INFX 547 Spring 2018

NRA Convention 2018

Sentiment Analysis

Barbara Williams, Naishan Xiang and Shreya Sabharwal

Introduction

In the wake of the American Civil War, the National Rifle Association (NRA) was founded, according to their website, to "promote and encourage rifle shooting on a scientific basis (National Rifle Association, 2018, para. 1)." The association states that the formation of the organization was due to the dismay of Union officers, who saw the lack of marksmanship in Union soldiers as something to be remedied. Civil War general Ambrose Burnside, former governor of Rhode Island, would become the first president of the association (Coleman, 2016). Today the NRA is most summed up to contemporary audiences by actor Charlton Heston's famous quote, where he claimed that anyone who took his gun away would do so "from my cold, dead hands (Eckstein, 2017)."

This correlation of freedom and gun ownership would permeate American culture for years, especially in reaction to tragic events. In December 14th, 2012, would be a turning point for both sides of the gun debate. The shooting at Sandy Hook Elementary School created a highly negative association to the organization in the American public's consciousness (Wozniak, 2017). However, the NRA would survive and even win the debate. By framing the debate not around gun ownership, but the attacker himself, the NRA created the idea of "the only thing that stops a bad guy with a gun is a good guy with a gun (Eckstein, 2017, para. 4)."

Earlier this February, another tragic shooting would awaken yet again public consciousness and ignite the once "dead" gun control debate. The shooting in Marjory Stoneman Douglas High School in Parkland, Florida would split the public's mindset on gun control. The NRA, along with President Trump, indicated that change was needed for the safety of children. The president and the NRA, however, utilized its strong skills in debate, suggesting that a solution would be to arm teachers, rather than impose any legislation (Scott, 2018).

Over the weekend of May 4th- 6th, 2018, President Trump spoke at the annual NRA convention in Dallas. This lead to the very recent **#NRAConvention** trending on Twitter (Dahlen, 2018). A quick scan of the tweets posted show a mixed reaction to the gathering in Dallas. Another high school shooting occurred in Texas, with 10 deaths on May 18th, which made the NRA and gun control debates trend once again on social media. According to the FOX news, the incoming NRA chief blamed school shootings on "culture of violence" and Ritalin drugs. After the accident, President Trump did send his condolences, but no actual plan or legal policy promised. It is nature

to see how public reacted on Social media about the NRA convention and a series of school shooting events. Our original hypothesis is that there is a negative emotional tone toward NRA on Twitter. However, President Trump's speech to the NRA last year is very positive (Bruke, 2017) and Mr. Trump is the first President since Ronald Reagan to speak on the NRA annual meeting. In addition, over five million NRA members did approve the positivity to some extent. So our research aims to gain an understanding of the public attitudes, opinions and emotions on this NRA convention 2018 through a series of words published on twitter. Our goal is to provide insight and a sentiment analysis to this trending statement.

Research Questions

According to the motivation, we have drilled down to two research questions:

1. Is the overall sentiment of the NRA positive or negative after a series of gun violence in the year of 2018?
2. Did the NRA Annual Convention 2018 raise a positive or negative sentiment towards President Donald Trump?

Data Collection

Plot Study

We did a pilot data extraction at first. We created a python script using Tweepy library and limited the number of tweets to be extracted to 10,000 at a time. For a maximum of 10,000 tweets, we observed that we got data only for 2 days (5/17/2018 and 5/18/2018). That gave us an idea about the number of tweets being generated with hashtag '#NRA' per day.

So, we continued with the same data collection methodology to collect more tweets from 5/17 to 5/23 with more trending hashtags like #guncontrol and #gunviolence. Due to the May 18th tragedy, we also collect the tweets specifically from 5/18 to 5/20 with at least two hashtags #NRA and #ifidieinaschoolshooting (which became viral in 2 days) being mentioned. By the end of 23rd May, we collected an overwhelming data of approximately 400,000 records. We removed duplicate records and identified 87,468 unique records.

The data was stored in JavaScript Object Notation (JSON) format. The JSON includes the metadata of the tweeter along with the tweet text. Below are the main attributes of the JSON object that we are interested in (Twitter Developer, 2018):

- *Created_at*: Timestamp when it was posted
- *Text*: Main content of the tweet
- *User*: metadata associated with the tweeter (only screen_name)
- *Urls*: Any urls used in the tweet
- *User_mentions*: any mentions in the tweet
- *Entities*:*Hashtags:* any hashtags in the tweet

Data Cleaning and Transformation

Since the data is unstructured, we cleaned and transformed the data into a more stable structured one before analysis. We used nltk library in Python by parsing the json and converting it to a dataframe. Below is the list of tasks we have completed in terms of data cleaning and transformation (Gull, Shoaib, Rasheed, Abid, Zahoor, 2016):.

- *Converting text to lowercase* – In order to make all the data consistent, we need to convert the tweet text into lowercase.
- *Removing emoticons and punctuations* – The emoticons might appear as garbage values for Python, hence we will get rid of all emoticons.
- *Removing URLs* – Since the URLs also so not provide any information about the sentiments of Twitter users, we have removed them as well.
- *Removing stop words* – Stop words such as ‘a’ and ‘the’ do not help us in our analysis, hence, we removed these words.
- *Performing Stemming/Lemmatization* – The words like ‘Sooooo’ and ‘happppppy’ need to be standardized and we have to stem or lemmatize each of the words so that all the words are converted to a common base form. This would helped us in determining the frequency of occurrence of each of the words used in the tweets.

