# School Shootings in the United States

Jonathan Brophy

*Abstract*— **The seemingly high increase of school shootings in the United States have garnered more attention by the media and policy makers in recent years. This report looks at the history of school shootings in the U.S. to identify overarching trends and possible insights into why these shootings are increasing, if at all. The data is collected from two primary sources of school reported shootings, *Wikipedia*[1] and *Everytown*[2]. This paper presents a visual representation of shootings geographically over the decades from 1774 - 2016. The shootings are also broken down by season, to identify any patterns related to the time of year. Following these initial statistics, this paper presents a more in-depth look into the worst school shootings in the U.S., where the worst shootings are defined as any shooting that has 10 or more deaths and injuries combined. Information regarding the shooters' background, types of guns used, mental state, as well as a similarity matrix between pairs of the worst shootings are reported. This paper employs a variety of techniques from data visualization to text mining and natural language processing to bring about a summary of the nation's school shootings.**

## I. INTRODUCTION

Anyone who has paid any attention to the news in recent years might be inclined to think that there are an epidemic of school shootings plaguing the United States. Combine this belief with the seemingly never-ending battle over gun control, and this can lead to confusion about how bad the situation really is. A clear overview and summary of the nation's past shootings compared to recent trends is needed for policy makers to be able to make informed and rational decisions regarding this topic.

The data collected for this report spans a time period from 1774 - 2016. The very first school shooting was in 1774, but most of the data reported actually occurs between 1840-2016. Thus, this paper covers approximately 175 years of school shootings where a number of different trends are analyzed. One example aspect to look at involves the location of these shootings. Looking at where these shootings take place can provide a clear understanding about which regions, or perhaps more specifically, which states or cities, are most affected by school shootings.

There are a number of different factors that contribute to a person or persons commiting a school shooting, and many questions can arise relating to this topic. This report attempts to look at as many contributing factors as possible to uncover meaningful and surprising patterns, or to possibly confirm societies' previously held beliefs about these events.

The following sections present information about the data collected for this project, basic statistics about the gathered data, proposed methods for extracting useful knowledge, and a summary of results painting a clear picture of the nation's past regarding school shootings as well as its projected direction for the future pertaining to these terrible events.

## II. BACKGROUND

A brief overview of the two primary data sources are outlined below. For both of these websites, the data was scraped and cleaned using various python scripts, and put into a mysql database[3] to perform queries on.

### A. Data

The data from both of these sites provide school shooting data using similar structures, but a lot of cleaning and pre-processing was still needed to be able to store them together and effectively mine their contents.

*1) Wikipedia:* The data from this site contains shootings dating back to the $18^{th}$ century, but the majority of the shootings reported there are from the $19^{th}$ century to the $21^{st}$ century.

The shootings are broken up by decade where each shooting contains the following: date, location, number of deaths, number of injuries, and a short description of the incident.

In order to get the data into a usable form, the raw HTML of the web page was downloaded and parsed using a standard python HTML parser [4]. Next a script pre-processed the data to be automatically inserted into

[1] https://en.wikipedia.org/wiki/List_of_school_shootings_in_the_United_States

[2] http://everytownresearch.org/school-shootings/

[3] https://www.mysql.com/

[4] https://docs.python.org/2/library/htmlparser.html

a shootings table in the mysql database. Any shootings added to the web page after the initial data dump was added manually to the database.

At this point, the database consists of one table of shootings that contains the date, location, deaths, injuries, and description of each shooting, but two more fields were added: U.S. population and the school type. The U.S. Population is recorded every 10 years by the U.S. census. This field can be helpful in determining the number of shootings as a function of the number of people in the country at that time. The school type represents what kind of school was affected by the shooting: elementary, middle, high school, or university. The addition of these two attributes can help discriminate the shootings even further.

*2) Everytown:* The second source for school reported shootings comes from the web page, *everytown*, where school shootings have been reported on that site since 2013. The time frame for reported shootings on this website are heavily geared towards more recent shootings, and must be taken into account when looking at all shootings dating back to the 1800's.

Each shooting on this site contains: date, location, school name, school type, and category of the shooting. The category refers to: an attack resulting in injury or death, unintentional firing resulting in injury or death, gun firing but no one is injured, or an attempted or completed suicide with no intent to injure anyone else.

The same technique used on the *wikipedia* page was used for this site to scrape, clean, preprocess, and insert into the database. The category from *everytown* was used for the description in the shootings table. Any additional shootings added to this site was then manually added to the database.

