Analyzing Political and Crime Events through Multi-Source Data Collection

Sridhar Madala
Computer Science
Binghamton university
Binghamton NY United States
smadala2@binghamton.edu

Durga Prasad Barla
Computer Science
Binghamton University
Binghamton NY United States
dbarla1@binghamton.edu

Murali Venkateswara Gopi Krishna Ponnada
Computer Science
Binghamton University
Binghamton NY United States
mponnad1@binghamton.edu

Harish Vodeyavargam Anand
Computer Science
Binghamton University
Binghamton NY United States
hvodeya1@binghamton.edu

Shreevara Andila
Computer Science
Binghamton University
Binghamton NY United States
sandila1@binghamton.edu

ABSTRACT

This report provides the analysis and visualization of diverse datasets extracted from YouTube and Reddit, encompassing posts, comments, videos, and political toxicity. The analysis includes the identification of patterns in the number of posts per hour and comments per video daily. The project also delves into the toxic comment classification within political subreddits, presenting insights into the prevalence of toxic comments. Additionally, provides a comparative analysis of video statistics by channel, highlighting the number of videos in crime and politics collections. The data is stored in a MongoDB database, and both data analysis and visualization techniques are implemented. The visual representations we generated provide crucial insights, serving as valuable resources for further research and in-depth analysis within the domain of online content and user behavior.

RQ's intended to unravel: Now it comes to this project which aims to solve two questions as follows:

1) Democrats VS Republicans who is more toxic? (using ModerateHateSpeech)
2) Comparing the nature of discussions of a crime on reddit vs how it is on YouTube for different topics (sentimental analysis)

KEYWORDS

Data collection, Data cleaning, Data analysis, Machine learning models, Visualization, Matplotlib, MongoDB, Python, Reddit API, YouTube Data API

INTRODUCTION

In the ever-evolving landscape of social media, this project delves into the analysis of discussions related to crime and politics on Reddit and the categorization of YouTube videos in these domains. Social media platforms, such as Reddit and YouTube, have become integral components of contemporary communication, serving as conduits for diverse perspectives and content sharing. The project aims to decode the sentiment patterns within Reddit discussions while employing natural language processing to classify YouTube videos, fostering a nuanced understanding of user engagement and content trends.

The methodology encompasses the integration of APIs, sentiment analysis pipelines, and advanced natural language processing techniques. The Reddit analysis involves real-time comment extraction from targeted subreddits, while YouTube analysis entails data retrieval from specified channels and subsequent categorization based on content. The systematic storage of analyzed data in MongoDB databases ensures a structured approach to information management.

The significance of this project lies in unraveling insights into online discourse, understanding sentiment dynamics in distinct social media communities, and contributing to a comprehensive MongoDB database for future research endeavors. By examining crime and politics discussions on these platforms, the project aims to offer valuable perspectives on user behavior, sentiment trends, and content preferences, ultimately contributing to the evolving landscape of social media analytics.

BACKGROUND RESEARCH

The advent of social media platforms has ushered in an era of unprecedented connectivity, transforming how individuals share information and engage in conversations. Researchers and analysts have increasingly turned their attention to the vast repository of data generated by users on platforms like Reddit and YouTube to discern patterns, preferences, and sentiments. Social media, once a medium for communication, has evolved into a multifaceted arena influencing public opinion, making the analysis of trends and topics paramount.

In existing research, themes such as the identification of prevalent topics, user engagement patterns, and the utilization of hashtags have garnered considerable attention. Scholars have delved into the content of social media posts, utilizing analysis techniques to unveil the most discussed themes and track how these topics evolve over time. Furthermore, understanding user engagement and activity patterns has been instrumental, with studies revealing variations in social media use during specific times or days of the week. Hashtags, serving as categorical markers, have been subject to scrutiny to discern popular discussion topics and user trends.

