Insights on US Mass Shootings

Sai Amar Nath Chintha
Dept of Computer Science. College of Arts and Sciences
Georgia State University
Atlanta, United States
schintha1@student.gsu.edu

Dhara Mandal
Dept of Computer Science. College of Arts and Sciences
Georgia State University
Atlanta, United States
dmandal1@student.gsu.edu

***Abstract*—The most heart broken event in the recent times is the Mass Shootings in Las Vegas.This course project is about deducing the insights on Mass Shootings in US from 1966-2017. Sentiment analyzing on how the audience reacted to the shootings and eventually estimating the mental health problems of the shooters and prediciting the criminals who are prone to commit the crime using their real time characteristics.The project also deals with analyzing the tweets related to the shootings. The data related to shootings in US are taken from kaggle. Twitter generates a huge amount of information which can be streamed using kafka. Streaming API gives a low latency access to twitter global stream of tweet data. The tweets are then processed and entities of the tweets are analyzed for the insights. The sentiments of the audience responses are evaluated to determine the reactions of the audience towards the shootings. Text summarization techniques will be applied to evaluate the summary of the shooters. Major insights include:**

- **How many people got killed and injured per year?**
- **Is there any correlation between shooter and his/her race, gender?**
- **What cities and states are prone to such attacks?**
- **How many shooters have some kind of mental health problem? Can we compare that shooter with general population with same condition.**
- **Predicting chances of person commiting a crime**
- **Evetually developing a model to predict the chances of a person to commit a crime given his personal features like mental illness, past crime history and mental status.**

Keywords—kafka; sentiment analysis; machine learning.

This rest of this paper is organized as follows. The introduction of our project followed by section II, Which explains the methods we used and later in section III, results were embedded and finally concluded.

I. INTRODUCTION

From 2009-2017 in the US, there have been 398 mass shootings incidents in which four or more people were shot and killed. These incidents resulted in 1,187 victims. In explanation 848 people were shot and killed, 339 people were shot and injured. In addition, 66 perpetrators killed themselves after a mass shooting and another 17 perpetrators were shot and killed by responding law enforcement. In majority of mass shootings- 54 percent of cases were related to domestic or family violence. Mass shootings significantly impacted children, 25 percent of mass shooting fatalities were children. Nearly half of the shootings, 42 percent of cases, the shooter exhibited warning signs before the shooting, indicating that they posed a danger to themselves or others. These red flags included acts, attempted acts, or threats of violence towards oneself or others. Violations of protective orders or evidence of ongoing substance abuse.

These findings reaffirm the value of gun violence prevention policies that address the circumstances underlying mass shootings. Strong domestic violence laws that keep guns away from abusers, mechanisms that allow for the temporary removal of guns from individuals who have exhibited dangerous recent behavior and background checks on all firearm sales to prevent people who are prohibited from having guns from buying them.

II. PROPOSED METHODOLOGY

A. Dataset Description

The dataset “shootings.csv” is taken from kaggle datasets. The dataset contains all the details of mass shootings in United States from 1966-2017. Data is collected in United States region over time period

1966-2017. It contains detailed information of 398 mass shootings in the United States of America that killed 1996 and injuried 2488 people.It have attributes like Title, Location, Data, Summary, Fatalities, Injured, Total, Victims, Menatl Health Issue, Race, Gender and Lat-Long information. Pandas library in python is used.

B. *Named Entity Recognition*

Named-entity recognition (NER) is a technique used to extract and locate the named entities into a set of categories like person, organization, location, quantities and time. Named Entity Recognition labels sequences of words in a text which are the names of things, such as person and company names or even protein names. It comes with well-engineered feature extraction for NER and many options for defining feature extractors. NER can be used to analyse the tweets in our model. We worked on 3 class classifier(PERSON, ORGANIZATION, LOCATION).