### B. Basic Statistics

Now that the data from both sites have been consolidated, it is useful to know some basic information about the data. First of all, the *wikipedia* site has approximately 380 recorded shootings, and the the *everytown* site has 170 recorded shootings. Since there is some overlap between the two sources, 493 school shootings are used for this report.

Table I shows how many shootings there have been for each school type dating back to 1774. The table suggests that high school shootings are the most popular, with universities following behind. If the elementary and middle school shootings were combined, then that number would be similar to the number of university school shootings. Thus overall, high school shootings are highly targeted, but surprisingly, there is quite a wide spread over all the different school types.

TABLE I
BREAKDOWN OF SHOOTINGS BY SCHOOL TYPE

| | Elementary | Middle | High School | Univ |
|---|---|---|---|---|
| # of Shootings | 57 | 53 | 259 | 124 |

According to Table II, there are more total injuries in all U.S. school shootings than deaths. Since there are 493 total shootings, this means there are an average of 1.067 deaths per shooting and an average of 1.434 injuries per shooting.

TABLE II
TOTAL DEATHS AND INJURIES FROM SHOOTINGS

| Deaths | Injuries |
|---|---|
| 526 | 707 |

Now that we have a better understanding of what the data looks like, the facts presented in this section are analyzed to a greater extent in the experiments section.

## III. METHODOLOGY

There are two main approaches taken to mine the data for this domain. The first is data visualization. The results from any data mining can often be tricky to analyze and make sense of, but data visualization helps combat this problem to convey as much knowledge as possible without overwhelming or confusing the observer [1]. This is the main approach for this report, where a variety of charts, maps, and graphs will be the driving force to explain the historical phenomena of school shootings in the United States. Data visualization is especially helpful in this domain where there is a time component to the data. It is useful to look at trends in shootings over time from different granularities: shootings by decade, seasonal shootings, and more recent shootings. By looking at the data in this way, we can view the shootings from various perspectives, providing more opportunities to uncover hidden patterns that may otherwise be hard to find.

The second approach involves some text mining and natural language processing to reveal more information about individual shootings. By mining the descriptions and reports cited by the websites in Section II, there is the possibility of uncovering similarities between shootings and common identifiers among them.

### A. Data Visualization

A few different techniques are used to display the data in a meaningful and educational way. The basic approach for this paper in using data visualization for mining is to query the data base and then use some combination of maps, charts, and graphs to concisely present what was found.

*1) Maps w/o Time:* There is the potential to learn a lot based on the date and location of school shootings in the United States. By first taking time out of the equation and compiling an aggregated list of all the shootings, we can mark those shootings onto a map of the nation. Just by visualizing this map, one can start to infer what regions, states, or cities are most affected by school shootings.

The first approach is to place a marker at every location where a shooting has taken place. The nice thing about this method is the fine granularity to see the exact location for every shooting.

Having the exact location of every shooting is certainly useful, but when multiple shootings start to overlap each other, it becomes hard to distinguish how many shootings have actually happened in one place. Thus, the second approach takes the same map of shootings, but creates a heat map of the nation using the states as separators. From this perspective, the observer can see from a lower granularity how individual states are affected from the shootings. This is especially nice when you have a number of states close together and cannot tell exactly which shootings belong to which states. With this view, policy makers can easily distinguish which states are having the most problems and start to act accordingly.

*2) Maps w/ Time:* After looking at this aggregated view of shootings, putting the time component back into the mix adds a new level of dimensionality and offers more detailed views of the data from different angles.

The first time component to look at would be by decade, where all shootings from a specific decade can be plotted on one map. Then once all shootings have been put onto their respective maps, these maps can be animated one after the other with respect to time. This gives a nice overview of how the location of shootings evolve over time. The observer can also start to tell if the number of shootings are increasing over time just by viewing this animation.

The second way to split up the shootings using time is by season. The idea is that perhaps some shootings occur more frequently during certain times of the year over others. Using this approach, we are able to view the number of shootings during the fall, winter, spring, and summer seasons and view their locations during these time periods. This can be useful if one season is heavily affected, then programs and policies could be put into place during that season to try and curb the possibility of any shootings.

The last time component this paper takes into account are the most recent shootings. Since these shootings are more recent in our past, they are more relevant and can help explain future trends of school shootings in the United States.

