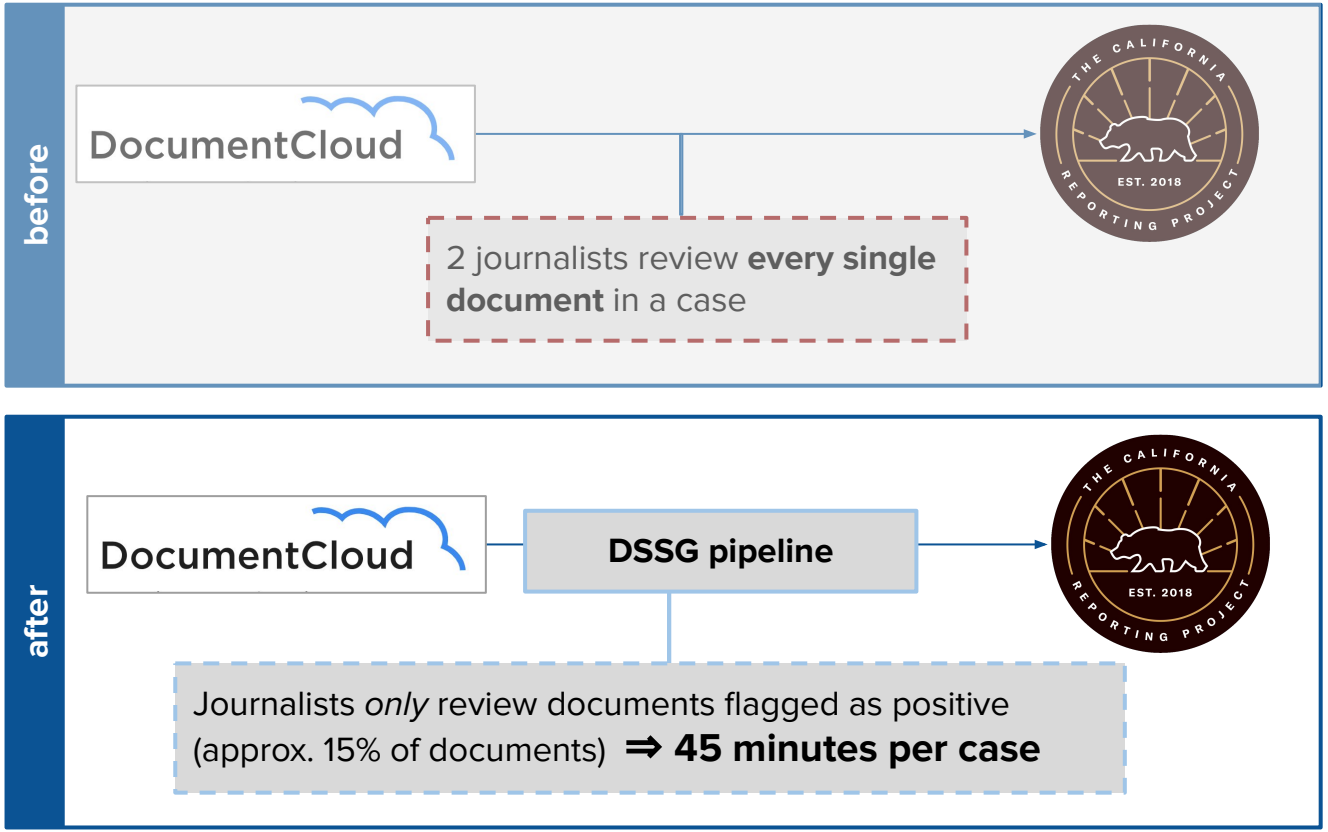
Mentions that subject's BAC was 0.07% and had cocaine in his system
⇒ classify as **True**

Mentions signs of chemical influence (Drugs/Alcohol)
⇒ classify as **True**

Automation reduces man hours required by 75%

75%

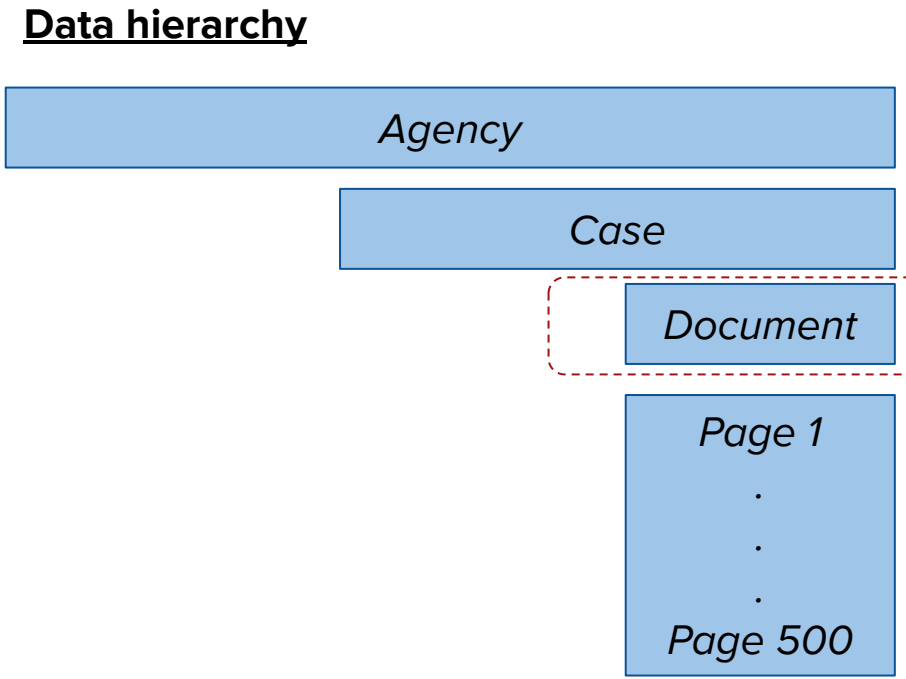
less time required for journalists to review a case with automation



Agenda

- **Project Background**
- **Data Hierarchy**
- **Our Pipeline**
  - Page Type Classification
  - Forms Pipeline
  - Narratives Pipeline
    - DistilBERT Model Justification
    - Text Cleaning
    - Training Labels
    - Model Performance
- **Conclusion and Future Work**

Data hierarchy spans 4 levels, from agency to page



Our training data

3 agencies

350+ cases

1600+ documents

Each document between one and several hundred pages

Recall:
We want to classify at the **Document Level!**

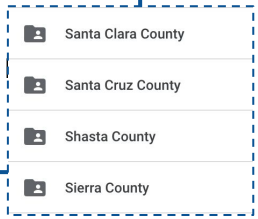
Agenda

- **Project Background**
- **Data Hierarchy**
- **Our Pipeline**
  - Page Type Classification
  - Forms Pipeline
  - Narratives Pipeline
    - DistilBERT Model Justification
    - Text Cleaning
    - Training Labels
    - Model Performance
- **Conclusion and Future Work**

Pipeline Overview

Retrieve Google Drive files

- **Input:** County name
- **Output:** File path to each PDF in given county folder



Page Type Classification

- **Input:** PDF file paths
- **Tools:** LayoutLMV2 Classifier
- **Output:** classification label per page *(form, narrative, interview, image, or other)*

rep_image	Label	doc	pg_num
../Shasta County/ima	narrative	2022 Shasta County	16
../Shasta County/ima	narrative	2022 Shasta County	48
../Shasta County/ima	other	2022 Shasta County	15

form pages

narrative pages

Forms Pipeline

- **Input:** file path to form pages
- **Tools:** Azure Form Recognizer
- **Output:** form label, label justification

label	rel_info
False	NaN
True	{'signs of drug impairment': 'Page #1: [936.0,...
False	San Jose Police Department Property & Evidence...

narrative text from forms

Narratives Pipeline

- **Input:** String of text
- **Tools:** Tesseract, NER with SpaCy, DistilBERT
- **Output:** narrative label, score

Final classification for each **document**

Page Type Classification

Retrieve Google Drive files

- **Input:** County name
- **Output:** File path to each PDF in given county folder

Page Classification

- **Input:** PDF file paths
- **Tools:** LayoutLMV2 Classifier
- **Output:** classification label per page *(form, narrative, interview, image, or other)*

rep_image	Label	doc	pg_num
../Shasta County/ima	narrative	2022 Shasta County	16
../Shasta County/ima	narrative	2022 Shasta County	48
../Shasta County/ima	other	2022 Shasta County	15

form pages

narrative pages

Forms Pipeline

- **Input:** file path to form pages
- **Tools:** Azure Form Recognizer
- **Output:** form label, label justification

label	rel_info
False	NaN
True	{'signs of drug impairment': 'Page #1: [936.0,...
False	San Jose Police Department Property & Evidence...