The above transformation helped us identify unique words and hashtags along with their counts used in the tweets.

We converted timestamp of tweets to datetime format in order to perform day wise analysis. WE removed unused columns and extracted hashtags from entities attribute.

Another thing we have observed in our collection is that some entities associated with the NRA (such as NRAtv) will always portray the organization as positive and the number of tweets made by NRAtv are substantial compared to other people, hence, we are removing these since they skew the data.

We also filtered tweets having hashtags associated with Trump to perform a sentiment analysis of those tweets separately in order to see what people's perception is about him in context of his speech at the NRA Convention.

Data Analysis

Sentiment Analysis

Sentiment Analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. This helped us understand the overall perception of tweeters towards NRA Convention 2018 and Donald Trump's speech. In light of the Santa Fe High School shooting on May 18th, 2018, we also collected data from hashtags such as #NRA, #guncontrol, #antigun, #ifidieinaschoolshooting, and #gunviolence. Each of these hashtags were considered whether the users reacted positively, negatively, or neutral to the hashtags. We were required to use a lexical approach as the output label wasn't known to us.

Hence, we imported VADER (Valence Aware Dictionary and sEntiment Reasoner) to perform sentiment analysis. One issue with sentiment analysis is to determine the difference between sarcasm and true positive sentiment. VADER has the capability to understand these. It utilizes qualitative analysis along with empirical validation using human raters from Amazon Mechanical Turk. It is sensitive to both polarity and intensity(strength) of emotion. It uses a dictionary to map lexical features to emotional intensities called sentiment scores. Additionally, colloquialisms get mapped to intensity. VADER also "examines the tri-gram before a sentiment-laden lexical feature to catch polarity negation" (C., 2017, para. 23).

Topic Modeling

Topic Modeling is used to find hidden topical patterns present in the textual material. It identifies a group of words from data that best represents the information in verbose text. Of the various

techniques used for topic modeling, we used Latent Dirichlet Allocation (LDA). LDA identifies the different topics the textual material contains, the words that belong to each topic and how much each of the topic is present in the text. It uses a Collapsed Gibbs sampling approach. It takes in as an input a bag of words where each document is represented as a row, with each column containing the count of words in the corpus. This is then reduced from a document to topic matrix and a topic to word matrix. These two matrices are then multiplied together to reproduce the bag of words matrix with the lowest error. In sum, LDA represents documents as mixtures of topics that spit out words with certain probabilities.

Results

Visual Exploratory Data Analysis

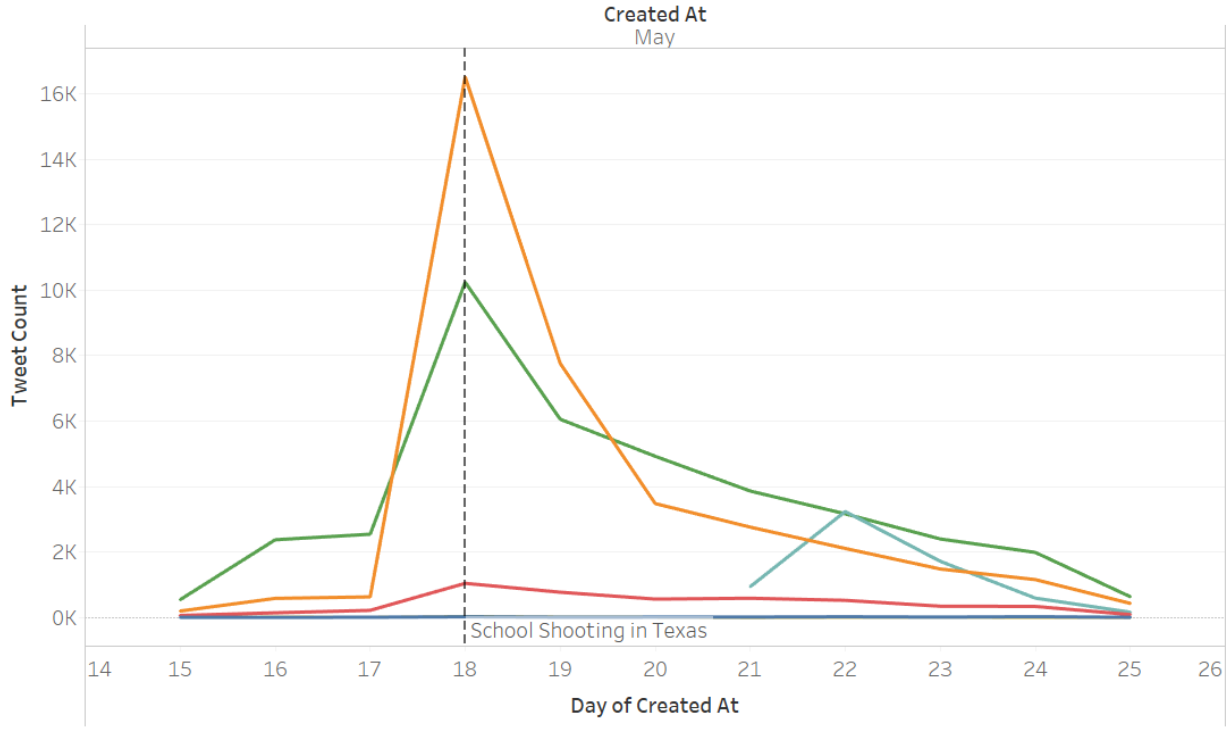
Below indicates our insightful findings from the Twitter dataset and the data visualization at Tableau.

https://public.tableau.com/profile/shreya.sabharwal#!/vizhome/NRAConventionSentimentAnalysis/Story1

How has the issue escalated after the school shooting?