However, the rapid evolution of social media platforms necessitates continuous exploration to keep abreast of the latest trends. New features, emerging platforms, and shifting user behaviors present ongoing challenges and opportunities for researchers in comprehending the dynamics of social media. As the project focuses on analyzing crime and politics discussions on Reddit and categorizing YouTube videos, it builds upon this background research to contribute insights into the ever-changing landscape of online interactions and content dissemination.

DESCRIBING DATASET

1) YouTube

This is designed to systematically retrieve and analyze data from YouTube channels, specifically focusing on the top 10 videos and their associated top 100 comments for each selected channel. This data is then stored in a MongoDB database, with individual tables dedicated to each YouTube channel, ensuring a structured and organized storage approach. Threading mechanisms are employed for periodic tasks, facilitating regular updates of the top videos and comments for the chosen channels.

Fig.1 depicts the plot of YouTube data in a period i.e., from Nov-1 to Nov-14 2023. X-axis represents the period and Y- axis represents the Number of Comments. And Table.1,2 represents the data of different channels with political & crime dataset information.

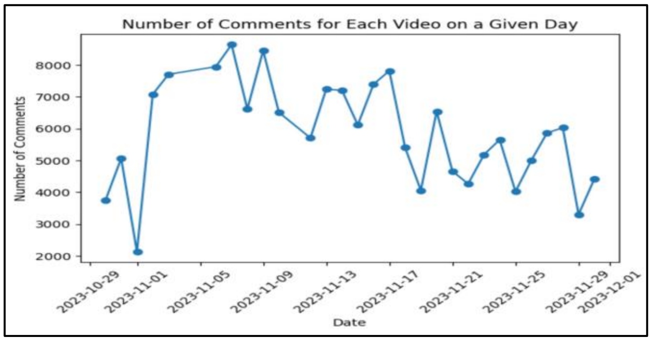
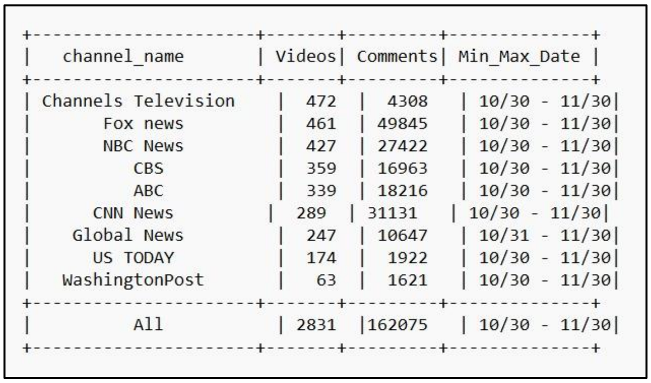
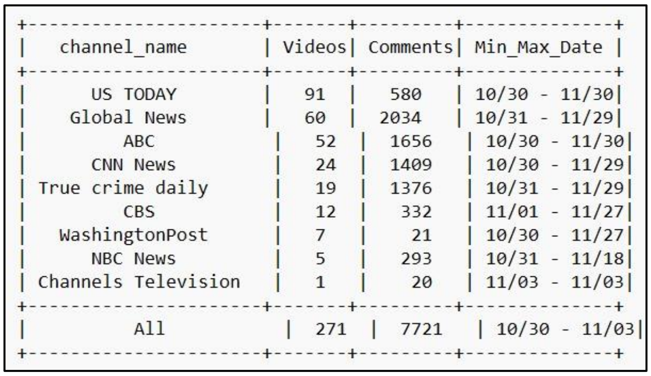


