SMU Data Science Review

Volume 6
Number 1 *Spring 2022*

Article 1

2022

Identifying Locations of Violent Injuries in Las Vegas to Implement the Cardiff Violence Prevention Model

Samuel VonPaays Soh
Southern Methodist University, svonpaayssoh@smu.edu

Kristi L. Herman
Southern Methodist University, kristih@smu.edu

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

Tim Musgrove
Callisto Media Lab, tmusgrove@callistomedia.com

Ying Zhang
Southern Nevada Health District, zhangy@snhd.org

Recommended Citation

VonPaays Soh, Samuel; Herman, Kristi L.; Papesh, Chris; Musgrove, Tim; and Zhang, Ying (2022) "Identifying Locations of Violent Injuries in Las Vegas to Implement the Cardiff Violence Prevention Model," *SMU Data Science Review*: Vol. 6: No. 1, Article 1.
Available at: https://scholar.smu.edu/datasciencereview/vol6/iss1/1

Identifying Locations of Violent Injuries in Las Vegas to Implement the Cardiff Violence Prevention Model

Kristi Herman[1], Samuel VonPaays Soh[1], Chris Papesh[2], Tim Musgrove[3], Ying Zhang[4],

[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
kristih@smu.edu
svonpaayssoh@smu.edu

[2] University of Nevada, 4700 S. Maryland Pkwy., Suite 335,
Las Vegas, NV 189119 USA
chris.papesh@unlv.edu

[3] Callisto Media Lab, 6005 Shellmound St, Suite 175,
Emeryville, CA 94608 USA
tmusgrove@callistomedia.com

[4] Southern Nevada Health District, P.O. Box 3902,
Las Vegas, NV 89127 USA
zhangy@snhd.org

Abstract. Public violence in the United States is a major health concern. Incidents involving violent crimes are often not reported to law enforcement. The Cardiff Model is a violence prevention program developed in the UK to identify and enable data sharing of violent injury locations between Emergency Rooms (ER) and local law enforcement to help identify areas for community improvements. The model is now in use in several major cities in the US to reduce violence. Las Vegas has seen an increase in public violence in recent years. As a result, researchers from the Southern Nevada Health District (SNHD) and University of Las Vegas (UNLV) believe the Cardiff Model is a viable solution to address this public health crisis. This research explores natural language processing and machine learning models to extract violent injury location information from ER records in preparation for introducing the Cardiff Violence Prevention Model in Clark County, Nevada.

1 Introduction

Violence is a major problem in the United States. While the overall number of violent incidents reported to law enforcement decreased between 1991 to 2014, the trend over the last six years has been steadily increasing. In 2020, violent crimes increased 5% from 2019 with 398 crimes per 100,000 people, totaling 1.3 million violent crimes (Statista, 2021). Handguns, firearms, and knives were the primary weapons used to commit violence. In addition to an increase in reported incidents, the Department of Justice estimated that in 2019 less than 50% of violent injuries were reported to law enforcement (Morgan et al., 2019). For example, when a violent injury involves drugs, gangs, or domestic violence, it will often go unreported.

Public Violence Prevention is an area of research to prevent violence from occurring in public places. The Route 91 Harvest Festival Shooting in Las Vegas is a tragic example of public violence that resulted in 58 deaths and over 800 injuries (History.com, 2018). It had the highest number of victims in a mass shooting in the US. In 2020, Nevada was the 12th highest state in number of violent crimes with 460.3 crimes committed per 100,000 residents (Statista, 2021). Almost 75% of Nevada residents live in Clark County which includes the cities of Las Vegas, North Las Vegas, Henderson, Boulder City, and Mesquite. After the mass shooting in 2017, the Clark County community identified a need for collaborative public health approaches to reduce violence.

The Cardiff Model is a public violence prevention program developed by an ER physician in the UK to enhance data collection and sharing between ERs and local law enforcement for the purpose of identifying community improvements to reduce violent injuries (Kollar et al., 2020). Two cities in the UK, Cardiff and Merseyside, experienced a reduction in violent injuries by 42% and 36% respectively after the introduction of the Cardiff Model. The Cardiff Model has been piloted in the US in Atlanta, Milwaukee, Philadelphia, and other cities. As a result, Atlanta and Milwaukee have identified neighborhoods to make community improvements and Milwaukee has expanded its program to address domestic violence, suicide, and opioid overdose (The Milwaukee Blueprint for Peace, n.d.).

Researchers at SNHD and UNLV in Clark County believe the Cardiff Model can help address this public health crisis in their community because it will enable data sharing between hospitals, law enforcement agencies, and other organizations interested in violence prevention through collaborative violence prevention strategies (Kollar et al., 2018). This data-sharing framework can help identify community improvements such as increased patrolling, improved lighting, security camera installation, and the creation of more youth programs.

In previous implementations of the Cardiff Model, ER nurses were trained to collect detailed violence-related injury information including the location of injuries (i.e., business name, street address, location description). Previous pilots also required the integration of Cardiff Model Screening Tools (CMSTs) into the hospital electronic medical record (EMR) systems. This research explores ways to circumvent system change requirements and to limit the training of hospital staff in Clark County hospitals by automating the collection of that data through machine learning.

Identifying injury location and hotspots is a key component of the Cardiff Model. This information is shared with law enforcement and public health agencies monthly to help determine what actions should be taken in the community to reduce violence. In the Atlanta pilot, a series of questions was documented by nurses such as “Was someone trying to hurt you?,” “Assault Method,” “Location Name,” “Street Address,” "Nearest Intersection (if address is unknown)," and "Location Description" (Kollar et al., 2020). With limited resources to train nurses and make system improvements in Clark County, other means of identifying locations are needed to effectively make use of the model.

When a patient is admitted to the hospital in Clark County, medical notes are made in a patient's medical record. There are not specific fields to identify the location of an injury, but the researchers believe that this data may be ascertained from some of the medical notes fields where nurses may choose to enter details about the injury.

Named Entity Recognition (NER) is a method of NLP to extract proper nouns, such as the names of people and locations, from unstructured text. This method can be used to automatically analyze medical notes and identify locations of violent injuries if they are available in the text. Once locations in Clark County have been identified, the data can be aggregated into hotspots to share with local law enforcement.

