School Shootings – News vs. Twitter

A3: Business Insight Report

Text Analytics - DAT-5317 - FMSBA3

Thomas Kurnicki

Deniz Gürcan

Student ID: 4927234

Master of Business Analytics

MSBA 3

HULT International Business School

Table of Contents

Introduction

Imagine your child goes to school to learn in the morning and never comes back. Sounds unreal? Unfortunately, it is not.

In the last 20 years more than 228.000 students in the US experienced gun violence in schools. 110 school shootings with 61 deaths in schools across the country within the last year. It could happen anytime anywhere (Brous & Lewis, 2019).

There are numerous of reasons regarding the occurrence of school shootings and gun violence occur almost on a daily basis in American schools. It is a very controversial topic, heavily discussed on the news and social media.

Common sense makes us think that the nature of these types of media are fundamentally differing in terms 5of expressing emotions, facts and sentiment. But in what manner does it really differ and where are similarities? The following will assess and evaluate public sentiment on school shooting within the USA comparing news, articles and journals to tweets.

Data Collection

In order to asses a large variety of journalism, in total 14 different news articles and journals from within the last years have been gathered and compiled in a text file. The social media data has been extracted from Twitter using the word combination “School shooting”.

Results

First both datasets were cleaned, transformed and tokenized in order to conduct further analysis. Words and word combinations that are directly linked to the topic such as school, shooting etc. have been removed to rather asses underlying patterns in the data.

In order to find out what the most common words used in regards with school shootings are on the news and twitter, word frequencies are displayed. It becomes obvious that the most frequent words on the news are linked to the occurrence of shootings and to the public in general, whereas frequent words on twitter are rather linked to single incidents, locations and responding times.

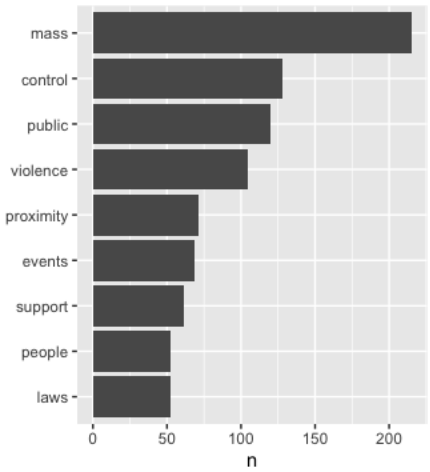

Figure 1: News word frequencies

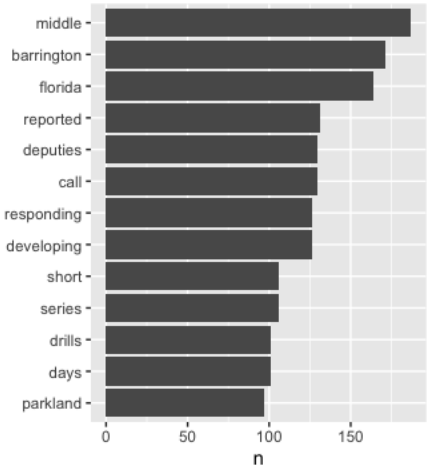

Figure 2: Twitter word frequencies

Regarding the overall sentiment on each type of media, one might assume that the sentiment on Twitter is more negative due a tendency of emotional and unresearched tweets compared to news. However, conducting an initial sentiment analysis wit *afinn*, news and twitter are almost equally neutral with slightly negative with a score of -0.414 and -0.577 respectively. Looking at how words are connotated it appears that negative words weigh higher than the positive ones which are more common but not as powerful. ‘Issue’, ‘Threat’, ‘Attacks’, as the most frequent negative words in the news empathize that the topic is approached holistically and objective. The sentiment in tweets can be seen similarly with ‘Difficult’ and ‘Threat’ as most frequent negatives. However, the emotionality in tweets can be seen by the occurrence of

swearwords and negative words such as ‘lost’ or ‘dark’ and ‘Destiny’, ‘Grateful’ or ‘Love’ on the positive side.

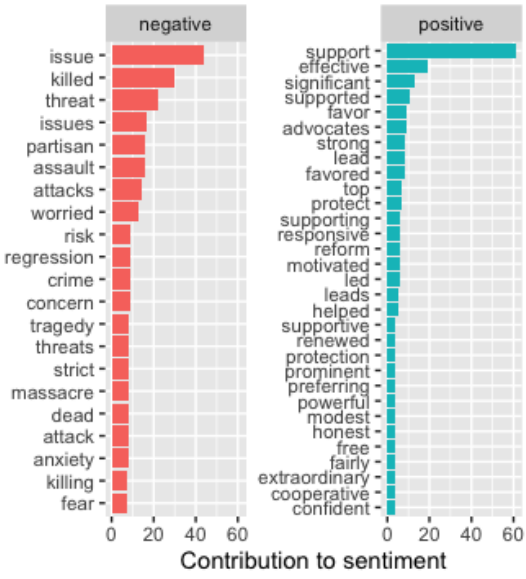

Figure 4: News sentiment frequencies (bing)

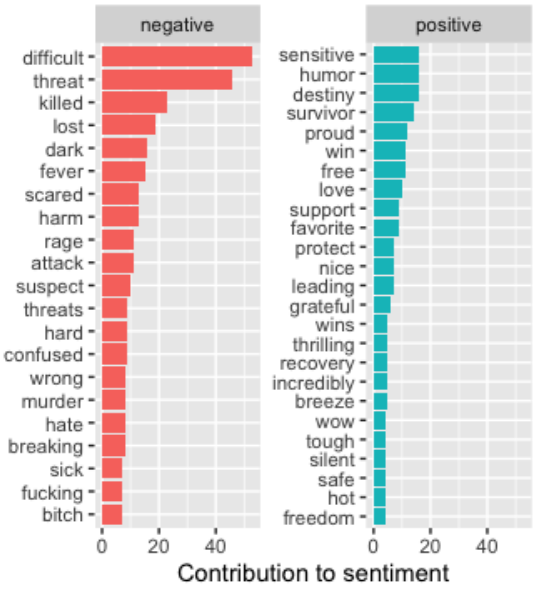

Figure 3: Twitter sentiment frequencies (bing)

