MWSUG 2018 - Paper BL-090

Analyzing YouTube Comments on Gun Violence using SAS® Viya and SAS® Enterprise Miner

Sridevi Loya, Oklahoma State University, Stillwater, OK

ABSTRACT

Gun violence is a major cause of premature death in the U.S. "Guns kill more than 38,000 people and cause nearly 85,000 injuries each year" according to the American Public Health Association [1].

This paper focuses on studying the public opinions and reforms of gun violence in schools within the United States of America. This research includes studying and understanding the views of people after the occurrence of the Parkland, Florida mass shooting. Text analytics is performed on the comments and replies written by people on YouTube. This paper is specially looking at two specific sites for textual data, the CNN news and the ABC news site. These comments were written by people after the incident occurred. This paper will be of benefit as the main aim of this paper is to come up with potential solutions by looking in detail for people's opinion, derive meaningful insights from the comments and analyze if people have any viable recommendations to mitigate such acts.

INTRODUCTION

Recently the world witnessed a tragic incident involving multiple victims of firearm-related violence. On February 14, 2018, a gunman opened fire at Marjory Stoneman Douglas High School in Parkland, Florida, killing seventeen students and staff members and injuring seventeen others [2]. Gun violence is a complex issue and it is spread throughout the country. No other developed nation comes close to the rate of US gun violence. Americans own an estimated 265 million guns, more than one gun for every adult. Data from the Gun Violence Archive reveals there is a mass shooting – defined as four or more people shot in one incident, not including the shooter – nine out of every ten days on average [3]. Every citizen of this country is endangered by this critical issue and it is important to provide more safety to the citizens and mitigate the threat caused by gun violence and mass shootings. This paper focuses on providing a deeper understanding of public beliefs on such incidents and identify any viable recommendations given by public to mitigate these incidents.

DATA EXTRACTION

The textual data is collected from YouTube. The comments are scraped using YouTube Comment Scraper and the sites from which the comments were scraped are:

- 3,317 comments from ABC News (https://www.youtube.com/watch?v=WeXOXeJHRqc)
- 9,554 comments from CNN News (https://www.youtube.com/watch?v=2K_rFcmrQqE)

The textual data for this analysis is prepared using the following steps. These steps include data extraction from YouTube, importing the textual data in SAS® Enterprise Miner to create a SAS data set using file import node and importing the data into Base SAS® for creating word-frequency data set to create word clouds in SAS® Viya.

Figure 1 shows the steps that were performed to obtain the final Excel file having all the comments and replies. These comments were downloaded as two csv files and the outcome of the below process was an Excel file with 12,871 text rows.

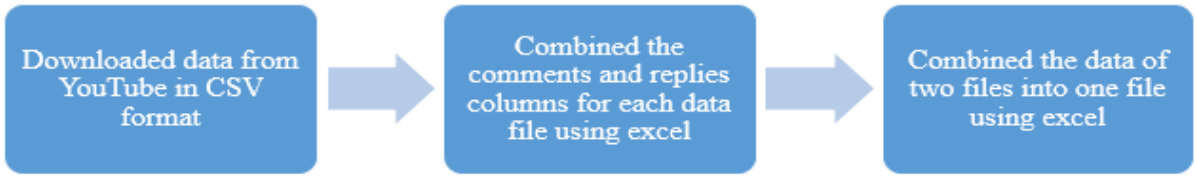


Figure 1: Process of Data Extraction

IMPORTING TEXTUAL DATA INTO SAS® ENVIRONMENT

The textual data residing in the Excel file was imported into the SAS Enterprise Miner using file import node and this node was run with default settings. In Base SAS®, the data was imported using Proc Import step.

METHODOLOGY:

Once textual data was available in SAS®, the following two-step methodology was used for text analytics.

1. Creating text clusters and topics for identifying meaningful categories and themes for the terms used in the comments using SAS® Enterprise Miner.
2. Generating word cloud based on the frequency of terms using SAS® Viya.