Because of the unexpected circumstance of the Santa Fe High School shooting, NRA mentions were spiked momentarily in frequency. Mentions of the keyword NRA occured over 10K on May 18th, 2018, but by May 19th, had gone down to 6K mentions. The keyword 'guncontrol' itself in regard to the NRA on May 18th spiked to about 16k mentions. The keyword 'NRAConvention' has little to no mentions over this period.

Tweet Counts by Date



What are most people concerned about regarding the NRA?

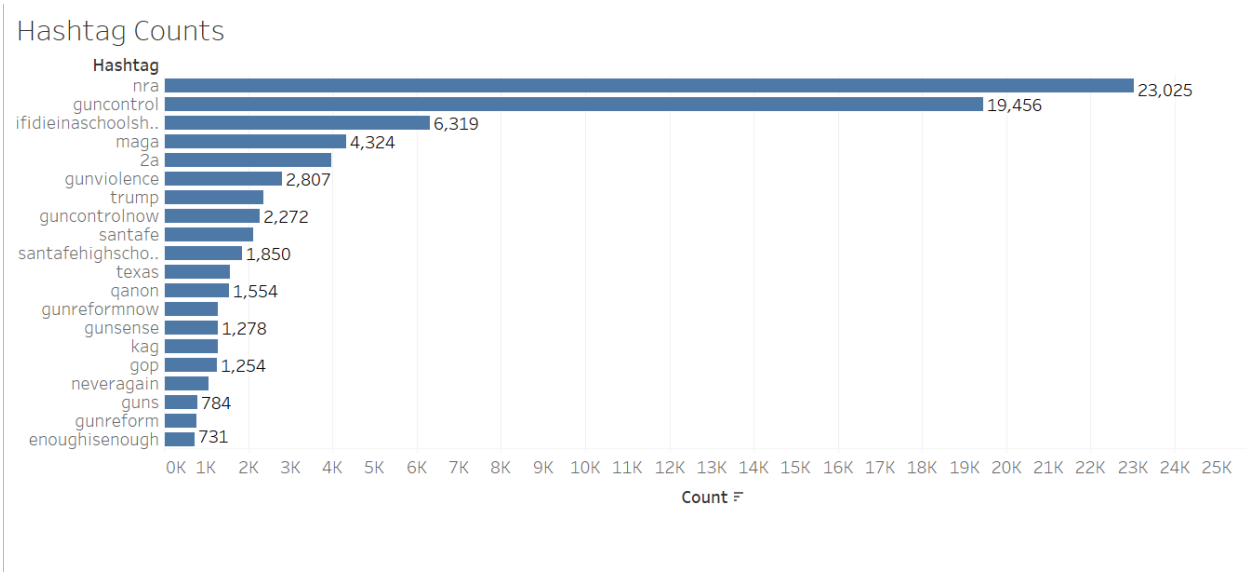
When it comes to what users are concerned with, in regards to the NRA, the number one keyword was 'gun.' This is not unusual, since the NRA is the National Rifle Association. However, the mentions of 'school' and 'shooting' were quite high, with the keyword 'shooting' having 7,407

Word Count

mentions and ‘school’ 9,733 mentions. With the results of the sentiment analysis, it is clear that many people associate school shootings and the NRA.

What is trending in the context of the NRA?

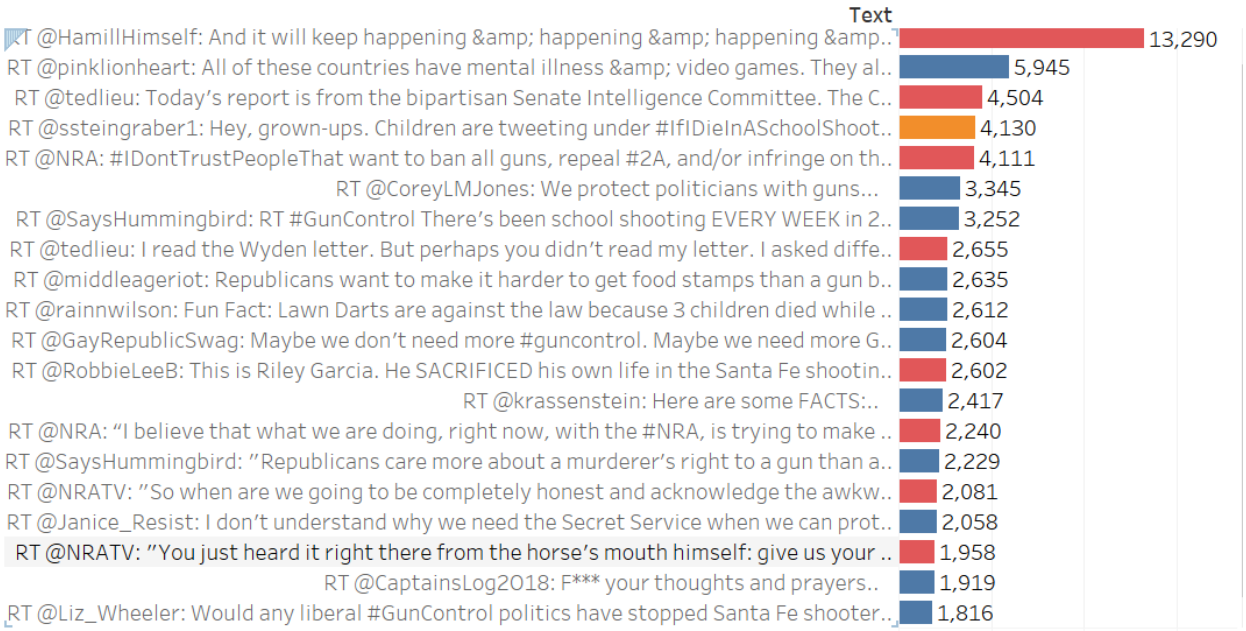
With a little more investigation into trending topics associated with the NRA, many of the topics included ‘gun control’ or ‘if I die in a school shooting’. Both have a negative sentiment in regards to the organization’s stance.



