Sentiment Analysis of Gun Control Using Twitter Data

Brandon Vinhnee, **Abhishek Nisha Anish**, **Tej Patel**

UC San Diego Halıcıoğlu Data Science Institute

Abstract

For this project, the one large topic we want to address is gun control. In recent years in the United States, there have been an outburst of several horrific events as a result of guns getting in the hands of the wrong people. In 2023, the number of mass shootings in the US has already reached triple digits. With these incidents, we believe it would be an interesting study to do further research into the sentiment and beliefs that people have toward these issues and study this using automation and machine learning. Twitter is a very vocal platform and with the use of our new knowledge regarding sentiment analysis, there are possibly very many interesting discoveries to be made, such as how user sentiment toward one topic may differ from another. We hypothesize that because of the many gruesome events that have occurred in recent years, we will observe a mostly positive sentiment toward gun control, in that people support more regulation of weapon distribution rather than less. Our proposed project period of 10 weeks can be attributed to the fact that we are looking to improve upon our previous work by expanding our knowledge on the Astra Streaming features, and working on more efficient implementation of our architecture to ensure that we receive enough data for sufficient analysis. We will also look to go further by integrating new features into our project, including displaying results from our visualizations on a website in addition to streaming data to a feature store or database. As we work towards these goals, we recognize that we may come across technical issues that may require us to be more flexible and adjust our project structure, so we would like to stay open minded in regards to what additional features will be added.

1. Introduction

In this ever growing period of chaos where gun violence seems to be spiraling out of control, it is important to consider the opinions of people who live in cities and communities affected by this epidemic. There is most likely no solution that would make everyone happy, but through sentiment analysis, the general opinions can be deduced and even used to influence these impactful decisions. Twitter is a microblogging and social media platform that has amassed an enormous platform of around 206 million daily active users that account for approximately 500 million tweets per day. This popularity can be attributed to the fact that users feel enabled to talk freely about virtually anything they want, including their daily lives or topics that they are passionate about. With this kind of presence, it is only natural that the controversial topic of gun control became a huge subject matter due to recent tragic events. Through sentiment analysis and the chosen machine learning models, the collected Tweets will be classified as either positive, negative, or neutral and through a majority voting manner, the common view will be identified.

Previous research has worked towards this goal as well but several studies encountered shortcomings as a result of the methods that were utilized. In one such study done by Kaur [2], sentiment

analysis was performed using a deep learning algorithm that designed a hybrid heterogeneous support vector machine. The limitation of this model was that it was not as effective for Tweets in different languages. Chakraborty [3] used the LSTM model on two types of rated Tweets, but some issues that were encountered included that it failed to find the most common words for polarity analysis and it was not able to reach the preferred validation accuracy with the given data. In this paper, the data goes through extraction, preprocessing, and classification so as to avoid these limitations. By taking the Twitter data from both the present and the time of the pandemic, this paper will also provide conclusions about the opinions from these two different periods to examine any differences. It is important that data is collected from both of these times as it is hypothesized that the sentiment towards gun control would be different during a time when events such as mass shootings were down compared to when there was a large spike in them.

The sample that will be used comes directly from the Twitter API. A Twitter firehose will be connected to a Pulsar Topic in DataStax Astra. The topic is then transformed with one Pulsar Function and the output is written to a second Pulsar Topic. Data will first be transformed using a Python script and then a machine learning model will be applied to measure sentiment and produce real-time analytics. This study will focus on gun control sentiment in the United States, so Twitter data will be filtered accordingly. Another important filter will be the time frame as we want to focus on Tweets from the time of the pandemic to now. Filtering in such a way makes the data suitable for the goal of this paper as this helps us target a certain area of Tweets while keeping a large enough dataset. Additionally, another part of the data that will be important in our analysis would be the labels that we choose to give our data. We could use this data along with TF-IDF and/or Word Embedding methods to figure out whether the particular data analyzed belongs to a positive, negative or neutral sentiment label.

