SMU Data Science Review

Volume 7 | Number 3 Article 8

2023

Identifying Locations of Drug Overdose in Las Vegas to Implement the Cardiff Violence Prevention Model

John Girard
Southern Methodist University, girardj@smu.edu

Shikha Pandey
Southern Methodist University, pandeys@smu.edu

Zack Bunn
Southern Methodist University, zackb@smu.edu

Chris Papesh
University of Nevada, Las Vegas, chris.papesh@unlv.edu

Jacquelyn Cheun PhD
Southern Methodist University, jcheun@smu.edu

See next page for additional authors

Identifying Locations of Drug Overdose in Las Vegas to Implement the Cardiff Violence Prevention Model

Authors

John Girard, Shikha Pandey, Zack Bunn, Chris Papesh, Jacquelyn Cheun PhD, and Ying Zhang

Identifying Locations of Drug Overdose in Las Vegas to Implement the Cardiff Violence Prevention Model

Jackie Cheun[2] John Girard[1,] Shikha Pandey[1], Zack Bunn[1], Jackie Cheun[1], Chris Papesh[2,] Ying Zhang[3]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
girardj@smu.edu
pandeys@smu.edu
zackb@smu.edu
jcheun@smu.edu
[2] University of Nevada, 4700 S. Maruland Pkwy., Suite 335,
Las Vegas, NV 189119 USA
Chris.papesh@unlv.edu
[3]Southern Nevada Health District, P.O. Box 3902,
Las Vegas, NV 89127 USA
zhangy@snhd.org

Abstract. This paper will provide an innovative approach to drug overdose prevention programs. Using data from Las Vegas emergency departments, this paper will analyze geospatial trends of drug overdoses. Leveraging the Cardiff Violence Prevention Model, the information is shared with local law enforcement agencies and decision makers to empower them to make evidence-based strategies. This paper highlights the efficacy of a data-driven model in addressing public health issues and underscoring its ability for even broader implementation in urban settings. Findings will suggest significant implications for policymaking, crime prevention, and public health initiatives, demonstrating a step towards a safer Las Vegas.

1 Introduction

Drug abuse remains a daunting challenge for communities nationwide. As per the U.S. Centers for Disease Control and Prevention, over 70,000 drug overdose deaths were reported in 2019. Notably, Las Vegas, the pulsing heart of Nevada and colloquially known as 'Sin City,' mirrors this alarming trend. In line with the FBI's 2020 statistics, Nevada registered a violent crime rate of 541.1 incidents per 100,000 residents—remarkably eclipsing the national average of 366.7. This concerning backdrop prompts an imperative inquiry: Can we harness data as a potent weapon against the pervasive blight of drug abuse in our society?

This paper explores the potential of the Cardiff Violence Prevention Model in addressing this crisis, particularly in Clark County's Las Vegas. The model stands out for its utilization of anonymized data from hospital emergency departments to pinpoint high-risk areas. Armed with such insights, local law enforcement and policy-makers can craft targeted prevention strategies, thus fostering safer communities. In essence, by unveiling the geographical nuances of violence and drug-related incidents, the Cardiff Model facilitates prompt and robust interventions.

Its successful track record in curbing violence in its original Welsh setting and its versatility in different global contexts has garnered commendations. By tapping into anonymized data from local emergency services, it has demonstrated prowess in identifying and addressing hotspots effectively. Its applications are not limited to violence prevention alone; geospatial data has proven instrumental in spotting drug overdose trends and, consequently, guiding precise interventions.

However, certain challenges mar the Cardiff Model's implementation. Its efficacy in unique sociocultural and economic terrains like Las Vegas is still under-explored. Furthermore, the model's dependence on comprehensive data is often met with hurdles as not all overdoses make it to emergency departments' records. Seamless execution also mandates robust inter-agency collaboration, which can get mired in bureaucratic intricacies and varied priorities.

This research embarks on a mission to bridge these gaps, by tailoring the Cardiff Model to Las Vegas's unique context. The intention is to map overdose and violence hotspots using anonymized emergency department data, guiding effective interventions. This study will also scrutinize the model's pertinence to drug overdoses and seek enhancements in data precision. Efforts will be directed towards strategizing to counteract inter-agency cooperation hindrances, thereby paving the way for the model's smooth adoption. By tackling these challenges head-on, this investigation hopes to finetune the Cardiff Model's deployment, ushering in enhanced violence and drug prevention measures in Las Vegas and potentially, other similar locales.

2 Literature Review

This section discusses the literature that was reviewed and has been helpful to gain more information related to the topic of research and machine learning tools and strategies that will help with solving the problem. The review focuses on four chief areas: Proliferation of the Cardiff Model, Rise of Drug Overdose Cases in Las Vegas, Data Science in Healthcare, and Location Based NER.

2.1 Proliferation of The Cardiff Model

In Cardiff, capital of Wales, UK a model was created to prevent public violence. Named for its city, The Cardiff Model, was incorporated using advanced data collection in collaboration between law enforcement and emergency departments utilizing the data to provide improvements to the infrastructure of the community (Shepard et al., 2022). Since its development in 1996 Cardiff, along with 14 similar cities, had a 32% decrease in police-recorded injuries and 42% decrease in hospital admissions due to violence-related crimes (Kollar et al, 2017). This model was further evaluated in other areas. A study in Australia applied the model to alcohol-related presentations and Laura Kollar replicated it in in Atlanta Georgia.