Analyzing the nrc sentiments, the data reveals an unexpected clustering of words around the factor ‘Joy’. This indicates that there is positive sentiment in regards to anti school violence incidents in the news and expression of empathy on Twitter. ‘Surprise’ and ‘Anticipation’ are dominant factors in both datasets. The twitter sentiment reveals strong emotional sentiment with

Figure 6: News sentiment factors (nrc)

Figure 5: Twitter sentiment factors (nrc)

‘Anger’, ‘Disgust’, ‘Sadness’, which contributes to the overall impression of the sentiment on twitter.

Conducting further analyses to see interconnection and potential causalities between words in the datasets reveals the following: It appears that the news discusses this topic in a reflected way referring to research putting violent incidents in context with politics at that time by connecting years with words such as ‘Election’ or ‘Congressional’. Furthermore ‘Policies’ are affecting the topic, being connected with ‘Rifles’, ‘NRA’ and ‘Weapons’. But also, being connected to ‘African American’, ‘Health’, ‘Voters’ to education. These connections indicate a holistic discussion of school shootings with other controversial topics in the USA in order to find underlying patterns of reasons which cause such events. This also can be seen reviewing the most frequent trigrams (Appendix A) which are centered around policies for prevention of shootings mentioning word combinations such as ‘Universal Background checks’.

Twitter reveals a discussion of politics in o more subtle way mentioning ‘Safety’ to ‘Gun’ and extremism. More remarkably and unlike the news a human factor plays a major role on twitter. Words centered around ‘Innocent’ are showing a focus on the victims and shooters. Mental health and illnesses are discussed as well, which shows a differentiated view and empathy. ‘Parkland’ as one of the major shootings in the recent past with 17 casualties has its 2 years anniversary and is therefore remembered and re-discussed on twitter, which causes many connections around that word (Andone, 2020). Thus, making twitter a platform to express opinions, feelings and remember. Politics are rather discussed briefly not in a reflected way being connected to buzzwords in that field.