Fig.1 Periodic entries of YouTube



```
+---------------------+-------+---------+--------------+
|   channel_name      | Videos| Comments| Min_Max_Date |
+---------------------+-------+---------+--------------+
| Channels Television   |  472  |  4308   | 10/30 - 11/30|
|       Fox news        |  461  | 49845   | 10/30 - 11/30|
|       NBC News        |  427  | 27422   | 10/30 - 11/30|
|          CBS          |  359  | 16963   | 10/30 - 11/30|
|          ABC          |  339  | 18216   | 10/30 - 11/30|
|       CNN News      |  289  | 31131   | 10/30 - 11/30|
|     Global News       |  247  | 10647   | 10/31 - 11/30|
|       US TODAY        |  174  |  1922   | 10/30 - 11/30|
|   WashingtonPost      |   63  |  1621   | 10/30 - 11/30|
+---------------------+-------+---------+--------------+
|         All           | 2831  |162075   | 10/30 - 11/30|
+---------------------+-------+---------+--------------+
```

Table.1 YouTube Crime Dataset



```
+---------------------+-------+---------+--------------+
|   channel_name      | Videos| Comments| Min_Max_Date |
+---------------------+-------+---------+--------------+
|       US TODAY      |   91  |   580   | 10/30 - 11/30|
|     Global News     |   60  |  2034   | 10/31 - 11/29|
|          ABC          |   52  |  1656   | 10/30 - 11/30|
|       CNN News        |   24  |  1409   | 10/30 - 11/29|
| True crime daily      |   19  |  1376   | 10/31 - 11/29|
|          CBS          |   12  |   332   | 11/01 - 11/27|
|   WashingtonPost      |   7   |    21   | 10/30 - 11/27|
|       NBC News        |   5   |   293   | 10/31 - 11/18|
| Channels Television   |   1   |    20   | 11/03 - 11/03|
+---------------------+-------+---------+--------------+
|         All             |  271  |  7721   | 10/30 - 11/03|
+---------------------+-------+---------+--------------+
```

Table.2 YouTube Politics Dataset

1.1 YouTube Comments Plot - Video Engagement Analysis

The "YouTube Comments Plot" serves as a visual representation of the engagement dynamics on YouTube videos over a 15-day period. This analysis focuses on the number of comments per day, providing insights into patterns of viewer interaction and highlighting days with heightened user engagement. YouTube's API was used to collect data on videos, specifically focusing on the number of comments each video received. The data was processed to aggregate the number of comments on videos daily, providing an overview of daily engagement levels. This plot was created using python visualization library matplotlib. This plot provides us with insights about identification of days with heightened engagement, indicating popular video releases or topics that generated increased viewer interaction, understanding how user engagement fluctuates across the 15-day period and providing insights into when viewers are most active in leaving comments.

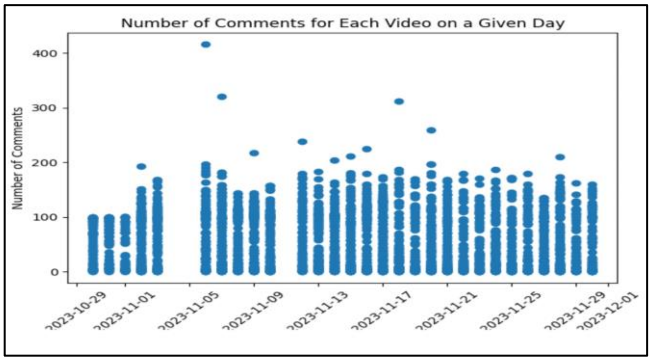


Fig.2 YouTube comments plot

Fig.2 depicts the plot of YouTube comments in a period i.e., from Nov-1 to Nov-14 2023. X-axis represents the period and Y- axis represents the Number of Comments.

1.2 YouTube Per Channel Plot - Crime vs. Politics Video Analysis

The "YouTube Per Channel Plot" offers a comprehensive overview of content distribution across various YouTube channels, specifically focusing on the number of videos related to crime and politics. This analysis spans multiple channels over a given period, providing insights into the content preferences and thematic focus of each channel. Data was collected from multiple YouTube channels to understand the diversity in content production and videos were categorized based on their thematic content, specifically focusing on crime-related and politics-related topics using a model called Multinomial Naive Bayes (MNB) classifier which is a text classification model. The x-axis represents the names of the YouTube channels under analysis and the y-axis quantifies the number of videos produced by each channel, with two bars per channel: one for crime-related videos and another for politics-related videos. This plot provides a visual representation of content distribution across different channels, highlighting the proportion of crime-related and politics-related videos while also identifying the channels with a predominant focus on crime or politics, as well as those that maintain a balance between the two themes.

