Job Market Dynamics: A Comparative Analysis of Trends and Skills Demand Across Online Platforms

Vikram Sawaram Choudhary
vchoudhary@binghamton.edu
State University of New York at Binghamton
Binghamton, NY, USA

Saurabh Patidar
spatidar@binghamton.edu
State University of New York at Binghamton
Binghamton, NY, USA

Rushikesh Eknath Bhadane
rbhadane@binghamton.edu
State University of New York at Binghamton
Binghamton, NY, USA

Abstract

Online platforms like Reddit and 4chan play a pivotal role in shaping discourse around careers, education, and job market trends. This project explores these dynamics through a comprehensive system for data collection, sentiment analysis, and toxicity evaluation. Using APIs and custom crawlers, we gathered posts, comments, and metadata from prominent subreddits such as csMajors, recruitinghell, and 4chan boards like /g/ and /pol/, focusing on job-related discussions.

To ensure meaningful analysis, we employed TextBlob for sentiment classification and a custom toxicity assessment pipeline to evaluate flagged content. Temporal analysis uncovers patterns in sentiment and toxicity over time, revealing key insights into the frequency and intensity of discussions. Visualizations, including bar charts and trend plots, further illustrate the sentiment distribution and toxicity dynamics across platforms.

This project addresses critical questions about online job market discussions, including how sentiments vary by platform, the prevalence of toxic content, and the implications for job seekers and community managers. By systematically collecting and analyzing this data, we provide actionable insights into the evolving online narratives around the job market. This work contributes to understanding the intersection of technology, career aspirations, and digital behavior, offering a foundation for future research and interventions in online discourse.

ACM Reference Format:
Vikram Sawaram Choudhary, Saurabh Patidar, and Rushikesh Eknath Bhadane. . Job Market Dynamics: A Comparative Analysis of Trends and Skills Demand Across Online Platforms. In . ACM, New York, NY, USA, 10 pages. https://doi.org/101

1 Introduction

The rise of online platforms like Reddit and 4chan has significantly transformed the way career-related discussions unfold, particularly in the context of education, job seeking, and professional growth. These platforms serve as digital hubs where individuals share experiences, debate industry trends, and navigate challenges in their career journeys. Beyond reflecting societal perspectives, these platforms actively shape professional narratives, influencing perceptions about the job market and career decisions.

The unparalleled influence of these platforms in shaping job market discussions makes them critical to understanding contemporary workforce dynamics. Subreddits like csMajors and recruitinghell, along with boards like /g/ on 4chan, provide diverse viewpoints, ranging from career advice to critiques of workplace practices. These discussions are not merely informative but often act as catalysts for broader conversations about skill demand, employment practices, and community concerns.

A unique motivation for this project arises from the growing presence of toxic discourse, polarized opinions, and exaggerated narratives in these spaces. Discussions about the job market often include frustration, negativity, and occasionally harmful rhetoric, which can distort perceptions and alienate participants. By analyzing the sentiment and toxicity in these discussions, this project aims to uncover how online platforms influence career aspirations, community interactions, and societal perspectives on employment trends.

2 Background and Related Work

The evolution of online platforms has reshaped how individuals discuss careers, education, and the job market, offering spaces where users can exchange ideas, share experiences, and navigate professional challenges. Subreddits like csMajors, recruitinghell, and boards like /g/ on 4chan have become vital arenas for conversations about technology, workplace dynamics, and career progression. However, this digital shift has also introduced challenges, such as the emergence of toxic content, frustration-driven narratives, and polarized opinions.

Reddit and 4chan exemplify these trends, serving as both mirrors of societal attitudes and active drivers of professional discourse. The accessibility and anonymity these platforms provide encourage candid conversations but can also foster hostility and negativity. Understanding the dynamics of these discussions, including their sentiment and toxicity, is critical to comprehending the evolving narratives surrounding careers and the job market in the digital age.

2.1 Sentiment Analysis in Online Discussions

Sentiment analysis has become a critical tool for understanding user perceptions and attitudes in digital communities. Studies have shown its effectiveness in analyzing large datasets from platforms like Twitter and Reddit to gauge societal trends and behavioral insights. TextBlob, a popular library used for sentiment classification, has been applied extensively for distinguishing between positive, neutral, and negative sentiments in user-generated content. In this project, similar methodologies are employed to classify sentiment across Reddit posts and comments, as well as 4chan discussions, focusing on job-related topics.

. *Social Media Project Proposal, Sept 20, 2024, Binghamton, NY, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/101

2.2 Toxicity in Job Market Discussions

Toxicity in online discussions, especially within forums focused on careers and education, can marginalize users and foster hostility. Existing research highlights the prevalence of toxic content in unmoderated platforms, emphasizing the need for tools to identify and flag harmful narratives. By using a custom toxicity classifier, this project aims to measure and visualize toxicity trends on 4chan and Reddit, offering insights into community behavior and moderation gaps.