In Australia, a study using the Cardiff Model was conducted on collection of alcohol consumption data in rural areas. In July 2017, individuals aged 18 and above visiting the Emergency Department (ED) were queried by the triage nurse about their alcohol consumption within the past 12 hours, their usual alcohol intake level, the place of most alcohol purchase, and the location of their last drink. Starting from April 2018, notifications were sent out quarterly to the five venues most frequently mentioned within the Emergency Department (ED). The data, which had been deidentified and aggregated, was disseminated to local authorities, encompassing the police, licensing bodies, and the local government. This information spotlighted the top five venues reported in the ED and summarized visits related to alcohol at the ED. To gauge the influence of the intervention on monthly injury rates and ED presentations linked to alcohol, interrupted time series analyses were employed. The results from the Time Series analysis revealed a gradual reduction in the rate of alcohol-related injuries (Baker et al, 2023).

The evaluation process in Atlanta used the same approach of incorporating support between emergency departments from hospitals and law enforcement agencies. The added challenge that it being in the US introduced is the US health system is largely privatized and local law enforcement has more accountability and control needing a joint effort across several agencies (Kollar et al, 2020). Even with the added difficulties of data access and multi-party cooperation the model still proved to be successful in Atlanta. Four Key areas were found to be vital to its success: collaboration between LE and hospitals, fortifying a capacity in hospitals for data collection, data aggregation and investigation, and developing then implementing violence prevention based on collected data.

These studies proved that the Cardiff Model can be used in multiple settings to prevent violence and other related hospital visits. It’s key in both that collaboration of

data collection and reporting between health providers and law enforcement is needed for success in pinpointing areas in need of closer surveillance and support.

The Alcohol study in particular shows that it can be applied to other areas of prevention. There is a clear precedence and demand for it to be used in drug overdose cases as there is less visibility among agencies other than the heath sector (Shepard et al. 2022). Data collection continues to be a hurdle in the US though the positive supporting research from areas that have incorporated the Cardiff Model have influenced many other agencies to invest in its use.

2.2 Rise of Drug overdose cases in Las Vegas

The use of Fentanyl and other synthetic opioids drugs in the greater US has been on a steady increase. In 2021, a significant rise in preventable deaths reached a record high of 224,935, which marks an 11.9% increase. Notably, drug overdose deaths surged by 17.6%, contributing significantly to this unprecedented number. Specifically, there were 98,268 fatalities from drug overdoses in 2021, a record high as well. This has been exacerbated by the COVID-19 Pandemic as there was a 20% increase in opioid deaths and another in 18% in 2021 (National Safety Council, 2023).

The rise of the overdose issues has led first responders and governing bodies to spread thin their resources across the US. In the area of interest of Las Vegas, there may be signs linked to more than just opioid related overdose. The study employed data from the Nevada State Emergency Department Databases to analyze all emergency department visits in 2018-2019 and 2020-2021, comparing the periods before and during the pandemic. The findings revealed that in both 2020 and 2021, several mental health conditions and emergency department visits related to substance use in Nevada exhibited significantly higher probabilities when compared to the pre-pandemic period. There were notable increases in instances of schizophrenia, suicide attempts and thoughts, cigarette smoking, and alcohol consumption during these years. Cannabis-related emergency department visits showed remarkably higher probabilities only in 2020, while there were no significant increases in opioid-related emergency department visits throughout the pandemic compared to the pre-pandemic time span. The research suggests that the pandemic had a more pronounced impact on mental health and substance use during its initial stages, with these effects potentially diminishing as the pandemic continued. This insight can assist policymakers in understanding the societal consequences of the pandemic at both local and national levels, enabling them to make informed decisions and develop effective policies and programs to address large-scale public health crises like the COVID-19 pandemic. (Mojtahedi et al., 2023).

Treatments for overdose are critical to mitigate and prevent deaths and other health concerns. A 2018 study comparing patients in Las Vegas and Tel-Aviv undergoing Methadone maintenance treatment (MMT) testing the long-term effectiveness. The study examined urine samples upon admission and succeeding one year, with long-term retention data up to 23 and 16 years in two clinics. Females were admitted at a younger age with shorter opioid usage compared to males. After one year, both genders had similar retention rates and opioid abstinence. Despite differences in admission

characteristics, both clinics showed comparable outcomes between genders, consistent with the known women's "telescoping effect" (Adelson et al, 2018). In this context means women begin using alcohol and drugs at lower levels than men do but escalates to addiction more quickly.

There is a clear demand for continued treatment and that it is effective overtime. Using the Cardiff Model has been a proactive approach to prevent overdose cases. First responders will know the critical areas where they occur, elucidating awareness for them and the surrounding community. These efforts hope to change the trends and not only avert overdose cases but stop them from happening in the first place.

2.3 Data Science in Healthcare

In the medical field, narrated records by medical professionals are the most important method to communicate in the healthcare domain. These records contain and relay patient-specific medical history and its assessments, which provides important data and information to make further medical treatment decisions for patients. However, usually, this type of medical information is unstructured but using Natural Language Processing (NLP) and Machine Learning, information that is relevant and meaningful can be obtained from massive amounts of narrated medical records.

The methodology used for one study was performing systematic reviews of 110 studies. The attributes of the data under consideration encompassed factors such as its size, origin, gathering techniques, annotations, and pertinent statistics (Spasic et al., 2020). NLP on clinical and medical narrated records was applied successfully to perform tasks such as text classification, Information Extraction (IE), Named Entity Recognition (NER), and Word Sense Disambiguation (WSD). Most of the datasets that were used to train machine learning models had only hundreds or thousands of documents. There were only ten studies that used tens of thousands of documents and very few studies that used more documents. Even though there were much larger datasets available, relatively smaller datasets were used to train models. In the studies that were reviewed, one main reason for such less data being used is the need for manual labeling in supervised machine learning algorithms. Healthcare data is sensitive, so, usually, the training data used to create models was from only one medical facility which then would not generalize well on test data from other medical facilities. Unsupervised machine learning would be efficient here because it doesn't require labeling.

