This document applies to midpoint submission. I am trying to capture information that is in the narrative text only (no additional free text usage) to identify new standard variable(s) that would be helpful for researchers.

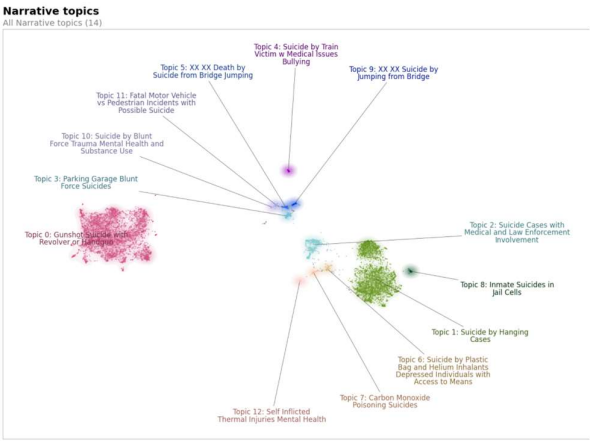
A new variable **ReportsDiscrepancy** is provided here to detect inconsistency between law enforcement (LE) and coroner/medical examiner (CME) reports. My motivation for this idea is that we expect both LE and CME reports to be similar. They should tell, more or less, the same story with different words and specific expertise. However, if they differ too much it could mean that the understanding of the suicide situation is not clear and might be reconsidered. Also, trust in narrative variables would become low in this case or make them difficult to use. Even, in rare case, it might not be a suicide or at least become suspicious.

My methodology to detect such reports is the following:

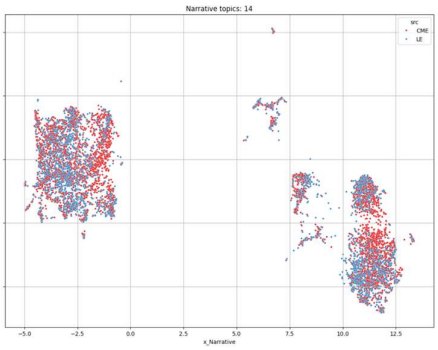
- Reformat train dataset: Stack CME and LE reports to move from 4000 rows to 8000 rows and add a source writer (src) column:

	uid	src	Narrative	tokens	words	multilabel	InjuryLocationType	WeaponType1
0	aaaf	CME	Victim (XX XX) shot himself in a motor vehicle.The Victim's mother called law enforcement and reported the Victim as missing and suicidal with a firearm.The Victim was located in a vehicle in a retail parking lot. When law enforcement approached the vehicle the Victim shot himself.There are no other circumstances.	64	50	4|17|19	2	5
1	aaaf	LE	Victim (XX XX) shot himself in a motor vehicle.The Victim's mother called law enforcement, stated the Victim had made suicidal comments, and she reported the Victim as missing and suicidal with a firearm.The Victim was located in a vehicle in a retail parking lot. When law enforcement approached the vehicle the Victim shot himself.There are no other circumstances.	74	58	4|17|19	2	5

- Apply the following topics modeling (unsupervised approach) to **detect similar Narrative reports** regardless to source writer:
    - Extract embeddings vector (dim=768) for each report from a pre-trained NLP model.
    - Reduce dimension from 768 to 5 with UMAP.
    - Cluster reduced vectors based on density with HDBSCAN algorithm.
    - Extract keywords (ngrams=2,3) and most representative reports in each cluster and feed a LLM model to generate a title/summary of each cluster.



Topics extracted



Topics with source writer (Red=CME, Blue=LE)

```python
# Extract embeddings
embedding_model = INSTRUCTOR('hkunlp/instructor-xl')
# Compute report embeddings (max_seq_length = 512, embeddings size = 768)
instruction = "Represent the report statement: "
embeddings = embedding_model.encode(documents, show_progress_bar=True, batch_size=32)
# UMAP
umap_model = UMAP(n_neighbors=15, n_components=5, min_dist=0.0, metric='cosine', random_state=42)
# Clustering
hdbscan_model = HDBSCAN(min_cluster_size=20, metric='euclidean', cluster_selection_method='eom', prediction_data=True) # We don't want small clusters
# Extract ngrams and keywords
vectorizer_model = CountVectorizer(ngram_range=(2,3), min_df=2, stop_words="english", strip_accents="unicode")
representation_model = KeyBERTInspired(top_n_words=25, nr_repr_docs=5, random_state=42)
topic_model = BERTopic(language="english", top_n_words=25, embedding_model=embedding_model, umap_model=umap_model, hdbscan_model=hdbscan_model,
                       vectorizer_model=vectorizer_model, representation_model=representation_model, calculate_probabilities=True, verbose=True)
topics, probs = topic_model.fit_transform(documents=essays, embeddings=embeddings)
```

Python related code sample

- For each report identify the ones that are not in the same cluster to feed the new binary variable (ReportsDiscrepancy).