*3) Charts/Graphs:* A lot of basic but potentially insightful information can be presented using charts and graphs. It is much easier to see an increasing or decreasing trend line on a graph than it is to read from a table. This paper uses both quite extensively in order to easily display shifting trends related to the number of shootings, deaths, and injuries over time. Also, the United States population at the time of each shooting is taken into account to see if the increasing number of shootings is caused by the rise in population, or if there is something more serious going on that needs urgent attention.

By utilizing these data visualization tools, the job of conveying important information is much easier and clearer.

### B. Natural Language Processing

There is a lot of potential for some aspects of natural language processing for this project since there are descriptions for every recorded shooting. Granted, some shootings have much more detailed descriptions than others, and the majority of shootings only have a bare minimum with regards to a description, indicating only the date, location, school type, and what kind of attack it was. Only a small subset of the shootings contain any information about the shooters, their backgrounds, the weapons used in the shootings, and other detailed information not easily found in the other shootings. Because of these facts, the following NLP methods are performed on a subset of the total number of shootings.

*1) TF-IDF:* The concept of Term Frequency Inverse Domain Frequency (TF-IDF) can be used to find the terms in a document that more or less uniquely identify that document [2]. Term Frequency is defined as:

$$\text{TF} = \frac{\#\ times\ term\ appears\ in\ doc}{\#\ total\ terms\ in\ doc}$$

Inverse Domain Frequency is defined as:

$$\text{IDF} = \log \frac{total\ \#\ docs}{\#\ docs\ with\ term\ in\ it}$$

The TF-IDF is defined as TF * IDF. The basic idea is that words appearing more frequently in a document, but also appearing in many documents is punished more heavily. This gives words like 'the' and 'and' low TF-IDF scores. Words that appear a lot in a certain document but not in many others are given higher TF-IDF scores.

This method of NLP can be used on documents to find similarities between pairs of documents. To do this, all documents are analyzed to create a vocabulary of words that contain all words seen across the documents. Then a TF-IDF score will be given to each word in the vocabulary for each document. Once this is done, taking the dot product between this matrix and itself gives similarity measures comparing pairwise documents. Scores are normalized between 0.0 and 1.0, 1.0 being the most similar, and 0.0 being the least similar.

TF-IDF can also be used for feature extraction for various classification tasks [3]. This paper explores the process of extracting select terms across multiple documents in order to classify other or future documents that might be similarly related. The terms resulting from this process can also give some insight into what patterns are similar among groups of documents.

## IV. Experiments

The results in this section come from performing the methodology outlined in section III on the data compiled for this domain. First, various data visualization tasks are applied to the shootings without considering time, then the time component of the shootings are put back in. Second, a similarity matrix using TF-IDF is created only for the worst shootings. Also, feature extraction using TF-IDF is performed on the worst shootings to find common terms among their reports.

### A. Shootings by Location

Figure 1 presents all 493 recorded school shootings in the U.S. dating back to the 1800s. The top image shows how many of the shootings are clustered together around major cities, and there seems to be a pretty clear divide down the middle of the U.S. where shootings become much more dense on the eastern half of the country. This can possibly be explained due to the more sparsely populated areas of the western half of the country.

The bottom image displays the shootings at much lower granularity, the state level. States colored yellow have had 2 or less shootings in their history, light orange: 5 or less, orange: 10 or less, dark orange: 15 or less, and red states have had more than 15 shootings in their states' history. This heat map of the U.S. makes it clear which states, such as California, Texas, and various states along the eastern seaboard, are most affected by school shootings.

