CPDBP: Injury Trends Checkpoint 5: Natural Language Processing

By: Hawkins Gay, Alex Leidner, Ramsey Wehbe

Data Science Seminar Fall 2020

Introduction:

1) Task: **Use NLP from the narratives within the CPDB to identify encounters that result in emergency medical care and if possible mode and outcome of that care – EMS (ambulance), hospital admission, emergency room.**

For this question we were interested in trying to identify direct contact between the CPDB allegations and the medical field. In order to do this, we attained the narrative fields out of the database. We utilized the lawsuit_lawsuit database to access individual complaint reports in the lawsuit filings. We created a summary table (Summary2) from this database containing filing 'id' and the complaint summary ('summary'). This table was then output to a CSV file for analysis purposes. In addition to this, in order to map to injury reports to police districts, a topic we discussed interest in checkpoint 2, id from data area and name of beat were linked to beat_id in data_allegation then convertred to district and linked from data_allegation on data_attachmentfile and linked again to data_attachmentnarrative.

For the analysis file, the following libraries we utilized for data processing, cleaning and NLP initiation: Gensim, spaCy, Keras (with TensorFlow). To initiate data cleaning, we dropped all empty summary cells from the CSV file. After dropping empty cells, there were 46,897 usable inputs. Using the spaCy NLP framework, we loaded the en_core_web_lg English model and replaced common punctuations with spaces, converted to lowercase text and lemmatized the summaries. Next we removed stopwords found frequently in the reports but likely offering little useful classification knowledge; examples include: chicago, officer, detective, accused. While they carry meaning they are ubiquitous and thus not useful for delineation in this domain. At the end of this step, sentences were composed of lowercase lemma with all standard stopwords and punctuation removed.

After completing the preprocessing and cleaning noted above, the sentences were tokenized into individual words (or tokens). The tokens were mapped to word vectors utilizing the Word2Vec word embedding model. This process allows for any word from the vocabulary to be represented as a weighted vector of similar words for model input purposes (see Figure 1).

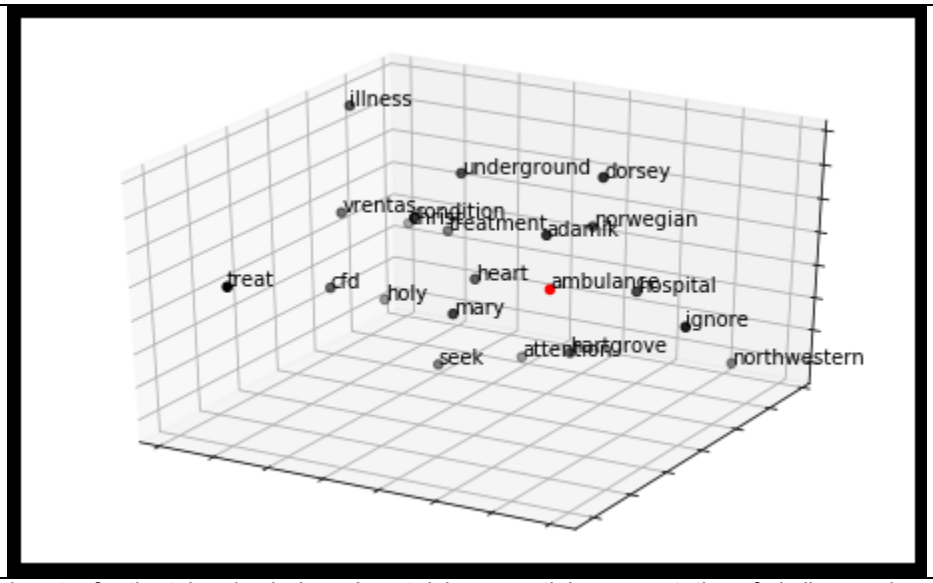
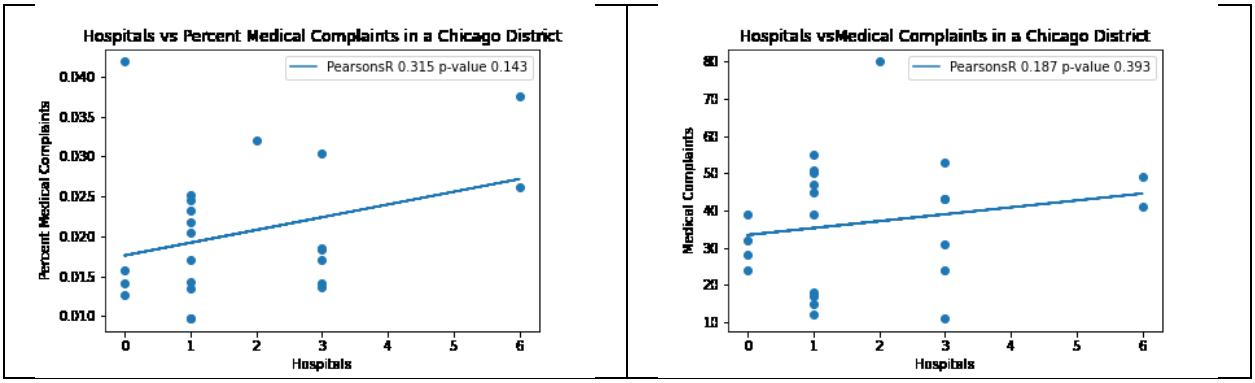