This research aims to use natural language processing and machine learning models to extract violent injury location information from ER records in preparation for implementing the Cardiff Violence Prevention Model in Clark County, NV.

2 Literature Review

The literature review focuses on four principal areas: The Cardiff Model, data science methods in clinical settings, NER in clinical data, and location-based NER.

2.1 The Cardiff Model

The Cardiff Model is a solution developed in the UK for enhancing data collection and sharing between ERs and local law enforcement for the purpose of identifying community improvements to reduce public violence (Kollar et al., 2020). Two cities in the UK, Cardiff and Merseyside, experienced a reduction in violent injuries by 42% and 36% respectively after utilizing the Cardiff Model. Kollar et al. (2020) did a process evaluation of the model in Atlanta, GA, and Boyle et al. (2013) reviewed the application of the model in Cambridge, England.

The Atlanta study focused on evaluating the partnerships between hospitals and law enforcement, data collection capabilities, data analysis, and planning interventions. One outcome was the identification of an area in Atlanta to improve lighting, install security cameras, increase patrols, and support youth programs. It concluded that the model is a feasible program for the US and that it can aid in data sharing, collaboration, and surveillance.

The Cambridge study focused on data collection, data sharing, and following the results over several years to see if the Cardiff Model resulted in fewer violent injuries. It found that there were fewer injuries reported to law enforcement, but not a statistically significant reduction in violent injury admissions to the ER. Even though there was not a targeted region to make community improvements, the data collected did inform various community decisions. For example, a liquor license was denied for an area of Cambridge that had a homeless shelter and high number of alcohol-related violent injuries (Boyle et al., 2013). The Cambridge study may not be as relevant to this research as the Atlanta study because it is a smaller city. Public violence problems may not be comparable to large cities like Las Vegas.

Both studies involved training hospital staff and system upgrades to support the collection of data. The Atlanta study used nurses to collect the data while the Cambridge study utilized receptionists. Both cities collected the date/time of the injury, location of the injury, the type of assault and weapon used.

The Cardiff Model provides an opportunity to make data-driven community improvements. Previous implementations included special training of staff and system enhancements to support the collection of data. Both are potential barriers for hospitals and public health agencies that want to take advantage of the Cardiff Model but do not have the resources to change existing processes. The current research aims to remove both barriers by collecting data in an automated way from existing medical notes fields.

2.2 Data Science in Medical Records

Medical records often contain unstructured data in the form of medical notes. NLP can be utilized to extract meaningful information from medical records. One study performed a systematic review of 110 research papers using NLP on clinical data for text classification, information extraction, NER, and word sense disambiguation (Spasic & Nenadic, 2020). Of the studies reviewed, a common problem was the need for manually labeled data to train and test the models. Due to the manual effort involved with labeling, most studies had to limit the size of the datasets from millions of available records to hundreds or thousands of labeled records. Because of the sensitive nature of health care data, the models trained often involved data from only one medical facility and did not generalize well when tested on other sources. The authors concluded that more exploration can be done in the areas of data augmentation, transfer learning, and distance learning to address the annotation problem. Unsupervised models could avoid the labeling problem altogether.

In addition to being a health concern, violence has a significant economic impact costing $671 billion per year. The Centers for Disease Control and Prevention (CDC) has acknowledged that data science, especially NLP, is a growing area of research that could help reduce or prevent injuries and violence (Ballesteros et al., 2020). Ballesteros et al. reviewed approaches that the CDC will take to use data science for the purpose of reducing violence:

- Making data available in real time
- Developing forecasting methods to predict violence
- Modeling space-time clusters for detecting outbreaks
- Improving the technical infrastructure for data sharing
- Using data visualization to improve analysis
- Using NLP for classification and automation

The researchers note that the geographic prediction of violent crime and injuries is a critical area for identifying health threats in communities. However, system limitations often result in stale data and analysis. They also point out that manual efforts to label data have been a barrier in the past. Data science provides an opportunity to overcome these issues.

Lai et al. developed a spell checker to auto-detect and correct long, free-text fields in medical records for the purpose of preventing medical errors (Lai et al., 2015). During preprocessing, the Stanford NER was used to detect and exclude named entities to avoid correcting proper nouns. A robust medical dictionary was created from multiple lexical resources to detect spelling errors. A word was classified as misspelled if it was not found in the dictionary. Once a word was identified as misspelled, the Aspell system, an open-source spell checker, was used to find a list of suggested replacements. They developed a custom scoring algorithm based on Shannon’s noisy channel to select the best replacement word for the misspelling. The detection and correction system performed well in terms of precision, recall, and F1-scores across three different corpora. Including NER to ignore proper names improved results by 12%. Because of a high number of errors on mistakenly corrected abbreviations, an abbreviation dictionary was added to improve the scores.

2.3 NER for Clinical Text

Conditional Random Fields (CRFs) and other supervised learning models like Support Vector Machines (SVMs), Structural Support Vector Machines (SSVMs), and Hidden Markov Models (HMMs) have been commonly used for NER in clinical texts. The primary downside of these models is that they require human feature engineering. Recurrent Neural Networks (RNNs) are a deep learning methodology that captures long-term dependencies in sequence data. It can provide an alternative for NER on clinical texts without the manual effort of feature engineering. Several studies of NER on clinical text have shown that RNN outperforms other models in terms of accuracy and F1-scores.

The research of Liu et al. (2017) and Wu et al. (2017) focused on entity recognition using Long Short-Term Memory (LSTM), a variant of RNN. In both studies, the models produced labels for clinical entities such as disease names, types of lab tests, treatments, and medication names. Liu’s study also identified protected health information (PHI). The LSTM models performed better than other models tested in both studies.

Liu’s research compared several LSTM models with CRFs and HMMs using clinical notes from i2b2 datasets (Liu et al., 2017). One model used only a word-level input layer. Another model used both word- and character-level inputs. Using both inputs resulted in higher scores. The best LSTM model contained three layers: an input layer consisting of a representation of every word at the token and character level; an LSTM output layer with the context of each word; and an inference layer that outputs a label. The token-based input layer uses a continuous bag-of-words (CBOW) and skip-grams while the character-based input layer uses a bi-directional LSTM that captures the past and future contexts of words. Within the LSTM layer there are three propagative gates: an input gate, a forget gate, and an output gate. The main function of these gates is to control the proportion of information transferred to a memory cell. The inference layer uses a CRF to estimate a label from a sequence of context resemblances.

