Example of Draft 1

Abstract. Public violence is a major health problem in the United States. Incidents involving violent crimes are often not reported to law enforcement. The Cardiff Model is a violence prevention program developed in the UK that combines violent injury information from Emergency Rooms (ER) and law enforcement. The model is now in use in several major cities in the US to reduce violence. Las Vegas has seen a significant increase in public violence in recent years. As a result, the Southern Nevada Health District (SNHD) and University of Las Vegas (UNLV) researchers 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 violence injury location information from ER records in preparation for implementing the Cardiff Violence Prevention Model in Clark County, Las Vegas.

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, there is an estimated 50% of violent injuries that are never reported to law enforcement. This could be caused by many reasons such a decrease in the number of law enforcement personnel since 2008 (Statista, 2021).

Public Violence Prevention is an area of research to prevent violence from occurring in public places. The Route 91 Harvest Festival Shooting four years ago in Las Vegas is a tragic example of public violence that resulted in fifty-eight (58) deaths and 546 injuries. It was the highest number of victims from 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). Clark County is a county in Nevada where almost 75% of Nevada residents live. It 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 that was 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 implementing the Cardiff Model. The Cardiff Model has been piloted in the US in Atlanta, Milwaukee, Philadelphia, and other cities. Results from the US pilot cities are not available yet, but Atlanta and Milwaukee have identified neighborhoods to make community improvements with Milwaukee expanding their implementation to include other types of violence including 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 because it will enable a data-sharing framework for hospitals, law enforcement agencies, public health agencies, community groups, and others interested in violence prevention to work together to develop collaborative violence prevention strategies. This data-sharing framework can help identify community improvements such as increased patrolling, better lighting, security cameras, and youth programs.

Previous implementations of the Cardiff Model required allocating resources to train ER nurses to collect detailed violence-related injury information including the location of injuries (e.g., business name and/or street address). With the frequent turnover of personnel in Clark County hospitals, training ER nurses is not financially feasible or sustainable. Previous implementations also required implementing special Cardiff Model Screening Tools (CMST) into the hospital electronic medical records (EMR) systems. Support for these types of changes is currently not available in the Clark County hospital systems.

Identifying injury location and hotspots is a key component of the Cardiff Model. The implementations in Atlanta, Milwaukee, and the UK included enhancements to hospital records systems to capture location information. 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. For example, in Atlanta a series of questions is filled out by nurses such as "Was someone trying to hurt you?," "Assault Method," "Location Name," "Street Address," "Nearest Intersection (if address is unknown)" and a free-text field for "Location Description." Without the resources to train nurses or make system improvements in the hospital system in Clark County, other means of identifying locations are needed to implement the model.

When a patient is admitted to the hospital in Clark County, medical notes are made in a patient's medical record. While there are not specific fields about the location, the researchers believe that location data may be ascertained from these unstructured medical notes.

Named Entity Recognition (NER) is a method of NLP to extract proper nouns such the names of people and locations from unstructured text. This method can be used to automatically analyze the medical notes to identify whether an injury is a result of violence and the location that the violence occurred. Once locations in Clark County have been identified, the data can be aggregated into hotspots to share with local law enforcement. An alternative method explored in this research is a deep-learning Recurrent Neural Network (RNN) to classify street names from the medical notes field.

Reiterate your problem statement.... This research aims to

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

Violence is a major problem in the United States. An estimated 50% of violent injuries are never reported to law enforcement. The Cardiff Model is a solution developed in the UK for enhancing data collection and sharing between ERs and local law enforcement to identify community improvements that could result in the reduction of 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 implementing the Cardiff Model. Several studies have reviewed other implementations such as Kollar et al. (2020) evaluated the implementation of the model in Atlanta, GA, and Boyle et al. (2013) explored the implementation of the model in Cambridge, England.

The Atlanta study focused on how the model was implemented, the impact on hospital staff collecting the data, and the results of sharing the data with local law enforcement. The study targeted one area in Atlanta to make community improvements. It identified businesses and public spaces to improve lighting, add security cameras, add patrols, and support youth programs. The study did not assess whether the changes identified in the communities were made or whether it had an impact of reducing violence (Kollar et al., 2020).