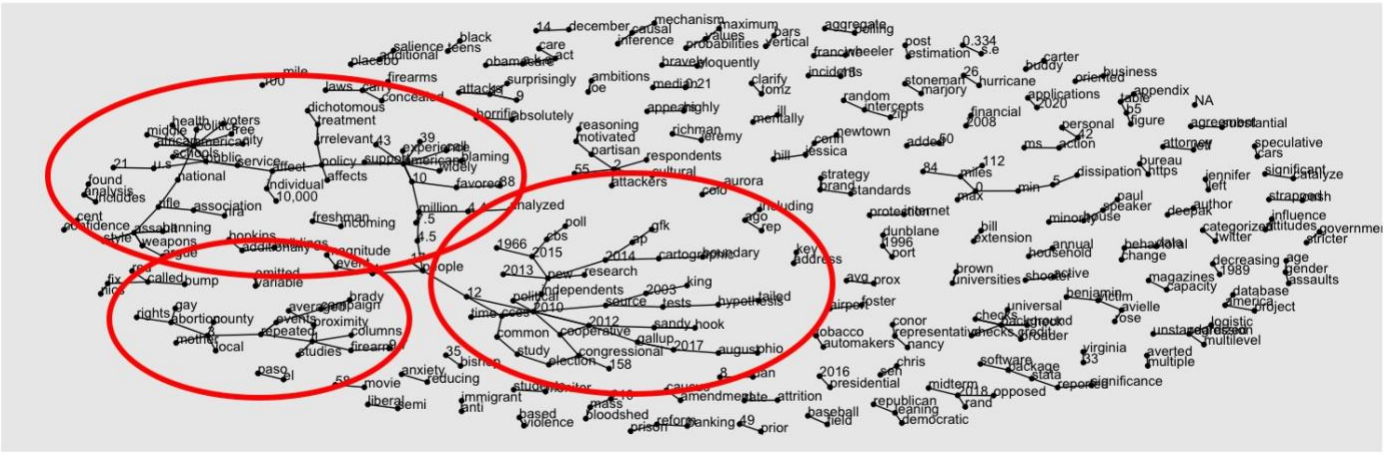

Figure 7: News interconnection of words

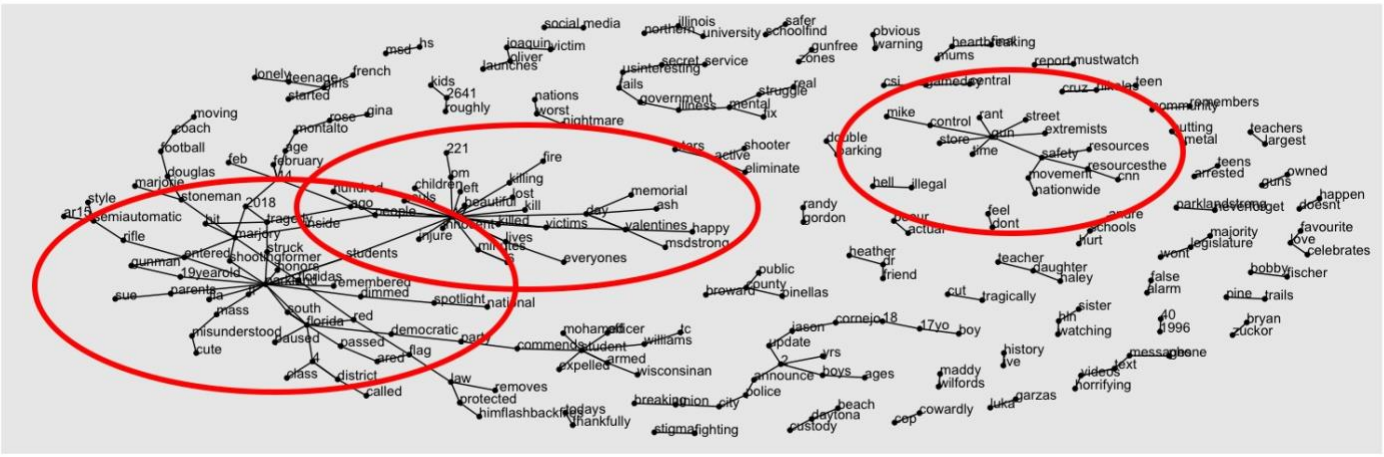

Figure 8: Twitter interconnection of words

Discussion and Conclusion

Comparing news, articles and journals to twitter is prawn to several biases. First one has to be aware of the length of text meaning that news are proportionally long in comparison to tweets. Second the authors. Professional journalists on the one hand and primarily normal people and semi-professional journalists or experts on the other hand. The depth of analysis and report is

also a factor to consider. Whereas news and articles are researched in depth tweets happen more intuitively and ad hoc.

Especially the last factor can be observed vividly through the analysis on hand. School shootings are reviewed, assessed and analyzed objectively in a holistic way putting the topic in a global context with differentiated weighting of differing opinions trying to find causation and patterns. Tweets instead are limited to particular incidents, reporting reactively with less in-depth analysis. Remarkably tweets are often dedicated to the victims with the expression of feelings emotions and empathy for those who are affected by gun violence in schools. Expressing emotions on twitter as a relief in the anonymity of the internet and social media is common in tweets. Emotional tweets with strong opinions also gain more reach though re-tweets and likes on this platform (Gantman, 2019).

Interestingly politics are rather under stated especially in the news, having in mind that many of the analyzed articles contain a vivid discussion about actions of democrats and republicans.

Overall this analysis shows that as expected news are more likely to offer a wholesome discussion of this very controversial topic in the US. Twitter is used to express emotions, to inform about possible incidents and to remember and help those who suffer. In the future there is a high necessity of further analysis especially of social media in general in order to eventually develop an intelligence that can detect people or areas that are prawn to gun violence based on shared content to prevent these horrible events. The ultimate goal should be to limit shootings to zero as it is expected from a highly developed country like the US, so that no one has to be scared by sending their kids to school in the morning.