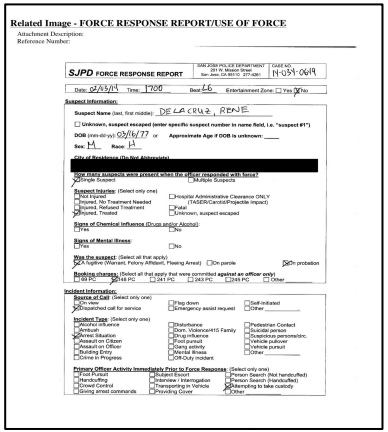
narrative text from forms

Narratives Pipeline

- **Input:** String of text
- **Tools:** Tesseract, NER with SpaCy, DistillBERT
- **Output:** narrative label, score

Final classification for each **document**

form



Related Image - FORCE RESPONSE REPORT/USE OF FORCE

SJPD FORCE RESPONSE REPORT

Suspect Information:

Suspect Name (last, first middle): DE LA CRUZ, RENE

DOB (mm-dd-yy): 09/16/77 or Approximate Age if DOB is unknown: ____

Sex: M Race: H

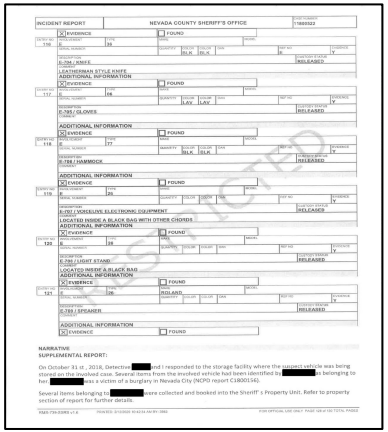
Incident Information:

Source of Call: On-view; Flag down; Self-initiated; Dispatched call for service; Emergency assist request; Other

Incident Type: Assault on Officer; Pedestrian Contact; Suicidal person; Vehicle pursuit

Primary Officer Activity Immediately Prior to Force Response: Foot Pursuit; Handcuffing; Crowd Control; Giving arrest commands; Subject Escort; Interview / Interrogation; Transporting in Vehicle; Providing Cover; Person Search (Not handcuffed); Person Search (Handcuffed); Attempting to take custody; Other

form

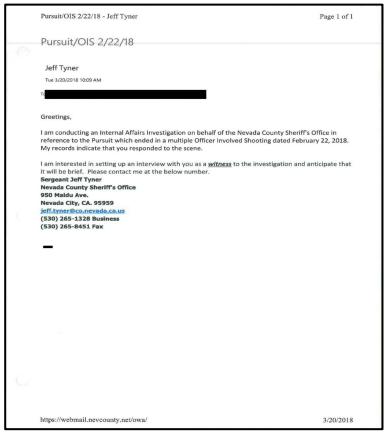


INCIDENT REPORT — NEVADA COUNTY SHERIFF'S OFFICE

NARRATIVE

SUPPLEMENTAL REPORT:

narrative



Pursuit/OIS 2/22/18 - Jeff Tyner Page 1 of 1

Pursuit/OIS 2/22/18

Jeff Tyner

Tue 3/20/2018 10:09 AM

Greetings,

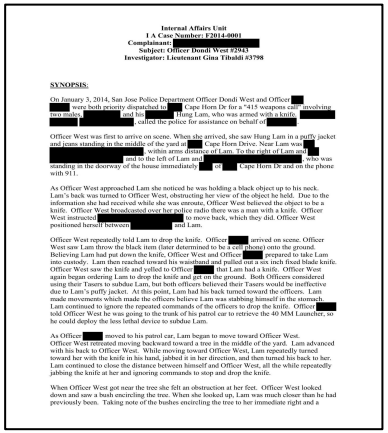
I am conducting an Internal Affairs Investigation on behalf of the Nevada County Sheriff's Office in reference to the Pursuit which ended in a multiple Officer Involved Shooting dated February 22, 2018. My records indicate that you responded to the scene.

I am interested in setting up an interview with you as a witness to the investigation and anticipate that it will be brief. Please contact me at the below number.

Sergeant Jeff Tyner
Nevada County Sheriff's Office
950 Maidu Ave.
Nevada City, CA. 95959
jeff.tyner@co.nevada.ca.us
(530) 265-1328 Business
(530) 265-8451 Fax

https://webmail.nevcounty.net/owa/ 3/20/2018

narrative



Internal Affairs Unit
1 A Case Number: F2014-0001
Complainant:
Subject: Officer Dondi West #2943
Investigator: Lieutenant Gina Tibaldi #3798

SYNOPSIS:

On January 3, 2014, San Jose Police Department Officer Dondi West and Officer were both priority dispatched to Cape Horn Dr for a "415 weapons call" involving two males and his Hung Lam, who was armed with a knife, called the police for assistance on behalf of

Officer West was first to arrive on scene. When she arrived, she saw Hung Lam in a puffy jacket and jeans standing in the middle of the yard at Cape Horn Drive. Near Lam was within arms distance of Lam. To the right of Lam and and to the left of Lam and who was standing in the doorway of the house immediately Cape Horn Dr and on the phone with 911.

As Officer West approached Lam she noticed he was holding a black object up to his neck. Lam's back was turned to Officer West, obstructing her view of the object he held. Due to the information she had received while she was enroute, Officer West believed the object to be a knife. Officer West broadcasted over her police radio there was a man with a knife. Officer West instructed to move back, which they did. Officer West positioned herself between and Lam.

Officer West repeatedly told Lam to drop the knife. Officer arrived on scene. Officer West saw Lam throw the black item (later determined to be a cell phone) onto the ground. Believing Lam had put down the knife, Officer West and Officer prepared to take Lam into custody. Lam then reached toward his waistband and pulled out a six inch fixed blade knife. Officer West saw the knife and yelled to Officer that Lam had a knife. Officer West again began ordering Lam to drop the knife and get on the ground. Both Officers considered using their Tasers to subdue Lam, but both officers believed their Tasers would be ineffective due to Lam's puffy jacket. At this point, Lam had his back turned toward the officers. Lam made movements which made the officers believe Lam was stabbing himself in the stomach. Lam continued to ignore the repeated commands of the officers to drop the knife. Officer told Officer West he was going to the trunk of his patrol car to retrieve the 40 MM Launcher, so he could deploy the less lethal device to subdue Lam.

As Officer moved to his patrol car, Lam began to move toward Officer West. Officer West retreated moving backward toward a tree in the middle of the yard. Lam advanced with his back to Officer West. While moving toward Officer West, Lam repeatedly turned toward her with the knife in his hand, jabbed it in her direction, and then turned his back to her. Lam continued to close the distance between himself and Officer West, all the while repeatedly jabbing the knife at her and ignoring commands to stop and drop the knife.