The Cambridge study focused on data collection, data sharing, and following the results over several years to see if data sharing resulted in fewer violent injuries. It was found that there were fewer injuries reported to law enforcement, but not a statistically significant reduction in violent injuries admissions to the ER. Even though there was not a targeted region to make community improvements or a specific action plan, the data did inform various community decisions. For example, a liquor license was denied for an area of Cambridge that had a homeless shelter and a high number of alcohol-related violent injuries (Boyle et al., 2013).

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 implementations 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 have required special training of staff and system enhancements to support the collection of data. Both are potential barriers for hospitals and 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.

and location of the injury. Ideally, the location information will contain street address, intersection, zip code, latitude, and longitude.

3 Methods

The data source for this research is an extract of ER records from Essence, a CDC surveillance database. This is provided by the Southern Nevada Health District (SNHD) and UNLV. The extract contains the date/time, location the EMS picked up the person (if applicable), a "Chief Complaints" text field, and an ICD-10 code to identify whether the hospitalization relates to violence.

Additional data sources include a street index for Clark County to help identify the names of streets. Spatial geolocation APIs may also be used to identify latitude, longitude, neighborhood, and zip code data.

To prepare the data for modeling, the "Chief Complaints' field will be pre-processing using tokenization to break the unstructured text field into words and sentences. Spell checking and correction will be applied using a spell checker such as Aspell or WordNet. Stemming and Lemmatization are methods of reducing words to their root forms. These methods can often interfere with properly identifying named entities so testing will be done with and without these methods. Parts of Speech (POS) tagging, breaking down sentences into parts of speech, will then be applied. NER will be used to create a list of proper nouns. Geoparsing will be used on the named entities to detect which ones are geographic locations. The locations will be compared to the street index to identify the name of the street that the violent injury occurred. The last step of pre-processing will be creating word embeddings to transform words into vectors. Two methods will be tested: Word2Vec and Glo-VE.

For modeling, an RNN will be trained and tested. This deep learning methodology captures long-term dependencies in sequence data and does not require feature engineering. Specifically, the LSTM will be tested. This model will be used to classify the street name.

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 address, intersection, zip code, latitude, and longitude.

4 Results

A. What you hope to find in your research? Accept or reject the hypothesis
**This Section is for statistical jargon and tables/Figures. Results are facts.

This research aims to identify specific locations that violent injuries occur in Las Vegas based on ER hospital records. The goal is to share the information with local law enforcement so that they can make community improvements to reduce violence.

5 Discussion

***Do not add New Results. This section is to apply and interpretate the results into lay terms.

*** Write questions you hope to answer in your research.

- Can location information be consistently extracted from ER records without requiring extra intervention, training, or processing by hospital staff?

- What location information is most relevant to LE in their efforts to reduce violence? There are different levels of granularity for location information (e.g., street name, full address, zip code, specific business name, etc.) and this research aims to identify the most useful information to law enforcement.

- What level of aggregation is useful to LE for identifying hot spots? For example, how many violent injuries occur within a 1-mile radius?

- What is the best way to extract accurate locations when there are misspellings and location ambiguity in the data. Examples of where mismatches may occur include: "avenue" vs. "street", "3010 Awesome Way" vs. "3001 Awesome Way", parking lot next to the gas station vs. parking lot next to 7-Eleven. What if the 7-eleven is out of business?

- Can other data sources like an index of street names in Las Vegas improve the accuracy/performance metrics?

A. Interpretations: What do the results mean?
B. Implications: Why do the results matter? How should the reader apply these findings?
C. What stood out as interesting/unique/unexpected?
D. Limitations
   a. What challenges occurred during analysis?
E. Ethics
   - Data used in this research will not contain any Personally Identifiable Information (PII). In cases where a violent injury location is identified to be a residential address, a method will be created to not disclose the address (e.g., zip code or neighborhood will be used as a proxy for address).

F. Future Research
   a. Are there areas of research where others can pick up and go deeper?

6 Conclusion

2 paragraphs max on the overall findings and summary of the research.

Acknowledgments. The heading should be treated as a 3rd level heading and should not be assigned a number.

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