What are people retweeting about the most?

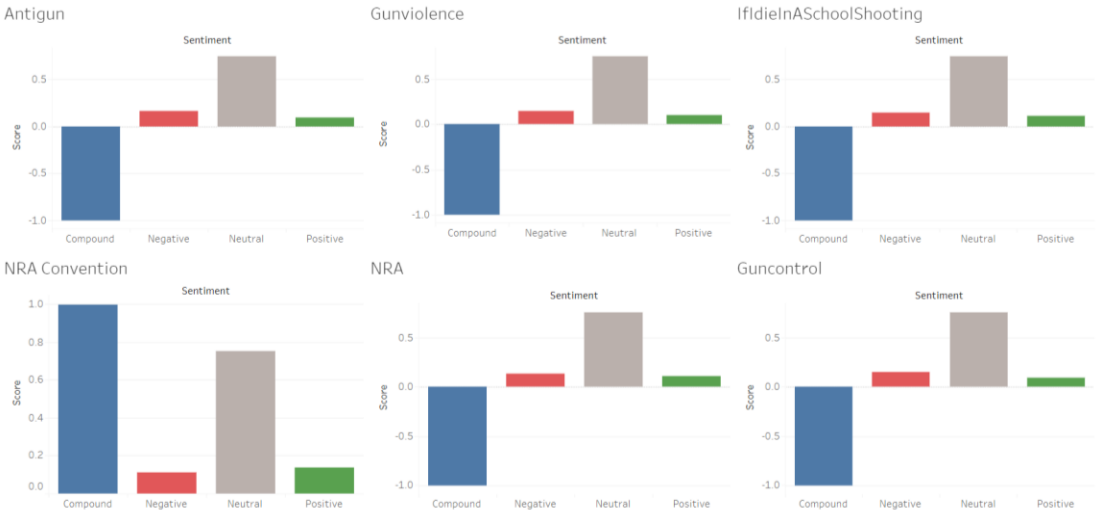
The number one retweet (over 13k) over this period (May 15-25th) was by actor Mark Hamill, who referenced the organization, by saying “And this will keep happening & happening & happening…”

Retweets by date and keyword

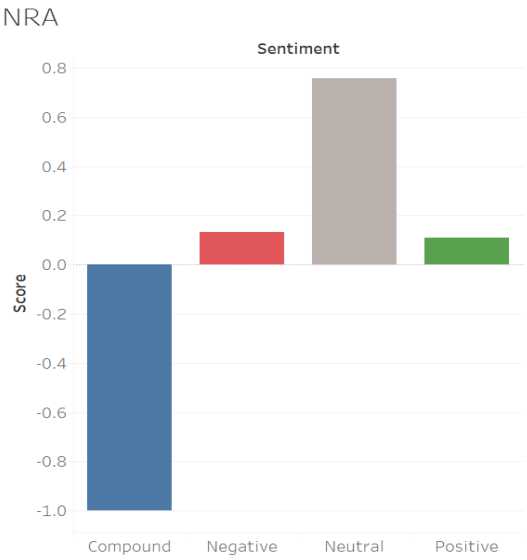


What are the sentiments of twitter users regarding the NRA and the NRA convention?

When looking at the overall polarity of several popular tweets between May 15-25th, all trending topics have an extremely negative sentiment (-1). The only exception to this would be the NRA convention, which had an overwhelmingly positive sentiment (+1).