When Officer West got near the tree she felt an obstruction at her feet. Officer West looked down and saw a bush encircling the tree. When she looked up, Lam was much closer than he had previously been. Taking note of the bushes encircling the tree to her immediate right and a

Page Type Classification

Goal: *Separate forms from narratives pages within a document*

Input — file path to Google Drive

Hugging Face — Page type classifier*

Output — document, page number, page path

page label

**created by Hellina Hailu Nigatu at UC Berkeley*

Forms Pipeline

Retrieve Google Drive files

- **Input:** County name
- **Output:** File path to each given county folder

Page Classification

- **Input:** PDF file paths
- **Tools:** LayoutLMV2 Classifier
- **Output:** classification label per page *(form, narrative, interview, image, or other)*

rep_image	Label	doc	pg_num
../Shasta County/ima	narrative	2022 Shasta County	16
../Shasta County/ima	narrative	2022 Shasta County	48
../Shasta County/ima	other	2022 Shasta County	15

form pages

narrative pages

Forms Pipeline

- **Input:** file path to form pages
- **Tools:** Azure Form Recognizer
- **Output:** form label, label justification

label	rel_info
False	NaN
True	{'signs of drug impairment': 'Page #1: [936.0,...
False	San Jose Police Department Property & Evidence...

narrative text from forms

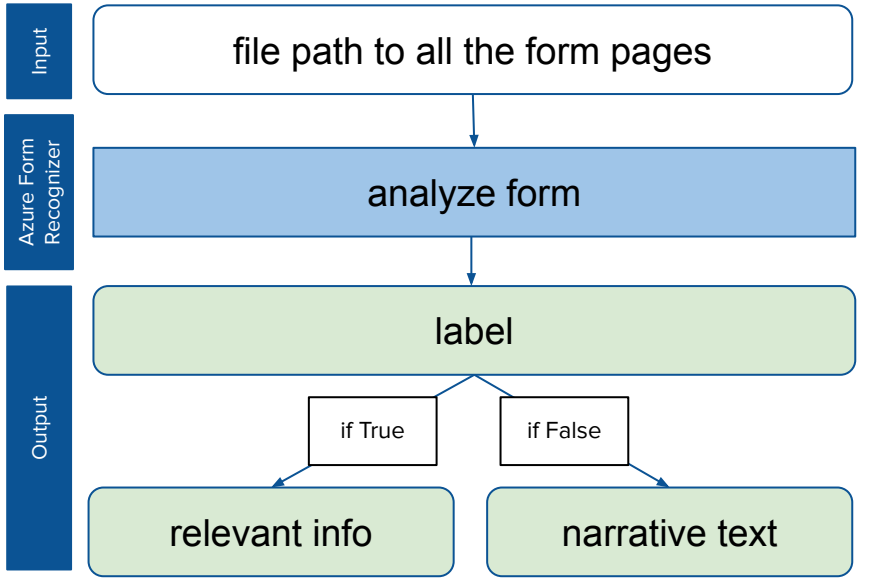
Narratives Pipeline

- **Input:** String of text
- **Tools:** Tesseract, NER with SpaCy, DistillBERT
- **Output:** narrative label, score

Final classification for each **document**

Forms Pipeline

Goal: *Find forms that mention behavioral health conditions + explain why*



For each form page…

1. Analyze checkboxes. Select the ones that are:
   - ❏ Selected
   - ❏ Relevant (contains some relevant word; e.g., “intoxication”, “influence”)

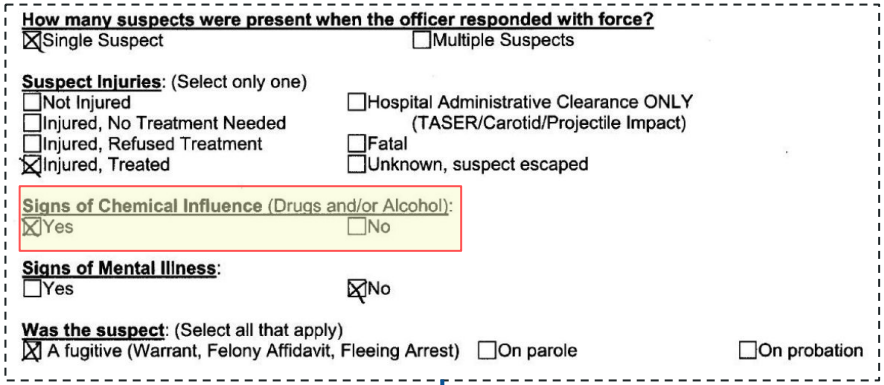
   Return True, checkbox title/location

2. If no relevant checkboxes ➔ retrieve any narrative text on page

   Return False, narrative text

Example outputs of forms

Example of **Positive Label**



How many suspects were present when the officer responded with force?
☒ Single Suspect ☐ Multiple Suspects

Suspect Injuries: (Select only one)
☐ Not Injured
☐ Injured, No Treatment Needed
☐ Injured, Refused Treatment
☒ Injured, Treated
☐ Hospital Administrative Clearance ONLY (TASER/Carotid/Projectile Impact)
☐ Fatal
☐ Unknown, suspect escaped

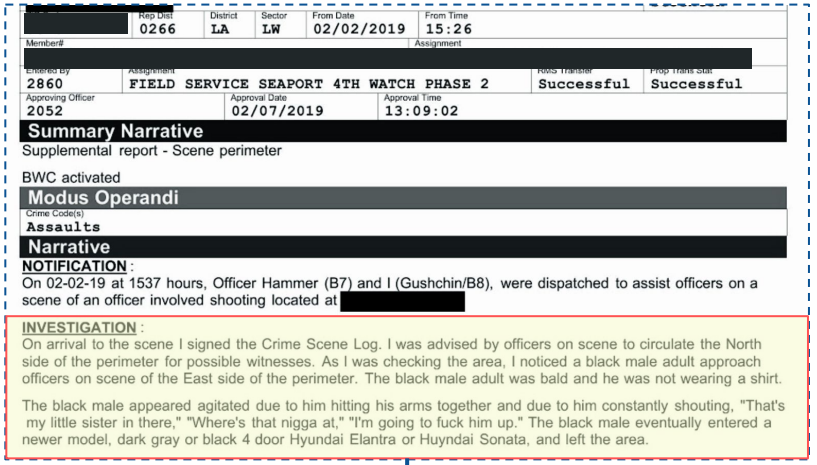
Signs of Chemical Influence (Drugs and/or Alcohol):
☒ Yes ☐ No

Signs of Mental Illness:
☐ Yes ☒ No

Was the suspect: (Select all that apply)
☒ A fugitive (Warrant, Felony Affidavit, Fleeing Arrest) ☐ On parole ☐ On probation

```
True
{'Signs of Chemical Influence (Drugs and/or Alcohol):':
'Page #1: [313.0, 904.0], [886.0, 904.0], [886.0, 930.
0], [313.0, 930.0]'}
```

Example of **Negative Label**



[redacted]	Rep Dist	District	Sector	From Date	From Time
	0266	LA	LW	02/02/2019	15:26

Member# [redacted] Assignment

Entered By	Assignment	RMS Transfer	Prop Trans Stat
2860	FIELD SERVICE SEAPORT 4TH WATCH PHASE 2	Successful	Successful

Approving Officer	Approval Date	Approval Time
2052	02/07/2019	13:09:02

Summary Narrative

Supplemental report - Scene perimeter

BWC activated

Modus Operandi

Crime Code(s)
Assaults

Narrative

NOTIFICATION:
On 02-02-19 at 1537 hours, Officer Hammer (B7) and I (Gushchin/B8), were dispatched to assist officers on a scene of an officer involved shooting located at [redacted]

INVESTIGATION:
On arrival to the scene I signed the Crime Scene Log. I was advised by officers on scene to circulate the North side of the perimeter for possible witnesses. As I was checking the area, I noticed a black male adult approach officers on scene of the East side of the perimeter. The black male adult was bald and he was not wearing a shirt.