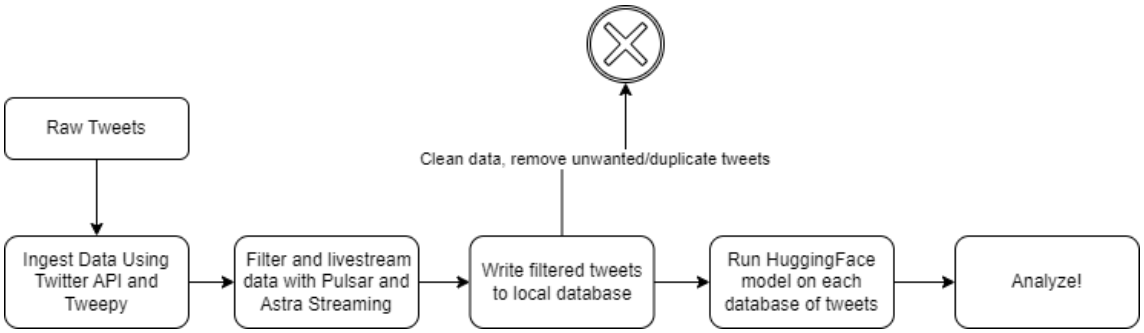

Figure 1. Pipeline Architecture

2. Methods

To live-stream our Twitter data, there were a wide variety of techniques to access Tweets in real time from the Twitter API. After thorough research of these techniques, we decided to use Tweepy, a well-documented python library with streaming and filtering capabilities that seemed like it would be a perfect fit in the scope of our project. The second crucial portion for developing the architecture of the project was setting up a streaming platform for the Tweets to be sent to for preprocessing and further analysis. For this portion, we chose to use Astra Streaming in conjunction with Pulsar, a pay-as-you go streaming service, which had three capabilities we needed to continue – the ability to handle input code to receive raw Tweet data as a consumer, pass this data through to a function to handle data preprocessing, and finally consume the final Tweets and apply a machine learning model to receive the sentiment of the message.

To go further in depth, our streaming architecture started with Tweepy, with the main streaming capabilities coming from Tweepy's StreamingClient Class. This class allowed for the raw tweets to be taken in from Twitter in real time. From here, we implemented the streaming class into our producer code, which produced the Tweets in real-time, and sent them to our Astra input topic. Instead of running our producer code directly from the producer file, we chose to implement a sort of "runner" file which created an instance of the producer class as well as the Tweepy streaming

class, added preliminary filters which got certain keywords we wanted within our queries, as well as removing Tweets with unnecessary media, tags, etc. This file was responsible for sending the tweets to our input topic. The next part of our streaming architecture is the function, which reads the tweets from input topic, and within our specific function, applied a regex to remove all links from the tweets we wanted to pull. From the function, the tweets would be passed to our output topic, ultimately to be read by our consumer. Within our consumer code, we feed in our processed Tweets to our pre-trained Natural Language Processing (NLP) model adapted from HuggingFace, which outputs a score and label based on the sentiment of the Tweet being negative, positive, or neutral. This model, called twitter-roberta-base-sentiment-latest by Cardiff NLP, has been trained on over 124 million tweets, and has been adjusted for sentiment analysis with the TweetEval benchmark. We decided to choose this model for our analysis as this model was specifically trained to analyze sentiments from tweets, which was the backbone of our analysis. With these sentiments, we chose to measure our output by counting the number of each sentiment in real time as Tweets are fed into our pipeline, and updated these counts continuously in a CSV. Once all our data was acquired for every keyword(s) query, we created visualizations such as pie charts to show the proportion of each sentiment per query and two histograms showing how the sentiment scores were distributed for two of our queries. By doing this, we were able to compare the results to our initial hypothesis and also understand the polarity of these sentiments, whether or not they were "extremely" leaning one way or just slightly.