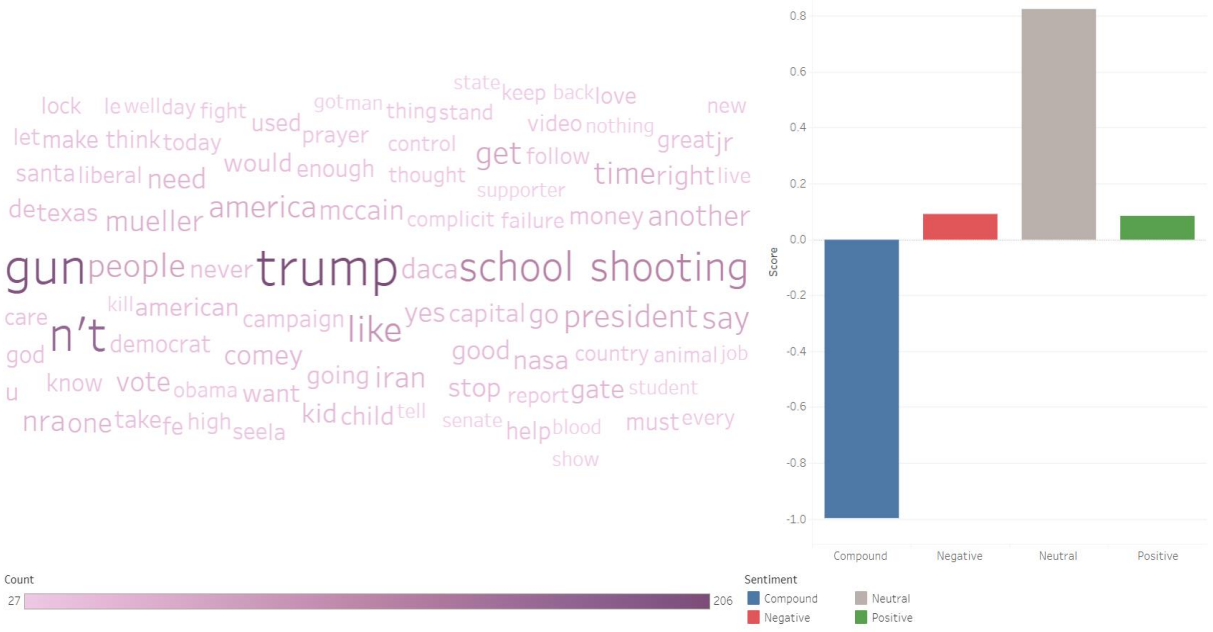


The object in which the sentiment or opinion is majorly expressed in the instance is the NRA. When looking into this question after our given data results, a neutral sentiment was considered the highest between positive or negative. The overall polarity, however, indicated for #NRA as extremely negative, with a -1.0 score.



What are people talking about Donald Trump with respect to the NRA?

President Trump was the first president to speak at the NRA convention since President Reagan. The goal of our analysis was to understand the sentiment of users on Twitter in regard to this event. President Trump is considered the object of this opinion analysis. Overall, the tweets of President Trump over May 15th- 25th period indicate a negative sentiment, with associations in tweets with the words ‘gun’ and ‘school shooting’.



What are common topics people are talking about?

From Topic Modeling using Latent Dirichlet allocation(LDA), we observed that the topmost topics/keywords for all the tweets include ‘school in shooting’, ‘2ndamendment’, ‘guncontrol’, ‘santafehighschool’ and ‘realdonaldtrump’.

Topic Modeling

Keyword	Topic
antigun	2adefenders is and antigun
	this nra antigun 2ndamendment
	to that so antigun
guncontrol	guncontrol guncontrolnow rt santafehighschool
	realdonaldtrump santafe guncontrol enough
	the school in shooting
gunviolence	gunviolence santafe guncontrolnow rt
	nra realdonaldtrump gunreformnow gop
IfIdieInASchoolShooting	ifidieinaschoolshooting my will be
	never to get will
	the is ifidieinaschoolshooting hashtag
nra	maga rt trump nra
	nra 2a gun new
	nra in the here
nraconvention	at the realdonaldtrump you
	nraconvention rt the to
	remember was do the

Limitations & Ethical Issues

In terms of data collection, there are a few limitations in our research. One is the use of Twitter REST API. It only allows free users to collect data for a week and one-week tweets may not represent the entire sentiment of all Twitter users. Therefore, we cannot extend our results to a broader population. The period is which this took place was a direct result after a school shooting. Thus, is caused an interesting spike in results.

Also, we were unable to use N-grams of text to create combinations of adjacent words and calculate the number of counts of the combinations to determine the sentiment. This was due to the vast data and hardware limitations which caused our systems to crash a number of times.

Lastly, the key hashtags we determined from Twitter. We believed that there are much more posts discussing about NRA convention without adding any hashtags mentioned above. We exclude these posts from our research and mainly focus on the one with the hashtags. It makes easier for reviewers to understand how we extracted data from Tweets and generated our conclusions.

As for possible ethical concerns, tweets we collected are not necessarily permanent. Previous work shows that many users believe their tweets to be inherently ephemeral (Proferesm, 2014) and complications therefore arise when we collect, store and analyze these data more permanently. The privacy of the user may be at risk, especially considering long term storage. In the reading, *Tweets Are Forever: A Large-Scale Quantitative Analysis of Deleted Tweets*, the research indicates that though the user has deleted the tweet, it still exists in Twitter. Some of the tweets we collected may be eventually deleted by the user, though we will still have access to this tweet (Almuhimedi, Wilson, Liu, Sadeh, & Acquisti, 2013). To us, username is insignificant, therefore we have no need to keep this long term to negate this problem. Moreover, we haven't utilized any other metadata of the users. Any research requires to take informed consent from users to utilize their data. We did not do so as Twitter takes informed consent from the users before they even create their social media accounts on the platform.

Conclusions

In conclusion, we find that the overall sentiment of tweets mentioning the NRA and President was overwhelmingly negative in association. It is important to note that the NRA convention, by the time of analysis, had little mention in tweets, though the convention took place in the same state as the Santa Fe High school shooting two weeks apart. Many trending hashtags associated with the NRA over this period are mostly about 'gun violence' or 'anti-gun' association. This strong sentiment of associating gun violence and the NRA are a good indication of the reasoning behind the negative sentiment.

The NRA convention itself had a positive sentiment analysis, which we figure was from actually attendees themselves. Because many who have a positive sentiment towards the organization would attend this convention, few tweets would be negative. However, Donald Trump's association with the organization has not bode well for twitter users' overall opinions. The president's attendance has brought about associations school shootings.

Though over this week period, less mentions occurred, and thus less associations with these negative sentiments. Time is the winner in regard to the overall sentiment of these organizations.

