Locked and Loaded: Analyzing Gun Culture Data and Strategies for Online Hate Speech Moderation

Kirthikraj Kamaraj
Computer Science
Binghamton University
Binghamton, New York, USA
kkamara1@binghamton.edu

Riddhi Patel
Computer Science
Binghamton University
Binghamton, New York, USA
rpate112@binghamton.edu

Sayoni Arup Nath
Computer Science
Binghamton University
Binghamton, New York, USA
snath2@binghamton.edu

Sathwik Krishtipati
Computer Science
Binghamton University
Binghamton, New York, USA
skrisht1@binghamton.edu

Anirudh Nori
Computer Science
Binghamton University
Binghamton, New York, USA
anori1@binghamton.edu

ABSTRACT

In this project, we aim to create a comprehensive understanding of gun violence trends and public sentiment surrounding this critical societal issue by collecting, analyzing, and interpreting data from various social media platforms, including Reddit, and 4chan. By leveraging the unique characteristics of each platform, we intend to gain insights into the frequency, volume, and emotional tones of discussions related to gun violence. This project's outcomes will contribute to informed decision-making, as well as foster a deeper understanding of public perceptions and reactions to gun violence in the digital age.

1 INTRODUCTION

In our data collection process, we utilize two prominent platforms: Reddit and 4chan, each offering distinct insights and perspectives within the realm of online discussions and communities. Initially, we considered including Twitter in our data collection efforts, but the limited number of tweets available through the free API tier precluded us from doing so.

2 DATA SOURCES

Reddit, as one of the largest and most diverse social platforms, offers a rich source of user-generated content. By tapping into the Reddit API, we gain access to a wide array of discussions, opinions, and experiences revolving around various topics, including but not limited to "gun politics', 'pro gun', 'CCW', 'Gun Culture', 'gun culture', 'Gun Politics',' Second Amendment', 'firearms', 'second amendment', 'NRA', 'Pro-gun', 'Pro-gun''.

For Data Points Collected, Title and self-text of posts, Author information, Post scores and number of comments, Upvote ratios, Comments (as an additional feature) For rationale, the aim of extracting this data is to perform in-depth analyses of trends, sentiments, and conversations in specific subreddits. By aggregating posts and their associated information, we gain valuable insights into user behavior and interests within these communities. Additionally, the moderation of hate speech is of paramount importance, prompting the inclusion of keywords that facilitate this endeavor.

4chan, an imageboard forum, provides a unique perspective on internet culture and unfiltered discussions. Utilizing the 4chan API allows us to capture the candid and often raw interactions taking place on the platform. For Data Points Collected, Post ID, username, and comment text Board and thread information. For timestamp of posts, By extracting and storing this data, we aim to delve into the unfiltered expressions and discussions taking place on 4chan. This includes a wide range of topics, from hobbies and entertainment to more niche and specialized interests.

3 IMPLEMENTATION

Reddit Crawler Implementation:
The Reddit crawler is a Python script designed to extract data from Reddit using the Reddit API. Here's a breakdown of its implementation:
For API Access and Authentication, the script utilizes the requests library to interact with the Reddit API. It employs OAuth2 authentication, which requires a client ID and client secret for authorization. To write API Endpoint, the crawler constructs a URL to the Reddit API endpoint using the desired subreddit and the number of top posts to retrieve. The endpoint is https://www.reddit.com/r/{subreddit}/top/.json?limit={limit}. For HTTP Requests and JSON Processing, the script sends an HTTP GET request to the Reddit API endpoint. The response, in JSON format, contains information about the top posts in the specified subreddit. For Data Extraction, the JSON response is parsed to extract relevant information such as post titles, self-text (post content), author names, scores, number of comments, upvote ratios, and associated comments.