Wu et al. (2017) also used an i2b2 dataset containing medical notes for discharge, radiology, electrocardiogram (ECG), and echocardiograms (ECHO). A Convolutional Neural Network (CNN) and an RNN model with word embeddings were tested against a CRF model. The word embeddings were pre-trained on the MIMIC II dataset. The RNN model with bi-directional LSTM had the best overall F1-score. The researchers noted that word embeddings inherent with deep learning models are better at identifying related concepts than CBOW models because related concepts often do not contain overlapping words (e.g., "mildly dilated right atrium" and "somewhat enlarged left ventricle").

Deep learning is a promising method for NER in clinical texts. It offers several advantages including replacing CBOWs with word embeddings, eliminating manual feature engineering, and addressing long-term dependencies. While this model outperformed traditional methods in the clinical domain, it does not always perform better in other domains. It also may not perform well for identifying location-based entities, which is the focus of the current research.

NER has been widely used to analyze unstructured text in medical records where other statistical tools have failed. A combination of feature engineering and standard ML algorithms such as CRF and SVM can effectively extract meaning from text. Feature engineering requires manual effort and is time consuming and inefficient. Deep learning algorithms, especially RNN, have proven to effectively eliminate these manual tasks by learning features automatically. Inigo et al. compared two RNN methods – bi-directional LSTM and bi-directional LSTM-CRF – against other RNN models and state-of-the-art systems for Drug Name Recognition (DNR) and Clinical Concept Extraction (CCE). Bi-directional LSTM uses the concatenation of both sides of generated input sentences to produce the final representation whereas bi-directional LSTM-CRF is the resulting network of joining decoded input sequences in the Viterbi path, an algorithm for analyzing a series of hidden states. Both bi-directional neural network methods show improvement over the baseline CRF model with high F1-scores for DNR and CCE (Inigo et al., 2017). Adding manually crafted features to these neural networks did not enhance performance because the neural networks are able to learn features automatically from pre-trained data eliminating the need for manual feature engineering.

Current ML-based approaches such as CRF and SVM have remarkable success in extracting information from clinical notes in EMRs. However, these ML methods need annotated corpora requiring manual labeling from domain experts such as nurses or physicians. This effort is time-consuming and expensive. Several active learning strategies have shown success in solving this issue. Active learning models may reduce cost and improve performance compared to passive learning approaches. Active learning would use NER to select informative sequences from a pool. To measure the informativeness of sequences, N-best sequence entropy, Part of Speech (POS) tagging, Noun Phrase (NP) chunking, and sentence similarity were used. Yukun et al. (2015) compared thirteen active learning models for clinical NER, six existing artificial intelligence (AI) algorithms, and seven new AI algorithms. Two newly created uncertainty-based sampling methods were developed: N-best sequence entropy and entity entropy. Both methods outperformed baseline models in terms of area under the learning curve (ALC) scores. Dynamic N-best sequence entropy takes the sum of the probability of the N-best sequence labels that are greater than 0.9. The entity entropy sums all the entropies of B-entity words (Yukun et al., 2015). These active learning algorithms perform better than passive learning for clinical NER tasks.

2.4 Location-Based NER

NER for location extraction is not a common area of research within the clinical domain. To evaluate methods for extracting location, other types of text corpora have been studied.

Location information is crucial during emergency situations or natural disasters. Twitter is one way to track the unfolding of a crisis in real time if the locations of user tweets can be identified. It has been observed that a user's location from Twitter data is often unknown or unreliable. One study predicted exact locations from tweets using a combination of word embeddings, a CNN to extract key features, and a layer for interpreting features (Kumar et al., 2019). Their method produced a high F1-score, which was credited to the n-gram features of the CNN.

Another approach to identify locations used a deep feedforward neural network to check whether a word or phrase was present in a pre-defined blacklist and whitelist created by Subject Matter Experts (SMEs) (Magge et al., 2018). The F1-score of this model was extremely high.

In another study, the accuracy of NER to pinpoint locations in Twitter data degraded as the radius of the predicted neighborhood increased. Collaboration between urban planners and ML models reduced the error of identifying the predicted neighborhood. This collaboration, along with cosine similarity between embedded neighborhoods, improved the accuracy within a 30-mile radius (Dutt et al., 2021).

Some research has looked at a chain or thread of tweets to extract location. While this provides more context and data than a single tweet, having long text may reduce accuracy due to the presence of multiple entities and an inability to distinguish them. Current NER systems such as ANNIE, Stanford NER, NERD-ML, YODIE, and Alchemy API were tested and all of them dropped in accuracy by 30-50% on long tweets (Derczynski et al., 2015). These NER systems were tested on three different datasets to minimize bias results. The NERD-ML system performed the best based on F1-score; the Stanford NER system was the second best.

Standard NER is often not enough for geography NLP tasks due to the geographic ambiguity of toponyms (name of places). Geoparsing is a method of translating free-text toponyms into geographic coordinates. It is a basic component of Geographic Information Retrieval (GIR), Geographic Information Extraction (GIE), and Geographic Information Analysis (GIA) to extract the topography of a document. A pragmatic taxonomy is used to evaluate geoparsing. This taxonomy of toponyms uses two different taxonomy types: literal (where something is physically located) and associative (an association with a location) (Milan et al., 2020). It outperformed Google Cloud NLP and Space NLP in terms of F1-score. This is an improvement upon existing NER taggers which are unable to identify locations due to their inability to extract and classify pragmatic toponyms.

Krdžalić-Korić and Yaman (2019) researched how to identify locations in unstructured data using NER in blog posts, news stories, and social media. A Tensorflow RNN LSTM model was trained on a dataset of approximately one million addresses in the United States. To preprocess the data, the addresses were broken into house number, road, state, city, and postal code. These were turned into word embeddings using Glo-Ve, an unsupervised learning algorithm for obtaining vector representations of words. The LSTM was trained and tested on the vectors, performing well on US addresses. When testing on addresses from other countries, the model was only able to correctly identify some fields. The model successfully distinguished cities and states with the same name (New York, New York). It was also able to understand and correctly label abbreviations such as St., Blvd. (Krdžalić-Korić & Yaman, 2019).