In figure1 a sample NER demo model is showed in which you can give your text as such. In our model tweet is given to our model as input. In figure2 we see the depression related words gets identified. In figure3, I shown the three class classifier demo model developed by Harward University. It identifies the Organization, Location and Person names

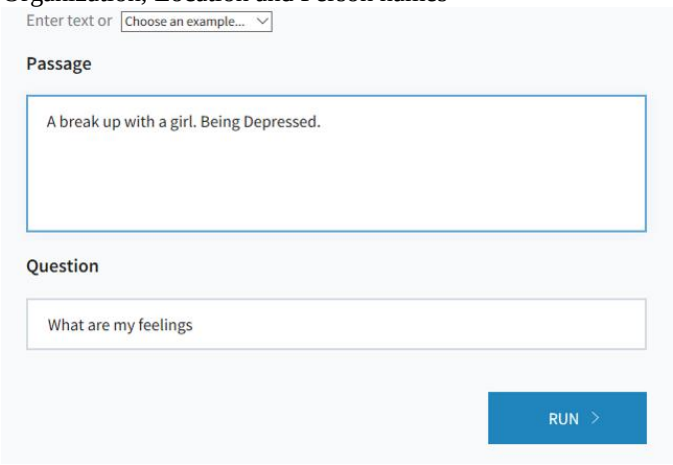


Figure 1: Shows the NER models demo

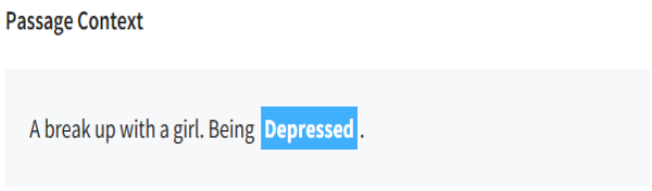


Figure 2: Sample output

Figure 3: Three class classifier

C. *Sentiment Analysis*

Sentiment Analysis is the process of computationally determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker.The real time implementations are explained below

Business: In marketing field,Companies use it to develop their strategies to understand customers feelings towards product or brand, how people respond to their campaigns or product launches and why consumers don't buy some products.

Politics: In political field, it is used to keep track of political view, to detect consistency and inconsistency between statements and actions at the government level. It can be used to predict election results as well.

Public Actions: Sentiment analysis is also used to monitor and analyse social phenomena for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.Opinion mining is a synonym for sentiment analysis. Classifying the polarity of the given document, text or feature based on the opinion in the documemt, text or feature as positive, negative or maybe neutral.

In our project we used sentiment analysis to find the people's opinion on mass shooting.

D. *Chucking:*

Chucking refers to an approach for making more efficient use of short-term memory by grouping information. Chucking breaks up long strings of information into units or chunks. The resulting chunks are easier to commit in memory than a longer uninterrupted string of information. chucking is the process of selecting a set of tokens to form a meaningful sense. Noun-Pharse chucking is used to select noun phrases in the corpus. In this project, We used NP- Chuncking which contains adjectives. This chuncks are analysed for sentiments.

D. *Graph Analysis*

Several plots have been generated using the data to list out some of the findings of the analysis like:

- Number of attacks per year
- Number of attacks by state
- Number of attacks by race
- Number of attacks by gender
- Number of attacks by the mental health of the shooter

E. *Random forest regressor*

Different kinds of models have different advantages. The random forest model is very good at handling tabular data with numerical features, or categorical features with fewer than hundreds of categories. Unlike linear models, random forests are able to capture non-linear interaction between the features and the target.

Decision tree is the basic building block of a random forest. It is a greedy technique and one of the most compelling machine learning model. There will be multiple trees in a random forest which consists of random subset of features. This has access to a random set of training data. This will lead to a overall powerful predictions. The random forest will take an average of all the individual tree estimates during the prediction phase. The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as:

$g(x)=f0(x)+f1(x)+f2(x)+...$

where the final model g is the sum of simple base models fi. Here, each base classifier is a simple decision tree. This technique of using multiple models for a better predictive performance is called as ensembling. In random forests, all the base models are constructed independently using a different subsample of the data.