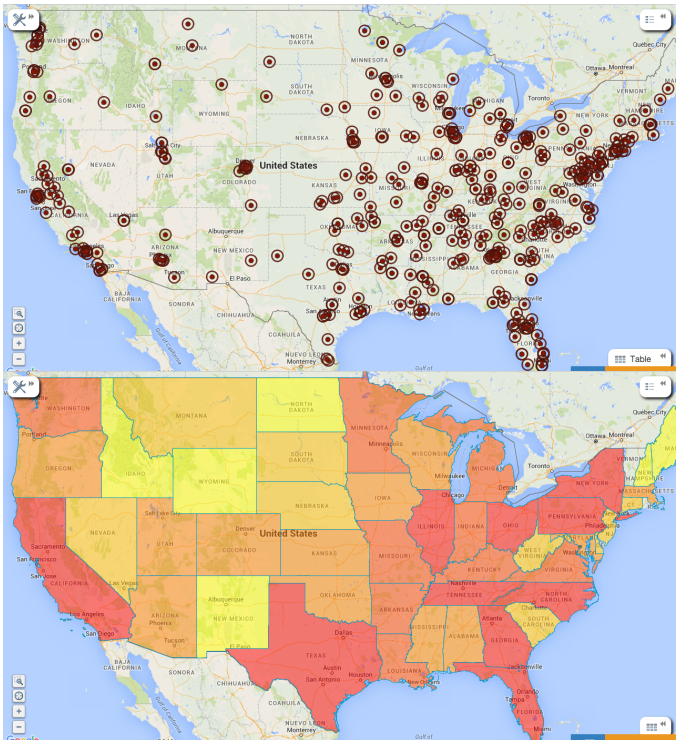


Fig. 1. All School Shootings From 1774-2016. Individually marked shootings (Top), Heat map of shootings (Bottom).

Alaska and Hawaii are not featured in Figure 1, but both states have had 1 shooting each in their respective histories.

### B. Shootings by Season

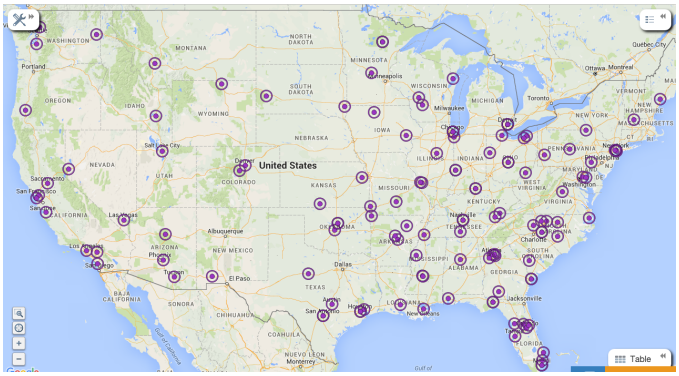


Fig. 2. Total Number of Fall Shootings: 130

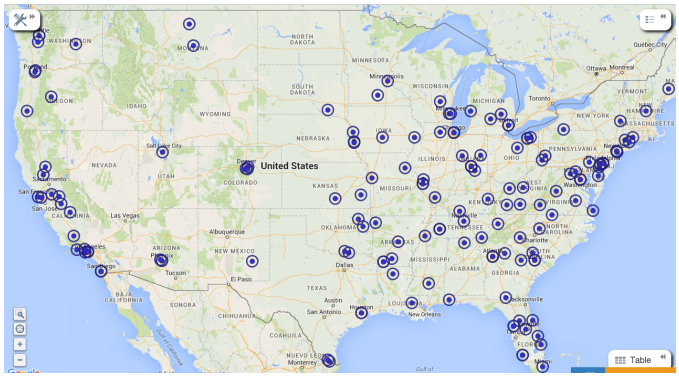


Fig. 3. Total Number of Winter Shootings: 164

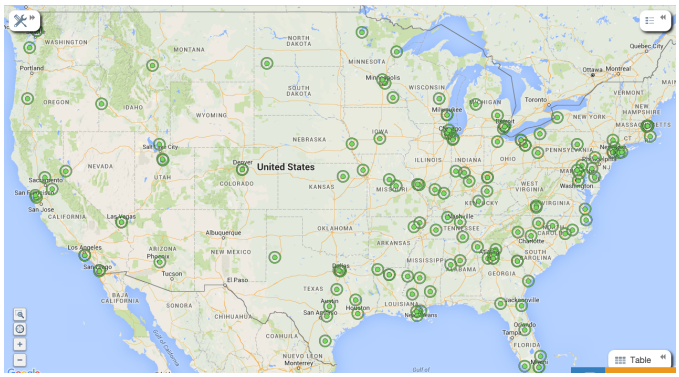


Fig. 4. Total Number of Spring Shootings: 143

Figures 2-5 show the breakdown for the number of shootings for each season and where those shootings were located. The shootings do not seem to be concentrated in any one season. Furthermore, winter, fall, and spring all have a relatively similar number of shootings, and the locations of those shootings have not changed very much as well. The only fact that is certain from analyzing these figures is that there seem to be less shootings during the summer. This makes intuitive sense since the majority of students are not enrolled in any classes during the summer.