There are several other use cases for NLP in healthcare including but not limited to clinical documentation to help medical professionals with patient health records, speech recognition to transcribe notes for electronic health records, data mining research in healthcare systems for knowledge discovery and in turn delivering better care for patients, clinical decision support to aid clinicians with diagnosis and symptom checking, predictive analysis to help identify patients that are at a higher risk of health inequality and to provide added monitoring. Data Science overall plays an increasingly important role in healthcare to analyze the vast amount of gathered data to improve patient outcomes, cost reduction, enhanced research, better resource allocation, and operational optimization to name a few.

2.4 Location Based NER

Named Entity Recognition (NER) is usually not used in medical data to extract location, but other domains of text datasets (corpora) have been used and studied to find methods to extract location.

Location information is a key factor when there are events like a natural disaster, an emergency incident, or a crisis, all of which involve public safety. In recent times, Twitter, now known as X, has been one of the examples where it was used as a main form of communication during crises to provide rescue and relief operations in real-time. In these fast-developing situations, quick and appropriate response is needed, and it becomes very important to have information about the geographical location of the incident and its users. However, it is often noticed that either user's location information can't always be gathered from tweets, or it isn't very reliable. It could be difficult to extract information about location from tweets because text from tweets is unstructured, often has improper English, poor grammar, incorrect spellings, abbreviations aren't standardized, and so on. In a particular study, researchers sought to extract geographic terms from tweets by employing a model based on a Convolutional Neural Network (CNN). The n-gram features of the CNN were attributed to the high F1-score achieved by their method. (Kumar et al., 2019).

In a different research study conducted by Gritta et al., 2020, Geoparsing was utilized with the objective of converting place names, known as toponyms, found in unstructured text into geographic identifiers (coordinates with latitude and longitude). Toponyms are essentially names of specific places or locations. Up until now, observational approaches in the field of geoparsing have lacked a standardized evaluation framework that clearly defines the task, specifies the metrics to be used, and outlines the data used for comparing the most advanced systems in this domain. This inconsistency in evaluation has further complicated matters, making the assessments less representative of real-world applications, primarily because it fails to distinguish between various types of toponyms. Addressing these shortcomings, the researchers in their publication introduce a new comprehensive framework that consists of three key components.

The first part of this framework focuses on Task Definition and aims to bring clarity through an in-depth analysis of linguistic corpora. This analysis has led to the proposal of a finely detailed Pragmatic Taxonomy of Toponyms. The second part of the framework delves into Metrics, discussing and reviewing them extensively to ensure a robust evaluation process. Additionally, it provides recommendations that are particularly valuable for practitioners working in the fields of Named Entity Recognition (NER) and Geoparsing. The third and final part of the framework deals with Evaluation Data. It introduces a novel dataset named GeoWebNews, designed to facilitate the testing and training of models. This dataset not only serves the purpose of fine-grained Geotagging and Toponym Resolution (Geocoding) but is also well-suited for the development and assessment of machine learning Natural Language Processing (NLP) models. Collectively, these three interconnected parts contribute significantly to the improvement and standardization of geoparsing. They serve as a valuable resource for researchers and practitioners in the fields of NER, Geoparsing, and beyond.

Another study focused the research on analyzing Tweets to build models for geo location prediction using NER. They used a very large number of tweet posts for this

research and began with latitude and longitude prediction. Given that a tweet is a short text which contains noise and is ambiguous in nature and only a minute percent is geo-tagged, there was some success in their predictions. They then adjusted their research goal, while still keeping the goal of granularity of location identification. Working along with domain experts (urban planners) they built deep neural network models using natural language processing (NLP) techniques using tweets that are not geo-tagged to predict geolocation at varying granular levels like neighborhood, zipcode, and longitude with latitudes (Dutt et al., 2021).

3 Methods

3.1 Data and Background

Using Machine Learning and Natural Language Processing, our study aimed to identify locations of drug overdose in Las Vegas to implement the Cardiff Violence Prevention Model. Based on preliminary data and information shared by Chris Papesh's team (Chris Papesh serves as the Principal Investigator on the UNLV Cardiff Grant), we expected to see a lot of drug overdoses around the resorts and casinos located on the Las Vegas Strip.

Data for this paper is sourced from emergency departments in Las Vegas hospitals. This data was shared by Chris Papesh and his team. Anonymized data was collected from a database that contains hospital ER records for drug overdoses and was uploaded in CSV file to a google drive that our team had access to. Data was anonymized to protect patient's personal identifiable information (PII) like gender, race, birthdate etc., because it contains vast amounts of incident reports that detail drug overdoses. The way this data was prepared and collected is not unusual of other Cardiff Model implementations.

This CSV file contains various types of information like date and time of admission in ER, name of the hospital, reason for patient's visit like injury and type of injury, discharge diagnosis medical codes, and two fields with unstructured text called "Chief Complaints". This dataset contained about 35,000 records for a period of about 6.5 years (2017-May 2023).

Our first figure demonstrates the amount of drug overdose cases reported in the past 6 years. 2023 has less cases as this data was pulled in late May. We can observe a steady uptick in reported cases from 2017 through 2022, however, indicating that drug

overdoses have become even more of a problem in the Las Vegas community in the most recent years.