However, a Twitter developer announced that on February 9th, 2023, support for free access to the Twitter API would no longer be available, limiting our options for continuation of the project. With this constraint in mind, a new task was decided upon, in that with the short timeframe left for free API access, we would collect as much data that we could in the form of tweets, and store them in databases based on the keywords we used to filter the livestream. The same streaming method as described above was used, however without passing in the tweets through the model, as this was a time consuming task that could be done later without the use of the Twitter API. Instead, the tweet data was simply written to the file, and the model was run on the database after all data was collected.

3. Results

We decided on six keywords and phrases closely related to the topic of gun control to use for filtering within our livestream. These six include, "gun", "gun control", "mass shooting", "school shooting", "second amendment", and "monterey park". We decided upon these as they are relevant to the topic, and wanted to experiment with the phrase "monterey park", as it was a recent event that we expected to have some conversation on twitter. After data filtering and cleaning, we were left with roughly 3000 tweets for "gun control", 5000 tweets for "gun", 500 for "mass shooting" and "school shooting", 1000 for "second amendment", and 200 for "monterey park".

With these keywords, we found that for "monterey park", 42.9 percent of users were seemingly pro gun control, 28.6 percent were seemingly against gun control, and 28.5 percent felt neutrally; for "gun", 8.1 percent of users were seemingly pro gun control, 24.9 percent were seemingly against gun control, and 67 percent felt neutrally; for "mass shooting", 90.9 percent of users were seemingly pro gun control, 4.4 percent were seemingly against gun control, and 4.7 percent felt neutrally; for "gun control", 11.8 percent of users were seemingly pro gun control, 31.2 percent were seemingly against gun control, and 57 percent felt neutrally; for "second amendment", 84.8 percent of users were seemingly pro gun control, 3.9 percent were seemingly against gun control, and 11.3 percent felt neutrally; and finally for "school shooting", 71.1 percent of users were seemingly pro gun control, 3.8 percent were seemingly against gun control, and 25.1 percent felt neutrally. With this data, we were met with some interesting results - overall, a strong "positive" sentiment toward gun control was observed, meaning many Twitter users talking about these topics are in favor of having more gun control. However, the percentage of negative sentiments in our research was not that much lower than the percentage of positives. Another interesting result was the large prevalence of neutral sentiments. A second statistic output from our model is the "sentiment score", or in other words a metric showing one's particular feeling toward their tweet. The score is measured between zero to one, with a small sentiment score (<0.5) would mean that the user's feeling wasn't

incredibly strong toward their tweet, and a large sentiment score (>0.5) would mean that the user had a strong feeling toward their tweet. For most of our keywords, we noticed most of our data had low sentiment scores, with most having scores around the 0.3 - 0.5 range.