The black male appeared agitated due to him hitting his arms together and due to him constantly shouting, "That's my little sister in there," "Where's that nigga at," "I'm going to fuck him up." The black male eventually entered a newer model, dark gray or black 4 door Hyundai Elantra or Huyndai Sonata, and left the area.

```
False,
 'On arrival to the scene I signed the Crime Scene Log. I was advised by officers on
scene to circulate the North side of the perimeter for possible witnesses. As I was
checking the area, I noticed a black male adult approach officers on scene of the
East side of the perimeter. The black male adult was bald and he was not wearing a
shirt. The black male appeared agitated due to him hitting his arms together and due
```

Narratives Pipeline

Retrieve Google Drive files

- **Input:** County name
- **Output:** File path to each given county folder

Page Classification

- **Input:** PDF file paths
- **Tools:** LayoutLMV2 Classifier
- **Output:** classification label per page *(form, narrative, interview, image, or other)*

rep_image	Label	doc	pg_num
../Shasta County/ima	narrative	2022 Shasta County	16
../Shasta County/ima	narrative	2022 Shasta County	48
../Shasta County/ima	other	2022 Shasta County	15

form pages

narrative pages

Forms Pipeline

- **Input:** file path to form pages
- **Tools:** Azure Form Recognizer
- **Output:** form label, label justification

label	rel_info
False	NaN
True	{'signs of drug impairment': 'Page #1: [936.0,...
False	San Jose Police Department Property & Evidence...

narrative text from forms

Narratives Pipeline

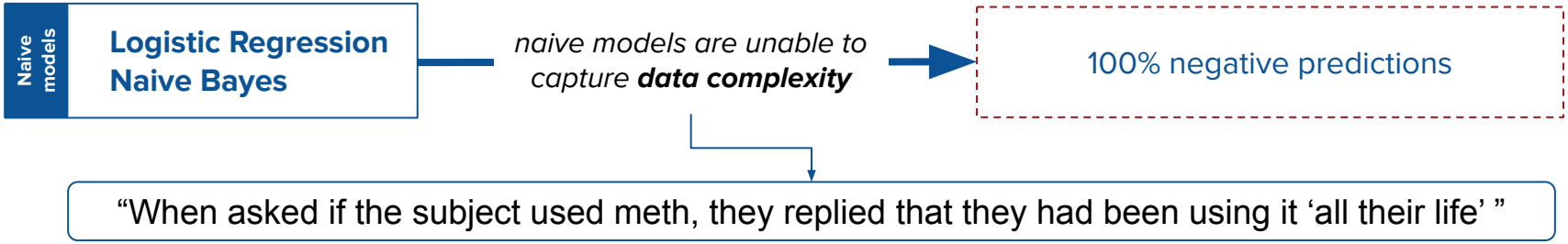
- **Input:** String of text
- **Tools:** pyTesseract, SpaCy, DistilBERT
- **Output:** narrative label, score

Final classification for each **document**

Agenda

- **Project Background**
- **Data Hierarchy**
- **Our Pipeline**
  - Page Type Classification
  - Forms Pipeline
  - Narratives Pipeline
    - DistilBERT Model Justification
    - Text Cleaning
    - Training Labels
    - Model Performance
- **Conclusion and Future Work**

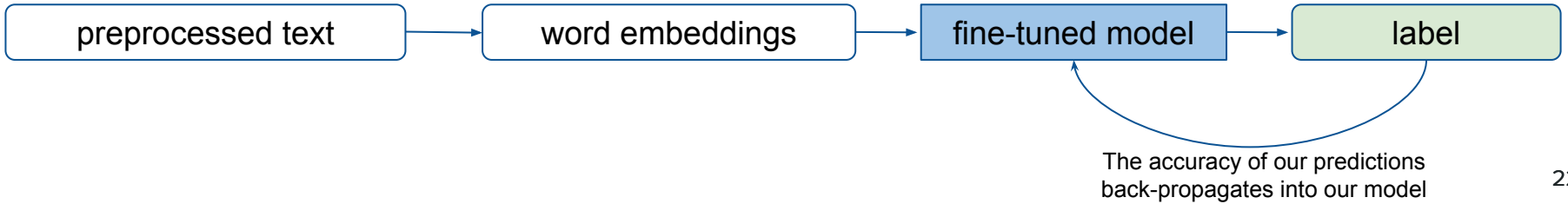
Complex model needed to learn language meaning



Recent advances in Natural Language Processing (NLP) have championed the use of Deep Neural Models which have shown tremendous performance on a variety of tasks

- Question Answering, Summarization, Translation, and also Text Classification

Training our model

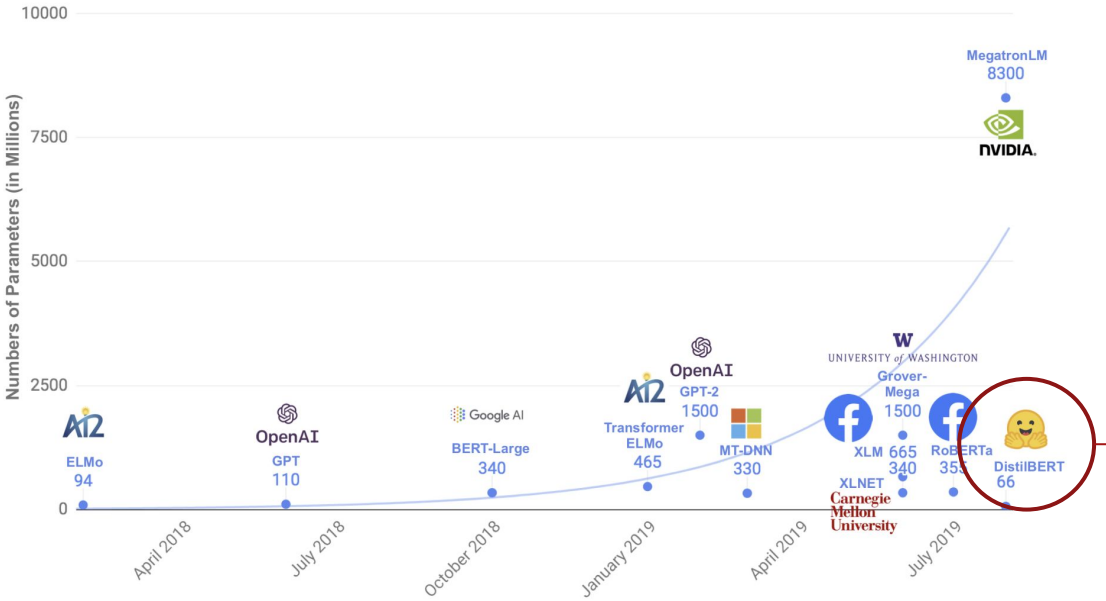


Pre-trained Transformer model fits our needs

- In 2017, Google released a new model architecture that became the model of choice for many NLP tasks: the Transformer
- Processes sequences at a time
- Utilizes a “self-attention” mechanism that can weight different parts of each sequence
- Is pre-trained on billions of texts from the web and Wikipedia, and can be fine-tuned for a variety of downstream tasks