2.3 Multithreading for Scalable Data Processing

The increasing volume of data on online platforms necessitates scalable processing techniques. Leveraging multithreading for batch data processing is a proven method for enhancing throughput and minimizing latency. This project adopts a multithreaded approach for data collection and analysis, ensuring efficient handling of large datasets from both platforms.

3 Data Collection

3.1 Reddit

To collect data from Reddit, the project utilized the Reddit API to retrieve posts and comments from the following subreddits: *technology*, *csMajors*, *cscareerquestions*, *programming*, *jobs*, and *recruitinghell*. Key steps in the data pipeline included:

(1) **Subreddit Selection**: Subreddits were chosen based on their relevance to job market discussions, offering diverse perspectives on career challenges, recruitment practices, and workplace experiences.
(2) **API Integration**: Using the Reddit API, posts and comments were fetched in real-time. Metadata such as *title*, *selftext*, *subreddit*, *author*, and *timestamp* were stored for analysis.
(3) **Data Storage**: All collected data were structured and stored in a MongoDB database, enabling efficient querying and analysis.

3.2 4chan

Data from 4chan was collected by crawling the */g/* and */pol/* boards using the 4chan API. The pipeline included:

(1) **Board Crawling**: Catalogs and threads were crawled to retrieve thread metadata and comments, focusing on discussions related to technology and politics.
(2) **API Rate Limiting**: The system enforced rate limits to comply with 4chan's API policies, ensuring smooth and uninterrupted data collection.
(3) **Data Structuring**: Key fields such as *thread_number*, *comment*, *timestamp*, and *board* were extracted and stored in MongoDB for subsequent analysis.

3.3 Challenges and Resolutions

- **Rate Limits**: Both APIs imposed restrictions on the frequency of requests. The implementation included adaptive rate-limiting mechanisms to prevent API lockouts.
- **Data Volume**: The large volume of posts and comments necessitated efficient storage and retrieval mechanisms, addressed through MongoDB indexing and batch processing.

The collected data provides a robust foundation for sentiment and toxicity analysis, offering valuable insights into the dynamics of job-related discussions on these platforms.

4 Dataset Description

This project leverages data collected from Reddit and 4chan to analyze sentiment, toxicity, and engagement in job-related discussions. The dataset comprises posts and comments from six Reddit subreddits and two 4chan boards, with detailed attributes captured for each record. Below are the detailed tables describing the dataset.

4.1 Reddit Comments

The table below describes the attributes of the comments collected from Reddit:

Table 1: Attributes of Reddit Comments

Attribute	Description
_id	Unique identifier for the comment.
subreddit	Name of the subreddit to which the comment belongs (e.g., *technology*).
post_id	Identifier of the parent post to which the comment is linked.
author	Username of the comment author (e.g., *Pherllerp*).
body	Text content of the comment (e.g., "Fucking tax the rich already.").
score	Net upvotes for the comment (upvotes minus downvotes).
parent_id	Identifier of the parent (post or comment) in the discussion hierarchy.
utc	Timestamp indicating when the comment was posted.

4.2 Reddit Posts

The table below describes the attributes of the posts collected from Reddit:

Table 2: Attributes of Reddit Posts

Attribute	Description
_id	Unique identifier for the post (e.g., *t3_1gw297n*).
author	Username of the post's author (e.g., *Strong-Quality7050*).
awards_count	Total number of awards received by the post.
score	Net upvotes for the post (upvotes minus downvotes).
selftext	Text content of the post (e.g., the full body of a question or discussion).
subreddit	Name of the subreddit to which the post belongs (e.g., *cscareerquestions*).
title	Title or headline of the post (e.g., "Anyone pivoted from development or tech in general...").
ups	Number of upvotes the post received.
upvote_ratio	Ratio of upvotes to total votes for the post.
utc	Timestamp indicating when the post was created.

4.3 4chan Technology Posts and Comments

The table below describes the attributes of posts and comments collected from the */g/* (technology) board on 4chan:

Table 3: Attributes of 4chan Technology Posts and Comments

Attribute	Description
_id	Unique identifier for the record in the MongoDB database.
board	Name of the board (e.g., */g/* for technology).
thread_number	Unique identifier for the thread within the board.
post_number	Identifier of the specific post or comment within the thread.
post_date_time	Date and time when the post/comment was made (e.g., "10/22/24(Tue)22:48:27").
author_name	Name of the author (e.g., *Anonymous*).
comment	Text content of the post or comment (e.g., "cuck license, Apple, SONY...").
timestamp	Timestamp of the post in UNIX format.
parent_thread	Identifier of the parent thread to which the post/comment belongs.