1. CREATING TEXT CLUSTERS AND TEXT TOPICS USING SAS® ENTERPRISE MINER

For identifying meaningful categories of the comments, the following text mining process flow was implemented. After importing the file, Text Parsing node was used for parsing the textual data to identify the term-document matrix, Text Filtering node was used to check for spelling errors using the dictionary. The output was then sent to Expectation Maximization Text Cluster node and Text Topic node as shown in figure 2.

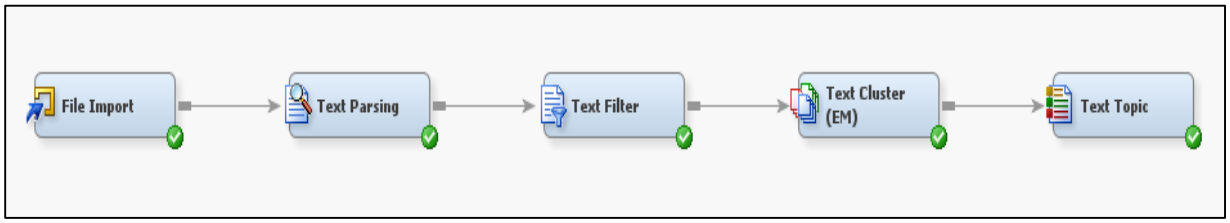


Figure 2: Modeling Diagram for Creating Text Clusters

TEXT PARSING

The textual data set generated by file import node was parsed to enumerate the terms contained in the document. It identified the word terms based on various parts of speech present in the document. The following properties were altered in properties panel of Text Parsing node.

- "Detect different parts of speech" was turned off which limits the terms with same parts of speech.
- "Find Entities" was set to "Standard".
- Parts of speech which filters prepositions, determinants, auxiliary verbs etc. were ignored as these generally contains very less information.
- Numeric and Punctuation attributes were ignored.
- Also entities such as Currency, Internet, Measure, Person etc. were ignored.

Property	Value
Parse Variable	Text
Language	English
Detect	
Different Parts of Speech	No
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULTI
Find Entities	Standard
Custom Entities	
Ignore	
Ignore Parts of Speech	'Abbr' 'Aux' 'Conj' 'Det' 'Interj'
Ignore Types of Entities	'Address' 'Company' 'Currency
Ignore Types of Attributes	'Num' 'Punct'
Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENGSYNMS
Filter	
Start List	
Stop List	SASHELP.ENGSTOP

Terms

Term	Role	Attribute	Freq
+ gun ...		Alpha	4048
+ have ...		Alpha	3295
+ people ...		Alpha	3053
s ...		Alpha	2654
+ shoot ...		Alpha	1903
+ school ...		Alpha	2026

Figure 3: Text Parsing Node Property Panel Settings and Output

As can be seen in figure 3 some of the terms such as “gun”, “school”, “shooting” etc. were the most occurring words which is obvious as we are analyzing gun violence in schools.

TEXT FILTERING

To reduce the number of terms used in the documents, text filter node was used. English dictionary was used to identify and correct the spell check errors; if not handled, spelling errors would result in keeping similar words as different terms expanding the term document matrix. Using filter viewer, all the documents containing a specific term were viewed and concept links based on those terms were created. Text filtering node with term weight property as “Inverse Document Frequency” was used. The “Check Spelling” option corrected wrong spellings of the word “definitely” as shown in figure 4.

Term	# Docs	Parent
definetly	1.0	definitely
definately	1.0	definitely
definetely	1.0	definitely

Figure 4: Text Filtering Node Output

Apart from the common English word synonyms identified using Text Parsing node, few custom synonyms list was created treating those terms as similar terms. Using interactive filter viewer, irrelevant terms were filtered. Some of the terms that were kept for further analysis are shown in figure 5.