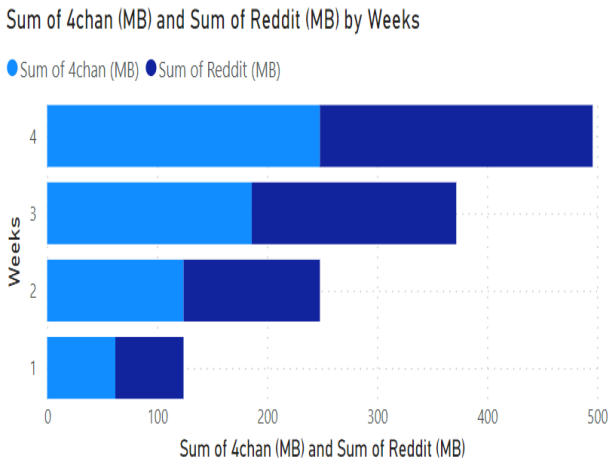
4chan Crawler Implementation:
The 4chan crawler is a Python script responsible for retrieving data from 4chan threads. Here's how it's implemented. For API Access, similar to the Reddit crawler, the 4chan crawler uses the requests library to communicate with the 4chan API. For API Endpoint, the crawler constructs a URL to the 4chan API endpoint, specifying the board and thread ID. The endpoint looks like this: https://a.4cdn.org/{board}/thread/{thread_id}.json. For HTTP Requests and JSON Processing: The script sends an HTTP GET request to the 4chan API endpoint. The response, in JSON format, contains information about the posts within the specified thread. For data extraction, the JSON response is parsed to extract pertinent information, including post IDs, usernames, comments, and timestamps.
For both Reddit and 4chan, Keyword Filtering: The script checks whether the title or self-text of each post contains any of the predefined keywords. If a match is found, the post is considered relevant and added to the dataset. For Data Storage, the collected data is stored in a PostgreSQL database. A dedicated table is created to hold Reddit data, ensuring structured and organized storage. Similarly, a table is created to store 4chan data.

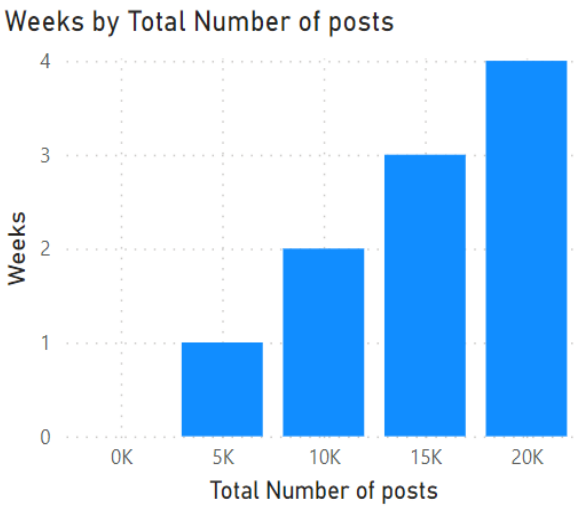
4 POSTGRESQL DATABASE

PostgreSQL was chosen as the database management system due to its seamless integration with Python, ensuring efficient data insertion and retrieval. It provides robust data integrity and reliability, which is essential for maintaining the accuracy and consistency of our dataset. Additionally, PostgreSQL's compatibility with third-party libraries and tools is vital for further processing and analysis, including text processing, sentiment analysis, and hate speech detection.

5 FAKTORY SCHEDULING

For this project, we have allocated 500MB of RAM for data collection purposes. To optimize data collection within this limited space, we performed a detailed capacity analysis. Based on this analysis, we have calculated that each Reddit post/thread consumes approximately 27KB, while each 4chan post occupies around 23KB. As a result, we have structured our data collection strategy to gather 2300 Reddit posts and 2700 4chan posts per week. This totals approximately 124MB of data collected per week. Over the course of four weeks, we anticipate accumulating a dataset comprising 9200 Reddit posts and 10800 4chan posts, resulting in a total of 20000 posts. This dataset is projected to consume approximately 498MB of storage space. This meticulous planning ensures that we gather a substantial amount of data for in-depth analysis while efficiently managing our allocated resources.





6 CHALLENGES FACED

API Rate Limits: We encountered instances where both Reddit and 4chan imposed rate limits on our API requests. This resulted in delays in data retrieval, especially when dealing with large volumes of posts. To address this, we had to implement a rate-limiting strategy and ensure that our data collection process adhered to the specified limits.

Handling Large Datasets: As the data collection progressed, the size of our dataset grew substantially. This presented challenges in terms of storage capacity and processing resources. We found that our system's resources were strained, leading to slower processing times. To mitigate this, we had to optimize our storage solutions and fine-tune our processing algorithms to handle the increasing dataset size efficiently.