References

Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N., & Acquisti, A. (2013). Tweets Are Forever: A Large-scale Quantitative Analysis of Deleted Tweet. In Proceedings of the 2013 Conference on Computer Supported Cooperative Work (pp. 897–908). New York, NY, USA: ACM. doi:10.1145/2441776.2441878

Burke, R. (2017, April, 29). NRA Speech Text and Sentiment Analysis. Retrieved May 25, 2018, from https://rstudio-pubs-static.s3.amazonaws.com/271941_a7985e3303b440589f3ef48c8b571a78.html

C., G. (2017, April 10). VADER Sentiment Analysis Explained. In *Data Meets Media*. Retrieved from http://datameetsmedia.com/vader-sentiment-analysis-explained/

Coleman, A. L. (2016, July 29). When the NRA Supported Gun Control. In *Time*. Retrieved May 5, 2018, from http://time.com/4431356/nra-gun-control-history/

Dahlen, D., & O'Connor, L. (2018, May 6). It Was Protests, Peddling And Presidents At This Year's NRA Convention. In *HuffPost*. Retrieved May 6, 2018, from https://www.huffingtonpost.com/entry/nra-annual-meeting-2018-photos_us_5aef69d7e4b0c4f19323d9d6

Eckstein, J., & Partlow Lefevre, S. T. (2017, Mar/Apr). Since Sandy Hook: Strategic Maneuvering in the Gun Control Debate. Western Journal Of Communication, 81(2), 225-242. doi:10.1080/10570314.2016.1244703

Gull, R., Shoaib, U., Rasheed, S., Abid, W., & Zahoor, B. (2016). *Pre processing of twitter's data for opinion mining in political context* doi:https://doi.org/10.1016/j.procs.2016.08.203

National Rifle Association. (2018). About the NRA. In *NRA: National Rifle Association*. Retrieved May 5, 2018, from https://home.nra.org/about-the-nra/

Proferes, N. (2014). What Happens to Tweets? Descriptions of Temporality in Twitter's Organizational Rhetoric. In iConference 2014 Proceedings (p. 76 - 87)

Scott, E. (2018, February 23). A big question in the debate about arming teachers: What about racial bias?. In *Washington Post*. Retrieved May 6, 2018, from https://www.washingtonpost.com/news/the-fix/wp/2018/02/23/a-big-question-in-the-debate-about-arming-teachers-what-about-racial-bias/?utm_term=.7a328daff177

Twitter Developer. (2018). Tweet Objects. In *Twitter Developer*. Retrieved May 6, 2018, from https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json.html

Wozniak, K. H. (2017). Public opinion about gun control Post–Sandy hook. *Criminal Justice Policy Review*, *28*(3), 255-278. doi:10.1177/0887403415577192

Appendix

Data Collection

```
from nltk import download
from html.parser import HTMLParser
import re
from nltk.corpus import stopwords
import string
from nltk.stem.wordnet import WordNetLemmatizer
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer
from itertools import groupby
import pandas as pd
import json
import numpy as np
import ast
from collections import Counter
import glob
import os
import warnings
from sklearn.feature_extraction.text import CountVectorizer
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
searchQuery = '#IfIDieInASchoolShooting'
maxTweets = 10,000
tweetsPerQuery = 100

# You can set intervals by specifying sinceID and max_id here.
sinceId = None
max_id = -1

tweet_count = 0
print("Downloading at least {} tweets".format(maxTweets))
with open('Tweets_ss_24.json','w') as f:
    while tweet_count <= maxTweets:
        try:
            if(max_id <= 0):
                if(not sinceId):
                    new_tweets = api.search(q = searchQuery, count = tweetsPerQuery)
                else:
                    new_tweets = api.search(q = searchQuery, count = tweetsPerQuery, since_id = sinceId)
            else:
                if(not sinceId):
                    new_tweets = api.search(q = searchQuery, count = tweetsPerQuery, max_id = str(max_id-1))
                else:
                    new_tweets = api.search(q = searchQuery, count = tweetsPerQuery, max_id = str(max_id-1), since_id =
sinceId)
            if not new_tweets:
                print('no more tweets found')
                break
            for tweet in new_tweets:
                json.dump(tweet._json, f, sort_keys = True)
```

```
            f.write('\n')
        tweet_count += len(new_tweets)
        print(tweet_count)
        max_id = new_tweets[-1].id
    except tweepy.TweepError as e:
        print('some error : ' + str(e))
        break
print('Done, tweets data stored in RESTtweets.json file.')

data = pd.read_json('json/Tweets_ss_24.json',lines=True)

df = pd.DataFrame(data)
df['keyword'] = searchQuery[1:]
print(df.head())
df.to_csv('Tweets_ss_24.csv', index = False)
```

Merge CSV files

```
all_files = glob.glob(os.path.join('data/',"*.csv"))
df_tweets = pd.DataFrame()
for file in all_files:
    df_file = pd.read_csv(file)
    df_tweets = df_tweets.append(df_file, ignore_index=True)
df_tweets.drop_duplicates(inplace=True)
df_tweets.to_csv('Tweets_retweet_count.csv', index=False)

print(df_tweets.shape)
df_tweets.drop_duplicates('text',inplace=True)
print(df_tweets.shape)

df_tweets.to_csv('Combined_Tweets_1.csv', index=False)
```