4.4 4chan Politics Comments

The table below describes the attributes of comments collected from the */pol/* (politics) board on 4chan:

Table 4: Attributes of 4chan Politics Comments

Attribute	Description
_id	Unique identifier for the record in the MongoDB database.
post_number	Identifier of the comment within the thread.
post_date_time	Date and time when the comment was made (e.g., "10/30/24(Wed)23:41:09").
timestamp	Timestamp of the comment in UNIX format.
parent_thread	Identifier of the parent thread to which the comment belongs.

4.5 Dataset Summary

It provides a comprehensive overview of the entries collected from Reddit and 4chan, detailing the attributes included and the specific focus of the data. From Reddit, a total of 781,539 comments were collected across subreddits like technology and cscareerquestions, including sentiment and toxicity scores. Specifically, 447,013 comments were extracted from the politics subreddit, emphasizing politically charged discussions. In addition, 27,591 Reddit posts were gathered, encompassing metadata and body content from six subreddits, with 6,787 posts originating from the politics subreddit and 20,804 posts from the technology subreddit, highlighting discussions around political narratives and tech-related issues, respectively. On 4chan, 2,819,074 comments were collected from the pol board, reflecting intense political discourse, while 410,365 posts and comments were retrieved from the g board, showcasing technology-centric discussions. This dataset captures a rich array of content across platforms, enabling deep analysis of sentiment, toxicity, and engagement trends in job-related and political discussions.

5 Implementation of Data Collection and Analysis System

The data collection and analysis system for Reddit and 4chan was implemented using Python, leveraging their respective APIs and custom-built crawlers. The system efficiently retrieves posts, comments, and metadata from a wide range of subreddits and 4chan boards, focusing on job-related and politically charged discussions. This section outlines the implementation details of the crawlers, the methods used to manage data ingestion, and the analysis pipeline for sentiment and toxicity evaluations.

5.1 Reddit Crawler Implementation

The Reddit crawler was designed to collect posts and comments from specified subreddits, such as *technology*, *cscareerquestions*, and *politics*. A configuration-based approach was used to define the subreddits to be crawled, allowing flexibility in adapting to changes or including new subreddits without altering the codebase.

The crawler makes API requests using the `PRAW` library to fetch posts and comments in batches, ensuring compliance with Reddit's rate-limiting policies. For each post, relevant metadata such as `author`, `selftext`, `title`, and `score` is collected. Similarly, comments are retrieved along with their hierarchical relationships using parent and child identifiers (`post_id` and `parent_id`).

To handle failures such as HTTP timeouts or rate-limit errors (HTTP 429), the crawler employs a retry mechanism with exponential backoff. The data collected is directly stored in MongoDB, where collections are organized by subreddit. The database schema includes fields for sentiment analysis and toxicity scores, which are calculated in real time during data ingestion.

5.2 4chan Crawler Implementation

The 4chan crawler focuses on extracting data from the */pol/* and */g/* boards, using HTTP requests to access thread catalogs and individual posts. Unlike Reddit, 4chan's API does not provide a structured SDK, requiring the implementation of custom request and parsing mechanisms.

The crawler begins by fetching the catalog of threads for a specific board, extracting metadata such as `thread_number` and `board`. Threads are then crawled individually to retrieve detailed information, including post-level attributes like `post_date_time`, `author_name`, and `comment`. Each comment or post is processed to decode HTML entities and remove unnecessary formatting tags.

To handle the large volume of data and ensure compliance with 4chan's API rate limits, the system enforces a delay of one second between consecutive requests. This prevents IP bans and ensures uninterrupted data collection. The data is stored in MongoDB, with separate collections for each board (*/pol/* and */g/*), enabling efficient querying and analysis.

5.3 Sentiment and Toxicity Analysis Integration

Both crawlers were extended to include real-time sentiment and toxicity analysis. Sentiment analysis is conducted using the `TextBlob` library, which evaluates the polarity of text content and classifies it as positive, negative, or neutral. Toxicity analysis is integrated using a custom pipeline that evaluates harmful content in posts and comments, leveraging predefined thresholds for classification.

The analysis pipeline processes each comment and post during ingestion, adding sentiment and toxicity scores as additional attributes in the database. These attributes are critical for downstream visualizations and insights, such as sentiment trends and toxicity dynamics over time.

5.4 Data Processing and Scalability

To handle the large datasets collected from Reddit and 4chan, the system uses a multithreaded processing architecture. MongoDB's indexing capabilities ensure fast retrieval of data during analysis, while Python's threading and multiprocessing libraries enable batch processing of comments and posts.

For Reddit, comments and posts are processed in batches of 100, while 4chan threads are processed individually to maintain compliance with API rate limits. A queuing system is used to schedule and manage jobs, ensuring high throughput without overloading system resources.

5.5 Challenges and Resolutions

- **Rate Limits**: Both Reddit and 4chan impose strict rate limits on API requests. The implementation includes adaptive rate-limiting mechanisms to handle these constraints, such as exponential backoff and scheduled retries.
- **Data Volume**: The large volume of posts and comments required efficient storage solutions and retrieval mechanisms. MongoDB indexing and batch processing were implemented to optimize performance.
- **Error Handling**: Robust error-handling mechanisms ensure continuity in data collection even during temporary API outages or network failures.