Several methods have been used successfully to identify location information in Twitter and other sources of data. These methods can be explored in medical text in the current research to identify the location of violent injuries.

Several NER methods and a Ktrain wrapper for deep neural networks will be tested and compared to determine which provides the best results for extracting violent injury location information from the textual notes in medical records. The final output will be a list of violent injuries with the date, time, and location of the injury. Ideally, the location information will contain street names, business names (if applicable), latitude, and longitude.

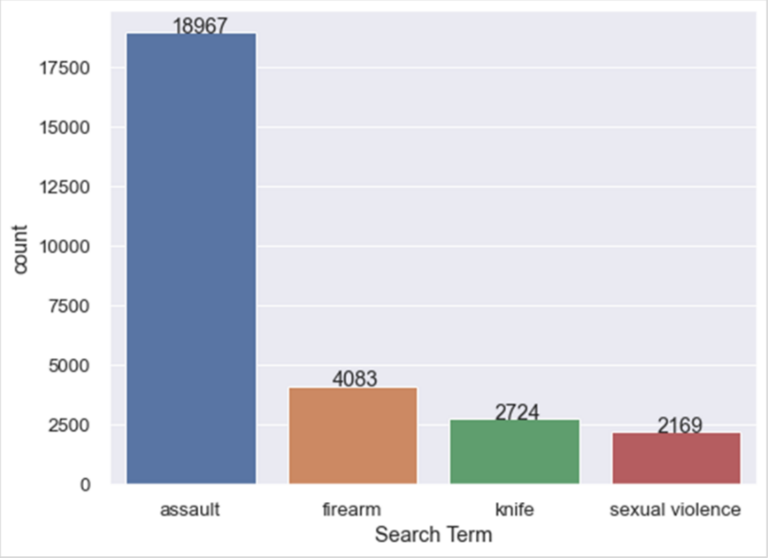
3 Methods

3.1 Data

The data source for this research was an extract of ER records for Clark County hospitals using a CDC surveillance database. SNHD and UNLV provided exported CSV files of records retrieved from searching the database for 'assault', 'firearm', 'sexual violence', and 'knife' across six years (2016 - 2021). The extracts contained date/time of hospital admission, several “Chief Complaints” unstructured text fields, hospital name, various categorical fields for type of injury, and characteristics of the patient like sex, gender, birth date, etc. Protected health information (PHI) was removed from the data prior to analysis. The full dataset contained about 28,000 records with most records coming from the ‘assault’ query. There is some duplication of records between categories because a single record could be retrieved using multiple search terms. Duplicates were left in the analysis.

Here is a breakdown of records by search term. ‘Assault’ has a significantly higher number of records than the other keywords.

Figure 1
Data By Type of Injury



Two "Chief Complaint" fields contain location information when available.

1. "ChiefComplaintOrig" contains the original text entered by the hospital staff when a patient is admitted.

   No location
   ```
   Pt states bullet grazed the left side of neck.
   ```

   Location present
   ```
   Pt was assaulted and woke up in and unknown location near
   Mandalay Bay flagged down a taxi and came here. Pt
   reports headache left jaw neck pain left rib pain. Pt
   does not remember how she was a
   ```

2. "ChiefComplaintParsed" is copy of the original text with all words capitalized, punctuation removed, and abbreviations expanded to their full form.

   No location
   ```
   PATIENT STATES BULLET GRAZED LEFT SIDE OF NECK
   ```

   Location present
   ```
   PATIENT WAS ASSAULTED WOKE UP UNKNOWN LOCATION NEAR
   MANDALAY BAY FLAGGED DOWN A TAXI CAME HERE PATIENT
   REPORTS HEADACHE LEFT JAW NECK PAIN LEFT RIB PAIN PATIENT
   DOES NOT REMEMBER HOW SHE WAS A
   ```

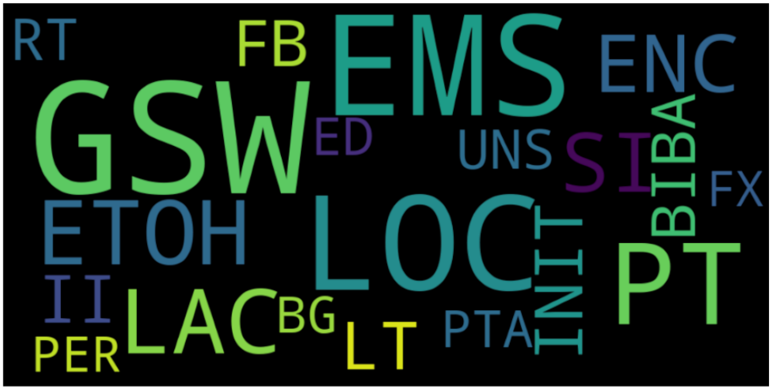
Additional data sources include a street index for Clark County to help identify street names. After streets or businesses are identified, Google Maps API is used to get the corresponding latitude and longitude.

To prepare the data for modeling, minor cleanup tasks were performed on the two "Chief Complaint" fields. A unique aspect of this data is the medical jargon and acronyms. In exploratory data analysis, the researchers observed that NER incorrectly identified many acronyms as locations. As a result, common acronyms and abbreviations were expanded prior to running the final NER models.

Below are some of the common acronyms and abbreviations in the dataset. Gunshot Wound 'GSW' was the most prevalent acronym.

Figure 2

Common Acronyms and Abbreviations in the Data



The models handle tokenization, breaking the unstructured text fields into words and sentences. Stemming and Lemmatization are methods of reducing words to their root forms. Because these methods can interfere with identifying proper nouns, they were not applied in pre-processing. Parts of Speech (POS) tagging, breaking down sentences into parts of speech, was also handled by the models as part of NER.

3.2 Model Selection

To evaluate the best model for identifying locations, a test/validation dataset was created with a random selection of 20% of the data. The rows were manually labeled with a '0' when there was no location present in the record and a '1' when a location was present. If an observation had a location, the geographic coordinates of the location were manually added to the validation data as well.

The NER and deep learning models were tested on the validation data against both "Chief Complaint" fields. If no location was identified by the model, the prediction was '0'. If a location was identified, then the prediction was '1'. Because NER can identify incorrect locations, an additional step was taken. For predictions that identified a location, the Google Maps API was called to get the coordinates of the named entity. If the predicted coordinates were within one mile of the actual coordinates, it was considered a true positive and a correct prediction. If the distance was greater than one mile, it was considered a false positive and an incorrect prediction.