Appendix A

Trigram Barplots

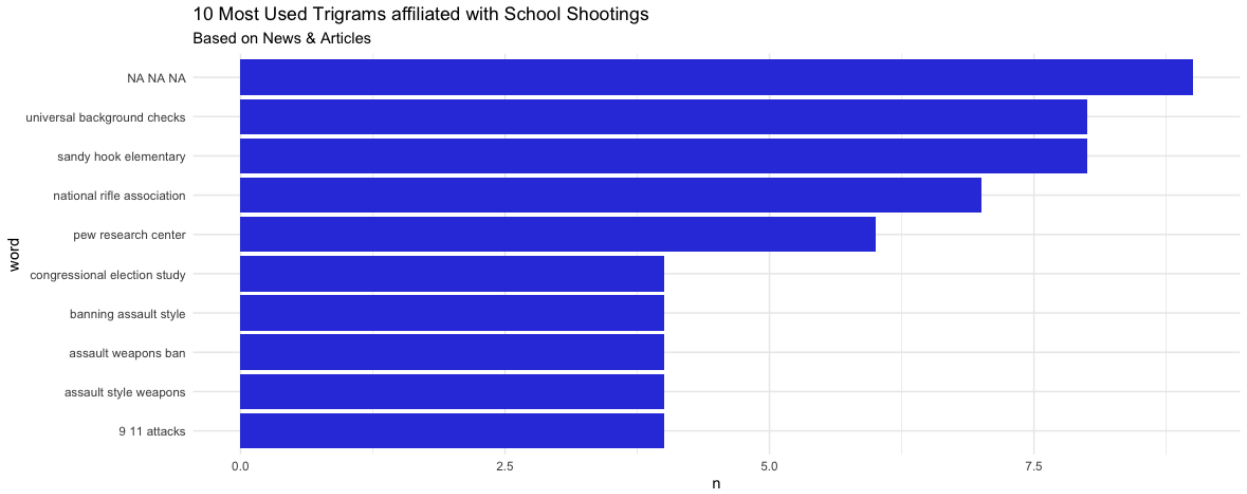

Figure 9: News Trigram Barplot

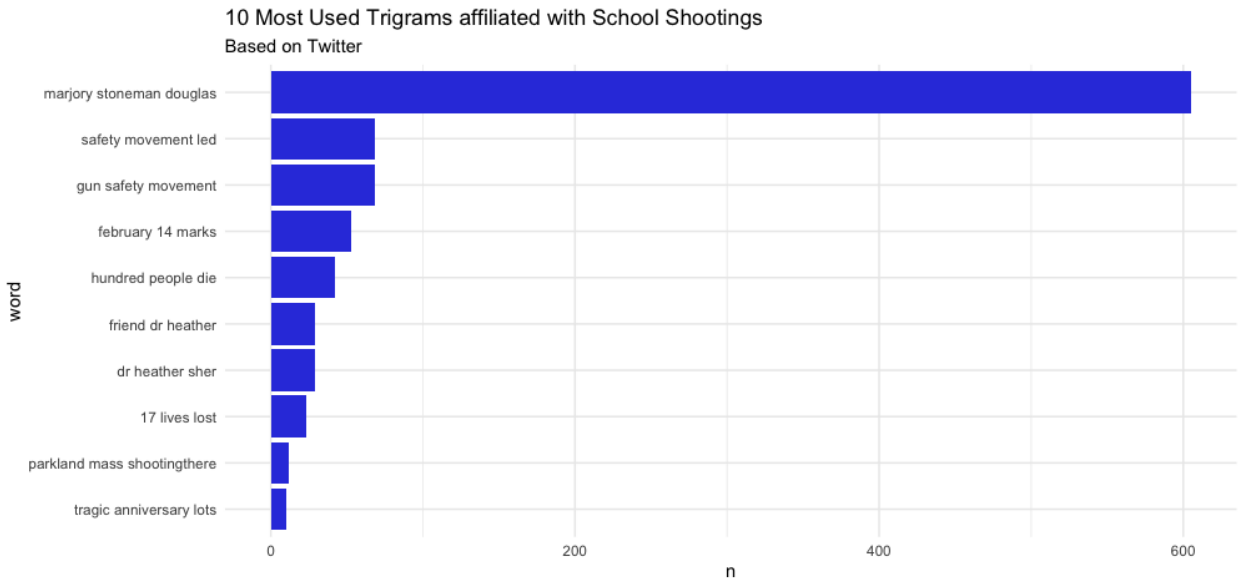

Figure 10: Twitter Trigram Barplot

References

Andone, D. (2020, February 14). It's been 2 years since the deadly shooting at a high school in Parkland, Florida. Retrieved February 14, 2020, from https://www.cnn.com/2020/02/14/us/parkland-shooting-marjory-stoneman-douglas-2-years/index.html

Beckett, L. (2019, October 23). Republicans propose mass student surveillance plan to prevent shootings. Retrieved February 11, 2020, from https://www.theguardian.com/world/2019/oct/23/republicans-mass-shootings-school-surveillance

Brous, S., & Lewis, J. J. (2019, September 18). A heart-stopping school shooting ad: No child should have to text last words to mom. Retrieved February 11, 2020, from https://www.usatoday.com/story/opinion/2019/09/18/sandy-hook-psa-gun-violence-school-shootings-not-inevitable-column/2354471001/

Cheadle, H. (2018, February 15). The Republicans' Plan to Stop Mass Shootings: Nothing. Retrieved February 11, 2020, from https://www.vice.com/en_us/article/wj4gqy/why-mass-shootings-keep-happening-republicans

Cochrane, E., & Stolberg, S. G. (2018, November 9). Another Mass Shooting, but This Time House Democrats Promise Action. Retrieved February 11, 2020, from https://www.nytimes.com/2018/11/09/us/politics/gun-control-california-shooting.html

Connolly, G. (2018, February 16). Democratic, Republican Responses to Parkland School Shooting Vary Wildly. Retrieved February 11, 2020, from https://rollcall.com/2018/02/16/democratic-republican-responses-to-parkland-school-shooting-vary-wildly/

Cornell University. (2020). Shootings, Guns and Public Opinion. Retrieved February 11, 2020, from https://ropercenter.cornell.edu/shootings-guns-and-public-opinion

Gantman, A. P. (2019, August 20). Why Moral Emotions Go Viral Online. Retrieved February 14, 2020, from https://www.scientificamerican.com/article/why-moral-emotions-go-viral-online/

Graf, N. (2018, April 18). Majority of teens worry about school shootings, and so do most parents. Retrieved February 11, 2020, from https://www.pewresearch.org/fact-tank/2018/04/18/a-majority-of-u-s-teens-fear-a-shooting-could-happen-at-their-school-and-most-parents-share-their-concern/

Ingraham, C. (2019, September 4). After mass shootings, GOP-led legislatures double efforts to loosen gun restrictions, data show. Retrieved February 11, 2020, from https://www.washingtonpost.com/business/2019/09/04/after-mass-shootings-republican-led-legislatures-double-efforts-loosen-gun-restrictions/

Jarvis, J., Dunwoodie, S., Hill, C., & Lancaster, J. (2019, October 3). Mass shootings have hit 158 House districts so far this year. Retrieved February 11, 2020, from https://thehill.com/homenews/house/463949-mass-shootings-have-hit-158-house-districts-so-far-this-year

Knowles, H. (2019, June 26). Democrats focus on victims, Republicans on perpetrators after mass shootings, study finds. Retrieved February 11, 2020, from https://www.washingtonpost.com/politics/2019/06/26/democrats-focus-more-victims-republicans-perpetrators-after-mass-shootings-study-finds/

McMaster, S. (2018, May 21). Things Republicans Have Blamed for School Shootings or a Top-200 Subreddit? Retrieved February 11, 2020, from https://www.mcsweeneys.net/articles/things-republicans-have-blamed-for-school-shootings-or-a-top-200-subreddit

Newman, B., & Hartman, T. (2019). Mass Shootings and Public Support for Gun Control. *British Journal of Political Science, 49*(4), 1527-1553. doi:10.1017/S0007123417000333

Plott, E. (2018, March 16). 'I Think We Have a Leadership Problem'. Retrieved February 11, 2020, from https://www.theatlantic.com/politics/archive/2018/03/republican-guns/555737/

Reinhart, R. J. (2018, January 23). In the News: School Shootings. Retrieved February 11, 2020, from https://news.gallup.com/poll/226202/news-school-shootings.aspx

Saad, L. (2019, November 15). Gallup's Guide to U.S. Public Opinion on Guns. Retrieved February 11, 2020, from https://news.gallup.com/opinion/gallup/262724/gallup-guide-public-opinion-guns.aspx