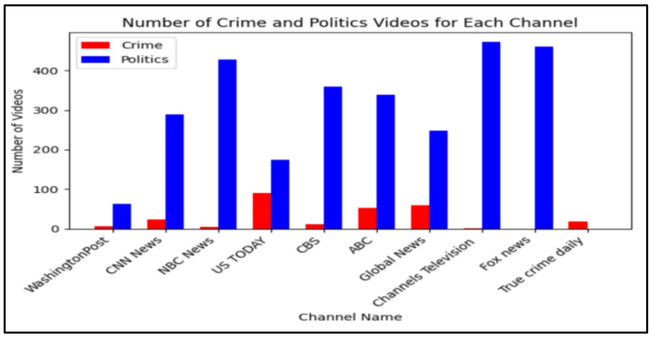


Fig.3 Crime vs Politics of YouTube

Fig.3 depicts the plot of crime and politics. X-axis represents the channel name and Y- axis represents the number of comments.

2) Reddit API

The data extracted from Reddit, focusing on two distinct categories—crime-related and political subreddits. This collects information about the latest posts and their associated comments, providing valuable insights into user engagement and content trends.

The script utilizes the Reddit API to systematically retrieve data from specified crime-related and political subreddits. For crime-related content, subreddits such as r/TrueCrime, r/SerialKillers, r/CrimeScene, and r/RedditCrimeCommunity are considered. Similarly, political subreddits include r/Republican, r/democrats, r/Ask_Politics, r/politics, and r/PoliticalDiscussion. MongoDB serves as the storage infrastructure for the collected data. Dedicated collections, namely **crime_collection** and **politics_collection**, house information related to crime and politics subreddits, respectively.

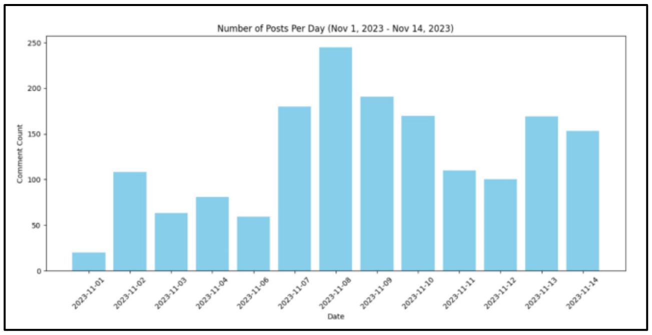


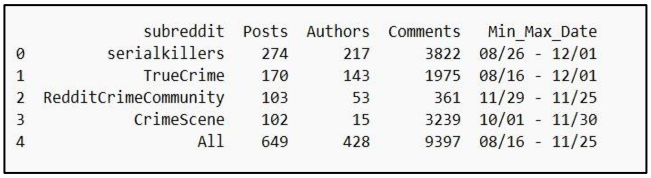
Fig.4 Periodic Entries of Reddit

Fig.4 represents the plot of periodic entries of Reddit. X-axis represents the period and Y- axis represents the Number of Comments.

2.1 Politics Post Plot – Reddit API Data Analysis

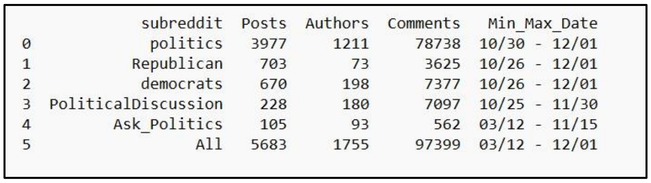
The "Politics Post Plot" is a visual representation of the comments count of Reddit posts related to politics over a 15-day period. The data for this plot was extracted from the Reddit API and stored in a MongoDB database. The aim was to analyze the engagement and discussion trends over time within the realm of political discussions on Reddit. The plot provides insights into how engagement in political discussions on Reddit fluctuates over the 15-day period and identification of specific days with heightened activity, potentially corresponding to significant political events or discussions while also understanding whether certain types of political posts or topics lead to more user engagement. This plot was created using the python visualization library Matplotlib.