3.2.1 Ktrain

Ktrain is a wrapper for deep learning libraries (e.g., TensorFlow Keras) to create a pipeline of neural networks for model creation. It is fast and simple, only requiring a few lines of code. Ktrain consists of two modules: image classification and text classification. It is well supported on learning rate finder, learning rate schedules, pre-canned models for text data, loading text and image data, data preprocessing, misclassification detection, and deploying models. An end-to-end Question-Answering system that uses Bidirectional Encoder Representations from Transformers (BERT) was used to extract location information. Four questions were asked of the model for this analysis: “What are the street names?", “What is the business name?”, “Where is the exact location?”, and “Where is the geopolitical location?”

3.2.2 spaCy NER

spaCy is an open-source NLP library with a method for NER. It is fast and works well on large volumes of text. There are four types of location entities that it can identify:

- "ORG" - Companies, agencies, and institutions
- "GPE" - Geopolitical entities, such as cities, states, and countries
- "FAC" - Buildings, airports, highways, and bridges
- "LOC" - Non-GPE locations like mountain ranges and bodies of water

A simple baseline model was created using spaCy NER filtered by the entity types above. A second spaCy model was created with the EntityRuler function to add a custom gazetteer of Clark County street names. This function applies a custom “STREET” label for matches in the gazetteer.

3.2.3 Custom NER with CRF

CRF algorithms use the context of neighboring words for classification. A CRF model with custom word features was built and trained on the Groningen Meaning Bank (GMB), a preprocessed corpus with POS and NER tags. The dataset contains over 45,000 geographical entities, 37,000 organizations, and 16,000 geopolitical entities. Word features were created such as lower- and upper-case versions of a word, POS tags, suffixes of the last two and three characters of a word, flags to indicate if a word is at the beginning or end of a sentence, and binary flags to indicate if a word is upper case, title case, or numeric. After training the model on the GMB dataset, it was used on the “Chief Complaint” fields. This model was modified from Sarkar’s custom NER (Sarkar, 2019).

3.3 Evaluation

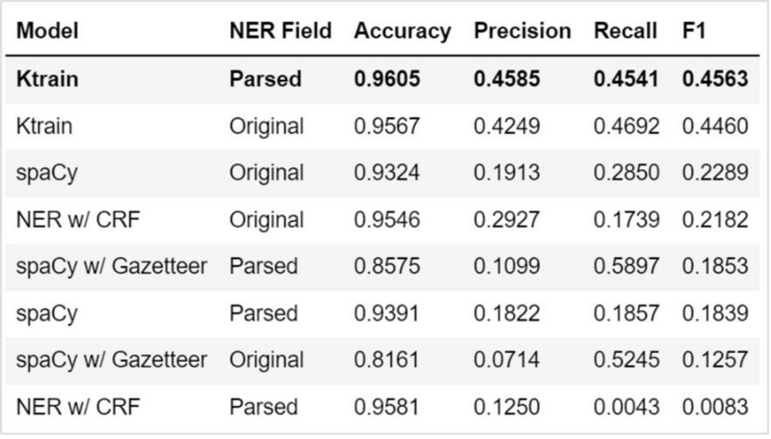
To evaluate the models, accuracy, precision, recall, and F1-scores were used. The researchers wanted to minimize both false positives and false negatives so the model with the highest F1-score was used to select the best model for this problem.

4 Results

4.1 Model Analysis

After running four models on both "Chief Complaint" fields, the Ktrain model had the highest F1-score for both fields. It performed slightly better on the "ChiefComplaintParsed" field. This may be because most acronyms are spelled out. Below are the performance metrics for each model sorted from highest F1-score to lowest.

Table 1
Model Comparison



Model	NER Field	Accuracy	Precision	Recall	F1
Ktrain	**Parsed**	**0.9605**	**0.4585**	**0.4541**	**0.4563**
Ktrain	Original	0.9567	0.4249	0.4692	0.4460
spaCy	Original	0.9324	0.1913	0.2850	0.2289
NER w/ CRF	Original	0.9546	0.2927	0.1739	0.2182
spaCy w/ Gazetteer	Parsed	0.8575	0.1099	0.5897	0.1853
spaCy	Parsed	0.9391	0.1822	0.1857	0.1839
spaCy w/ Gazetteer	Original	0.8161	0.0714	0.5245	0.1257
NER w/ CRF	Parsed	0.9581	0.1250	0.0043	0.0083

The Ktrain model produces similar precision and recall scores with an F1-score of 0.4563. This means that false positives and false negatives are similar in value and 45% of the time the correct location was identified when a location was present. The high accuracy of 0.9605 can be attributed to the volume of records in the dataset that do not have a location. The model detected that no location is present most of the time.

The basic spaCy model on the "ChiefComplaintOrig" performed the best after the Ktrain model, however the F1-score of 0.2289 was much lower than the Ktrain models. When the gazetteer was added to the spaCy model, it produced the highest recall score of 0.5897, meaning that it was best at predicting locations that are present in the data. However, the low precision score of 0.1099 indicates a high number of false positive predictions. The Custom NER with CRF model on the

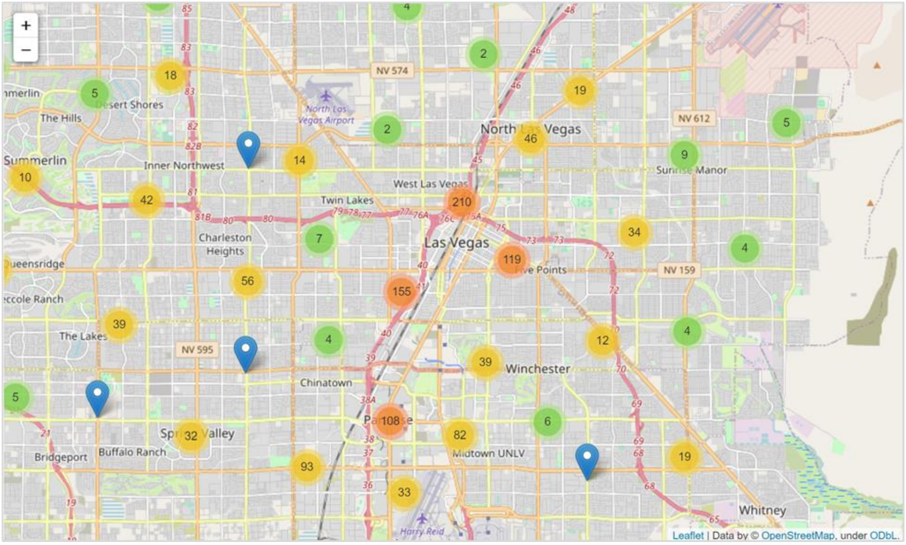
"ChiefComplaintParsed" field performed the worst based on F1-score due to its inability to recognize entities in capitalized text. The performance of the model on the "ChiefComplaintOrig" field was much better with a 0.2182 F1-score.