Sipes, L. (2018, April 3). Public Opinion on School Shootings and Violence. Retrieved February 11, 2020, from https://www.lawenforcementtoday.com/public-opinion-school-shootings-violence/

Timm, J. C. (2018, May 19). After Santa Fe school shooting, Democratic lawmakers slam GOP for inaction on gun control. Retrieved February 11, 2020, from https://www.nbcnews.com/politics/politics-news/after-santa-fe-school-shooting-democratic-lawmakers-slam-gop-inaction-n875466

R Code

```
################################

#Individual Assignment - Business Insight Report

#Deniz Guercan

#MSBA3

################################


library(tidyverse)

library(tidytext)

library(dplyr)

library(textreadr)

library(stringr)

library(twitteR)

library(tm)

library(wordcloud)

library(scales)

library(wordcloud2)

library(RColorBrewer)

library(reshape2)

library(igraph)

library(ggraph)

library(plotly)

library(widyr)
```

```
#Importing the .txt file with the news, journals and articles (in the following as news)

setwd("/Users/deniz/Desktop/HULT_Stuff/DUAL DEGREE MBAN/SPRING/TEXT
ANALYSIS")

#using read document to acces the file and store it as a vector
ndata <- read_document(file="/Users/deniz/Desktop/HULT_Stuff/DUAL DEGREE
MBAN/SPRING/TEXT ANALYSIS/Shootings_text_data.txt")

#Transorming to a data frame
ndata <- data.frame(line=1:length(ndata),text=ndata, stringsAsFactors = FALSE)

#################
#Accessing Twitter through TwitteR to acces tweets about shool schootings to compare news to
social media
#################
consumer_key <- 'FMG3XjyXuRkEiOH9C0Q275ZK0'
consumer_secret <- 'kUbLSFvIxNPquviM9dLv8hJR8ss9hxIM2MJ9FX0Y3FghGetkDL'
access_token <- '1217606322838855680-Al87L2vaDuwxLpsBQ9bTkbwTY1ZFQx'
access_secret <- '2PRbCDmmXjgQ8cSDSVdbkoQbg4R7Fo95WWljAjwlwpnu2'
```

```
shoosag <- twitteR::searchTwitter('school shooting', lang = 'en', n = 1500, since = '2000-06-01',
retryOnRateLimit = 1e3)

View(shoosag)

ss = twitteR::twListToDF(shoosag)
ss <- as.data.frame(ss)

view(ss)
#clean the twitter data
ss$text <- gsub("https\\S*", "", ss$text)
ss$text <- gsub("@\\S*", "", ss$text)
ss$text <- gsub("amp", "", ss$text)
ss$text <- gsub("[\r\n]", "", ss$text)
ss$text <- gsub("[[:punct:]]", "", ss$text)
ss$text <- gsub("http[^[:space:]]*","",  ss$text)
ss$text <- gsub("http[^[:space:]]*","", ss$text)

#Create own stop_word list
my_junk_twit <- data_frame(word=c("t.co", "1        o:it", "brianna", "ta", "msnbc", "tony",
"cnns","trump","iqykldai5y", "top", "sumoh7", "ii", "iii", "i'm", "y'all'", "celestljade","like",
"https", "rt", "shooting", "shoot", "school", "im", "hey", "cus" ), lexicon="my_junk_twit")
```

```
my_junk_news <- data_frame(word=c("shooting","trump","shoot","na","gun","john", "guns",
"school", "shootings"), lexicon= "my_junk_news")

#################
#Tokenizing both news and twitter data
#################

news_tokens <- ndata %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_news) %>%
  count(word, sort=T) %>%
  ungroup()

view(news_tokens)

ss_tokens <- ss %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_twit) %>%
  count(word, sort=T) %>%
  ungroup()
```

```
view(ss_tokens)


############
#inspect frequencies
###########


#frequencies news


frequ_tokens_news <- ndata %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_news) %>%
  count(word, sort=TRUE)
print(frequ_tokens_news)


frequ_tokens_news %>%
  filter(n > 50) %>%
  mutate(word = reorder(word, n, color = "blue")) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip()
```

```
#frequencies twitter


frequ_tokens_twit <- ss %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_twit) %>%
  count(word, sort=TRUE)
print(frequ_tokens_twit)


frequ_tokens_twit %>%
  filter(n > 70) %>%
  mutate(word = reorder(word, n, color = "blue")) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip()


######################
#Sentiment Analysis
######################


get_sentiments("bing")
get_sentiments("afinn")
```

```
get_sentiments("nrc")


#Sentiments - bing
news_senti <- ndata %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_news) %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

news_senti

news_senti %>%
  group_by(sentiment) %>%
  top_n(20) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="Contribution to sentiment", x=NULL)+
  coord_flip()
```

```
twit_senti <- ss %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_twit) %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

twit_senti

twit_senti %>%
  group_by(sentiment) %>%
  top_n(20) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="Contribution to sentiment", x=NULL)+
  coord_flip()

#Sentiment afinn
```

```
news_senti_num <- ndata %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_news) %>%
  inner_join(get_sentiments("afinn")) %>%
  count(word, value, sort=T) %>%
  summarize(mean(value)) %>%
  ungroup()

print(news_senti_num)

twit_senti_num <- ss %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_twit) %>%
  inner_join(get_sentiments("afinn")) %>%
  count(word, value, sort=T) %>%
  summarize(mean(value)) %>%
  ungroup()

print(twit_senti_num)
```

```
#NRC and Comparison cloud


news_senti_nrc <- ndata %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_news)%>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()


#Pizza
news_senti_nrc %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(colors = c("grey20", "gray80"),
                   title.colors=c("red","blue"),
                   max.words=100, fixed.asp=TRUE,