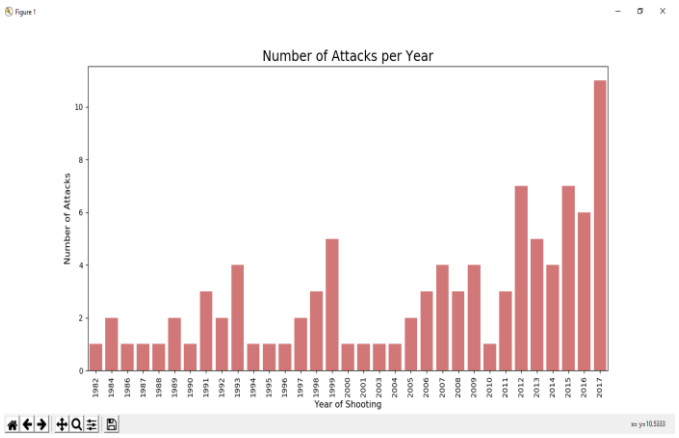
Random forest Regressor is a supervised algorithm as both the features and the targets that we want to predict is available.

During training, we give the random forest both the features and targets and it must learn how to map the data to a prediction. Moreover, this is a regression task because the target value is continuous.

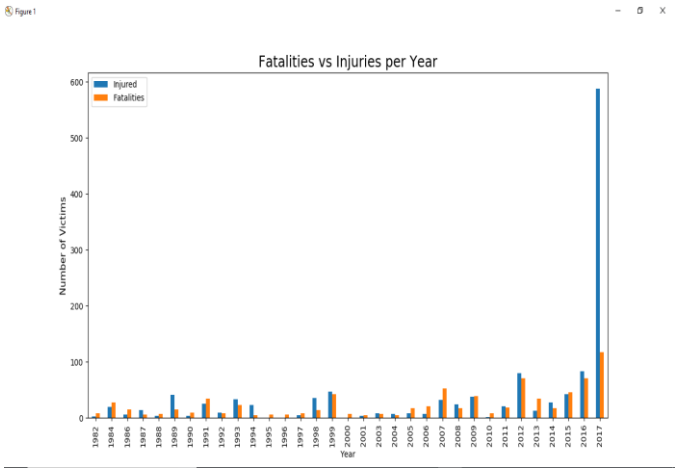
III. RESULTS

Using the dataset, We generated the graphs to analyze the key causes of the Mass Shootings. The basic idea is, Analysis the key causes will help to train a model by giving them as features.

The following are the graphs generated:

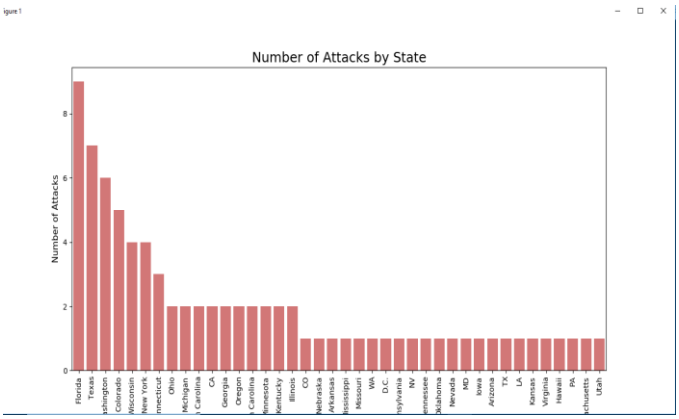


The above graph is generated as attacks vs year of shooting.We see that as the years passes the incidents are drastically increasing. So before it becomes a serious problem we need to avoid these.

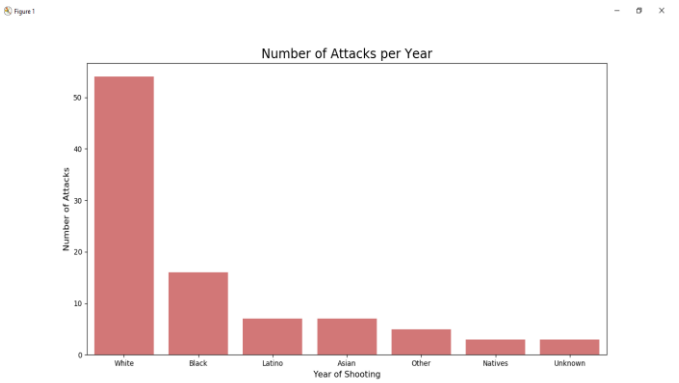


Above graph represents the Fatalities vs Injuries per Year.By this we can conclude that, With the advancement of weapons, every year the deaths are surging. By this we can