Figure 1: word vector for the token 'ambulance' containing a spatial representation of similar words.

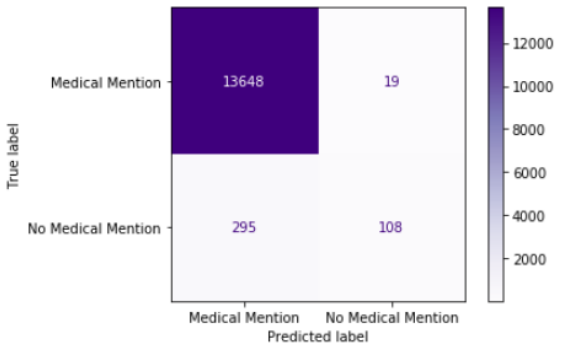
Unfortunately the core of the data for our question does not have labels. Our first language task was to create such labels, using semi-active learning. We utilized Gensim's Word2Vec to create word embedding. From there we chose some common medical terms like ambulance, hospital, doctor and medical to find similar words with similar embedding given their cosine similarities. Next, we then looked at the vectors for these additional words and decided if they indicated interaction with the medical field. For example, consciousness could just mean someone who woke up, which is not something to help us answer our question. However if we inspected further and saw that it in this corpus it was most similar to words like surgery, contusion, and other words that indicate interaction with the medical field or severe injury we included it in our list of meaningful words. Finally we took our cleaned corpus and attached a label of 1 to complaints that mentioned interaction with the medical field, or a severe injury requiring medical attention, and 0 for others. Ideally we would have funding to get appropriate labels so that a classifier could be made to do this automatically in the future with some reliability. Next we wanted to correlate this to physical location of reports.

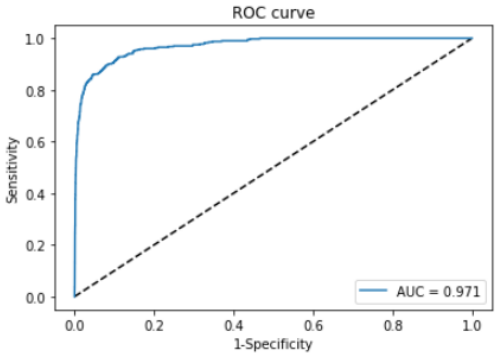
We looked for an association between hospital locations and the medical complaints by plotting the percent of narratives containing medical complaints in each district by the number of hospitals in that district. We also compared this as a total instead of a percent. As an objective measure, the Pearson's coefficient was calculated this relationship. The results indicated a trend towards correlation with a coefficient with 0.315 and a p value of 0.143. While this is not statistically significant it is possible with gold standard labels or curated data that we may find a correlation. This could indicate that people in or closer to a medical district have more medical problems, are seeking medical care, or that police exhibit less restraint in these districts.