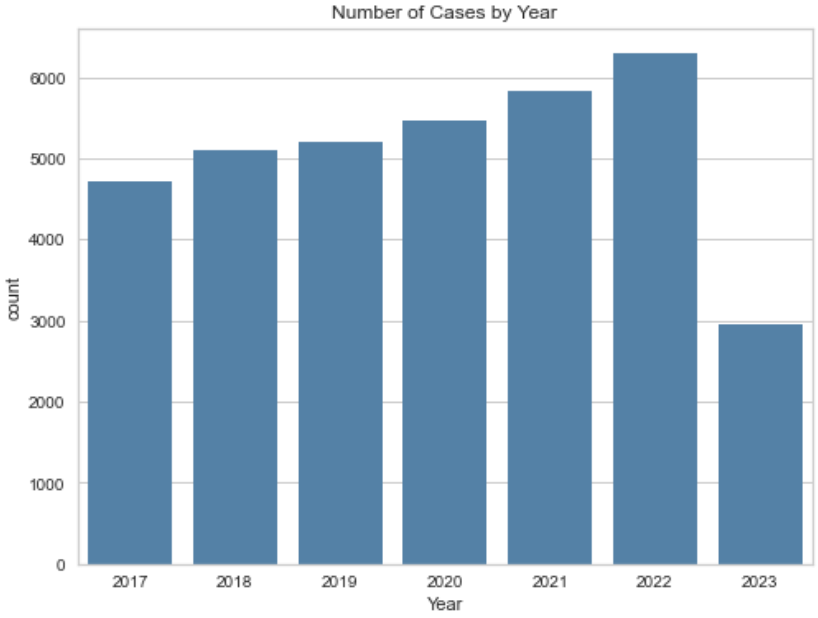

Fig. 1, Cases by *Year*. Count is in the thousands on the y axis and *Year* labeled on the x axis.

The two "Chief Complaints" fields contain unstructured text are from ER medical notes and may have location information embedded in those notes if it was available.

- 'ChiefComplaintOrig' is the medical note that is taken down originally by the hospital personnel upon a patient's admission. An example from the data is:

  No Location
  SI with a plan to overdose on seroquel

  Location
  Patient was picked up at Arby's on Flamigo and Jones in the bathroom heroin ovedose he was awake

- 'ChiefComplaintParsed' is the copy of original notes with all letters in upper case, all acronyms expanded in their full form, and any punctuation removed. An example from the data is:

  No Location
  PATIENT OVERDOSE D AS SUICIDE ATTEMPT ON DONEPEZIL 10 MILLIGRAM STATED STRESS RIGHT T LIVING AT HOME WITH MULTIPLE FAMILY MEMBERS WHO STEAL MONEY PERSONAL ITEMS TOOK 8 10 TABS UNKNOWN HOW LONG AGO

Location

PATIENT PICKED UP AT DENNY S ON FREMONT FOURTH SORE THROAT PER EMS PATIENT CALLED COMPLAINS OF SUICIDAL THOUGHTS WITH PLAN TO OVERDOSE ON HER MUSCLE RELAXANTS STATES THAT SHE IS HAVING AUDITORY VISUAL HALLUCINATIONS PATIENT

As the first step in any data modeling process, data is cleaned, prepped and an exploratory data analysis (EDA) is done to understand the data. We did some data cleaning for null values and duplicate data. There were 896 (2.5%) instances where both 'ChiefComplaintOrig' and 'ChiefComplaintParsed' fields were either found null/blank or had a special character that didn't hold any pertinent information to be valuable in predicting a location, and so they were removed. There were also 10 duplicate rows of data found and out of those only original rows of data were kept.

Our next figure illustrates a word cloud taken from the original chief complaint field. As we can see a frequency in words such as overdose, suicide, Narcan, unresponsive, and words referring to various drugs are represented in this field. All words in this field come from an origination point where someone typed in what the patient was experiencing.

Fig. 2, Frequency of words found in *ChiefComplaintOrig* and *ChiefComplaintParsed* fields within the data. Larger words shown here are ones that occur the most often.

Research engaged key stakeholders including representatives from healthcare, law enforcement, community members, and other policy makers. Interviews were set up with some stakeholders to understand more about issues that are impacting them directly.

The Cardiff Model is implemented based on findings from data analysis and stakeholder engagement. The impact of predictive interventions using collected data and assessing changes in identified hotspots and trends over time.

Throughout these methods, the research aims to not only apply the Cardiff Model to a new context, but to also enhance its implementation by addressing identified gaps and problems.

3.2 Model Selection

As mentioned above to best test a model for selection, a test/validation dataset was generated by randomly selecting 20% of our full dataset. Each record in said validation dataset was labeled with a "0" by researchers when no location information was present. When a location was identified, a "1" was manually put in. If a location was found in the observation, the latitude and longitude coordinates of that location were input into the dataset manually.

The NER and deep learning models underwent testing on the validation dataset, encompassing both "Chief Complaint" fields. When the model failed to identify a location, it resulted in a "0" prediction. Conversely, if a location was successfully identified, it was marked as "1". However, since NER models may occasionally extract incorrect locations, an additional step was incorporated. For predictions where a location was identified, the Google Maps API was employed to retrieve the coordinates of the named entity. These predicted coordinates were subsequently compared to the actual coordinates, and if they fell within one mile of each other, the prediction was classified as a true positive, denoted as "TP", signifying a correct prediction. In cases where the distance exceeded one mile, it was categorized as a false positive, indicating an incorrect prediction.

3.2.1 Ktrain

Ktrain stands out as a versatile tool that functions as a wrapper for deep learning libraries like TensorFlow Keras. Its primary aim is to simplify the intricate process of establishing neural network pipelines for model development, making it a user-friendly and efficient choice, with minimal lines of code required for implementation. The software comprises two fundamental modules dedicated to image and text classification, covering a wide array of essential functionalities, including learning rate optimization, ready-made models for text data, data preprocessing, misclassification detection, and model deployment. In this particular analysis, a comprehensive Question-Answering system was harnessed, utilizing Bidirectional Encoder Representations from Transformers (BERT) to extract location information. The model was subjected to a battery of four specific inquiries: "the street names?", "the business name?", "the exact location?", and "the geopolitical location?"