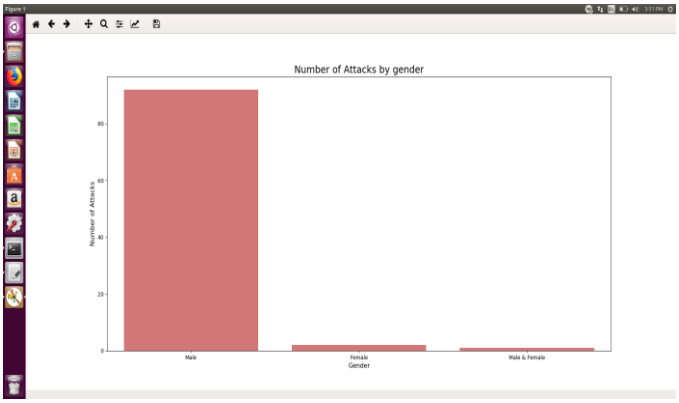
Protest sellings guns to the people who had exhibited dangerous signs earlier.



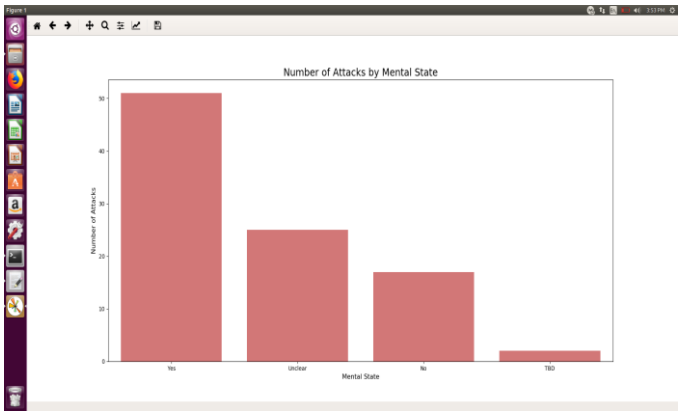
This explains the number of attacks by state. If we see the predictions generated in the above graph, Flourida is in top position and interesting fact is recently there is an incident happened. This also turns out to be interesting fact and if we can increase the security in highly crime rated cities we can avoid the crimes.



The graph illustrates the number of attacks by race. The whites followed by Blacks and latino were turned out to be top rated people to commit a crime.

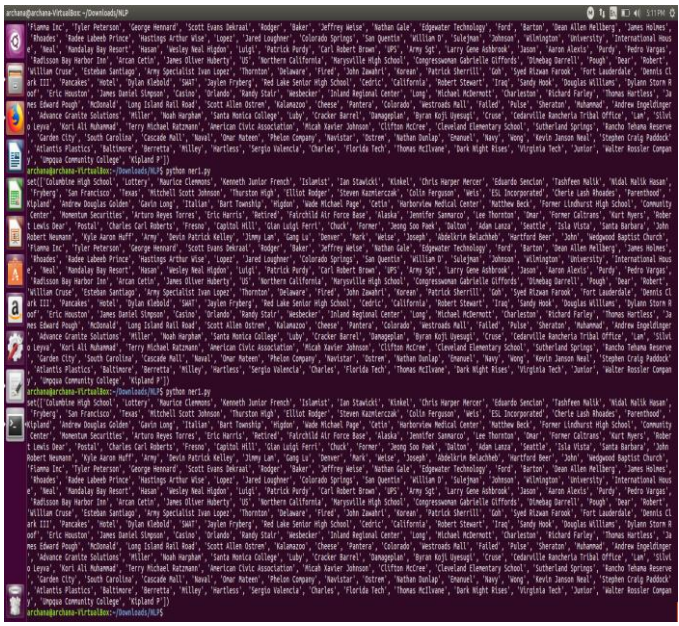


The above figure demonstrates the number of attacks vs gender. Males are turned out to commit more Mass Shootings.