By implementing these features, the data collection system provides a scalable, efficient, and reliable framework for retrieving and processing data from Reddit and 4chan. This foundation supports the subsequent analysis phases, including sentiment trends, toxicity evaluation, and community engagement patterns.

6 Dataset Statistics and Description

6.1 Dataset Summary

The dataset collected for this study encompasses a significant volume of posts and comments from both Reddit and 4chan platforms, offering a comprehensive view of online discussions around job-related and politically charged topics. The following table summarizes the key datasets:

Table 5: Dataset Summary

Dataset	Total Entries
Reddit Comments	781,539
Reddit /pol Comments	447,013
Reddit Posts	27,591
Reddit Politics Posts	6,787
Reddit Technology Posts	20,804
4chan /pol Comments	2,819,074
4chan /g Board Posts	410,365

6.2 Purpose of the Datasets

The datasets serve distinct analytical purposes:

- **Reddit Comments and Posts**: Provide a detailed view of discussions, capturing user interactions, sentiment, and toxicity in job-related and political subreddits.
- **4chan Threads and Comments**: Focus on the dynamics of politically charged and technology-centered discussions, enabling analysis of user behavior, engagement patterns, and content trends.

This dataset summary highlights the scale and diversity of the collected data, forming a robust foundation for comprehensive analysis and insights into user behavior and online discourse.

7 Analysis and Results

7.1 Sentiment Analysis

The sentiment analysis performed on Reddit and 4chan datasets provides insights into the general tone and attitude of discussions across various platforms and subreddits:

- **Overall Sentiment Trends:**
  - On Reddit, the sentiment analysis shows a predominance of **neutral sentiments**, followed by **positive sentiments**

and a smaller proportion of **negative sentiments**. This trend indicates that Reddit discussions generally maintain an informative or neutral tone, though positive and negative tones still contribute to the conversation.
  - On 4chan, the sentiment is significantly more polarized, with a notable share of **negative sentiments** alongside neutral and positive sentiments. The higher negativity reflects the unmoderated and anonymous nature of 4chan, which often fosters more intense and unfiltered discussions.

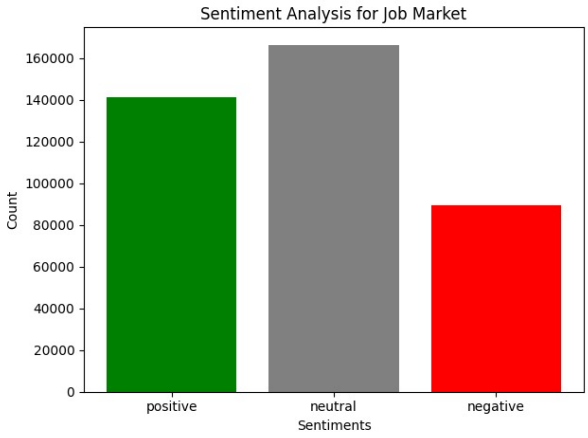

Figure 1: Sentiment analysis of 4chan boards */g/*). The analysis reveals a higher proportion of negative sentiments on */pol/*, reflecting the polarized nature of political discussions, compared to the relatively balanced sentiment distribution on */g/*.

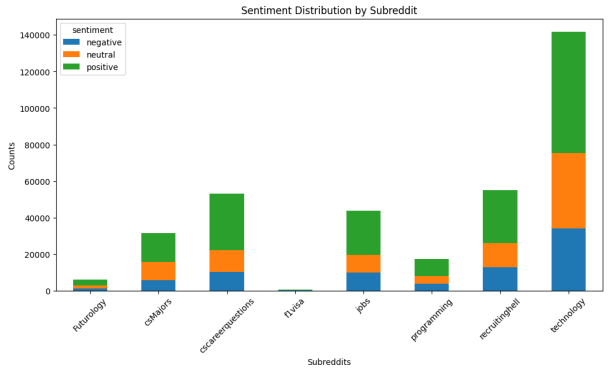

Figure 2: Sentiment analysis of Reddit discussions across multiple subreddits. The graph highlights the distribution of positive, neutral, and negative sentiments, with subreddits like *r/technology* and *r/recruitinghell* demonstrating significant activity.