3.2.2 spaCy

spaCy is a NLP library that includes a feature for NER. It is known for it's efficiency and ability to handle large volumes of text data. spaCy's NER module can identify four main types of location entities:

- "ORG" for companies
- "GPE" for cities, states, and countries
- "FAC" for buildings, or structures
- "LOC" for non-geopolitical locations

A fundamental baseline model was created using this method using the entity types above. A second spaCy model was developed as well, which integrated the EntityRuler function to introduce a custom gazetteer containing street names in the Clark County area. The function assigns a street label and matches to anything found in the gazetteer enhancing the model's ability to recognize street names accurately.

3.2.3 Custom NER with CRF

The Custom NER model using Conditional Random Fields (CRF) uses contextual information from neighboring words for classification. This CRF model was constructed and trained on the GMB. The model's architecture uses a range of words features, POS tags, suffixes based on characters of words, indicators for word position, and binary markers for uppercase, title case, and numeric words. Once the model was trained on the GMB data, it was applied to both of the "Chief Complaint" fields.

3.3 Evaluation

To best understand and assess the performance of the models, several metrics were employed: Accuracy, precision, recall, and F1-Scores. The primary aim was to have as little of false positives and false negative results. The model with the highest F1-Score was to be the model selected as the most appropriate solution for this particular problem.

4 Results

4.1 Model Analysis

We selected to run 4 models on both "Chief Complaint" fields which contain the text from ER medical notes and potentially have location information embedded in those notes. Among all the models, Ktrain model performed with best F1-score on both "Chief Complaint" fields. Ktrain model scored and performed only slightly better on "ChiefComplaintParsed" field probably because this field had a lot of acronyms in their expanded in their full form.

A table comparing the metrics as to how each of these various models performed is shown here sorted by highest to lowest F1-score:

Table 1 *Model Comparison* across 4 types being tested on *ChiefComplaintParsed* and *CheifComplaintOrig* each.

Model	Field	Accuracy	Precision	Recall	F1
Ktrain	ChiefComplaintParsed	0.9383	0.4536	0.4213	0.4369
Ktrain	ChiefComplaintOrig	0.9338	0.4175	0.4550	0.4354
Custom NER w/ CRF	ChiefComplaintOrig	0.8996	0.1962	0.2732	0.2284
spaCy w/ Gazetteer	ChiefComplaintParsed	0.8268	0.1305	0.5424	0.2104
Custom NER w/ CRF	ChiefComplaintParsed	0.9129	0.2200	0.1887	0.2032
spaCy	ChiefComplaintOrig	0.8556	0.1302	0.3305	0.1868
spaCy w/ Gazetteer	ChiefComplaintOrig	0.8543	0.1267	0.3285	0.1829
spaCy	ChiefComplaintParsed	0.9027	0.1514	0.1422	0.1467

An observation is that precision and recall scores generated by the Ktrain model are similar with an F1-score of 0.4369. This indicates that false negatives and false positives are identified similar times and when a location was present, the correct location was identified around 44% of the time. This model's accuracy metric at 0.9383 was high and could be chalked up to a high number of records in the data that do not have a location, so, it identified that accurately majority of the time.

Custom NER with CRF on 'ChiefComplaintOrig' field performed best after the Ktrain model with F1-score of 0.2284 following closely by spaCy with the street name Gazetteer on 'ChiefComplaintParsed' field with F1-score of 0.2104 but they both were much lower than Ktrain models. spaCy with the street name Gazetteer on the 'ChiefComplaintParsed' field had the highest recall with 0.5424 which implies it performed best at predicting locations that were there in the data, but it had a low Precision with 0.1305 indicating towards high false positive predictions due to the overidentification of streets.

It is important to reiterate that while the observations provide insight into the relative performance of the models, the ultimate determination of the best model should be made upon acquiring comprehensive results, which will offer a more conclusive assessment of their effectiveness.

5 Discussion

5.1 Application and Findings

Out of all the models we found that our Ktrain Model in conjunction with using our location extraction tool, BERT, was the best performer on the 'ChiefCompliantParsed'

field though only marginally than the 'ChiefCompliantOrig' field. To convey our model in a practical sense we wanted to be able to give a view of the data. Shown below are the findings plotted out onto a street map of Las Vagas made with the Folium Python package and Google Maps' API. Each bubble is a representation of a reported overdose case. They are separated by color that shows the density of cases with green being the smallest number and orange as the largest. This is only a screenshot of it, as it's able to be used like Google Map's web app. As you zoom in the bubbles separate to smaller sections all the way until there's individual cases visible.