This is very interesting and high motivational fact for us to further proceed in this project because above 50% of Mass shooters are suffering with some kind of mental illness. Using this as a avoidance method. if we can say a person is suffering with some kind of mental illness and also have previous crime history there is high probability for him to commit a mass shooting.

Further, we processed tweets and applied sentimental analysis on them to predict the society's opinion on the mass shootings. The results are, Most of them have negative thoughts about the incidents.



The next step was creating a dataset with top five selected features. This dataset was used to predict a model.

First step : Data pre-processing

We read the excel file which consists of categorical values. The header value is set for all the columns. After the headers are set, the categorical values are converted to numerical ones.

Secound step: Splitting the dataset

Once the dataframe is prepared, the data is split into training data and testing data.

Third step: Training the data and predicting

Random forest Regressor algorithm is now applied to train the data and then predict with the test labels.

```
In [210]: print(model.score(train_features, train_labels))
0.791964287553162
```

```
# Import the libraries
import numpy as np
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor

df  = pd.read_excel('book1.xlsx',sep=",",header=None)
new_header = df.iloc[0] #grab the first row for the header
df = df[1:] #take the data less the header row
df.columns = new_header #set the header row as the df header

data = df.iloc[:,:5]
label = df.iloc[:,5:]
number = LabelEncoder()
data['prior sign'] = number.fit_transform(data['prior sign'].astype('str'))
data['mental illness'] = number.fit_transform(data['mental illness'].astype('str'))
data['Race'] = number.fit_transform(data['Race'].astype('str'))
data['weapon possessed'] = number.fit_transform(data['weapon possessed'].astype('str'))

train_features, test_features, train_labels, test_labels = train_test_split(data, label,
                                                                            test_size = 0.3,
                                                                            random_state = 42)

train_labels=train_labels.astype('int')
test_labels=test_labels.astype('int')

#applying Random forest regressor
model = RandomForestRegressor(n_jobs=-1)
estimators = np.arange(10, 200, 10)
scores = []
for n in estimators:
    model.set_params(n_estimators=n)
    model.fit(train_features, np.ravel(train_labels))
    scores.append(model.score(test_features, test_labels))

print(model.score(train_features, train_labels))
```

IV. CONCLUSION

This project touches the surface of a huge topic. The data itself did have some telling features. Of note to the authors was the gun type used, prevalence of previously diagnosed mental illness among the shooters and seasonality to the occurrence of mass shootings events. Our hope is that this type of information will lead to practical and balances reform, allowing all parties to be satisfied with the outcome. Few topics are as emotionally charged as gun control. The horror of mass shootings seems to ring in our ears as an emotionally charged as gun control. The horror of mass shootings seems to ring in our ears as an everyday occurrence. America has 5% of the world's population, yet 31% of its mass shooters. Sadely, resolution of mass shootings and alignment on gun control is at a partisan deadlock. Discussions are emotionally charged and avoid facts. To encourage people to find common ground between civil liberty and the social contract we developed an application to explore relevant data. We hoped to create meaningful dialogue by including links to national gun control forums as well. The following are the analysis of this project:

- Number of victims are gradually increasing and maximum in the year 2017.
- Number of injured are more than number of deaths per each year
- Whites are prone to be most among the number of attacks
- Flourida tops among the most number of attacks
- Male attackers are dominantly more in number
- Majority of the attackers are suffering from some mental health issues

The random forest model is very good at handling tabular data with categorical features which consists of fewer than hundreds of categories. Random forests algorithm is able to capture non-linear interaction between the features and the target.

REFERENCES

[1] https://www.kaggle.com/carlosparadis/last-50-years-us-mass-shootings

[2] http://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data/#

[3] http://www.nltk.org/book/ch07.html

[4] https://nycdatascience.com/blog/student-works/mass-shootings-in-america/

[5] http://nlp.stanford.edu/software/CRF-NER.shtml