Terms

	TERM	FREQ	# DOCS	KEEP ▼	WEIGHT
⊞	gun	4064	2426	☑	3.322
⊞	people	3065	2143	☑	3.501
⊞	shoot	1924	1541	☑	3.977
⊞	school	2055	1514	☑	4.002
⊞	white	1415	961	☑	4.658
⊞	kill	1200	926	☑	4.711
⊞	kid	1154	914	☑	4.73

Figure 5: Text Filtering Node High Frequency Terms

Concept Linking

Concept linking helps in understanding the relationships between words based on the co-occurrence of words in the documents. The hub and spoke structure of the concept link represents the association between those terms and width of the link represents the strength of association. The thicker the link is between two terms, the stronger the association is between the terms. Some of the important terms related to mass shooting such as “Gun” and “Stop” were analyzed as seen below.

1. Gun: - As per the concept link shown in figure 6, among all the words associated with “Gun”, “Owner” is the most frequent term occurring together (43/49 i.e. 43 out of 49 documents in which the term owner occurs). This sounds reasonable, as Owner is the most important parameter to understand the gun violence. Also we can see, black market (46/55) and access (72/91) as other terms strongly related to gun which possibly identifies having access to black markets to buy guns.

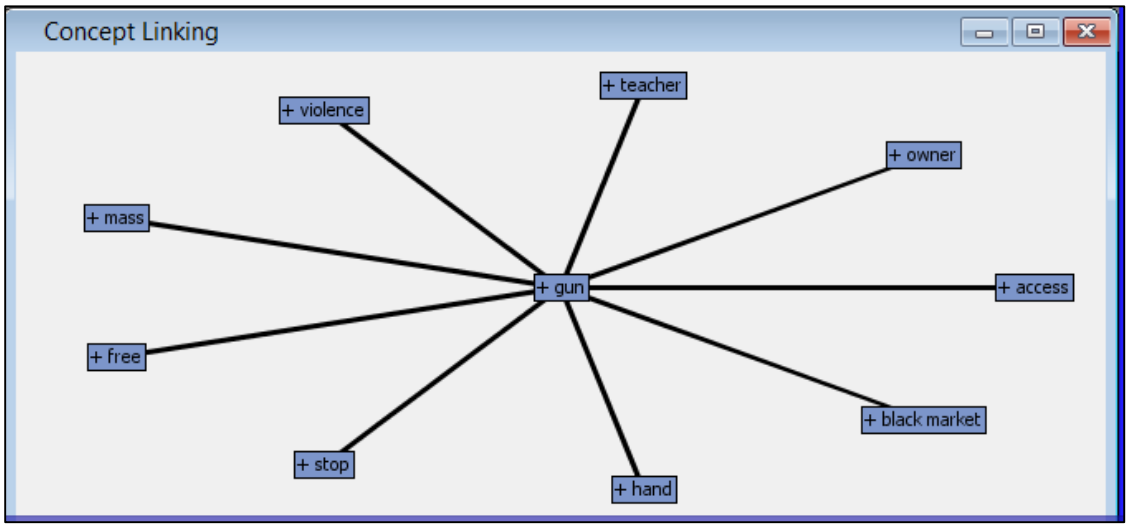


Figure 6: Concept Links for Gun

2. Stop: - As per the concept link shown in figure 7, among all the words associated with “Stop”, “Problem” is the most frequent term occurring together (73/406 i.e. 73 out of 406 documents in which the term Problem occurs). This sounds reasonable as gun violence is a problem and people are using stop indicating they want to stop the problem.