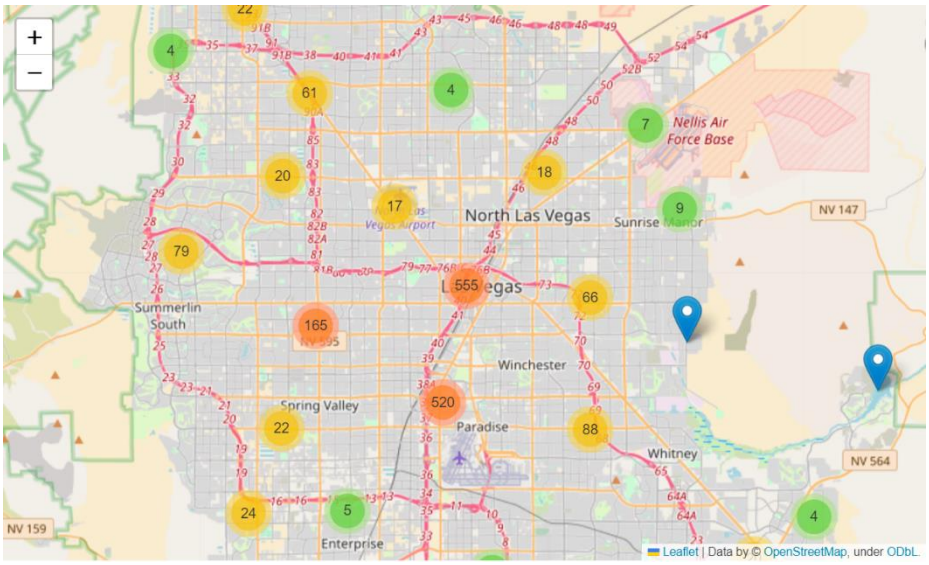

Fig. 4, Model map of the greater Las Vegas with Overdose Location *Hotspots*

To accompany that first map another was created to display our K-Means clustering. The means are set to k = 4 with letters labeled A through D. This provides a direct view of city zones having the greatest density of overdose cases. Interestingly, C isn't visible which suggests the cluster very close inside another one. Given that the A cluster is the most highly concentrated one, we believe C may be somewhere located in there.

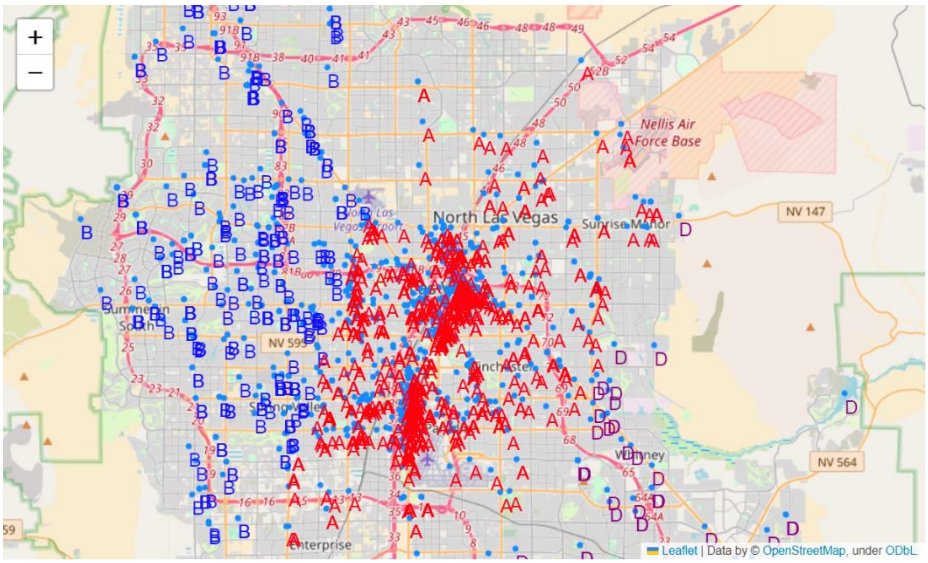

Fig. 5, Model map for Las Vegas organized by Overdose Location *Clusters*

Both of these maps are made available to Chris Papesh and his team to utilize as they see fit. It should be taken attention however that these models may not have correctly identified all locations of cases so we cannot advise the use of these for the carrying out of professional judgements and conclusions.

5.1.1 Quality of Data

The research conducted was aimed to accurately identify areas in which drug overdoses would occur using NER. Despite our Ktrain model having the most success, there are areas in which this research could have been improved. Perhaps our largest issue lies within the inconsistency of the two "Chief Complaint" fields. This data is incredibly unstructured, and there is seemingly no structure or standardization to this. Due to this, there is difficulty in assessing any location data. A row from the data was selected at random:

"*Pt found outside on the sidewalk unresponsive 911 was called. Pt received 1 mg of Narcan by fire and became more responsive. Pt is now alert enough to answer questions. Dex 58*"

From this selection, no location is identified, not even an intersection. This example details a large challenge in efforts to accurately identify and categorize location information. This is consistent with the data that was provided as there was not many locations found within it.

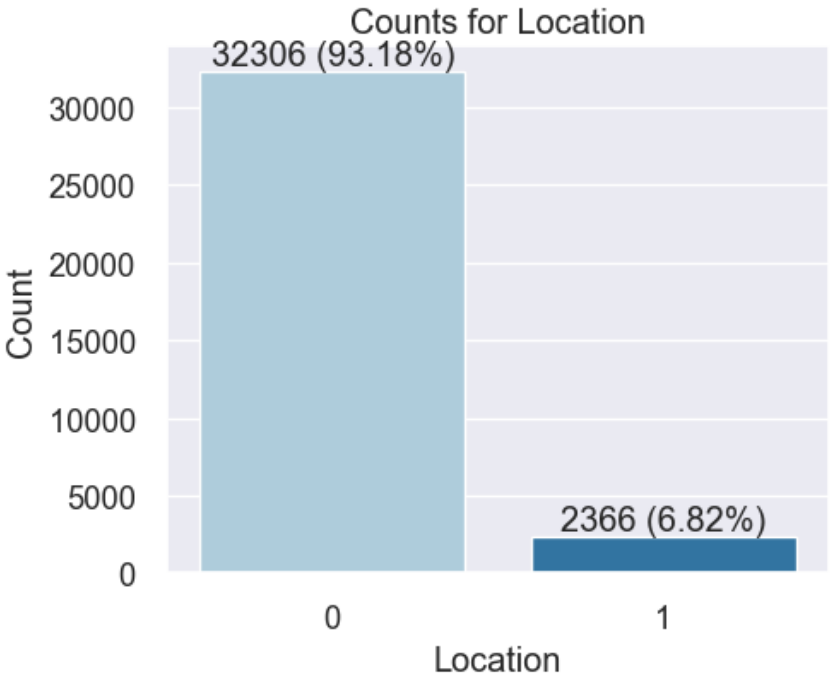

Fig 6., Plot showing distribution of *Location* availability in the data.