Source: “Attention is All You Need,” Google Brain

DistilBERT balances training simplicity and performance

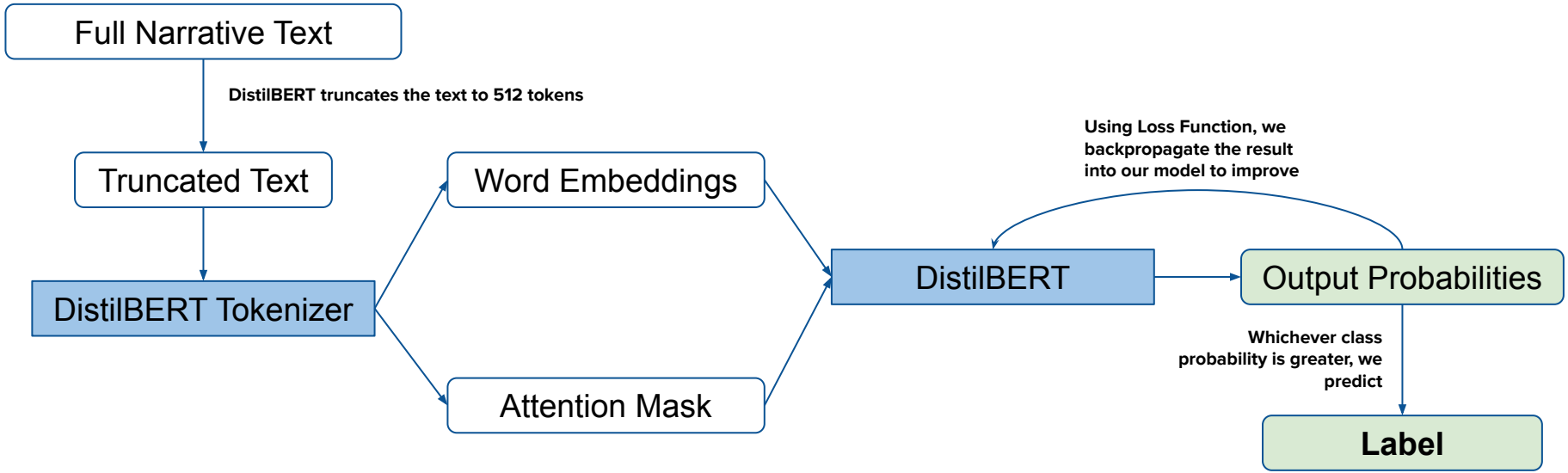


Our choice model, DistilBERT, is a variant of Google's Transformer model—**but mimics performance with far fewer parameters**

⇒ faster training + faster inference, matching our need to classify documents at scale

Source: [“DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”](), HuggingFace

DistilBERT Pipeline



Agenda

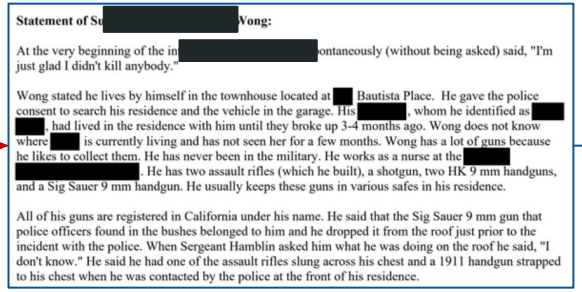
- **Project Background**
- **Data Hierarchy**
- **Our Pipeline**
  - Page Type Classification
  - Forms Pipeline
  - Narratives Pipeline
    - DistilBERT Model Justification
    - Text Cleaning
    - Training Labels
    - Model Performance
- **Conclusion and Future Work**

pyTesseract used to extract text from narrative pages

page	label	link
1	narrative	https://www.documentcloud.org/documents/21067706-15-315-0038-ia-report_redacted
17	image	https://www.documentcloud.org/documents/21067706-15-315-0038-ia-report_redacted
6	narrative	https://www.documentcloud.org/documents/21067706-15-315-0038-ia-report_redacted
12	narrative	https://www.documentcloud.org/documents/21067706-15-315-0038-ia-report_redacted
4	narrative	https://www.documentcloud.org/documents/21067706-15-315-0038-ia-report_redacted
5	narrative	https://www.documentcloud.org/documents/21067706-15-315-0038-ia-report_redacted
15	narrative	https://www.documentcloud.org/documents/21067706-15-315-0038-ia-report_redacted
23	form	https://www.documentcloud.org/documents/21067706-15-315-0038-ia-report_redacted

Output of page type classifier

Narrative page



Statement of Su [redacted] Vong:

At the very beginning of the in [redacted] ontaneously (without being asked) said, "I'm just glad I didn't kill anybody."

Wong stated he lives by himself in the townhouse located at [redacted] Bautista Place. He gave the police consent to search his residence and the vehicle in the garage. His [redacted], whom he identified as [redacted], had lived in the residence with him until they broke up 3-4 months ago. Wong does not know where [redacted] is currently living and has not seen her for a few months. Wong has a lot of guns because he likes to collect them. He has never been in the military. He works as a nurse at the [redacted]. He has two assault rifles (which he built), a shotgun, two HK 9 mm handguns, and a Sig Sauer 9 mm handgun. He usually keeps these guns in various safes in his residence.

All of his guns are registered in California under his name. He said that the Sig Sauer 9 mm gun that police officers found in the bushes belonged to him and he dropped it from the roof just prior to the incident with the police. When Sergeant Hamblin asked him what he was doing on the roof he said, "I don't know." He said he had one of the assault rifles slung across his chest and a 1911 handgun strapped to his chest when he was contacted by the police at the front of his residence.

pyTesseract

Input: Image of page

Output: Text extracted from the image

Raw text after extraction (pyTesseract output)

```
tement. Refer to the General Offense report for additional information. Statement of Suspect [redacted]: At
the very beginning of the interview, [redacted] spontaneously (without being asked) said, "I'm just glad I didn't k
ill anybody." [redacted]stated he lives by himself in the townhouse located at Bautista Place. He gave the police consent
to search his residence and the vehicle in the garage. His , whom he identified as , had lived in the residence with
him until they broke up 3-4 months ago. Wong does not know where is currently living and has not seen her for a few m
onths. [redacted]as a lot of guns because he likes to collect them. He has never been in the military. He works as a nurs
e at the . He has two assault rifles (which he built), a shotgun, two HK 9 mm handguns, and a Sig Sauer 9 mm handgun.
He usually keeps these guns in various safes in his residence. All of his guns are registered in California under his
name. He said that the Sig Sauer 9 mm gun that police officers found in the bushes belonged to him and he dropped it
from the roof just prior to the incident with the police. When [redacted]asked him what he was doing on the ro
of he said, "I don't know." He said he had one of the assault rifles slung across his chest and a 1911 handgun strapp
ed to his chest when he was contacted by the police at the front of his residence.
```

Named Entity Recognition (with spaCy)

NER classifies named entities mentioned in unstructured text into predefined categories (e.g., person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.)

Raw Text

'KERN COUNTY DISTRICT ATTORNEY\n\nO F F I C E\n\nM E M O R A N D U M\n\nTO: District Attorney Lisa Green\nFROM: DDA R AM\nDATE: January 20, 2017\nSUBJECT: OIS near Feliz Dr; BPD Officers Shaff, Mayberry, Mueller, Pace, and\nBarber; Sus pect [redacted] eal\n\nBrief Factual Synopsis:\nOn August 18, 2016, just before 3:30am, F[redacted]al ra n from a car that had just ended a\nlengthy vehicle pursuit with BPD. He told the occupants in the car with him that he was "not\ngoing back to prison. [redacted] armed with a loaded .380 semi-automatic handgun.\nOfficer Ary saw Villarreal with a gun, but as he a[redacted] him another male passenger\nexited the car and assaulted Ary. Ary yelled "gun" but was not heard by the other officers. Five\nofficers were in pursuit [redacted] foot through an apartment complex that is between Feliz\nDr and McNew Ct. None of the officers in pursuit [redacted] was armed. As Officer Pace\nclosed in [redacted] pointed his gun behind him and shot at her from roughly 10 feet away.\nSh e yelled "gun" but was not heard by the other officers as they were approaching Villarreal\nfrom different direction s. Other officers mistook the sound of the gunshot for Officer Pace's\ntaser being deployed. Officer Shaff then close d in and began to attempt to t[redacted] o the\nground. [redacted] ed, pointed the gun at Shaff's head and f ired. Shaff dropped to the ground\nin an effort to avoid being shot. Officers saw Villarreal attempt to shoot Shaff. All five officers\nthen shot at [redacted] various location[redacted] tinued to run as they shot at him fo r\n15 to 20 feet and then fell. Officers Shaff, Pace, and Mueller approached Villarreal from behind\nas he was lying on the ground with his back to them. Officer Barber and Mayberry came from\nthe side to appro[redacted] As Barbe r came within five to ten feet[redacted] sed his arm\nand shot at Barber's head. Barber dropped to the ground. H e and the other four officers [redacted] fire. After the shooting stopped, Officer Shaff walked ahead of the other officers and as\nhe did he[redacted] s arm and body "twitching" in a similar movement to when he fired at\nBarbe