Data Preparation and Cleaning

```
import warnings
import pandas as pd
import ast
warnings.filterwarnings('ignore')
df_tweets_data = pd.read_csv('Combined_Tweets_1.csv')

df_tweets = df_tweets_data[['created_at', 'entities', 'retweet_count', 'retweeted', 'text', 'user', 'keyword']]

df_tweets.reset_index(drop=True,inplace=True)
df_tweets.entities.fillna("{'hashtags':[]}", inplace=True)
df_tweets.entities = df_tweets.entities.apply(lambda x : dict(eval(x)) )
df_tweets['hashtags'] = '0'
for i in range(df_tweets.shape[0]):
```

```
    list_hashtags = df_tweets.entities.iloc[i]['hashtags']
    l_hashtags = []
    if(len(list_hashtags)>0):
        for j in range(len(list_hashtags)):
            l_hashtags.append(list_hashtags[j]['text'])
        df_tweets.at[i, 'hashtags']= str(l_hashtags)

df_tweets.user.isnull().sum()
df_tweets.user.fillna("{}", inplace=True)

df_tweets.user = df_tweets.user.apply(lambda x : dict(eval(x)) )

df_tweets['location'] = ''

for i in range(df_tweets.shape[0]):
    if(i in df_tweets.user.index):
        if(df_tweets.user[i].get('location', False)):
            df_tweets.at[i, 'location']= df_tweets.user[i]['location']

for i in range(df_tweets.shape[0]):
    if(i in df_tweets.user.index):
        if(df_tweets.user[i].get('screen_name', False)):
            df_tweets.at[i, 'screen_name']= df_tweets.user[i]['screen_name']

df_tweets.drop(['user'], axis=1,inplace=True)

df_tweets['created_at'] = [pd.to_datetime(x).date() for x in df_tweets.created_at]

filtered_tweets = df_tweets[df_tweets.screen_name != 'NRATV']

filtered_tweets.to_csv('Cleaned_Tweets.csv', index=False)

trump_tweets_df = pd.DataFrame(columns = filtered_tweets.columns);
for i in range(filtered_tweets.shape[0]):
    if('trump' in filtered_tweets.iloc[i].hashtags or 'Trump' in filtered_tweets.iloc[i].hashtags or 'TRUMP' in
filtered_tweets.iloc[i].hashtags):
        trump_tweets_df=trump_tweets_df.append(filtered_tweets.iloc[i])

trump_tweets_df.to_csv('Trump_Tweets.csv', index=False)

df_guncontrol = filtered_tweets[filtered_tweets.keyword == 'guncontrol']
trump_tweets_df.to_csv('Trump_Tweets.csv', index=False)
```

Word Counts

```
def extract_words(str_of_words):
    """
    return the list of words in the string
```

```
    """
    new_word_list = []
    word_list = word_tokenize(str_of_words)
    word_list = [word.lower() for word in word_list if word[0].isalpha() and len(word)>1]

    return word_list

def lemmatize_words(list_words):
    """
    count number of easy words from the list of words
    """
    lemmatized_words = []
    lemmatizer = WordNetLemmatizer()
    for word in list_words:
        word_n = lemmatizer.lemmatize(word, 'n')
        word_v = lemmatizer.lemmatize(word, 'v')
        lemmatized_words.append(word_n)
        if (word != word_n and word_n not in word_v and word_v not in word_n):
            lemmatized_words.append(word_v)
    return lemmatized_words

def perform_stemming(list_words):
    ps = PorterStemmer()
    word_list = [ps.stem(word) for word in list_words]
    return word_list
# standardize words
def standardize_words(word_list):
    standarized_words = []
    for word in word_list:
        word_l= [x[0] for x in groupby(word)]
        word_l = (").join(word_l)
        standarized_words.append(word_l)
    return standarized_words

RE_EMOJI = re.compile('[\U00010000-\U0010ffff]', flags=re.UNICODE)
# remove emojis
def strip_emoji(text):
    return RE_EMOJI.sub(r'', text)

#print(strip_emoji('baba black sheep 😍🤧💝'))

df_tweets = pd.read_csv('Cleaned_Tweets.csv')

df_keyword_counts = pd.DataFrame(pd.value_counts(df_tweets.keyword)).reset_index(level=0)
df_keyword_counts.columns = ['keyword', 'count']
df_keyword_counts.to_csv('keywords_counts.csv')

df_tweet_totals = df_tweets.groupby(['created_at',
```

```
    'keyword'])['text'].count().reset_index().sort_values(by = ['text'],
                                                          ascending = False)

df_tweet_totals.to_csv('Tweet_Totals.csv')

text_data = df_tweets['text']
print('before, number of nan:', text_data.isnull().sum())
df_tweets.text.fillna('', inplace=True)
print('after, number of nan:',text_data.isnull().sum())

list_text = list(df_tweets.text)

def preprocessing(list_text):
    word_list = []
    i=0
    for tweet in list_text:
        # remove urls
        tweet = re.sub(r"http\S+", "", tweet)

        df_tweets.at[i,'text'] =  tweet
        i=i+1
        #remove hashtags
        tweet = re.sub(r"#\S+", '', tweet)
        #remove mentions
        tweet = re.sub(r"@\S+", '', tweet)
        #tweet = " ".join(re.findall('[A-Z][^A-Z]*', tweet))
        tweet = re.sub('/', '', tweet)
        #remove emojis
        tweet = strip_emoji(tweet)
        # convert to lowercase and list of words
        word_list.extend(extract_words(tweet))

    #remove stop words
    word_list = remove_stop_words(word_list)

    #standardize words
    #std_words = standardize_words(word_list)
    # stemming
    lemmatized_words = lemmatize_words(word_list)
    return lemmatized_words

lemmatized_words = preprocessing(list_text)

w_counts = Counter(lemmatized_words)
df = pd.DataFrame(w_counts.most_common(100), columns=['Word', 'Count'])
df.to_csv('word_counts.csv')
```