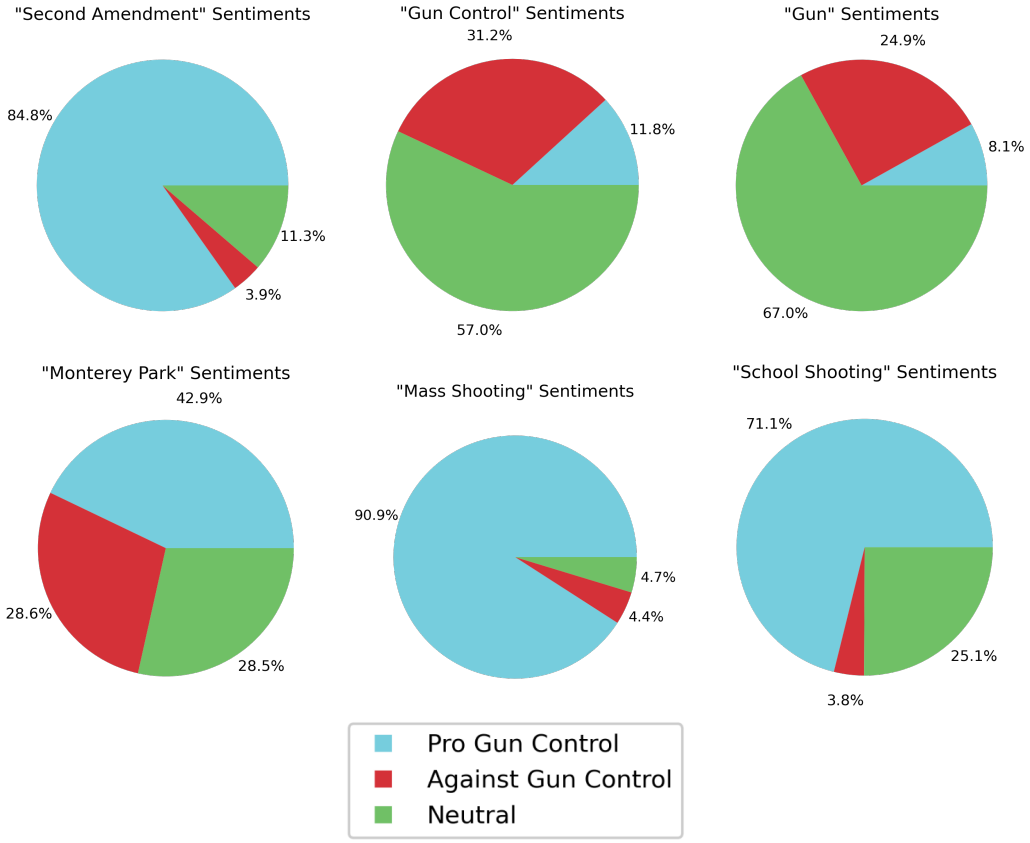

Figure 2. Sentiment Results

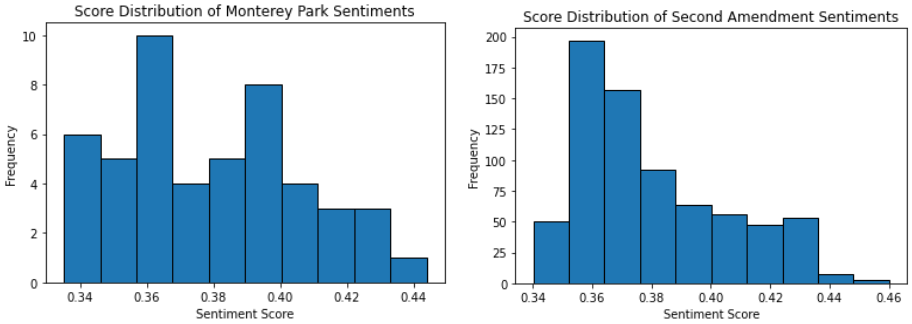

Figure 3. Sentiment Scores

4. Discussion

Seeing that the majority sentiment was “positive” toward gun control in most of our results, we can support and accept our hypothesis in saying that the population on Twitter would mostly be in support of gun control. For the keywords "second amendment", "school shooting" and "mass shooting", we can an overwhelming support for this, with pro-gun control sentiments consisting of over 70 percent of our collected data. However, we can also recognize that the sentiment scores received from our model do not show strong polarity. While sentiments for “gun” are relatively neutral, with an interesting 67 percent of tweets showing neutral feelings, sentiments toward more specific/emotion inducing keywords such as “mass shooting” and “gun control” show overwhelming support for more gun control within the Twitter community.

However, with these results, there is possible room for improvement. With the constraints from the Twitter API, a strong and lengthy database was not able to be achieved. In having a larger database, custom model training for the topic of pro/against gun control could have taken place, in addition to the simple increase in analysis trustworthiness as a result of having more data ingested into our results. For further research, we would have also liked to obtain more tweet data from different keywords, in order to obtain viewpoints from different topics as well.