5 Discussion

5.1 Applications

The best performing model, Ktrain on the "ChiefComplaintParsed" field, was run on the full dataset of 28,000 records. Below are the results mapped onto a street view of Las Vegas using the Folium package in Python. The circles represent the number of violent injuries identified in that area. The colors indicate the concentration of injuries, orange circles having a high number, and green a small number. The map is interactive and zooms in to a single injury location. Figure 5 is a snapshot of the interactive map. This can be used by SNHD and UNLV to further analyze the locations identified by the model. However, it should be noted that not all locations identified are correct.

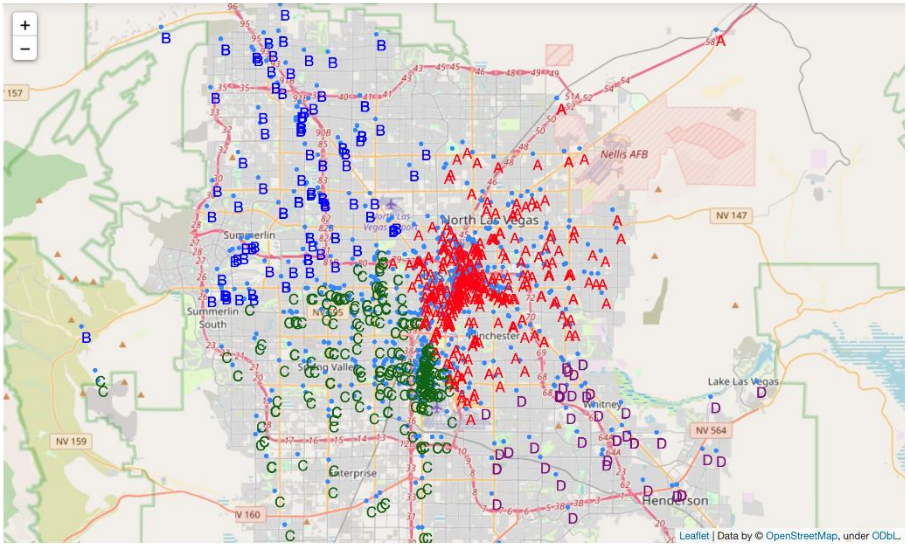
Figure 3
Map of Las Vegas with Injury Location Hotspots



Below is another view of injury locations using K-Means clustering where k=4. The purpose of this map is to get a high-level view of which zones or regions have the highest concentration of violent injuries. Outliers located outside of the Las Vegas region were removed based on three sigma calculations. The locations are shown in four zones with red (A) and green (C) having a higher concentration of injuries than blue (B) and purple (D).

Figure 4

Map of Las Vegas with Injury Location Clusters

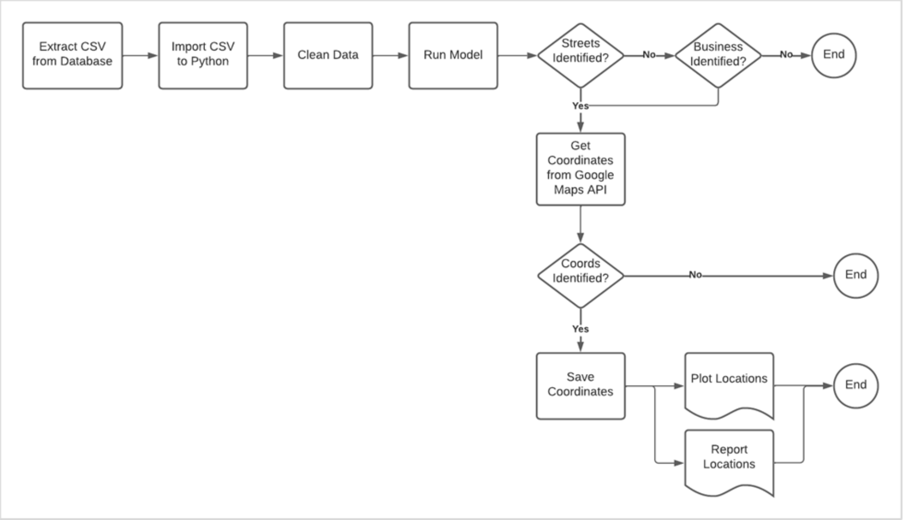


In addition to a map view, a report of the full dataset was created with all the streets, addresses, businesses, and geographic coordinates that were identified. This is intended to help SNHD and UNLV further analyze the locations identified by the model.

5.1.1 Location Identification Pipeline

The full process from extracting CSV files, running the model, identifying locations and coordinates, and creating maps and reports is outlined in Figure 5.

Figure 5
Location Identification Pipeline

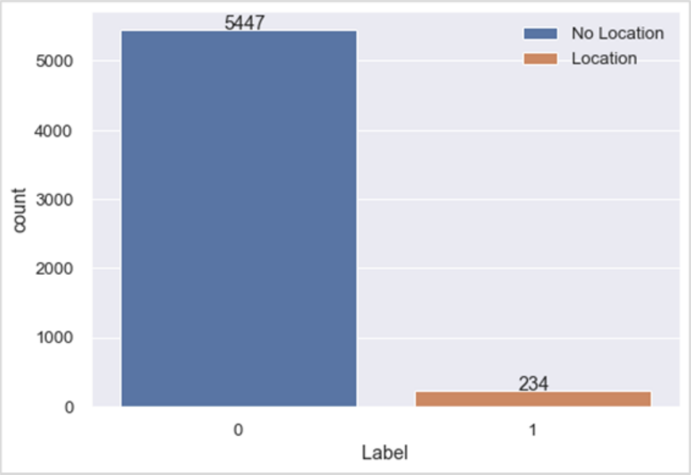


5.2 Limitations

This research aimed to accurately identify violent injury locations from ER hospital records in an automatic way using NER. One of the models had some success, but there is a significant opportunity for improvement.