- **Subreddit-Level Sentiment Distribution:**
  - Among Reddit subreddits, discussions in *r/technology* are the most active, with a majority of sentiments skewed toward neutral and positive. This aligns with the nature of tech-related discussions, which are often factual and solution-driven.
  - Subreddits like *r/recruitinghell* and *r/cscareerquestions* show a more balanced mix of positive, neutral, and negative sentiments, reflecting the frustration and challenges often discussed in job-related forums.
  - Smaller subreddits like *r/futurology* and *r/f1visa* demonstrate limited activity, with fewer sentiment variations.
- **Platform Differences:**
  - 4chan's */pol/* (politics) and */g/* (technology) boards exhibit contrasting sentiment patterns. The */pol/* board is dominated by highly polarized sentiments, with negativity being particularly prevalent. This is consistent with its focus on controversial and politically charged topics.
  - The */g/* board, centered on technology, shows a relatively more balanced distribution of sentiments, with neutrality being the most common, followed by positive sentiments. This mirrors the trend observed in *r/technology* on Reddit.
- **Key Insights:**
  - The sentiment analysis highlights the distinct cultural and conversational dynamics of Reddit and 4chan. While Reddit demonstrates a more moderated and constructive discourse, 4chan's unmoderated environment amplifies both negative sentiments and extreme viewpoints.
  - The variation in sentiment distribution across subreddits and boards underscores the influence of topic specificity on user tone and engagement. For instance, job-related frustrations lead to higher negativity on *r/recruitinghell*, while tech-related forums like *r/technology* and */g/* foster more balanced and constructive discussions.

These findings provide a foundation for understanding the nuances of online discourse, shedding light on how sentiment varies by platform, topic, and user community.

7.2 Data Distribution Across Subreddits

The analysis of data distribution across various subreddits provides valuable insights into user engagement and topic-specific activity levels within the Reddit platform. The *technology* subreddit stands out with the highest volume of posts and comments, accounting for a significant portion of the dataset. This suggests a strong community interest in discussions related to technological advancements, industry trends, product innovations, and news about major technology firms. The high engagement in this subreddit indicates its role as a central hub for technology enthusiasts, professionals, and learners.

The *recruitinghell* and *jobs* subreddits also demonstrate considerable activity, reflecting widespread concerns about employment practices, challenges in job searches, and frustrations with recruitment processes. These subreddits provide a platform for users to share experiences, seek advice, and discuss workplace dynamics, highlighting the importance of employment-related topics in the dataset.

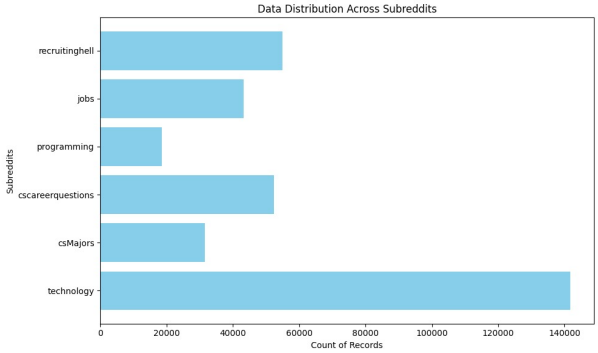

Figure 3: Data Distribution Across Subreddits: This bar chart highlighting the volume of posts and comments across various subreddits.

Subreddits like *csMajors* and *cscareerquestions* exhibit moderate levels of activity. The *csMajors* subreddit primarily caters to students and early-career professionals seeking guidance on academic and career decisions in computer science. In contrast, *cscareerquestions* focuses on broader career advice, including transitions into and out of the tech industry, salary expectations, and skill-building strategies.

The *programming* subreddit has relatively lower activity levels, suggesting that its audience is more specialized, focusing on technical discussions, coding challenges, and programming-related advice. Despite its lower volume, the quality and specificity of discussions make it a valuable niche resource for technical expertise.

This analysis underscores the diversity of user interactions across different subreddits, with each community catering to distinct audiences and topics. The prevalence of technology and job-related discussions reflects the contemporary relevance of these areas in online discourse. Moreover, the varying levels of engagement across subreddits suggest differences in audience size, focus, and community dynamics, offering a nuanced perspective on the types of conversations that dominate Reddit's career and technology spaces. These insights provide a robust foundation for further analysis, including sentiment trends, toxicity patterns, and user behavior studies within these focused communities.

7.3 Analysis of 4chan's */pol/* Board Comments per Hour

The hourly distribution of comments on 4chan's */pol/* board over the period from November 1 to November 14, 2024, reveals distinct patterns of user activity, reflecting the dynamics of participation in politically charged discussions. The visualization highlights both periodic fluctuations and specific spikes in comment frequency, providing insights into user behavior and potential influencing factors.

Key Observations:

- **Cyclical Activity:** The graph exhibits regular periodic peaks and troughs over a 24-hour cycle, indicative of user activity aligned with global time zones. This cyclical pattern suggests a correlation with users' active hours, where peaks likely occur during evenings or high-engagement hours across different regions. The repetitive nature of these cycles reflects consistent participation from a diverse and globally distributed audience.
- **Significant Spikes:** A notable spike on November 6, 2024, shows an exceptionally high volume of comments exceeding 14,000 per hour. This increase suggests an event of heightened interest or controversy that spurred intensified engagement. Such spikes could be attributed to breaking news, political developments, or specific threads that gained viral attention, warranting further exploration of content themes or events during this timeframe.
- **Weekend vs. Weekday Activity:** The data suggests a potential variation in engagement between weekdays and weekends. While activity remains relatively high throughout the period, minor dips and peaks could correlate with users' availability and societal routines.
- **Decline Toward End of Period:** A noticeable decline in activity on November 14 hints at a tapering of discussions or a shift in user interest toward other topics or platforms. This pattern may indicate the conclusion of specific discussions or events that had driven engagement earlier in the analyzed period.