                   scale=c(0.8,0.8), title.size=2, rot.per=0.25)



twit_senti_nrc <- ss %>%
  unnest_tokens(word, text) %>%
```

```
  anti_join(stop_words) %>%
  anti_join(my_junk_twit)%>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

twit_senti_nrc %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(colors = c("grey20", "gray80"),
                   title.colors=c("red","blue"),
                   max.words=100, fixed.asp=TRUE,
                   scale=c(0.8,0.8), title.size=2, rot.per=0.25)

bing_and_nrc <- bind_rows(
  news_tokens %>%
    inner_join(get_sentiments("bing")) %>%
    mutate(method = "Bing et al."),
  news_tokens %>%
    inner_join(get_sentiments("nrc") %>%
                 filter(sentiment %in% c("positive", mutate(method = "NRC")) %>%
"negative"))) %>%
```

```
  count(method, index = linenumber %/% 80, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)



#################
#Ngrams
#################
news_ngrams <- ndata %>%
  unnest_tokens(bigram, text, token = "ngrams", n=3) %>%
  separate(bigram, c("word1", "word2", "word3"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!word3 %in% stop_words$word) %>%
  filter(!word1 %in% my_junk_news$word) %>%
  filter(!word2 %in% my_junk_news$word) %>%
  filter(!word3 %in% my_junk_news$word) %>%
  count(word1, word2, word3, sort = TRUE)

ggraph(slice(news_ngrams, 1:300), layout = "fr") +
  geom_edge_link()+
  geom_node_point()+
  geom_node_text(aes(label=name), vjust =0.0001, hjust=0.0001)
```

```
twit_ngram <- ss %>%
  unnest_tokens(bigram, text, token = "ngrams", n=3) %>%
  separate(bigram, c("word1", "word2", "word3"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!word3 %in% stop_words$word) %>%
  filter(!word1 %in% my_junk_twit$word) %>%
  filter(!word2 %in% my_junk_twit$word) %>%
  filter(!word3 %in% my_junk_twit$word) %>%
  count(word1, word2, word3, sort = TRUE)

ggraph(slice(twit_ngram, 1:300), layout = "fr") +
  geom_edge_link()+
  geom_node_point()+
  geom_node_text(aes(label=name), vjust =0.0001, hjust=0.0001)

#######
#Additional ngrams for testing purposes - not used in analysis
########
twit_ngrams
twit_senti_ngram <- twit_ngram %>%
  inner_join(get_sentiments('afinn'), by=c(word2="word")) %>%
```

```
  count(word1, word2, value, sort=TRUE) %>%
  summarise(mean(value)) %>%
  ungroup()

twit_senti_ngram

set.seed(2017)
ggraph(twit_ngram, layout = "fr") +
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1)



set.seed(2016)
a <- grid::arrow(type = "closed", length = unit(.15, "inches"))
ggraph(twit_ngram, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1) +
  theme_void()

###############
```

```
#Trigramm barplots - not used because of NA values
###############
ndata %>%
  unnest_tokens(word, text, token = "ngrams", n = 3) %>%
  separate(word, c("word1", "word2", "word3"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!word3 %in% stop_words$word) %>%
  filter(!word1 %in% my_junk_news$word) %>%
  filter(!word2 %in% my_junk_news$word) %>%
  filter(!word3 %in% my_junk_news$word)%>%
  unite(word,word1, word2, word3, sep = " ") %>%
  count(word, sort = TRUE) %>%
  slice(1:10) %>%
  mutate(word = reorder(word,n ))%>%
  ggplot() + geom_bar(aes(word, n), stat = "identity", fill = "#3344de") +
  theme_minimal() +
  coord_flip() +
  labs(title = "10 Most Used Trigrams affiliated with School Shootings",
       subtitle = "Based on News & Articles",
       caption = "")


ss %>%
```

```
  unnest_tokens(word, text, token = "ngrams", n = 3) %>%
  separate(word, c("word1", "word2", "word3"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!word3 %in% stop_words$word) %>%
  filter(!word1 %in% my_junk_twit$word) %>%
  filter(!word2 %in% my_junk_twit$word) %>%
  filter(!word3 %in% my_junk_twit$word)%>%
  unite(word,word1, word2, word3, sep = " ") %>%
  count(word, sort = TRUE) %>%
  slice(1:10) %>%
  mutate(word = reorder(word,n ))%>%
  ggplot() + geom_bar(aes(word, n), stat = "identity", fill = "#3344de") +
  theme_minimal() +
  coord_flip() +
  labs(title = "10 Most Used Trigrams affiliated with School Shootings",
       subtitle = "Based on Twitter",
       caption = "")



#################
#Wordclouds - not used because of redundance with prior sentiments
#################
```

```
news_tokens %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(colors = c("grey20", "gray80"))

ss_tokens %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(colors = c("grey20", "gray80"))



############################
#Correlogram - not used because of a lack of words
############################

frequ <- bind_rows(mutate(news_tokens, author="news"),
                   mutate(ss_tokens, author="twitter")) %>%
  mutate(word=str_extract(word, "[a-z']+")) %>%
  count(author, word) %>%
```

```
  group_by(author) %>%
  mutate(proportion = n/sum(n))%>%
  select(-n) %>%
  spread(author, proportion) %>%
  gather(author, proportion,`twitter`)%>%
  filter(proportion, 0.0001)



ggplot(frequ, aes(x=proportion, y=`news`,
                  color = abs(`news`- proportion)))+
  geom_abline(color="grey40", lty=2)+
  geom_jitter(alpha=.1, size=0.6, width=0.4, height=0.4)+
  geom_text(aes(label=word), check_overlap = TRUE, vjust=.4) +
  scale_x_log10(labels = percent_format())+
  scale_y_log10(labels= percent_format())+
  scale_color_gradient(limits = c(0,0.001), low = "darkslategray4", high = "gray75")+
  facet_wrap(~author, ncol=3)+
  theme(legend.position = "none")+
  labs(y= "news", x=NULL)
```