Hashtag Counts

```
l_hashtags = []

for entity in df_tweets.entities:
    d = ast.literal_eval(entity)
    if d.get('hashtags', False):
        list_hashtags = [dict_ht['text'].lower() for dict_ht in d['hashtags']]
        l_hashtags.extend(list_hashtags)

# most common hashtags
counts = Counter(l_hashtags)

df = pd.DataFrame(counts.most_common(20), columns=['Hashtag', 'Count'])
df.to_csv('hashtag_counts.csv')

hashtag_df = pd.DataFrame.from_dict(list(dict(counts).items()))
hashtag_df.columns = ['keyword', 'count']
sorted_hashtag_df = hashtag_df.sort_values(by='count', ascending=False)
```

Retweets

```
df = pd.read_csv('Tweets_retweet_count.csv')

df.text.isnull().sum()
df.text.fillna('', inplace=True)
df_rt = df[df.text.str.contains('^RT')]
df_rt.head()

rt_count = df_rt[['text', 'keyword']].groupby(['text', 'keyword']).size().reset_index()
rt_count.columns = ['text', 'keyword', 'count']
rt_count.sort_values(by = ['count'], ascending = False, inplace = True)

rt_count.head(20).to_csv('retweets.csv')

rt_count_1 = df_rt[['keyword']].groupby(['keyword']).size().reset_index()
rt_count_1.columns = ['keyword', 'count']
rt_count_1.sort_values(by = ['count'], ascending = False, inplace = True)

rt_count_1.to_csv('retweets_by_keywords.csv')
```

Trump Tweets

```
df_trump_tweets = pd.read_csv('Trump_Tweets.csv')
text_data = df_trump_tweets['text']
text_data.fillna('', inplace=True)
```

```
list_text = list(text_data)

trump_words = preprocessing(list_text)

t_counts = Counter(trump_words)

df = pd.DataFrame(t_counts.most_common(100), columns=['T_Words', 'Count'])
df.to_csv('t_word_count.csv')
```

Polarity of Tweets

```
analyzer = SentimentIntensityAnalyzer()
def sentiment_scores(keyword):
    sentence = ' '.join(df_tweets[df_tweets.keyword == keyword].text.values)
    snt = analyzer.polarity_scores(sentence)
    return(snt)

sentiment_scores('gunviolence')
sentiment_scores('guncontrol')
sentiment_scores('nra')
sentiment_scores('nraconvention')
sentiment_scores('ifidieinaschoolshooting')
sentiment_scores('antigun')
```

Polarity of Trump Tweets

```
df_trump_tweets = pd.read_csv('Trump_Tweets.csv')
text_data = df_trump_tweets['text']
text_data.fillna('', inplace=True)

sentence = ' '.join(df_tweets.text.values)
snt = analyzer.polarity_scores(sentence)
print(snt)
```

LDA

```
def topic_dataframe(keyword,feature,words, model):
"""
Takes in a model, the features names, top words and the keyword
Returns a dataframe with topic for that keyword
"""
  list_of_topics = [] # Define an empty list to append results

   for ix, topic in enumerate(model.components_): # Index for components to iterate

       a = [feature[i] for i in topic.argsort()[:-word - 1:-1]]
```

```
        list_of_topics.append(topic_idx + 1)
        list_of_topics.append([" ".join(a)])

    # Define a dataframe to store results
    topic_dataframe = pd.DataFrame(topics, columns=["topic", "topic_numer"])
    topic_dataframe["keyword"] = keyword

    return topic_dataframe

def latent_dirichlect(keyword,topic,words, n, df_tweets):
    """
    Takes in a datframe, the keyword, topics, words and features - uses tf-idf
    Returns a dataframe with the LDA results.
    """
    # Intialize LDA model
    lda = LatentDirichletAllocation(n_components=topic)
    # Intialize tf-idf
    tfid = TfidfVectorizer(max_features = n)

    tfidf_text = tfid.fit_transform(df_tweets[df_tweets.keyword == keyword].text)
    # Define model
    text = lda.fit(tfidf_text)
    # Get features to display
    features = tfid.get_feature_names()
    # Call the method to make a dataframe
    data_frame = topic_dataframe(keyword, features, words, text)
    return data_frame

lda_nra_convention = latent_dirichlect ('nraconvention', 4, 4, df_tweets)
print(lda_nra_convention)

lda_gunviolence = latent_dirichlect ('gunviolence', 4, 4, df_tweets)
print(lda_gunviolence)

lda_antigun = latent_dirichlect ('antigun', 4, 4, df_tweets)
print(lda_antigun)

lda_school_shooting = latent_dirichlect ('IfIdieInASchoolShooting', 4, 4, df_tweets,)
print(lda_school_shooting)

lda_guncontrol = latent_dirichlect ('guncontrol', 4, 4, df_tweets)
print(lda_guncontrol)

lda_nra = latent_dirichlect ('nra', 4, 4, df_tweets,)
print(lda_nra)

lda_df = pd.concat([lda_nra_convention, lda_gunviolence, lda_antigun, lda_school_shooting, lda_guncontrol,
lda_nra])
lda_df.to_csv('lda_data.csv')
```