### C. *Shootings by Decade*

Figure 6 suggests a slowly increasing trend in school shootings over the decades since 1774. There is a sharp increase in school shootings in the decade labeled 2010. This is caused mainly because of the shootings reported in the *everytown* source. There are 170 reported shootings in that report, meaning the 2010 decade alone has at least 170 shootings in it, and this is the cause of the massive spike in shootings in Figure 6.

Here are two possible explanations for the rise in reported school shootings over the years: The first being a rise in the country's population, and the second being the fact that over time, the ability to report and maintain a record of school shootings has gotten much easier with new and emerging technology.


Fig. 5. Total Number of Summer Shootings: 56

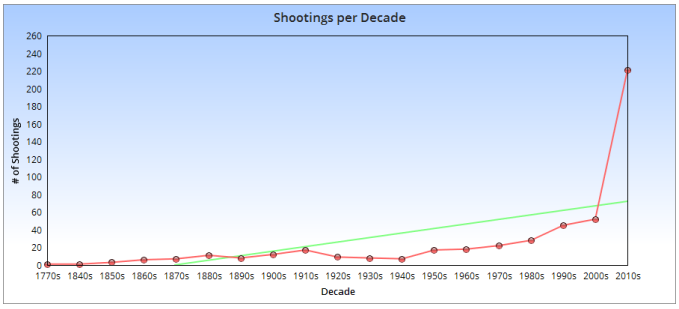


Fig. 6. Number Shootings per Decade

The trend line in Figure 6 looks disconcerting, but this graph does not take the population at the time of each decade into account.

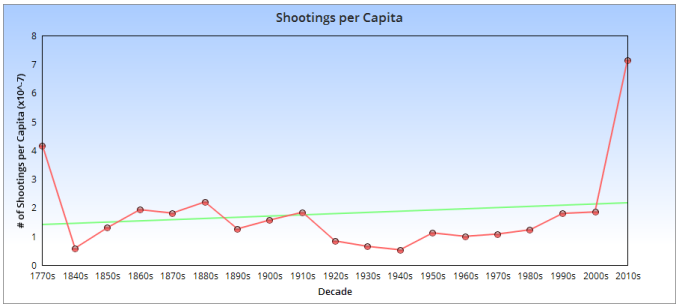


Fig. 7. Number of Shootings per Capita over Time

The trend line in Figure 7 looks much less threatening and suggests that the number of shootings per decade has been going up proportionately with the population of the United States. The only exception is the current decade, in which there is a sharp increase in shootings even after taking the population into account, in which this might be cause for concern for the future.

### D. Deaths and Injuries

We have already seen the total and average number of deaths and injuries from Section II, but we did not take time into consideration.

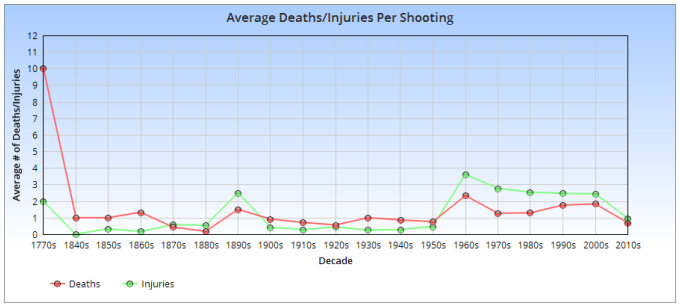


Fig. 8. Average Number of Deaths and Injuries per Decade

Figure 8 captures the trends of average deaths and injuries over the years, where there appears to be a big spike in deaths in the 1770s. This is because there was only one shooting in the 1770s, and it happened to be a particularly bad one, making the average values for that decade based off that single shooting.

After the 1770s, the average death and injury rate stayed relatively low and constant from the 1840s to the 1960s. The 1960s was a decade where gun violence started to rise and 2 out of 3 homicides were committed by guns, which brought much attention to the White House where they attempted to combat the rising gun violence [4]. This is the point where the graph suggests a sustained increase in both average deaths and injuries.

The final thing to note about Figure 8 is the significant decrease in average deaths and injuries during the current decade, the 2010s. This seems odd, since Figures 6 and 7 both indicate a major increase in school shootings during that time. It is not easy to see from the graph, but there is an average of 0.69 deaths and 0.96 injuries per shooting in this decade. This means that even though there have been many shootings in this decade alone (222 so far), each shooting has not been very dangerous in regards to not many people getting killed or injured in each one. One reason this could be the case is that *everytown* defines a school shooting as any time a gun goes off on school campus. There were a significant number of shootings on their site where the event involved a gun going off but no one was killed or injured. Another significant portion of the shootings reported there involved a suicide of an individual who decided to attempt taking their life on a school campus but never had any intention of hurting anyone else.