R Output

```
################################

#Individual Assignment - Business Insight Report

#Deniz Guercan

#MSBA3

################################


library(tidyverse)

library(tidytext)

library(dplyr)

library(textreadr)

library(stringr)

library(twitteR)

library(tm)

library(wordcloud)

library(scales)

library(wordcloud2)

library(RColorBrewer)

library(reshape2)

library(igraph)

library(ggraph)

library(plotly)

library(widyr)
```

```
#Importing the .txt file with the news, journals and articles (in the following as news)

setwd("/Users/deniz/Desktop/HULT_Stuff/DUAL DEGREE MBAN/SPRING/TEXT
ANALYSIS")

#using read document to acces the file and store it as a vector
ndata <- read_document(file="/Users/deniz/Desktop/HULT_Stuff/DUAL DEGREE
MBAN/SPRING/TEXT ANALYSIS/Shootings_text_data.txt")

#Transorming to a data frame
ndata <- data.frame(line=1:length(ndata),text=ndata, stringsAsFactors = FALSE)

#################
#Accessing Twitter through TwitteR to acces tweets about shool schootings to compare news to
social media
#################
consumer_key <- 'FMG3XjyXuRkEiOH9C0Q275ZK0'
consumer_secret <- 'kUbLSFvIxNPquviM9dLv8hJR8ss9hxIM2MJ9FX0Y3FghGetkDL'
access_token <- '1217606322838855680-Al87L2vaDuwxLpsBQ9bTkbwTY1ZFQx'
access_secret <- '2PRbCDmmXjgQ8cSDSVdbkoQbg4R7Fo95WWljAjwlwpnu2'
```

```
shoosag <- twitteR::searchTwitter('school shooting', lang = 'en', n = 1500, since = '2000-06-01',
retryOnRateLimit = 1e3)

View(shoosag)

ss = twitteR::twListToDF(shoosag)
ss <- as.data.frame(ss)

view(ss)
#clean the twitter data
ss$text <- gsub("https\\S*", "", ss$text)
ss$text <- gsub("@\\S*", "", ss$text)
ss$text <- gsub("amp", "", ss$text)
ss$text <- gsub("[\r\n]", "", ss$text)
ss$text <- gsub("[[:punct:]]", "", ss$text)
ss$text <- gsub("http[^[:space:]]*","",  ss$text)
ss$text <- gsub("http[^[:space:]]*","", ss$text)

#Create own stop_word list
my_junk_twit <- data_frame(word=c("t.co", "1        o:it", "brianna", "ta", "msnbc", "tony",
"cnns","trump","iqykldai5y", "top", "sumoh7", "ii", "iii", "i'm", "y'all'", "celestljade","like",
"https", "rt", "shooting","shoot", "school", "im", "hey", "cus" ), lexicon="my_junk_twit")
```

```
my_junk_news <- data_frame(word=c("shooting","trump","shoot","na","gun","john", "guns",
"school", "shootings"), lexicon= "my_junk_news")

#################
#Tokenizizing both news and twitter data
#################

news_tokens <- ndata %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_news) %>%
  count(word, sort=T) %>%
  ungroup()

view(news_tokens)

ss_tokens <- ss %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_twit) %>%
  count(word, sort=T) %>%
  ungroup()
```

```
view(ss_tokens)


############
#inspect frequencies
###########


#frequencies news


frequ_tokens_news <- ndata %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_news) %>%
  count(word, sort=TRUE)
print(frequ_tokens_news)


frequ_tokens_news %>%
  filter(n > 50) %>%
  mutate(word = reorder(word, n, color = "blue")) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip()
```

```
#frequencies twitter


frequ_tokens_twit <- ss %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_twit) %>%
  count(word, sort=TRUE)
print(frequ_tokens_twit)


frequ_tokens_twit %>%
  filter(n > 70) %>%
  mutate(word = reorder(word, n, color = "blue")) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip()


######################
#Sentiment Analysis
######################


get_sentiments("bing")
get_sentiments("afinn")
```

```
get_sentiments("nrc")


#Sentiments - bing
news_senti <- ndata %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_news) %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

news_senti

news_senti %>%
  group_by(sentiment) %>%
  top_n(20) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="Contribution to sentiment", x=NULL)+
  coord_flip()
```

```
twit_senti <- ss %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_twit) %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

twit_senti

twit_senti %>%
  group_by(sentiment) %>%
  top_n(20) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="Contribution to sentiment", x=NULL)+
  coord_flip()

#Sentiment afinn
```

```
news_senti_num <- ndata %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_news) %>%
  inner_join(get_sentiments("afinn")) %>%
  count(word, value, sort=T) %>%
  summarize(mean(value)) %>%
  ungroup()

print(news_senti_num)

twit_senti_num <- ss %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_twit) %>%
  inner_join(get_sentiments("afinn")) %>%
  count(word, value, sort=T) %>%
  summarize(mean(value)) %>%
  ungroup()

print(twit_senti_num)
```

```
#NRC and Comparison cloud


news_senti_nrc <- ndata %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(my_junk_news)%>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()


#Pizza
news_senti_nrc %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(colors = c("grey20", "gray80"),
                   title.colors=c("red","blue"),
                   max.words=100, fixed.asp=TRUE,
                   scale=c(0.8,0.8), title.size=2, rot.per=0.25)



twit_senti_nrc <- ss %>%
  unnest_tokens(word, text) %>%
```

```
  anti_join(stop_words) %>%
  anti_join(my_junk_twit)%>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