List of Named Entities (NE) in the raw text

[redacted] PERSON
DDA RAM ORG
January 20, 2017 DATE
OIS ORG
Feliz Dr PERSON
BPD Officers Shaff ORG
Mayberry [redacted]
Mueller P[redacted]
Barber PE[redacted]
Suspect F[redacted]
August 18, 2016 DATE
3:30am CARDINAL
[redacted] PERSON
BPD ORG
[redacted] PERSON
[redacted] PERSON
Ary PERSON
Five CARDINAL

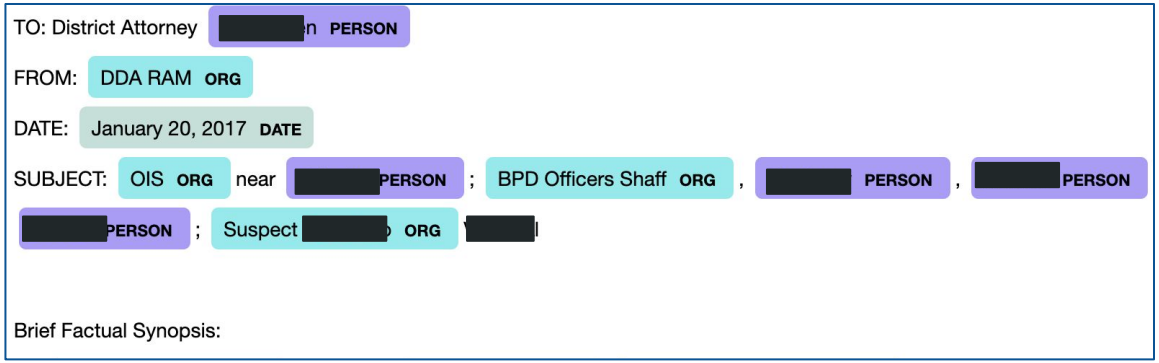
```
spacy.explain("ORG")
```

```
'Companies, agencies, institutions, etc.'
```

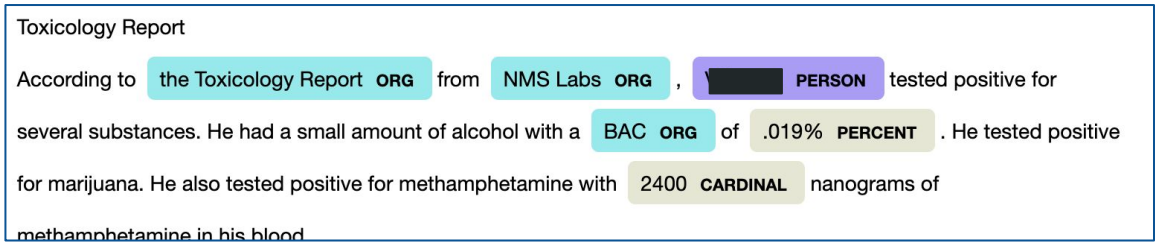
```
spacy.explain("CARDINAL")
```

```
'Numerals that do not fall under another type'
```

6 of 18 entities are useful, remaining 12 are masked



Noisy Named Entities

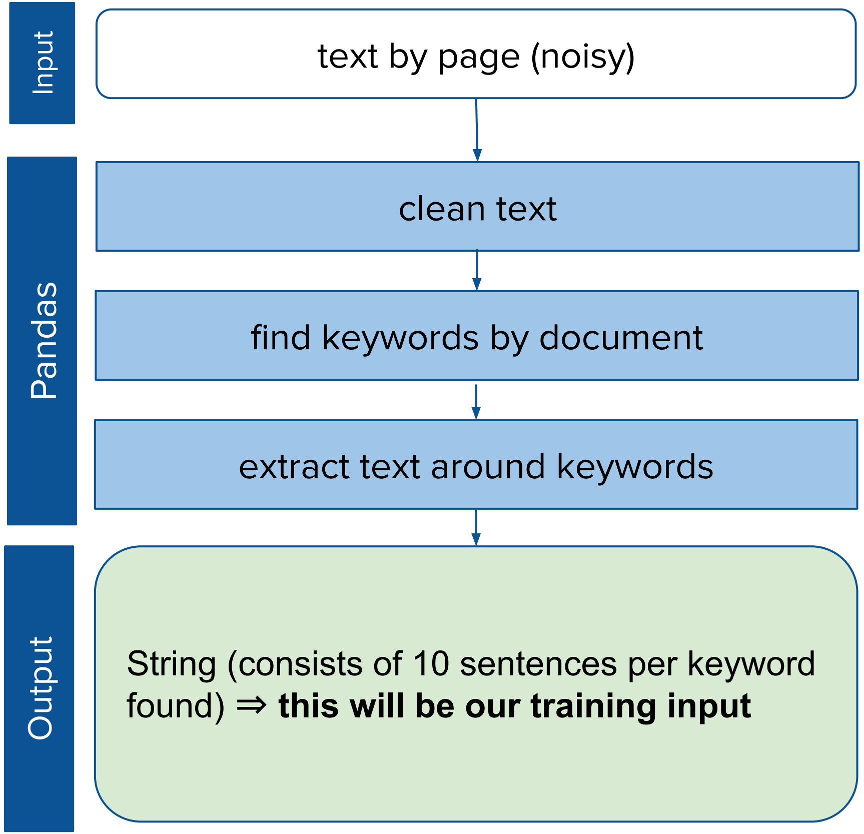


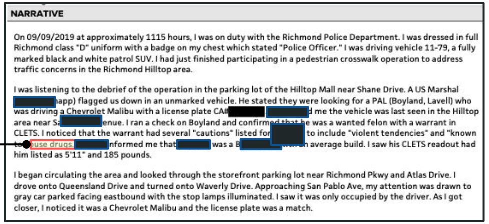
Relevant Named Entities

After exploring NER, we found that for our purposes:

- 6 categories: **['ORG', 'CARDINAL', 'ORDINAL', 'PRODUCT', 'PERCENT', 'QUANTITY']** may have relevant information
- 12 of 18 categories: noise
- Masked the noisy 12, kept the useful 6 as is

3 text preprocessing steps to prepare text for input to model




NARRATIVE

On 09/09/2019 at approximately 1115 hours, I was on duty with the Richmond Police Department. I was dressed in full Richmond class "D" uniform with a badge on my chest which stated "Police Officer." I was driving vehicle 11-79, a fully marked black and white patrol SUV. I had just finished participating in a pedestrian crosswalk operation to address traffic concerns in the Richmond Hilltop area.

I was listening to the debrief of the operation in the parking lot of the Hilltop Mall near Shane Drive. A US Marshal [redacted]app) flagged us down in an unmarked vehicle. He stated they were looking for a PAL (Boyland, Lavell) who was driving a Chevrolet Malibu with a license plate CA#[redacted] me the vehicle was last seen in the Hilltop area near Sa[redacted]enue. I ran a check on Boyland and confirm[redacted]e was a wanted felon with a warrant in CLETS. I noticed that the warrant had several "cautions" listed for[redacted] to include "violent tendencies" and "known to use drugs.[redacted]nformed me that[redacted] was a B[redacted]n average build. I saw his CLETS readout had him listed as 5'11" and 185 pounds.