Tables 3 and 4 represents the table with different political & crime subreddits, including the number of posts, unique authors, total comments, and the date range during which the data was collected. The "All" row aggregates data across all subreddits for a comprehensive overview.



```
             subreddit  Posts  Authors  Comments    Min_Max_Date
0        serialkillers    274      217      3822   08/26 - 12/01
1            TrueCrime    170      143      1975   08/16 - 12/01
2  RedditCrimeCommunity   103       53       361   11/29 - 11/25
3           CrimeScene    102       15      3239   10/01 - 11/30
4                  All    649      428      9397   08/16 - 11/25
```

Table.3 Reddit Crime Dataset



```
            subreddit  Posts  Authors  Comments   Min_Max_Date
0            politics   3977     1211     78738  10/30 - 12/01
1          Republican    703       73      3625  10/26 - 12/01
2            democrats    670      198      7377  10/26 - 12/01
3  PoliticalDiscussion    228      180      7097  10/25 - 11/30
4        Ask_Politics    105       93       562  03/12 - 11/15
5                 All   5683     1755     97399  03/12 - 12/01
```

Table.4 Reddit Politics Dataset

2.3 Politics Comments Plot – Reddit API Data Analysis

The "Politics Comments Plot" is a visual representation of the relationship between the number of Reddit comments related to politics and the corresponding number of comments over a 15-day period, analyzed on an hourly basis. This analysis aims to uncover patterns in user engagement within political discussions on Reddit, offering insights into when discussions are most active and garner the highest number of comments. The data was processed to aggregate the number of comments on political posts on an hourly basis for each day. This plot provides the insights into the identification of specific hours throughout the day with heightened activity, providing insights into when political discussions on Reddit attract the most comments, analysis of temporal patterns to determine if there are recurring trends in user engagement during specific hours across the 15-day period, and understanding optimal times for posting political content to maximize engagement and comment interaction. This analysis facilitates the identification of peak hours for user interaction, aiding in the development of strategic posting approaches for those contributing to political discourse on the platform.

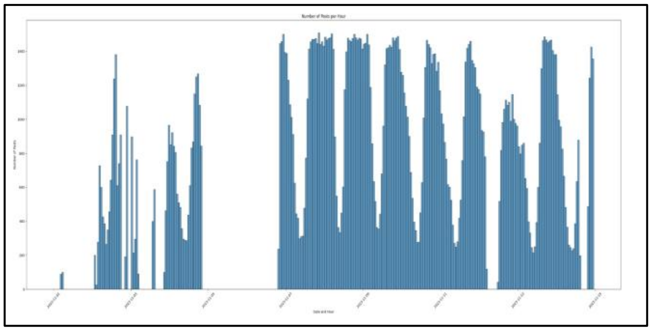


Fig.5 Posts per Hour of Reddit

DISCUSSION

From the Data Analysis we observed and were able to answer the proposed RQ's.

1) In order to find Democrats VS Republicans who is more toxic. The "Toxicity Comparison Plot" offers a comparative analysis of toxic comments within the Democrats and Republicans subreddits. Toxicity levels were determined using the Moderate Hate Speech API, providing insights into the percentage of toxic comments within each political subreddit. The comments were collected from two distinct political subreddits, each representing a different ideological perspective. The Moderate Hate Speech API was utilized to evaluate the toxicity levels of each comment, categorizing them as either toxic or non-toxic. This Fig.6 was created using matplotlib. The x-axis represents the two political subreddits under analysis: Democrats and Republicans and the y-axis quantifies the number of comments in each subreddit. The plot reveals that 3.43% of the total comments within the Democrats subreddit were classified as toxic by the Moderate Hate Speech API and in contrast, the Republicans subreddit exhibited a higher percentage, with 6.48% of the total comments identified as toxic. This plot facilitates a direct comparison of toxicity levels between the Democrats and Republicans subreddits, and provides identification of notable differences in toxicity percentages, providing insights into the online discourse dynamics within each political community.