The above plot indicates that less than 7% of the data had an identifiable location present. A standardization of these intake forms with location information would have drastically benefited the model's performance.

The implementation of well-structured admission forms featuring designated data input fields is highly likely to result in a significant enhancement in data quality comprehension. At present, it appears that the text input within these fields during triage is not undergoing thorough scrutiny or analysis, potentially leading to a lack of in-depth understanding. The integration of a data validation approach, particularly one encompassing address information (or intersections), could have substantially elevated the performance of our models.

Another large inconsistency with the data was references to other hospitals. The NER model would capture these as named entities, despite the assault not transpiring at the particular institution.

Addressing such issues programmatically, for instance, through rule-based Named Entity Recognition (NER), could provide some solutions. Nevertheless, enhancing the data's quality offers a potential avenue to enhance model performance, especially in the context of drug-related cases. Similar to other municipalities that have adopted the Cardiff Model, Clark County might contemplate introducing dedicated data fields for specific drug-related incidents within their medical systems. These structured fields could serve to bolster the accurate identification of drug-related events and minimize the incidence of erroneous associations. It's crucial to acknowledge that even with

structured data fields, the possibility of typographical errors and ambiguities in location entries remains a consideration that warrants attention.

In conclusion, the quality of data is paramount to the success of models that rely on it for decision-making. Inconsistent and incomplete data pose formidable challenges that, with structured data entry fields and validation procedures, can be effectively mitigated, leading to improved model performance and more reliable outcomes.

5.2 Future Research

Our exploration of location identification in medical records illuminated numerous avenues for future research, particularly in the context of addressing data quality challenges and further advancing the capabilities of machine learning models. Several intriguing directions for future investigations have emerged from our work.

5.2.1 Leveraging Medical Datasets

One avenue for further studies lies in the exploration of leveraging comprehensive medical datasets or dictionaries for improving location identification. Rather than manually creating rules, as was done in this study, incorporating medical terminologies and acronyms can enhance the accuracy and consistency of identifying and cleaning such abbreviations. An approach such as this could significantly streamline data preparation and further boost the overall performance of location prediction models.

5.2.2 Customization and Flexibility

The Ktrain model's approach to simplicity was both beneficial and a limiting factor. Future research should focus on introducing customization options that allow for more tailored approaches. A potential ability to incorporate whitelisting or blacklisting in a firewall fashion could allow users to further fine-tune the model for specific contexts, which would allow for more accurate and relevant predictions.

5.2.3 Location-Centric Datasets

Expanding the horizon of location prediction research, future investigations can delve into the creation of location-specific datasets. Training models on datasets that are tailored to the geographical area of interest, such as Clark County, can yield superior results. Location-specific datasets would contain local nuances, regional terminologies, and specialized location references that may not be adequately captured by general datasets like the Groningen Meaning Bank (GMB).

5.2.4 Automated Labeling Methods

The manual labeling of data for model evaluation is a time-consuming process and inefficient process. Research conducted in the future could utilize the implementation of automated labeling methods. Exploring some techniques such as RNN's with LTSM can provide much more efficient data labeling processes and accelerate the development and evaluation of location prediction models.

In conclusion, the research conducted has illuminated several opportunities for future exploration, driven by the imperative need to address data quality and unlock the full potential of machine learning models for location prediction in medical records. These potential avenues of research hold promise of not only enhancing the accuracy and robustness of such models but broadening their applicability to diverse contexts and domains.

5.3 Ethics

When working with emergency room records, the ethical considerations of data privacy and patient confidentially is large. Patient data contained within the ER records is inherently sensitive, and thus, the utmost care was taken to preserve the privacy of patients and the principles of data ethics throughout this process.

This research approach began with the meticulous exclusion of all fields containing PHI information, such as patient names, DOB, and home addresses. This information was excluded before being turned over. The deliberate omission of aforementioned data fields was a foundational step to ensure that sensitive information remained secure and was not exploited in the research.

In tandem with PHI exclusion, the research adheres to a robust data governance framework. The established framework outlines comprehensive guidelines and practices for the responsible and ethical handling of patient data. The actual data will also be unavailable from the researchers after the culmination of the project to ensure no information could be subsequently leaked.

In summary, the ethical commitment to patient data privacy was taken seriously throughout the course of the research. The study diligently adhered to any and all principles of patient confidentiality, data security, and HIPAA compliance while using ethical data governance principles to protect patient information throughout the research process.

6 Conclusion

The primary aim of this research was to assess the hotspots of where drug overdoses were occurring. To aid in this goal we assessed the feasibility and performance of extracting location information from hospital admissions. Unfortunately, the outcomes of the machine learning models have demonstrated a lack of consistency and reliability in predicting locations. Consequently, the models current state is deemed inadequate on making informed and investment decisions to aid in community improvement, which is a goal of the Cardiff Model.

In response to these challenges and concerns, several strategies have been proposed to enhance models and address the quality of data concerns. Specifically, and most

importantly, integrating mandatory location data into the entry fields of the Electronic Medical Records is HIGHLY recommended. This strategic investment holds the potential to improve data consistency and accuracy which will best enhance model performance. Further, the research suggests that a cross-city model assessment could be completed after mandatory location field capture is the most prudent way to gauge adaptability and effectiveness. In summary, addressing data quality issues and augmenting machine learning models for location prediction in medical records is imperative. These recommendations lay the foundation for the development of more reliable tools, and ultimately supporting informed decision-making for community improvement initiatives and healthcare systems.