Next we used a naïve bayes classifier to create a predictive algorithm for classifying complaint reports as 'medical complaint' or not. The goal of such a model would be to classify complaints that came in the future as medical or non-medical. We believe this is important because complaints have themes of serious injury resulting in hospital utilization, or ignoring medical problems that should be attended to before police protocol. To develop input features for this model, we used a term frequency–inverse document frequency (TF-IDF) bag-of-words model to build vectors from the corpus vocabulary. After splitting the preprocessed data into a train set and test set, the vectorizor was applied to the training data to create a feature matrix. After training, we applied our tuned model to the testing set for assessment. Measures of accuracy, precision, recall and AUC were calculated in this assessment. The naïve bayes classifier was very accurate, at 98%. There were very few false positives, which is reassuring as we did not want to misclassify complaints inappropriately as seeing medical attention. AUROC shows the model's results for sensitivity and specificity, with strong performance in both.

Naïve Bayes Classification Model:



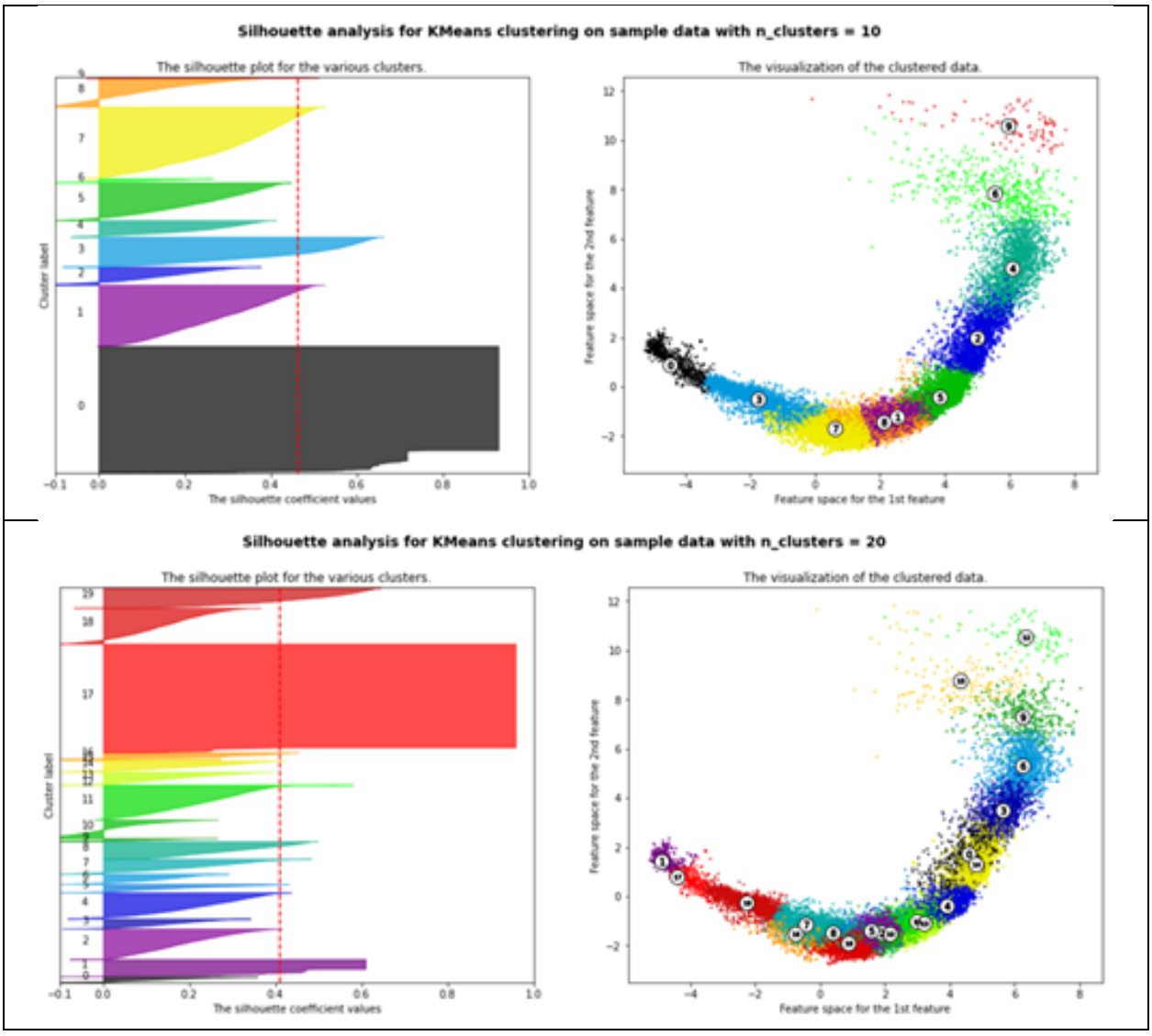


	precision	recall	f1-score	support
0	0.98	1.00	0.99	13667
1	0.85	0.27	0.41	403
accuracy			0.98	14070
macro avg	0.91	0.63	0.70	14070
weighted avg	0.98	0.98	0.97	14070