twit_senti_nrc %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(colors = c("grey20", "gray80"),
                   title.colors=c("red","blue"),
                   max.words=100, fixed.asp=TRUE,
                   scale=c(0.8,0.8), title.size=2, rot.per=0.25)

bing_and_nrc <- bind_rows(
  news_tokens %>%
    inner_join(get_sentiments("bing")) %>%
    mutate(method = "Bing et al."),
  news_tokens %>%
    inner_join(get_sentiments("nrc") %>%
                 filter(sentiment %in% c("positive", mutate(method = "NRC")) %>%
"negative"))) %>%
```

```
  count(method, index = linenumber %/% 80, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)


#################
#Ngrams
#################
news_ngrams <- ndata %>%
  unnest_tokens(bigram, text, token = "ngrams", n=3) %>%
  separate(bigram, c("word1", "word2", "word3"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!word3 %in% stop_words$word) %>%
  filter(!word1 %in% my_junk_news$word) %>%
  filter(!word2 %in% my_junk_news$word) %>%
  filter(!word3 %in% my_junk_news$word) %>%
  count(word1, word2, word3, sort = TRUE)

ggraph(slice(news_ngrams, 1:300), layout = "fr") +
  geom_edge_link()+
  geom_node_point()+
  geom_node_text(aes(label=name), vjust =0.0001, hjust=0.0001)
```

```
twit_ngram <- ss %>%
  unnest_tokens(bigram, text, token = "ngrams", n=3) %>%
  separate(bigram, c("word1", "word2", "word3"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!word3 %in% stop_words$word) %>%
  filter(!word1 %in% my_junk_twit$word) %>%
  filter(!word2 %in% my_junk_twit$word) %>%
  filter(!word3 %in% my_junk_twit$word) %>%
  count(word1, word2, word3, sort = TRUE)

ggraph(slice(twit_ngram, 1:300), layout = "fr") +
  geom_edge_link()+
  geom_node_point()+
  geom_node_text(aes(label=name), vjust =0.0001, hjust=0.0001)

#######
#Additional ngrams for testing purposes - not used in analysis
########
twit_ngrams
twit_senti_ngram <- twit_ngram %>%
  inner_join(get_sentiments('afinn'), by=c(word2="word")) %>%
```

```
  count(word1, word2, value, sort=TRUE) %>%
  summarise(mean(value)) %>%
  ungroup()

twit_senti_ngram

set.seed(2017)
ggraph(twit_ngram, layout = "fr") +
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1)



set.seed(2016)
a <- grid::arrow(type = "closed", length = unit(.15, "inches"))
ggraph(twit_ngram, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1) +
  theme_void()

###############
```

```
#Trigramm barplots - not used because of NA values
###############
ndata %>%
  unnest_tokens(word, text, token = "ngrams", n = 3) %>%
  separate(word, c("word1", "word2", "word3"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!word3 %in% stop_words$word) %>%
  filter(!word1 %in% my_junk_news$word) %>%
  filter(!word2 %in% my_junk_news$word) %>%
  filter(!word3 %in% my_junk_news$word)%>%
  unite(word,word1, word2, word3, sep = " ") %>%
  count(word, sort = TRUE) %>%
  slice(1:10) %>%
  mutate(word = reorder(word,n ))%>%
  ggplot() + geom_bar(aes(word, n), stat = "identity", fill = "#3344de") +
  theme_minimal() +
  coord_flip() +
  labs(title = "10 Most Used Trigrams affiliated with School Shootings",
       subtitle = "Based on News & Articles",
       caption = "")
```

```
ss %>%
  unnest_tokens(word, text, token = "ngrams", n = 3) %>%
  separate(word, c("word1", "word2", "word3"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!word3 %in% stop_words$word) %>%
  filter(!word1 %in% my_junk_twit$word) %>%
  filter(!word2 %in% my_junk_twit$word) %>%
  filter(!word3 %in% my_junk_twit$word)%>%
  unite(word,word1, word2, word3, sep = " ") %>%
  count(word, sort = TRUE) %>%
  slice(1:10) %>%
  mutate(word = reorder(word,n ))%>%
  ggplot() + geom_bar(aes(word, n), stat = "identity", fill = "#3344de") +
  theme_minimal() +
  coord_flip() +
  labs(title = "10 Most Used Trigrams affiliated with School Shootings",
       subtitle = "Based on Twitter",
       caption = "")
```

```
################

#Wordclouds - not used because of redundance with prior sentiments

################


news_tokens %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(colors = c("grey20", "gray80"))


ss_tokens %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(colors = c("grey20", "gray80"))
```

```
############################
#Correlogram - not used because of a lack of words
############################


frequ <- bind_rows(mutate(news_tokens, author="news"),
                   mutate(ss_tokens, author="twitter")) %>%
  mutate(word=str_extract(word, "[a-z']+")) %>%
  count(author, word) %>%
  group_by(author) %>%
  mutate(proportion = n/sum(n))%>%
  select(-n) %>%
  spread(author, proportion) %>%
  gather(author, proportion,`twitter`)%>%
  filter(proportion, 0.0001)



ggplot(frequ, aes(x=proportion, y=`news`,
                  color = abs(`news`- proportion)))+
  geom_abline(color="grey40", lty=2)+
  geom_jitter(alpha=.1, size=0.6, width=0.4, height=0.4)+
  geom_text(aes(label=word), check_overlap = TRUE, vjust=.4) +
  scale_x_log10(labels = percent_format())+
  scale_y_log10(labels= percent_format())+
```

```
  scale_color_gradient(limits = c(0,0.001), low = "darkslategray4", high = "gray75")+
  facet_wrap(~author, ncol=3)+
  theme(legend.position = "none")+
  labs(y= "news", x=NULL)
```