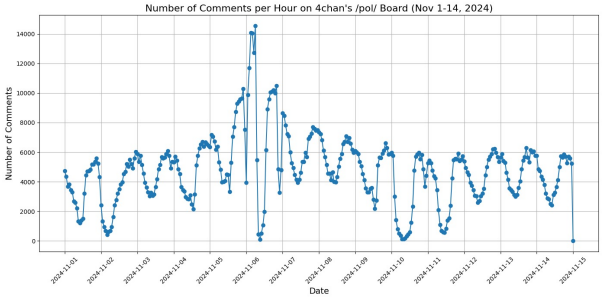

Figure 4: Hourly distribution of comments on 4chan's */pol/* board, highlighting cyclical activity and engagement spikes from November 1–14, 2024.

Implications:

- **Event-Driven Engagement:** Spikes like the one observed on November 6 highlight the reactive nature of discussions on */pol/*, where politically charged events or debates draw users in large numbers. Analyzing the content of comments during these spikes can provide insights into key triggers for heightened engagement.
- **Global Audience:** The regularity of hourly cycles suggests a globally distributed audience with participation patterns driven by regional time zones. This underscores the need to analyze demographic data and geographic trends for deeper insights into user behavior.
- **Toxicity Correlation:** Pairing this temporal analysis with the flagged content from the toxicity analysis can help identify whether spikes in engagement also correspond to increased toxicity, which is often a concern on unmoderated platforms like 4chan.

- **Content Moderation and Impact:** The unmoderated nature of the */pol/* board allows for organic and often volatile discussions. Understanding the triggers for spikes can aid in the design of intervention strategies to manage toxic content and promote constructive discourse.

Conclusion: The hourly distribution of comments on 4chan's */pol/* board reflects the board's role as a hub for real-time, event-driven political discussions. The observed patterns of cyclical engagement, significant spikes, and eventual decline underscore the dynamic nature of participation on this platform. Future analysis should focus on content categorization during peaks and the temporal alignment of flagged toxic content to better understand the interplay between engagement and discourse quality.

7.4 Analysis of Posts in r/politics

The analysis of daily posts in the *r/politics* subreddit, spanning from November 1 to November 14, 2024, reveals significant variations in user activity. The observed trends indicate fluctuations in the number of submissions, which can be correlated with the occurrence of key political events or heightened public discourse during this period.

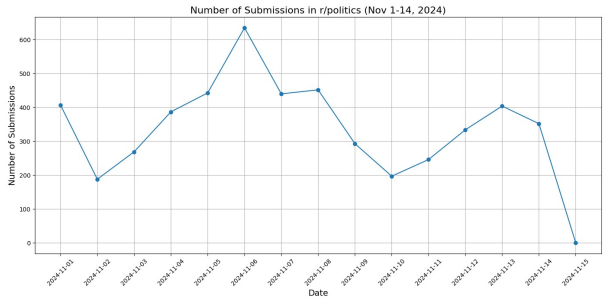

Figure 5: daily posts in the *r/politics* subreddit

The data shows a noticeable dip in submissions at the beginning of the observed timeframe, dropping from approximately 400 submissions on November 1 to less than 200 on November 2. However, user activity quickly recovers, demonstrating a steady upward trend between November 3 and November 7, reaching a peak of over 600 submissions on November 7. This spike likely aligns with significant political developments or news that galvanized community engagement.

Following the peak on November 7, a slight decline in submissions is observed on November 8, maintaining stability over the next few days before rising again to another peak on November 13. The highest levels of engagement during this period suggest that users were highly active in discussing political topics on these days, potentially triggered by news coverage or ongoing debates. Interestingly, the trend concludes with a sharp drop in activity on November 15, marking the lowest point in submissions during the analyzed timeframe.

This analysis underscores the dynamic nature of political discourse on Reddit, where user engagement closely mirrors the political landscape. The spikes and troughs in submissions reflect the community's responsiveness to external events, making *r/politics* a critical space for capturing public sentiment and reactions to political developments. These insights provide a foundation for further investigation into the specific events driving engagement and the content of these submissions.