I began circulating the area and looked through the storefront parking lot near Richmond Pkwy and Atlas Drive. I drove onto Queensland Drive and turned onto Waverly Drive. Approaching San Pablo Ave, my attention was drawn to gray car parked facing eastbound with the stop lamps illuminated. I saw it was only occupied by the driver. As I got closer, I noticed it was a Chevrolet Malibu and the license plate was a match.

Keyword: "Drugs"
Output: 5 sentences before keyword appears, 5 sentences after keyword appears

Reduced text from **125 sentences → 10 sentences**

KEYWORDS

```
['erratic', 'alcohol', 'mental', 'influence', 'strange', 'odd', 'drug',
 'narcotics', 'crazy', 'toxicology', 'bizarre', 'PCP', 'agitated',
 '5150', 'drunk', 'cocaine', 'involuntary', 'psychiatric', 'methamphetamine',
 'amphetamine', 'marijuana', 'narcotic', 'impairment', 'impaired', 'substance']
```

Agenda

- **Project Background**
- **Data Hierarchy**
- **Our Pipeline**
  - Page Type Classification
  - Forms Pipeline
  - Narratives Pipeline
    - DistilBERT Model Justification
    - Text Cleaning
    - Training Labels
    - Model Performance
- **Conclusion and Future Work**

Existing training data is noisy due to **case-level labels**

We were given case-level labels.

Case Label: True

Case ID	*document*	*actual label*	*label we have*
080622	incident_report	False	Case Label = True
080622	tox_report	False	Case Label = True
080622	witness_summary	True	Case Label = True

Problem: noisy labels

Some documents in a case may not mention behavioral health conditions, but the case will still be labeled True

⇒ Need document-level labels for document-level classifications

Annotations offer a way to create document level labels

We were given case-level labels.

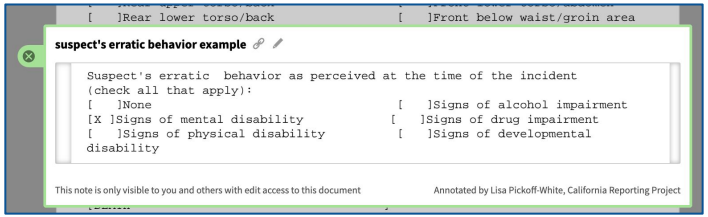
Case Label: True

Case ID	*document*	*actual label*	*label we have*
080622	incident_report	False	Case Label = True
080622	tox_report	False	Case Label = True
080622	witness_summary	True	Case Label = True

Problem: noisy labels

Some documents in a case may not mention behavioral health conditions, but the case will still be labeled True

We decided to generate new document-level labels using annotations for our training data.

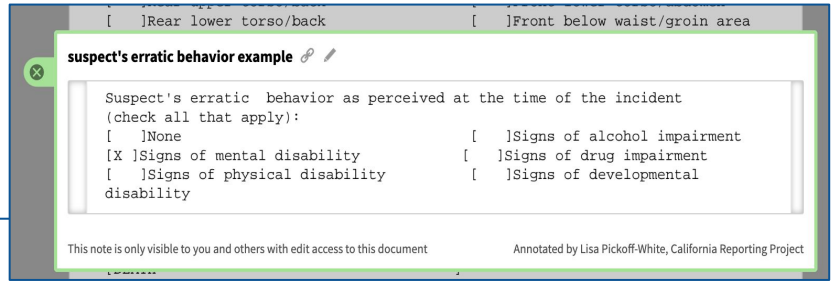


Example of a Annotated File on DocumentCloud

Case ID	***document***	***annotation_label***
080622	incident_report	**True if (relevant annotation), else False**
080622	tox_report	**True if (relevant annotation), else False**

Annotations as labels

We decided to generate new document-level labels using annotations for our training data.

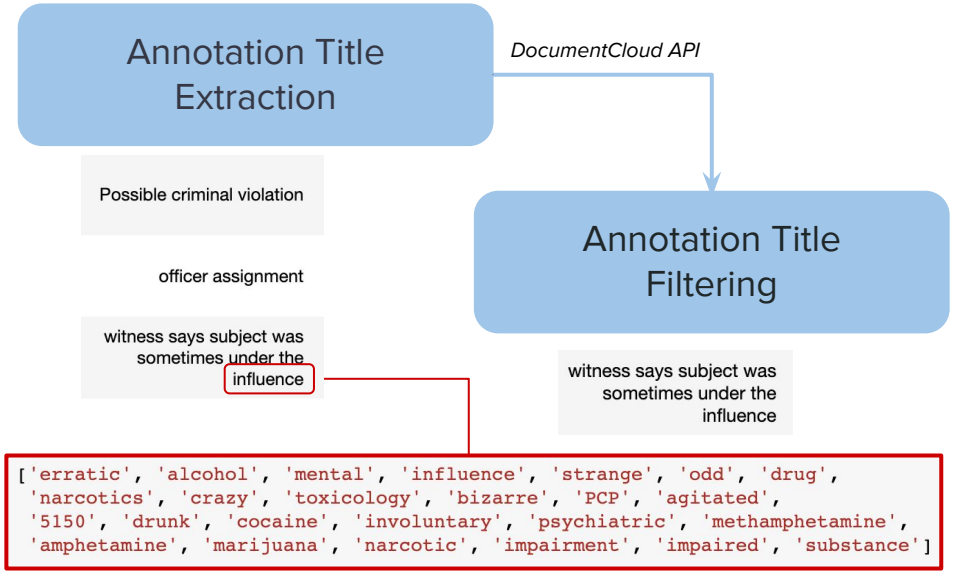


KEYWORDS

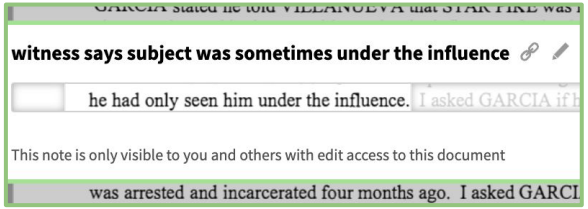
```
['erratic', 'alcohol', 'mental', 'influence', 'strange', 'odd', 'drug',
 'narcotics', 'crazy', 'toxicology', 'bizarre', 'PCP', 'agitated',
 '5150', 'drunk', 'cocaine', 'involuntary', 'psychiatric', 'methamphetamine',
 'amphetamine', 'marijuana', 'narcotic', 'impairment', 'impaired', 'substance']
```

Example of a Annotated File on DocumentCloud

Annotation relevance determined based on title content



KEYWORDS



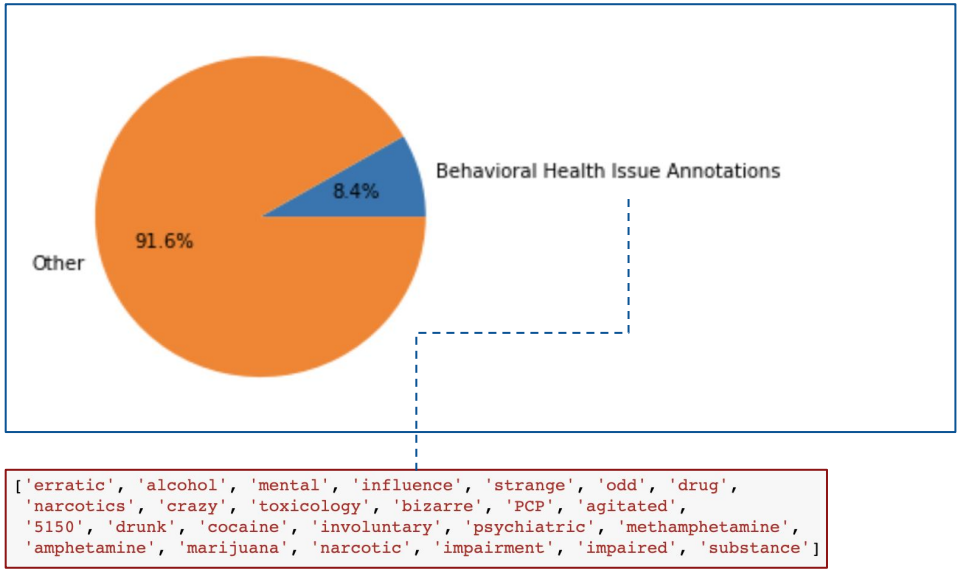
Relevant Annotation: **mentions "under the influence"**