Given the difficulties in reliably labeling such a large corpus, we attempted an unsupervised approach to topic modeling of the complaint reports with the goal of identifying topic clusters that might identify interactions or mentions of medical care. We used 4 different approaches to unsupervised clustering:

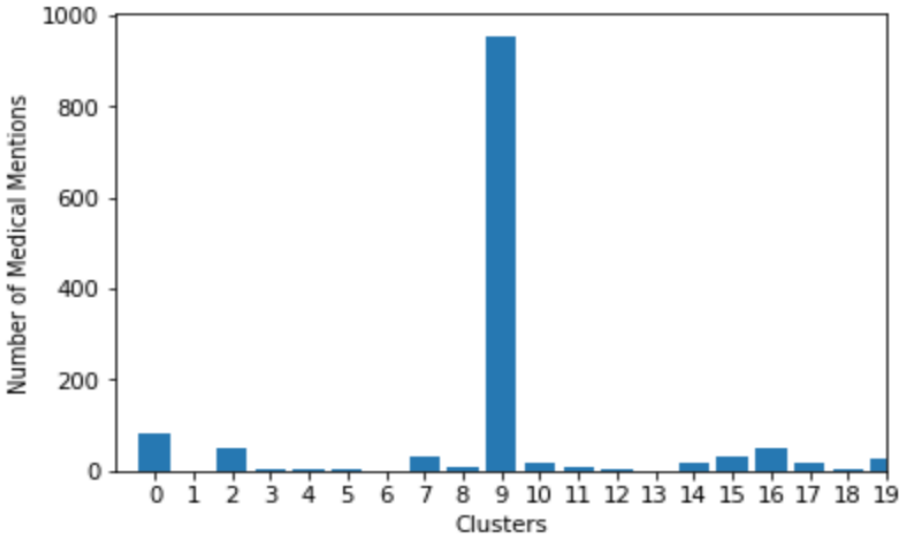
I. Latent Dirichlet allocation (LDA) for topic modeling
II. Taking the mean of word2vec embeddings (trained on complaints corpus) in a complaint report to derive document embeddings, followed by PCA for dimensionality reduction and K-means clustering
III. Doc2vec embeddings (trained on complaints corpus), followed by PCA for dimensionality reduction and K-means clustering
IV. Pre-trained BERT embeddings for the [CLS] token as a form of paragraph embedding for the complaint report, followed by PCA for dimensionality reduction and K-means clustering

For all clustering algorithms except for LDA, we performed a silhouette score analysis in order to determine optimal number of clusters. For example, the following silhouette score analysis for approach IV using pretrained BERT word embeddings suggests that 10 clusters were optimal (*images on next page*).



For LDA we varied k (the number of topics) until the perplexity score and log-likelihood metric were maximized. In order to define topics within each cluster we identified the top 10 words (by TF or TF-IDF) in each cluster. Additionally, we compared clusters to labels derived from regular expressions in the prior step to determine whether certain topic clusters seemed to correspond to medical mentions.