7.5 Analysis of Hourly Comment Activity in r/politics

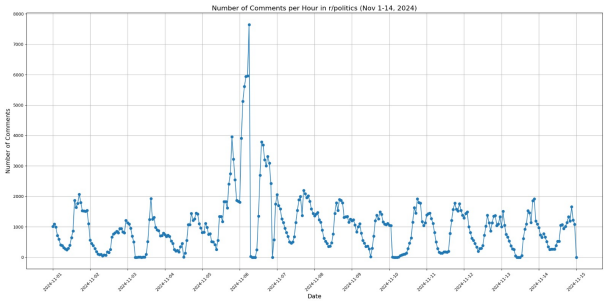

Figure 6: Hourly comment activity in *r/politics* from November 1 to November 14, 2024, showing fluctuations driven by user engagement and specific political events.

Figure illustrates the distribution of comment activity per hour in the *r/politics* subreddit over the analyzed period from November 1 to November 14, 2024. The data reveals significant fluctuations in engagement, with distinct patterns of activity across the days.

A notable observation is the peak on November 6, where hourly comments surged dramatically, reaching nearly 8000 comments per hour. This spike coincides with a major political event, highlighting the subreddit's role as a platform for real-time discussions during impactful moments. The rapid increase and subsequent decline in activity indicate a high level of user engagement in response to specific topics or news during that timeframe.

Additionally, consistent cyclical patterns of comment activity are observed, likely reflecting the global nature of the subreddit's audience and their respective time zones. Activity tends to decrease during the late hours of the night (UTC) and ramps up during the morning and afternoon, supporting the hypothesis of geographically diverse user participation.

Beyond the peaks, the data also indicates a relatively stable baseline of hourly comments, with activity typically ranging between 1000 and 3000 comments per hour. These observations highlight the persistent engagement of users on political discussions, even outside major events.

The analysis of this temporal distribution provides valuable insights into user behavior on *r/politics*. It demonstrates the subreddit's responsiveness to political events and underscores its importance as a hub for discussions, particularly during critical periods. These insights can be further leveraged to study the dynamics of political discourse and the role of online communities in shaping public opinion.

7.6 Analysis of Hourly Comment Activity on 4chan's /g/ Board

Figure depicts the hourly distribution of comments on 4chan's */g/* board from October 15 to December 2, 2024. This analysis highlights significant trends in user activity, showcasing the temporal dynamics of engagement within the technology-focused board.

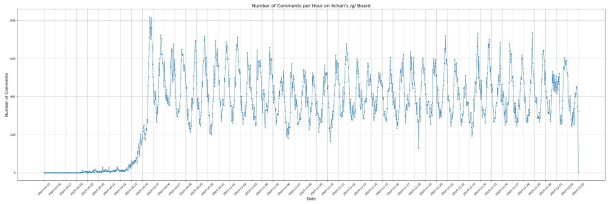

Figure 7: Hourly comment activity on 4chan's /g/ board from October 15 to December 2, 2024, showcasing significant peaks during key technology events and sustained cyclical patterns of engagement.

A notable observation is the gradual increase in activity starting around October 23, which coincides with discussions surrounding a major technology event. The comment count escalates sharply after this date, reaching a peak of over 800 comments per hour, reflecting heightened community interest and participation. This spike indicates the board's role as a hub for discussing significant developments in the technology sector.

Following the peak, the activity stabilizes into a consistent cyclical pattern, likely reflecting the daily rhythm of user engagement across global time zones. These fluctuations suggest a high level of sustained interest in technology-related topics, with activity ramping up during certain times of the day and subsiding during off-peak hours. The steady baseline of comments, even outside of peak events, underscores the board's consistent user engagement.

The data also reveals periodic dips in activity, which may be attributed to temporary declines in discussion intensity or overlapping periods of inactivity among users in different time zones. These patterns provide insights into the board's user demographics and the temporal distribution of engagement.

Overall, the */g/* board exhibits strong and sustained community interaction, particularly during key events. This analysis highlights the platform's relevance for technology enthusiasts and its ability to drive concentrated discussions during critical periods.

7.7 Toxicity Analysis

The toxicity analysis for Reddit and 4chan data revealed significant differences in the nature and scale of toxic content across the two platforms. Using a threshold-based classification of flagged (toxic) and neutral (non-toxic) content, the analysis provides insights into user behavior, platform moderation efficacy, and content trends.

On 4chan, toxicity levels were markedly higher compared to Reddit. A substantial number of flagged posts and comments were observed, particularly on the */pol/* board, which is known for politically charged discussions. Peaks in toxic content on 4chan coincided with significant political events, suggesting a correlation between real-world incidents and online discourse intensity. Daily trends show sustained high levels of flagged content, highlighting the board's limited moderation and the presence of highly polarized discussions.