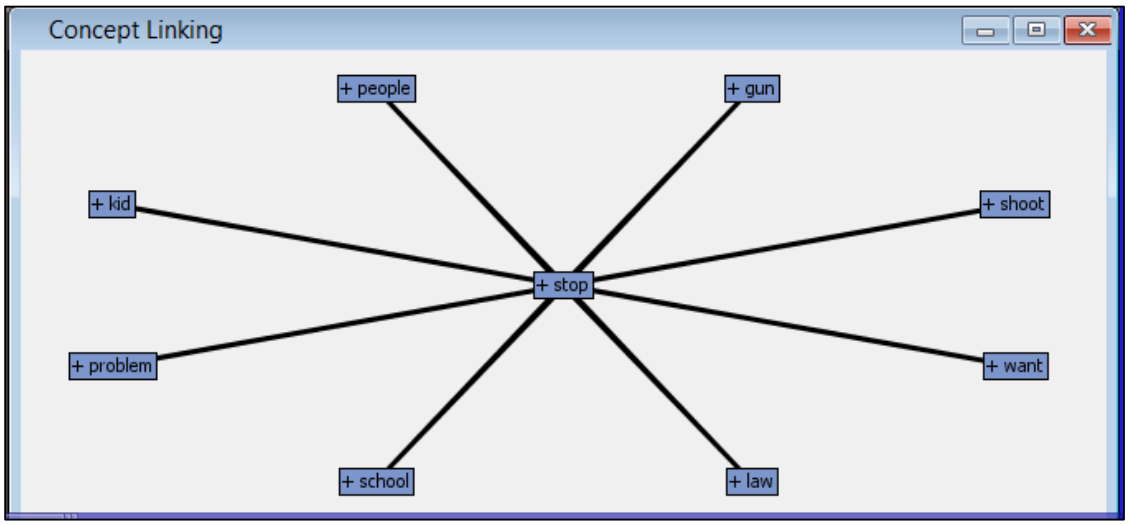
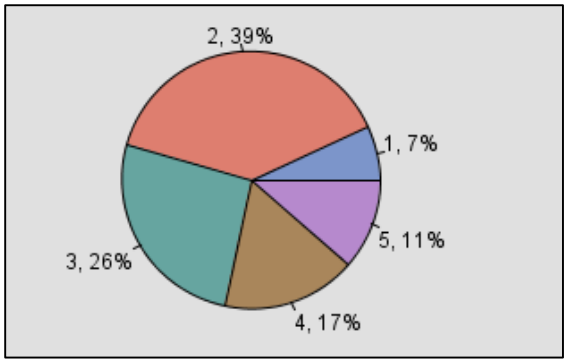


Figure 7: Concept Links for Stop

TEXT CLUSTERING

After filtering the irrelevant terms and combining similar terms, Text Cluster Node was used for clustering the comments into meaningful categories based on the terms present in them. The Expectation - Maximization cluster algorithm was used. Initially with default properties, many clusters with overlapping terms were developed. The number of clusters in the property panel of Text Cluster Node were restricted to arrive at fewer reasonable clusters (E.g. 4, 5 and 6 clusters). Among those, a 5 cluster solution was selected based on the least overlap among the descriptive terms. The cluster output is as follows. The descriptive terms in this 5-cluster solution and their terms are shown in figure 8.



Cluster ID	Descriptive Terms
1	+good +long +lose +pack +'threaten post' +aid +drink +hero +inform +move
2	+day +fake +american +live +news +trump +love +world +sad +funny
3	+gun +school +control +law +arm +weapon +teacher +'gun control' +problem +mental
4	+shoot +good +kid +life +guy +mass +lose +death real +bad
5	+white +people +kill +black +racist +race +hispanic +always +crime +blame

Figure 8: Text Cluster Descriptive Terms

TEXT TOPICS

After clustering the documents into 5 clusters, Text Topic node was used to identify the topics or terms in each cluster.

Topic ID	Topic
1	+gun,+control,+gun control,+law,+ban
2	+white,+black,+racist,+people,+hispanic
3	+school,+shoot,+school shooting,+kid,+mass
4	+people,+kill,+shoot,+die,+innocent
5	+kid,+good,+day,+life,+family

Figure 9: Text Topics

After Analyzing these words, clusters were categorized into more meaningful categories representing the clusters as listed below.