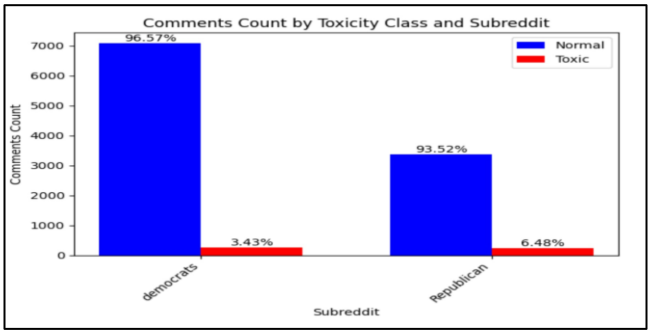


Fig.6 Politics Toxicity Plot using ModerateHateSpeech API

2) In order to compare the nature of discussions of a crime on reddit vs how it is on YouTube for different topics (sentimental analysis). In two dataset comparison plots, we have plotted the sentimental analysis output between Reddit and YouTube datasets. The Fig.7 contains the source type on x-axis and 'Sentiment Score' on y-axis, differentiating the data based on 'Sentiment' category column. The values fall under -0.5 to -1.0 is Negative and values from +0.5 to +1.0 is Positive and rest considered to be Neutral Sentiment

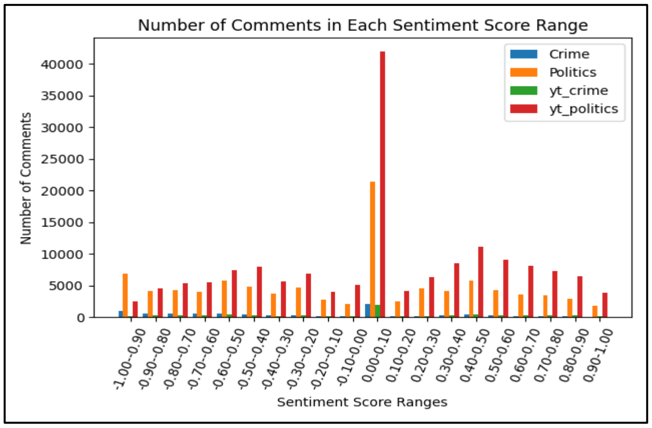


Fig.7 Sentiment Analysis

3) The modified version of our sentiment analysis code for Project 3 represents a significant improvement over its predecessor by incorporating user-defined thresholds for positive, neutral, and negative comments, thus providing a more interactive and customizable experience. In the above Fig.7, sentiment analysis was performed on Reddit and YouTube comments data, and a predefined set of thresholds was used to categorize sentiments. In contrast, the enhanced version now allows users to input their desired threshold values, empowering them to tailor the sentiment classification according to their specific preferences or the nature of the dataset being analyzed. This dynamic approach ensures greater flexibility and adaptability, enabling users to fine-tune the sentiment analysis results based on the context of the investigation. Fig.8 & Fig.9 shows the sentiment analysis output for different threshold values. Fig.8 threshold values have been set between -0.5 to 0.5 for neutral with sentiment score above 0.5 to be positive and negative for below -0.5. Similarly, in Fig.9 the threshold values are set between -0.25 to 0.25 for neutral.