Acknowledgments. Jacquelyn Cheun-Jensen, Ph.D. – Capstone Advisor

References

1. Jonathan Shepherd (2022), 2022-Report_Cardiff-Model-for-Violence-Prevention https://www.cardiff.ac.uk/__data/assets/pdf_file/0006/2621418/2022-Report_Cardiff-Model-for-Violence-Prevention.pdf
2. Mercer Kollar, L.M., Jacoby, S.F., Ridgeway, G., Sumner, S.A. (2017). Cardiff Model Toolkit: Community Guidance for Violence Prevention. Atlanta, GA: Division of Violence Prevention, National Center for Injury Prevention and Control, Centers for Disease Control and Prevention. https://www.cdc.gov/violenceprevention/pdf/cardiffmodel/cardiff-toolkit508.pdf
3. Butchart A, Phinney A, Check P, Villaveces A. (2004). Preventing violence: A guide to implementing the recommendations of the World report on violence and health. Department of Injuries and Violence Prevention, World Health Organization, Geneva.
4. Carter, P. M., Cranford, J. A., Buu, A., Walton, M. A., Zimmerman, M. A., Goldstick, J., Ngo, Q., & Cunningham, R. M. (2020). Daily patterns of substance use and violence among a high-risk urban emerging adult sample: Results from the Flint Youth Injury Study. Addictive Behaviors, 101, 106127. https://doi.org/10.1016/j.addbeh.2019.106127
5. Spasic, I., & Nenadic, G. (2020). Clinical Text Data in Machine Learning: Systematic Review. JMIR Medical Informatics; JMIR Med Inform, 8(3), e17984. https://doi.org/10.2196/17984
6. Neill, D. B., & Herlands, W. (2018). Machine Learning for Drug Overdose Surveillance. Journal of Technology in Human Services, 36(1), 8–14. https://doi.org/10.1080/15228835.2017.1416511

7. Wells, K. B., Sherbourne, C. D., Sturm, R., Young, A. S., & Audrey Burnam, M. (2002). Alcohol, Drug Abuse, and Mental Health Care for Uninsured and Insured Adults. Health Services Research, 37(4), 1055–1066. https://doi.org/10.1034/j.1600-0560.2002.65.x
8. Duncan, D. F., Ellis-Griffith, G., Nicholson, T., & Nimkar, S. (2023). Health care administration and drug policy. The International Journal of Health Planning and Management, 38(3), 735–746. https://doi.org/10.1002/hpm.3621
9. Wagner, K. D., Harding, R. W., Kelley, R., Labus, B., Verdugo, S. R., Copulsky, E., Bowles, J. M., Mittal, M. L., & Davidson, P. J. (2019). Post-overdose interventions triggered by calling 911: Centering the perspectives of people who use drugs (PWUDs). PloS One, 14(10), e0223823–e0223823. https://doi.org/10.1371/journal.pone.0223823
10. Gritta, M., Pilehvar, M.T. & Collier, N. A pragmatic guide to geoparsing evaluation. Lang Resources & Evaluation 54, 683–712 (2020). https://doi.org/10.1007/s10579-019-09475-3
11. Mercer Kollar, L. M., Sumner, S. A., Bartholow, B., Wu, D. T., Moore, J. C., Mays, E. W., Atkins, E. V., Fraser, D. A., Flood, C. E., & Shepherd, J. P. (2020). Building capacity for injury prevention: a process evaluation of a replication of the Cardiff Violence Prevention Programme in the Southeastern USA. Injury Prevention, 26(3), 221–228. https://doi.org/10.1136/injuryprev-2018-043127
12. Baker, T., Taylor, N., Kloot, K., Miller, P., Egerton-Warburton, D., & Shepherd, J. (2023). Using the Cardiff model to reduce late-night alcohol-related presentations in regional Australia. The Australian Journal of Rural Health, 31(3), 532–539. https://doi.org/10.1111/ajr.12983
13. Kumar, A., & Singh, J. P. (2019). Location reference identification from tweets during emergencies: A deep learning approach. International Journal of Disaster Risk Reduction, 33, 365-375. https://doi.org/10.1016/j.ijdrr.2018.10.021
14. Dutt, F., & Das, S. (2021). Fine-grained Geolocation Prediction of Tweets with Human Machine Collaboration. ArXiv, abs/2106.13411.
15. National Safety Council, (2023) Drug Overdoses. Injury Facts. https://injuryfacts.nsc.org/home-and-community/safety-topics/drugoverdoses/#:~:text=The%20number%20of%20preventable%20deaths
16. Mojtahedi, Z., Guo, Y., Kim, P., Khawari, P., Ephrem, H., & Shen, J. J. (2023). Mental Health Conditions- and Substance Use-Associated Emergency Department Visits during the COVID-19 Pandemic in Nevada, USA. International Journal of Environmental Research and Public Health, 20(5), 4389–. https://doi.org/10.3390/ijerph20054389
17. Yaman, E., & Krdžalic-Koric, K. (2019). Address Entities Extraction using Named Entity Recognition. 2019 7th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), 13-17. doi: 10.1109/FiCloudW.2019.00016.
18. Sarkar, D. (2019). Text Analytics with Python A Practitioner's Guide to Natural Language Processing (2nd ed.). Apress. https://doi.org/10.1007/978-1-4842-4354-1
19. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python (1st ed.). O'Reilly Media, Inc.
20. Adelson, M., Linzy, S., & Peles, E. (2018). Characteristics and Outcome of Male and Female Methadone Maintenance Patients: MMT in Tel Aviv and Las Vegas. Substance Use & Misuse, 53(2), 230–238. https://doi.org/10.1080/10826084.2017.1298619