Topic 1: Law and Control
Topic 2: Diversity
Topic 3: School shooting
Topic 4: Innocent dying

Topic 5: Family and life

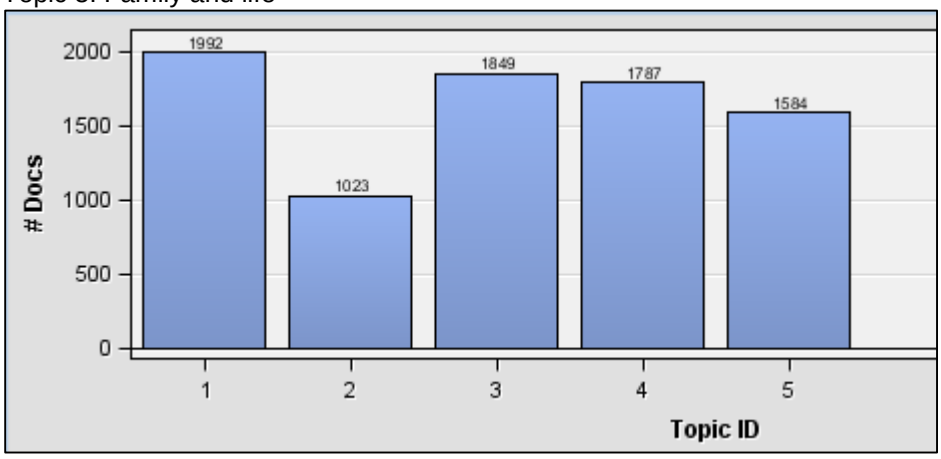


Figure 10: Number of Documents by Topics

Figure 10 is a bar graph of number of documents or comments in each topic. It can be seen that topic 1 has the highest comments followed by topic 3, topic 4, topic 5 and topic 2. Topic 1 which represents law, control, ban has occurred in most of the documents followed by school shooting and innocent being killed. People are also empathetic and have used topic 5 which represents life and family.

2. GENERATING WORD CLOUD USING SAS® VIYA:

In order to create word cloud in SAS® Viya, a word frequency data set was created by using the below code in Base SAS®.

Firstly, for importing the combined comments file into Base SAS®, the following proc import code was used.

```
proc import out= work.data1
            datafile= "c:\data\combined_comments.xlsx"
            dbms=excel replace;
     range="sheet1$";
     getnames=yes;
     mixed=no;
     scantext=yes;
     usedate=yes;
     scantime=yes;
run;
```

To read the text and break them into words, the following code was used.

```
data temp1(keep=b);
    set data1;
    delims = ' 0123456789,.!?';
    i=1;
    do while (scan(text, i) ne "");
        b=scan(text,i);
        j+1;
        output;
        i+1;
    end;
run;
```

All the words were converted to lower case and all the special characters, numbers and other unwanted characters were omitted using the below code.

```
data temp2(keep= new);
    set temp1;
    x=lowcase(b);
    new=compress(x,"abcdefghijklmnopqrstuvwxyz" , "kis");
run;
```

A stop word data set is created in order to remove all the stop words from the text. Not all words were included here.

```
data stopwords;
    input words $;
    datalines;
there
that
/* All stop words /
;
run;
```

A data set ‘word’ was created having all the words excluding stop words and the frequency of their occurrence.

```
proc sql;

    create table word as
          select new as word, count(*)as frequency
          from temp2
          where new not in(select words from stopwords)
          group by new
          order by 2 desc;
quit;
```

This word data set is exported using Export wizard in Base SAS® and imported in SAS® Viya. The word cloud is listed under objects in SAS® Viya as shown in figure 11.

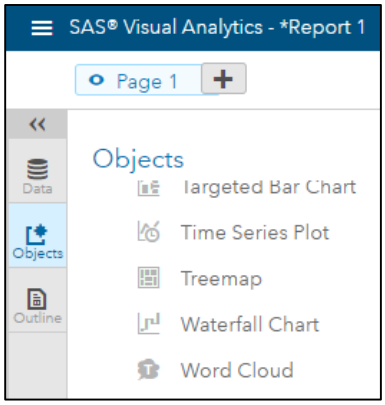


Figure 11: SAS® Viya Word Cloud Object

After setting up the roles of the variables in word cloud object, the word cloud for all the words used in comments was created as can be seen in figure 12.