Of all of the above models, we surprisingly found that approach I using LDA performed the best. The optimal number of topics appeared to be 20 for LDA (perplexity score = 439). This is a bar plot showing the number of documents with medical mentions derived from regex labels in each topic cluster:



And finally, top words from each topic cluster are shown below.

```
Topics in LDA model:
Topic #0: fail provide vehicle traffic return license driver phone inven
tory accident
Topic #1: offender weapon situation type armed batter flee authorization
discharge service
Topic #2: state regard case respond fail unknown threaten telephone time
manner
Topic #3: false use direct profanity bag proper robbery murder statement
pant
Topic #4: remove order child action court contact school landlord docume
nt effect
Topic #5: record motorist photo lane change mr ago ms motorcycle thompso
n
Topic #6: area time present september january james basement john lllino
is duty
Topic #7: citation witness issue state incident justification fuck super
visor tell ticket
Topic #8: file complaint state information steal girlfriend obtain occur
pocket sell
Topic #9: tell place arrest home handcuff gun charge arrive battery year
Topic #10: search residence warrant district justification th damage apa
rtment property door
Topic #11: arrest falsely justification member department cause possessi
on probable family burglary
Topic #12: daughter rd father comment pretense mail unwarranted ts cite
demeanor
Topic #13: enter officer home business discover check tenant computer bu
rglarize letter
Topic #14: car drive strike street pull stop sign squad light traffic
Topic #15: hour service approximately unit duty itis employee march fail
july
```

```
Topic #16: vehicle male stop white unknown harass uniformed black state
subject
Topic #17: number fail star request complete attempt inventorie referenc
e behalf affidavit
Topic #18: location work leave department hour june february november oc
tober employment
Topic #19: refuse rude unprofessional state tell allow want ass leave re
quest
```

Notably, topic 9 which corresponded most to our regex labels of medical mentions appeared to involve violent acts such as “handcuff”, “gun”, “battery” which would be more likely to result in medical attention for complainants.

The following complaints from topic cluster 9 show the success of the model in identifying complaints related to injury resulting in medical care.

```
after work day construction contractor cadle visit friend home south cad
le park car friend home begin talk friend garage soon approach ask car g
arage cadle reply demand license registration cadle ask question falsely
traffic violation deny allegation cadle provide license registration fri
end ask cadle cooperate cadle allow search car sobriety test pass find c
ontraband cadle car falsely open alcohol backseat grab cadle slam nearby
gate shove ground strike rib cadle feel pain chest difficulty breathe ha
ndcuff cadle shackle leg throw backseat car drive cadle station lock des
pite repeat request medical care hour transport cadle paddy wagon hospit
al cadle diagnose puncture lung break rib fractured collarbone cadle fal
sely charge resist arrest traffic citation include illegal transportatio
n alcohol month charge citation cadle drop

mccambry purse vehicle bailey pursue mccambry foot   bailey marked vehic
le east  th street south calumet avenue bailey overtake mccambry drive s
idewalk strike injure bailey vehicle mccambry tell hip break need hospit
al bailey ignore plea yell mccambry beat bailey drag foot squad car lean
o nichols arrive drive mccambry station mccambry continue request care
tell   shut   mccambry arrive station walk nichols leano drag station ha
ndcuff wall mccambry tell watts extreme pain need medical care watts tel
l mccambry shut hospital threaten gun case allow hospital hour custody m
edical treatment mccambry hospital diagnose multiple fracture follow mon
th nichols falsely testify mccambry injure fall fence squad car
```

Although these results are promising, even our best unsupervised model was unable to extremely reliably identify topic clusters that corresponded to medical mentions. This is a difficulty in extracting such specific information using unsupervised machine learning and simpler approaches such as the regex based approach we presented earlier seem to be more useful for this task. However, one limitation of our approach was treating each complaint report as a single document and thus a single vector representation. Given each report may cover several topics, it may be more useful to provide sentence level topic clustering and then identify complaint reports that contain sentences related to interaction with medical care. Moreover, the feature space for the pre-trained word embeddings appeared to be oblong, which can present

difficulties for the standard K-means algorithm, and gaussian mixture models may be more useful for this purpose. These are directions for future investigation.