Irrelevant Annotation: **describes administrative detail**

Ratio of relevant to irrelevant annotations

8.4%

of annotations proved “relevant,” or had words in their title belonging to our list of keywords



KEYWORDS

Final train/test set comprises 936 documents

- **Documents:** Originally 766 documents, oversampled positive documents to result in **936** total
- **Text:** Narrative text extracted from document classification, cleaned by NER and then filtered by keyword search
- **Labels**: Presence of useful annotations as a label
- **Test Set:** Randomly sampled 10% of the total dataset to use for evaluation

link	label	text	pages_with_keywords
https://www.documentcloud.org/documents/205175...	0	BAKERSFIELD POLICE DEPARTMENT PRESS RELEASE Ch...	[]
https://www.documentcloud.org/documents/205176...	0	BAKERSFIELD POLICE DEPARTMENT PRESS RELEASE Ch...	[]
https://www.documentcloud.org/documents/205189...	0	They could not find it. said ok you got me and...	[1, 2, 17]
https://www.documentcloud.org/documents/205189...	0	BAKERSFIELD POLICE DEPARTMENT PRESS RELEASE Ch...	[]
https://www.documentcloud.org/documents/205936...	1	immediately drew department issued firearm and...	[10, 20]
https://www.documentcloud.org/documents/206029...	1	See Oficer Beatties supplemental report for fu...	[7, 8, 11]
https://www.documentcloud.org/documents/206045...	1	i stated observed the vehicle flee but was una...	[1, 5, 15, 18, 20, 28, 29, 35, 41, 42]

Agenda

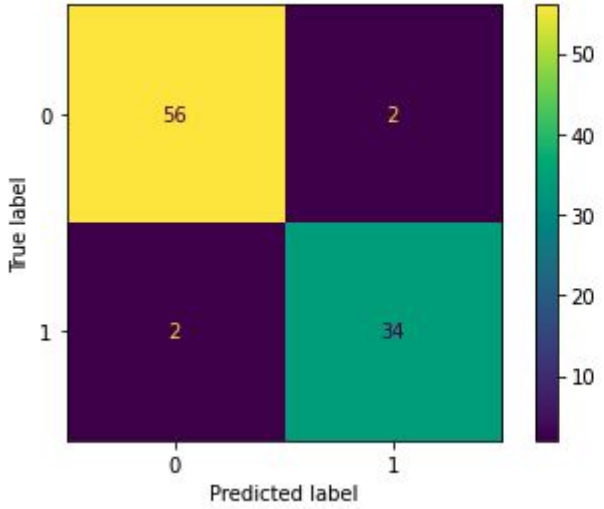
- **Project Background**
- **Data Hierarchy**
- **Our Pipeline**
  - Page Type Classification
  - Forms Pipeline
  - Narratives Pipeline
    - DistilBERT Model Justification
    - Text Cleaning
    - Training Labels
    - Model Performance
- **Conclusion and Future Work**

Narrative model yields strong test accuracy at 96%

Using our document-by-document dataset, we were able to get performance of **~96% accuracy** on our test data (94 documents)!

	precision	recall	f1-score
0	0.97	0.97	0.97
1	0.94	0.94	0.94
accuracy			0.96

Accuracy: 0.957
Precision: 0.944
Sensitivity/Recall: 0.944
Specificity: 0.965



Final Label Aggregation

Retrieve Google Drive files

- **Input:** County name
- **Output:** File path
  given county fold

Santa Clara County

form pages

Forms Pipeline

- **Input:** file path to form pages
- **Tools:** Azure Form Recognizer
- **Output:** form label, label justification

label	rel_info
False	NaN
	1: [936.0,...
	Evidence...

True if either forms or narratives pipeline output True

Along with relevant information to indicate why the label is True

False if both forms and narrative labels were False

Final classification for each **document**

Page Classification

- **Input:** PDF file pa
- **Tools:** LayoutLM
- **Output:** classification label per page *(form, narrative, interview, image, or other)*

rep_image	Label	doc	pg_num
../Shasta County/ima	narrative	2022 Shasta County	16
../Shasta County/ima	narrative	2022 Shasta County	48
../Shasta County/ima	other	2022 Shasta County	15

narrative pages

Narratives Pipeline

- **Input:** String of text
- **Tools:** Tesseract, NER with SpaCy, DistillBERT
- **Output:** narrative label, score

Agenda

- **Project Background**
- **Data Hierarchy**
- **Our Pipeline**
  - Page Type Classification
  - Forms Pipeline
  - Narratives Pipeline
    - Training Data
    - Text Pre-processing
    - Model Justification + Selection
- **Conclusion and Future Work**

Strong results but avenues for future work remain

75%

less time required for journalists to review a case with automation

Now journalists can better prioritize which documents to look at first, fast-tracking the annotation process

Areas for future work:

More granular classification	Can we classify at the *page* level, instead of document-level?
Multi-label classification	Does a document mention “51-50”? What about a specific substance?
Latency reduction in Azure	Currently takes 10 seconds to analyze a form page—moving documents onto Azure storage containers will reduce latency

Thank You! Questions?

Team BigLocal

Stanford DSSG

August 2022

Retrieve Google Drive files

- **Input:** County name
- **Output:** File path to each given county folder

Page Type Classification

- **Input:** PDF file paths
- **Tools:** LayoutLMV2 Classifier
- **Output:** classification label per page *(form, narrative, interview, image, or other)*

rep_image	Label	doc	pg_num
../Shasta County/ima	narrative	2022 Shasta County	16
../Shasta County/ima	narrative	2022 Shasta County	48
../Shasta County/ima	other	2022 Shasta County	15

form pages

narrative pages

Forms Pipeline

- **Input:** file path to form pages
- **Tools:** Azure Form Recognizer
- **Output:** form label, label justification

label	rel_info
False	NaN
True	{'signs of drug impairment': 'Page #1: [936.0,...
False	San Jose Police Department Property & Evidence...

narrative text from forms

Narratives Pipeline

- **Input:** String of text
- **Tools:** Tesseract, NER with SpaCy, DistilBERT
- **Output:** narrative label, score

Final classification for each **document**

Case Files (Raw Data) (%cd ‘content/gdrive/MyDrive/Case Files (Raw Data)’

- dssg_biglocal
  - client_facing_pipeline.ipynb
  - model
  - mylib_dssg
    - form_recognizer.py
    - model.py
    - narrative_pipeline.py ➜ for inference
    - narrative_training_pipeline.py ➜ for training
    - label_agg.py
  - mylib (Hellina modified)
- Shasta County
  - 2022 Shasta County DA
    - Pdf1
    - Pdf2
  - Anderson PD
    - AndersonPDEricHaynes
      - **DIL report.pdf**
    - **CA News Coalition.pdf**
  - images
    - 2022 Shasta County DA__Pdf1-01.jpg
    - 2022 Shasta County DA__Pdf1-02.jpg
    - Anderson PD__AndersonPDEricHaynes__**DIL report.pdf**
    - Anderson PD__**CA News Coalition.pdf**