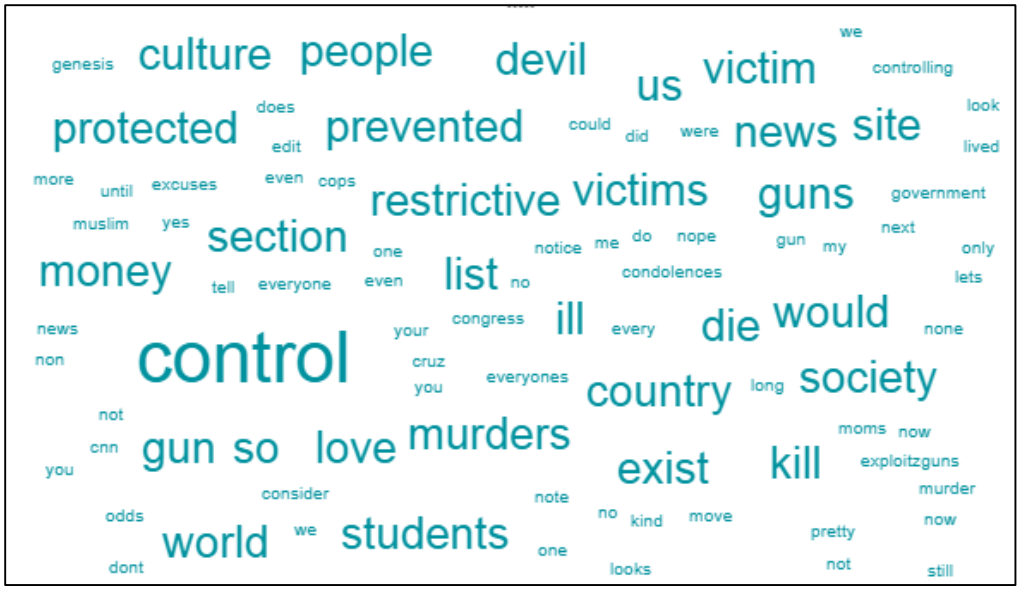


Figure 12: Word Cloud

From the word cloud generated in SAS® Viya, we can see that people have been using a lot of different words related to Gun violence. The bigger the word, the more frequently it was used while commenting. Some of the words that are bigger in size are control, restriction, prevention, protection, government, world, culture, students and many more as seen in figure 12.

SUMMARY

From the results, it can be seen that the comments from most of the people were based on topics such as law and control, diversity, school shootings, innocent dying, family and life. Of all these topics, the most frequent topic used was law and control. The word cloud also indicated the most frequently used words by people such as control, prevent, restrict, protect, students, country, society, world and many more.

This analysis was based on very limited data gathered only from the comments posted by public on CNN and ABC YouTube news sites after school shooting in Florida, therefore, a very comprehensive study has to be carried out before implementing any laws or taking any action on such an intricate subject. But, some possible solutions inferred from public views were around gun control, education, prevention and protection.

REFERENCES

[1] https://www.apha.org/topics-and-issues/gun-violence
[2] https://en.wikipedia.org/wiki/Stoneman_Douglas_High_School_shooting
[3] https://www.theguardian.com/us-news/ng-interactive/2017/oct/02/america-mass-shootings-gun-violence
[4] https://support.sas.com/documentation/onlinedoc/txtminer/14.1/tmref.pdf
[5] https://support.sas.com/documentation/onlinedoc/va/

ACKNOWLEDGEMENT

I thank Dr. Goutam Chakraborty, Ralph A. and Peggy A. Brenneman professor of marketing, Dr. Miriam McGaugh, lecturer at Oklahoma State University for their guidance and support throughout the research.

CITATION

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sridevi Loya
Oklahoma State University
Email: sridevi.loya@okstate.edu

Sridevi Loya is a Business Analytics Graduate Student at Oklahoma State University. She is SAS® certified Base Programmer, Advance Programmer and Predictive Modeler. She has two years of experience using SAS® tools with a focus on Data Mining, Text Analytics, Database Marketing and Business research.