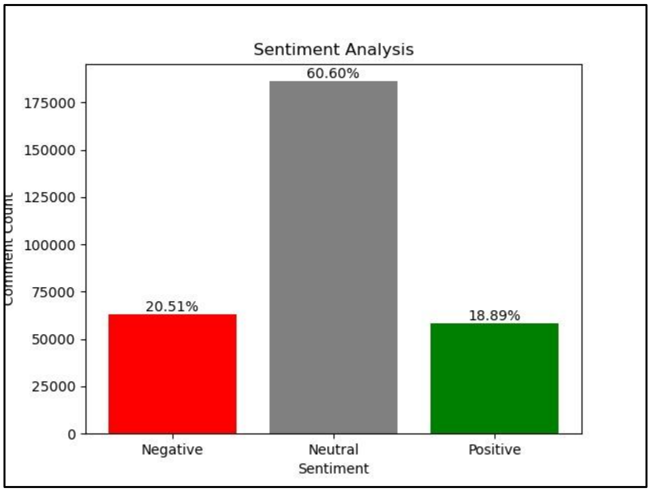


Fig.8 Sentiment Analysis for -0.5 to 0.5

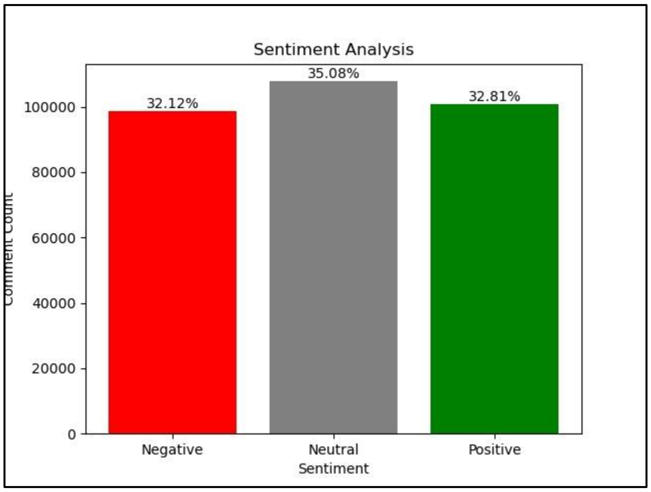


Fig.9 Sentiment Analysis for -0.25 to 0.25

CONCLUSION

After collecting some data from Reddit and YouTube in as much real time as possible, we see that there is more interest in politics than crime regardless of the platform. Also, in politics subreddits conceding the two major parties of the USA, we see that democrats are more active on their subreddit whereas Republicans are less likely to comment. Also, people in Republican subreddit seem to be nearly 2 times toxic online than democrat subreddit. While there are limitations to the amount and the way we collected the data, it still offers some valuable insights in the online community.

Our sentiment analysis of Reddit and YouTube comments brought out an interesting pattern – there are way more positive comments than negative ones, with a big chunk falling into the neutral category. This suggests that the overall vibe in the online community we studied is mostly positive. The abundance of neutral comments implies a fair amount of discussions that aren't strongly emotional or biased. Recognizing these sentiment trends is vital for grasping the online conversations' tone. These findings offer a straightforward yet insightful look into the prevalent attitudes and opinions within our dataset.

LIMITATIONS

Our study has some limitations. First, we only looked at posts from a certain time, so our findings might not reflect what all social media users are discussing. Second, because people choose to participate in social media, the data might not represent everyone, introducing some bias. Lastly, because social media changes a lot, what we found might not apply to future tech trends on these platforms.

REFERENCES

[1] Jeremy Blackburn, Utkucan Balcı, Chen Ling, Emiliano De Cristoforo Megan Squire, Gianluca Stringhini. Beyond Fish and Bicycles: Exploring the Varieties of Online Women ’ s Ideological Spaces, 2023.

[2] Multinomial Naive Bayes classifier: https://towardsdatascience.com/multinomial-naïve-bayes-for-documents-classification-and-natural-language-processing-nlp-e08cc848ce6,https://www.geeksforgeeks.org/applying-multinomial-naive-bayes-to-nlp-problems

[3] https://developers.google.com/youtube/v3