So the fact that there are a growing number of school shootings in this decade is disconcerting, but Figure 8 suggests that the average school shooting is not very dangerous.

### E. Worst School Shootings

When society thinks about a school shooting, most people think of the worst ones, such as Columbine, Sandy Hook, Virginia Tech, as well as several others. These shootings tend to receive the most attention from the media.

This section of the experiments is dedicated to analyzing the worst shootings in the nation's history for two reasons: First, since these shootings are the most notorious, there is much more detailed information about each event itself, as well as information about the shooter(s) and their backgrounds. Second, I believe that these shootings have had the greatest impact on our society, and studying them further may be more beneficial in preventing them in the future.

This report defines a shooting as being one of the worst shootings if the event contains more than a combined number of 10 or more deaths and injuries. After querying the database for the shootings fitting this description, a list of 23 shootings came up. Out of those 23 shootings, 21 of them have occurred after 1960. The two that occurred before them were in 1774 and 1891, respectively. The shooting in 1774 was actually an Indian attack on colonial settlers in Greencastle, Pennyslvania [5]. This shooting is the very first recorded school shooting in the nation's history. The second oldest worst shooting was in 1891 in Liberty, Mississippi, where an unknown assailant fired a double barreled shotgun into a crowded schoolhouse auditorium, where a concert was being held that night.

TABLE III
SEX OF SHOOTERS INVOLVED IN WORST SHOOTINGS

| Male | Female |
|---|---|
| 19 | 1 |

Table III shows the vast discrepancy of male shooters over female shooters involved in these types of shootings. The reason only 20 shooters are mentioned in Table III even though there are 23 worst shootings is because 3 of those shootings were caused from police and crowd control, where it was not one or two shooters linked to the event.

Table IV suggests that the worst shootings are targeted slightly more towards high schools and universities, but elementary and middle schools are certainly

TABLE IV
BREAKDOWN OF WORST SHOOTINGS BY SCHOOL TYPE

| | Elementary | Middle | High School | Univ |
|---|---|---|---|---|
| # of Shootings | 4 | 2 | 9 | 8 |

represented as well, especially famous shootings such as Sandy Hook, which took place in an elementary school in Connecticut.

The sources cited in *wikipedia* were followed to create a document report for each of the worst shootings. Those reports include detailed information about each shooting, who the shooters were and their backgrounds, and any extra information relevant to that particular shooting. These text documents are then processed using the TF-IDF method proposed in Section III to create a similarity matrix comparing pair-wise documents.

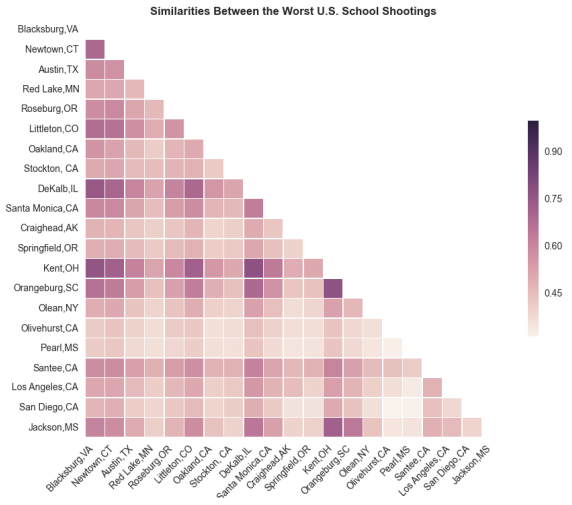


Fig. 9. Pairwise-Similarity Between Worst Shootings

Using the documents for each shooting, the similarity matrix in Figure 9 shows how similar the worst shootings are to each other. The darker a cell is, the more similar those two shootings are to one another. The similarity measure is a normalized value between 0.0 and 1.0.

Following up the y-axis relative to Kent, OH, we can see that Jackson, MS and Orangeburg, SC are both similar to the Kent, OH shooting. This is most likely because all three of those shootings were protests or riots in one form or another in which police authorities stepped in and were the cause of those school shootings.