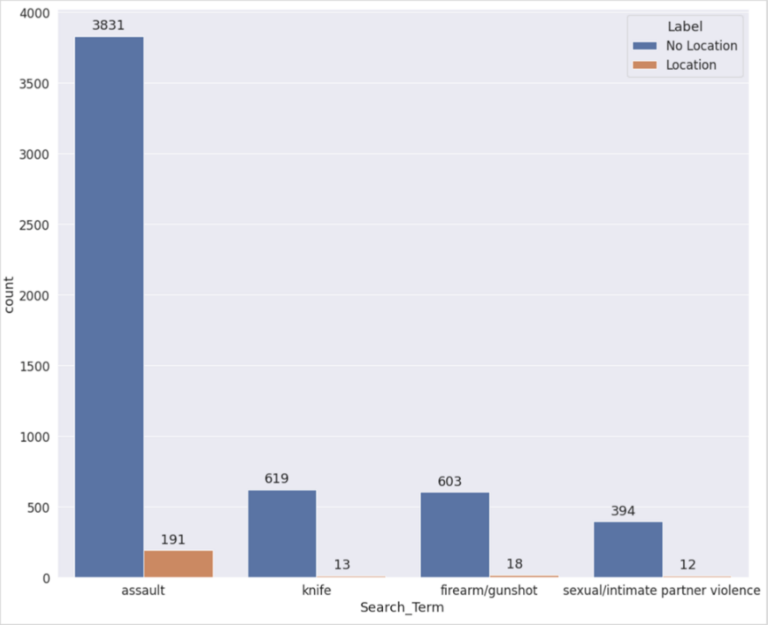
Ktrain, the best performing model, identified the correct location 45% of the time when a location was available. The accuracy of 96% seems high, but this is because most records in the dataset did not have a location. In the manually labeled validation data, only 4% of the records had a location as shown in Figure 7.

Figure 6
Uneven Distribution of Manually Labeled Data



Here is another view of the same data broken down by type of injury.

Figure 7
Number of Records By Search Term and Location Label



A model that is better at predicting locations when a location is available would yield more confidence.

One area that is likely to help improve model performance is improving the quality of the data. Below are some examples of data quality issues.

5.2.1 Inconsistent Data

The "Chief Complaint" field is unstructured, and nurses have not been trained to enter locations or follow any standards. One example is the numerous ways nurses entered Las Vegas Boulevard. A named entity was detected in some records and not in others.

Location not identified
```
pt was walking down the BLVD
```

Location identified
```
Patient picked up on Las Vegas Blvd
```

5.2.2 Incomplete Data

There seems to be a limit to the number of characters in the "Chief Complaint" fields. Some records stop mid-sentence.

```
Patient has complaint of neck and back pain.  Patient states
that he was assaulted 2 days ago and was seen by EMS but
refused to go to the hospital.  Patient picked up on Las Vegas
Blvd in front of
```

For records that contain a business name with multiple locations, if NER identifies the business (e.g., McDonald’s), Google Maps API will return location data for that business in Las Vegas. However, there is no way of knowing if it is the correct one because additional details are lacking (there are over 110 McDonald’s in Las Vegas).

```
Patient Assaulted at McDonald's
```

5.2.3 Inconsistent Usage of Capital Letters and Punctuation

Data entered in all caps or initial caps often resulted in false positives in the predictions. Here are records that the model incorrectly predicted as locations:

```
ASSAULT LOSS OF CONSCIOUSNESS
```

```
Pregnant Vaginal Bleeding;Assault Victim
```

5.2.4 Referencing Other Medical Facilities

When a “Chief Complaint” field references other hospitals or medical centers, they are often identified as a named entity even though the assault did not occur at that facility. For example, in this record the location of the assault is not stated, but “UMC” (University Medical Center) is incorrectly identified as the location:

```
pt was seen at UMC a few hours ago for an assault and he has
stitches in his face.
```

There are programmatic ways (e.g., rule-based NER) to address some of these issues but improving the quality of the data may help improve model performance. Other cities using the Cardiff Model have added location fields to their medical systems. Clark County should consider doing the same. Having specific, structured data entry fields may increase the correct identification of injury locations while reducing the number of incorrect locations identified. However, even with these fields available, nurses may still make typos and enter ambiguous locations.

5.3 Ethics

The data source for this research contains PHI including patient name, date of birth, and home address in some cases. Those fields were excluded from this analysis, but occasionally an address or name appears in the “Chief Complaint” fields. For future use of these models, Data Governance is needed to ensure that patient information is protected and secure. The public health agency and anyone working with this data must adhere to Health Insurance Portability and Accountability Act (HIPAA) rules and procedures.

5.4 Future Research

There were many interesting discoveries and surprises in this research. Medical terminology and abbreviations were an interesting challenge. For this analysis, a rule-based method was used for identifying and cleaning acronyms. One area for further exploration would be to use medical datasets or dictionaries instead of manually creating rules.

The Ktrain model was surprisingly simple to use. It can take a list of questions, such as “What are the street names?” or “What is the business name?” and attempt to find the answer in the data. This means the model is heavily dependent on asking precise questions. Generic questions will produce poor results. The downside of the model is that there are no customization options available. For example, introducing a blacklist or whitelist of terms is not possible.

Another surprise was that the spaCy model with the custom street name gazetteer did not improve results. This is because the model creates a lot of incorrect predictions by over-identifying streets. For example, Bullet Rd is a street name. The model incorrectly identified “Bullet” as the street location:

```
Bullet is lodged in left hip
```

There are currently over 35,000 streets in the gazetteer. This could be refined in future research.

The custom NER with CRF allowed the researchers to create custom word features such as lower- and upper-case versions of a word, POS tags, suffixes, etc. Additional word features could be explored in future research as well as using different datasets to train the model. A location-specific dataset may produce better results than the GMB.