Once executed, I have identified **171 reports** over 4000 with **discrepancy**. After a quick review of them we can notice that main root causes of discrepancies are:

- Story is different: Contrariness and disagreement about the cause of death or victim's history. For example, report *uid=adiw* could make the situation suspicious about the brother:
  - CME reported: *[…] The brother stated the Victim made **no threats of suicide** and was not been bullied at school […]*
  - LE reported: […] *The brother stated the Victim (sister) **attempted suicide** by overdosing in the early year. The brother stated the Victim was dealing with some type of conflict with her friends* […]

uid	src	Narrative	topic_Narrative	topic_title_Narrative	ReportsDiscrepancy
adiw	CME	The Victim, an XX XX, was discovered by her brother on the property of her parents residence. The Victim lives with her mother and brother, she stayed home from school because she has a headache. When the Victim's brother came home he looked for his sister (Victim) and located the Victim laying on the ground a few feet from the house, a revolver was located under her left leg and her cell phone was located under her right leg and a gun shot wound was located to the right temple, 911 was called. Emergency Medical Services arrived and performed an EKG strip which showed asystole, the Victim was pronounced dead on the scene. The brother stated the Victim made no threats of suicide and was not been bullied at school. The Victim obtained a gun that belong to her father. The Victim is currently receiving psychiatric services and was diagnosed with depression. The Victim attempted suicide with prescription medication early in the year. The Victim had friends, but would isolate herself from them. Mother reports that the Victim posted an ominous messages on social media saying "you won't see me again". Alcohol was located in the Victim's room 3/4 empty.	0	Gunshot Suicide with Revolver or Handgun	1
adiw	LE	The Victim, an XX XX, was discovered by her brother on the property of her parents residence deceased, 911 was called. Police arrived and located the Victim, no pulse was located. A large amount of blood was on the ground around her head, a 2 inch revolver was partially under her left leg and a phone was under her right leg. Emergency Medical Services placed a monitor on the Victim, the monitor showed no signs of life, the Victim was dead on scene. The Victim was located on the northern exterior of the home, blood could be seen on the right temple area. The brother stated the Victim (sister) attempted suicide by overdosing in the early year. The brother stated the Victim was dealing with some type of conflict with her friends and she posted something on social media about her friends "regretting this". The Victim said she had a headache and stayed home, her brother came by and discovered her dead. Social media statements were located saying "I can't do this anymore", alcohol was located in the Victim's room.	2	"Suicide Cases with Medical and Law Enforcement Involvement"	1

- Much more details provided either by CME or LE (e.g. report *uid= ajao* with LE content almost empty). For instance *uid=btkc*, LE reports "unknown Cause Of Death" but CME reports "Intentional Drowning as Cause Of Death". In this case WeaponType1 variable is "Unknown" which is wrong. It should be "Drowning".

uid	src	Narrative	topic_Narrative	topic_title_Narrative	ReportsDiscrepancy
btkc	CME	Victim XX XX died from an Intentional Drowning in a body of water. A camper saw a body in the water and reported it to the Department of Natural Resources. Victim was known to heavily consume alcoholic beverages since her boyfriend committed suicide a year ago. She was known to use mollies, N bombs and cocaine. No further information was provided.	9	"XX XX Suicide by Jumping from Bridge"	1
btkc	LE	Victim XX XX died of unspecified causes with an unspecified weapon at an unspecified location. Per the victim's aunt, the victim had been depressed since graduating college. The victim had also been having confrontations with her parents, so she had asked her aunt to come get her. The victim's aunt did notice that an extension cord was missing from their camp site. The victim had been missing for two days at the time of this report. No further details at this time.	-1	"Redacted Suicide Case Reports with Blunt Force Injuries"\n\n(This label captures the key concepts of redacted case reports, suicide, blunt force injuries, and the involvement of law enforcement.)	1

- Different point of views (i.e. witnesses in either LE or CME report).

I plan to compute the cosine distance between LE and CME to weight the discrepancy. Also, my next plan is to identify the root cause of discrepancies automatically by running a local LLM that will feed another variable **ReportsDiscrepancyRootCause**. The prompt will contain role, general instruction and response expected and some examples optionally:

> You are police inspector. You need to summary the root cause of the discrepancy between law enforcement (a.k.a LE) and coroner/medical examiner (a.k.a CME) reports:
> LE report: "[…]"
> CME report: "[…]"
> Provide a discrepancy status such as Contrariness, Disagreement, Details missing, Different point of view …

References:

NLP model: hkunlp/instructor-xl with prompt="Represent the report statement:"
Local LLM: mistralai/Mixtral-8x7B-Instruct-v0.1
Topics modeling Python package: Bertopic