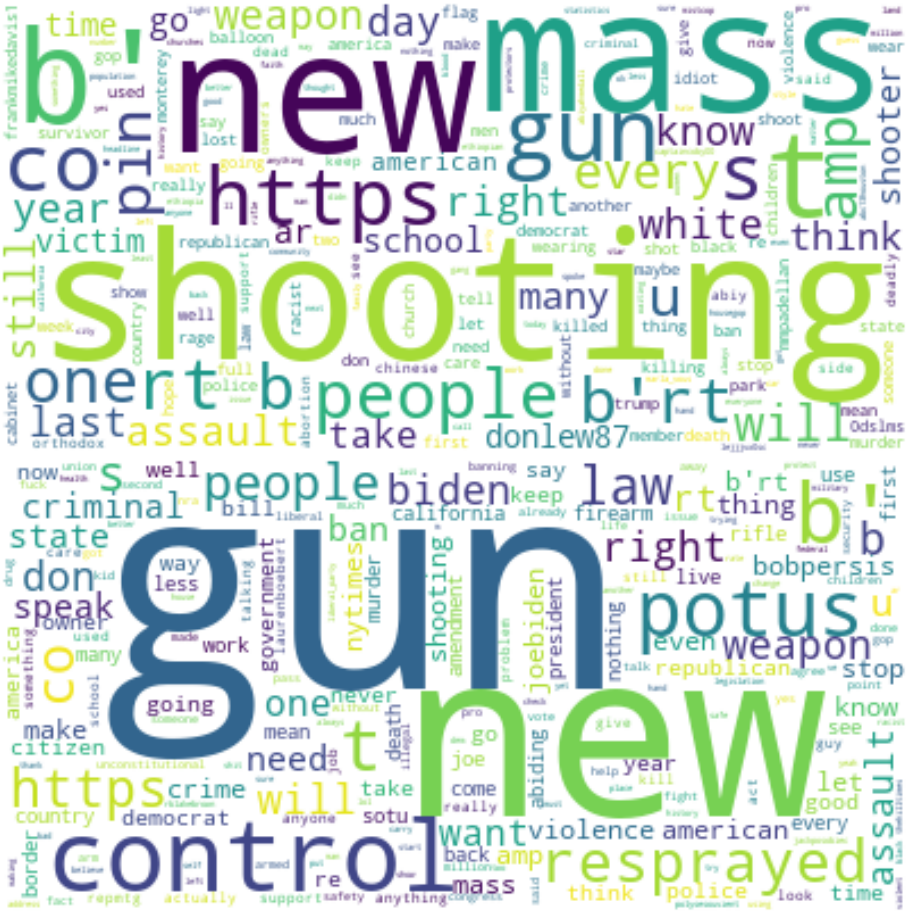

Figure 4. Word Clouds for "Mass Shooting" and "Gun Control"

References

[1] Cach Dang, N., N. Moreno-García, M., De la Prieta, F. (2021, August 12). Sentiment Analysis Based on Deep Learning: A Comparative Study. Retrieved October 28, 2022.

[2] H. Kaur, S. U. Ahsaan, B. Alankar, and V. Chang, "A Proposed Sentiment Analysis Deep Learning Algorithm for Analyzing COVID-19 Tweets", Information Systems Frontiers, pp. 1-13, 2022.

[3] A. K. Chakraborty, S. Das, and A. K. Kolya, "Sentiment Analysis of Covid-19 Tweets Using Evolutionary Classification-Based LSTM Model", In: Proc. of Research and Applications in Artificial Intelligence, Springer, Singapore, pp. 75-86, 2022.

[4] Martin Müller, Marcel Salathé, and Per E. Kummervold. "COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter." In: arXiv preprint arXiv:2005.07503 (2020).

[5] Hegde, Nagaratna, et al. Employee Sentiment Analysis Towards Remote Work during COVID-19 Using Twitter Data, 12 Aug. 2021.