Another area of exploration is predicting the probability of the location of a chain restaurant or store (e.g., McDonald’s) when no other identifiers are available. A Bayesian algorithm could predict the location using prior knowledge of the distance between each chain and the hospital the patient was admitted to.

Lastly, the researchers manually labeled data to evaluate which method was best at predicting locations. This process is time consuming and inefficient. Additional methods such as RNN with LSTM could be explored in future research.

6 Conclusion

This research was to determine whether location information could be extracted from ER records without requiring additional intervention, training, or processing by hospital staff. Based on the data available, the models did not identify reliable predictions consistently. As a result, the researchers do not recommend any of the models in their current state for making investment decisions for community improvements for the Cardiff Model. The mitigating methods proposed (Bayesian approach for chain stores, creating more word features, better handling of medical transfers) may create a more reliable tool.

In addition to exploring other methods to improve the models, the researchers recommend that Clark County hospitals integrate specific data entry fields for locations into their EMRs, like Milwaukee and Atlanta have done. If the quality of the data improves, the researchers believe model performance will improve. One way to test the effectiveness of adding location fields before making any system changes is to run the models on data from other cities. The models can then be re-evaluated and tuned further based on those results.

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

References

1. Ballesteros, M. F., Sumner, S. A., Law, R., Wolkin, A., & Jones, C. (2020). Advancing injury and violence prevention through data science. *Journal of Safety Research; J Safety Res, 73*, 189-193. 10.1016/j.jsr.2020.02.018

2. 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. 10.1016/j.ijdrr.2018.10.021

3. Magge, A., Weissenbacher, D., Sarker, A., Scotch, M., & Gonzalez-Hernandez, G. (2018). Deep neural networks and distant supervision for geographic location mention extraction. *Bioinformatics; Bioinformatics, 34*(13), i565-i573. 10.1093/bioinformatics/bty273

4. Dutt, F., & Das, S. (2021). Fine-grained Geolocation Prediction of Tweets with Human Machine Collaboration.

5. Spasic, I., & Nenadic, G. (2020). Clinical Text Data in Machine Learning: Systematic Review. *JMIR Medical Informatics; JMIR Med Inform, 8*(3), e17984. 10.2196/17984

6. Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., & Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. Information Processing & Management, 51(2), 32-49. 10.1016/j.ipm.2014.10.006

7. Kollar, L. M. 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 Program in the Southeastern United States. Injury Prevention, 26(3), 221-228. 10.1136/injuryprev-2018-043127

8. Liu, Z., Yang, M., Wang, X., Chen, Q., Tang, B., Wang, Z., & Xu, H. (2017). Entity recognition from clinical texts via recurrent neural network. BMC Medical Informatics and Decision Making; BMC Med Inform Decis Mak, 17, 67.

10.1186/s12911-017-0468-7

9. Wu, Y., Jiang, M., Xu, J., Zhi, D., & Xu, H. (2017). Clinical Named Entity Recognition Using Deep Learning Models. AMIA ...Annual Symposium Proceedings; AMIA Annu Symp Proc, 2017, 1812-1819

10. Milan, G., Pilehvar, M. T., & Nigel, C. (2020). A pragmatic guide to geoparsing evaluation. Language Resources and Evaluation, 54(3), 683-712. 10.1007/s10579-019-09475-3

11. Boyle, A. A., Snelling, K., White, L., Ariel, B., & Ashelford, L. (2013). External validation of the Cardiff model of information sharing to reduce community violence: natural experiment. Emergency Medicine Journal: EMJ; Emerg Med J, 30(12), 1020-1023. 10.1136/emermed-2012-201898

12. Lai, K. H., Topaz, M., Goss, F. R., & Zhou, L. (2015). Automated misspelling detection and correction in clinical free-text records. Journal of Biomedical Informatics; J Biomed Inform, 55, 188-195. 10.1016/j.jbi.2015.04.008

13. Chen, Y., Lasko, T. A., Mei, Q., Denny, J. C., & Xu, H. (2015). A study of active learning methods for named entity recognition in clinical text. Journal of Biomedical Informatics; J Biomed Inform, 58, 11-18. 10.1016/j.jbi.2015.09.010

14. Jauregi Unanue, I., Zare Borzeshi, E., & Piccardi, M. (2017). Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. Journal of Biomedical Informatics; J Biomed Inform, 76, 102-109. 10.1016/j.jbi.2017.11.007

15. Krdžalić-Korić, K., & Yaman, E. (2019). Address entities extraction using named entity recognition. International Journal of Computers, 13(1998-4308), 97-101.

16. Statista Research Department (2021, Sept 28) Total violent crime reported in the United Stated from 1990 to 2020 (per 100,000 of the population). Statista. https://www.statista.com/statistics/191129/reported-violent-crime-in-the-us-since-1990/

17. Statista Research Department (2021, Sept 29) Reported violent crime rate in the United States from 1990 to 2020 (per 100,000 of the population). Statista. https://www.statista.com/statistics/191219/reported-violent-crime-rate-in-the-usa-since-1990/

18. Statista Research Department (2021, Sept 29) Reported violent crime rate in the United States in 2020, by state (per 100,000 of the population). Statista. https://www.statista.com/statistics/200445/reported-violent-crime-rate-in-the-us-states/

19. The Milwaukee Blueprint for Peace (n.d.) retrieved November 7, 2021 from https://city.milwaukee.gov/414Life/Blueprint

20. Kollar, Laura M., Kurnit, Molly, Summer, Steve A., Jacoby, Sara F., Ridgeway, Greg (2018) Cardiff Model Toolkit: Community Guidance for Violence Prevention. https://www.cdc.gov/violenceprevention/pdf/cardiffmodel/cardiff-toolkit508.pdf

21. Sarkar, D. (2019). Text Analytics with Python A Practitioner's Guide to Natural Language Processing (2nd ed.). Apress. 10.1007/978-1-4842-4354-1

22. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python (1st ed.). O'Reilly Media, Inc.

23. Morgan, R. E., & Truman, J. L. (2020). Criminal Victimization, 2019. Bureau of Justice Statistics. https://permanent.fdlp.gov/lps23993/2019/2019_cv19.pdf

24. History.com (2018, October 1). Gunman opens fire on Las Vegas concert crowd, wounding hundreds and killing 58. History.com. Retrieved February 20, 2022, from https://www.history.com/this-day-in-history/2017-las-vegas-shooting