Looking up the y-axis relative to Blacksburg, VA, we can see it that it is similar to the Kent, OH, Dekalb, IL, Littleton, CO, and Newtown, CT shootings. These are described as some of the deadliest school shootings even compared to the other worst shootings, so it makes sense that some of these would be similar to each other.

Table V shows the terms with the highest sum TF-IDF scores after adding up each terms' TF-IDF score for each document. All english stop words have been taken out because even though their TF-IDF score might be low for each document, summing them up can still overwhelm other more important terms.

TABLE V
TOP TERMS ACROSS WORST SHOOTING REPORTS

| Term | Total TF-IDF Score |
|---|---|
| school | 2.34 |
| shooting | 1.91 |
| students | 1.39 |
| police | 1.24 |
| state | 1.14 |
| shot | 1.02 |
| high | 0.97 |
| student | 0.94 |
| university | 0.93 |
| said | 0.93 |
| killed | 0.90 |
| campus | 0.87 |
| hall | 0.74 |
| year | 0.73 |
| mother | 0.71 |
| time | 0.70 |
| old | 0.69 |
| later | 0.68 |
| shootings | 0.64 |
| classroom | 0.62 |
| people | 0.60 |
| tech | 0.59 |
| father | 0.58 |
| olean | 0.56 |
| trial | 0.56 |

The terms in Table V are the top 25 words that collectively summarize the reports concerning the worst shootings in the United States. These words can possibly be used as features in a classification model to classify unlabeled or future reports.

TABLE VI
WEAPONS USED IN WORST SHOOTINGS

| Rifles | Pistols | Shotguns | Other |
|---|---|---|---|
| 23 | 22 | 14 | 4 |

Table VI presents the breakdown of weapons used over all the worst shootings. Rifles and pistols are the

most common choice, with shotguns not far behind. The guns counted in Table VI are a mix of automatic, semi-automatic, and non-automatic weapons. The weapons labeled 'other' contain knives, batons, weapons of arson, and other items.

One final thing to note about these shooters is that the average age is 20.85. Also, almost every one of them either had some type of mental illness, was the victim of a form of childhood abuse or bullying, or had some sort of trauma earlier in their lives.

TABLE VII
COMMON SYMPTOMS AMONG SHOOTERS' BACKGROUNDS

| Mental Illness | Abuse | Anger/Revenge | Suicidal |
|---|---|---|---|
| 9 | 8 | 3 | 3 |

Table VII shows the number of shooters that exhibited negative traits documented in the reports for these shootings. There is the possibility that even more traits remain undiscovered about the shooters, but the majority of them were slain during the act of their shooting.

## V. CONCLUSIONS

It is now easy to see that school shootings have continued to affect all regions of the United States and does not appear to be slowing down. The good news is that the majority of these shootings are not very dangerous. The real shootings to be wary of are those that involve the deaths and injuries of many people.

This report gives an overview of school shootings in this nations' past, and basic statistics about where these shootings have taken place, what type of schools they affect, and how dangerous they are. I think more effort needs to be taken to further investigate the shootings deemed as most dangerous. These are the shootings that heavily impact peoples' lives and learning more about them can hopefully prevent future attacks of similar nature. Looking more into the shooters' childhoods and how they obtained their weapons could also give more insight into possibly identifying certain people as potential threats before these type of events occur.

Looking beyond the U.S. can also be of great benefit. Not all developed countries have this problem, so looking at what they do differently can inspire people in this country to adopt similar principles and practices. We also need to see how school shootings in the U.S. compare to the rest of the world. If we turn out to be one of the worst countries in this regard, then we need to seriously reconsider how we handle gun ownership and availability.

## REFERENCES


[1] Segel; Edward; and Heer, J. 2010. Narrative Visualization: Telling Stories with Data. IEEE Transactions on Visualization and Computer Graphics, v.16. n.6, pp.1139-48.

[2] Juan, R. 2003. "Using tf-idf to determine word relevance in document queries." Proceedings of the first instructional conference on machine learning.

[3] Van Zaanen; Menno; and Kanters, P. 2010 "Automatic Mood Classification Using TF* IDF Based on Lyrics." ISMIR.

[4] Zimring, F. 1975. "Firearms and federal law: the Gun Control Act of 1968." The Journal of Legal Studies 4.1, pages 133-198.

[5] Strait, M. 2010. "Enoch Brown: A Massacre Unmatched". http://pabook2.libraries.psu.edu/palitmap/Enoch.html.