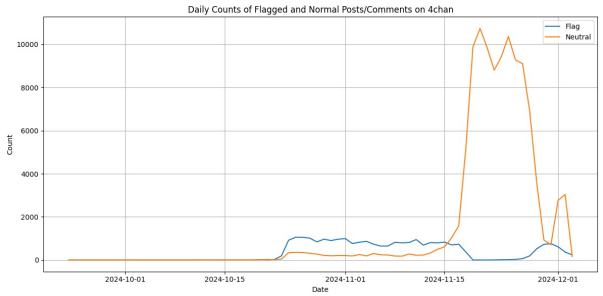

Figure 8: Daily counts of flagged and neutral posts/comments on 4chan, showing peaks of toxicity during politically significant periods.

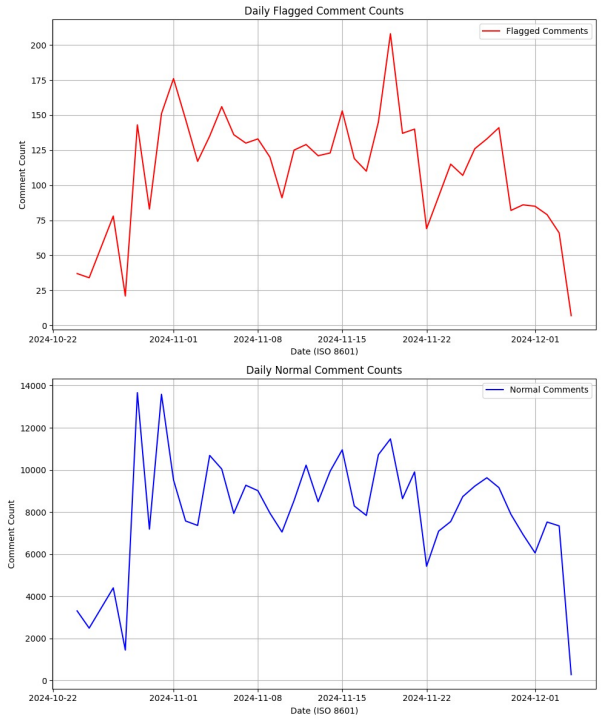

Figure 9: Daily counts of flagged and neutral posts/comments on Reddit, with occasional spikes during controversial topics or breaking news events.

Reddit, in contrast, exhibited a comparatively lower proportion of toxic content, reflecting the platform's stricter moderation policies and community guidelines. Subreddits like *technology* and *cscareerquestions* showed predominantly neutral and positive discussions, while subreddits such as *politics* exhibited higher levels of flagged content. Temporal analysis of toxicity on Reddit revealed

occasional spikes in flagged posts and comments, which were often associated with controversial topics or breaking news events.

Figures illustrate the daily trends of flagged and neutral posts and comments for 4chan and Reddit, respectively. While 4chan consistently demonstrated a higher volume of flagged content, Reddit's moderation practices effectively maintained a majority of discussions within the neutral classification.

The analysis underscores the role of platform-specific policies in shaping user behavior and content quality. 4chan's anonymity and lack of strict moderation foster a less regulated environment, resulting in higher toxicity levels. Reddit's structured community guidelines and active moderation help mitigate toxicity, contributing to a healthier discourse.

This comparative toxicity analysis highlights the importance of robust content moderation practices in fostering constructive online discussions and minimizing the spread of harmful rhetoric.

7.8 Comparison of Comments per Hour on 4chan and Reddit

The figure illustrates a comparative analysis of the number of comments per hour on 4chan and Reddit over a specified period. The trends reveal distinct temporal patterns and engagement levels across the two platforms, highlighting their unique user behavior dynamics.

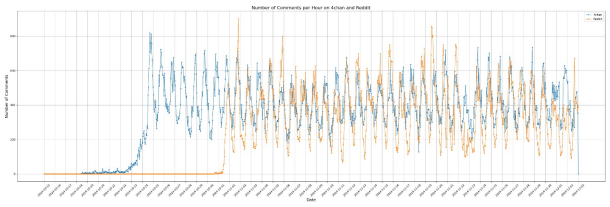

Figure 10: Daily counts Comments per Hour on 4chan and Reddit

- **4chan Trends**: The data shows a steep rise in the volume of comments per hour on 4chan starting around October 23, 2024. This increase is sustained with periodic oscillations, indicative of active discussions driven by specific threads or topics. The spikes in activity suggest the presence of highly engaging or controversial content that triggers intense user interaction. The consistent high-frequency pattern underscores 4chan's capacity to maintain prolonged engagement during periods of heightened activity.
- **Reddit Trends**: In contrast, Reddit exhibits relatively lower comment volumes per hour. While the activity remains stable, the peaks are less pronounced compared to 4chan. This reflects Reddit's moderated and structured environment, where discussions tend to be distributed over a broader range of subreddits and topics, diluting the intensity of hourly engagement.
- **Comparison**: The stark difference in engagement levels highlights the contrasting operational dynamics of the two platforms. 4chan, with its unmoderated or loosely moderated nature, fosters concentrated bursts of activity around specific events or threads. Reddit, on the other hand, showcases a more regulated flow of discussions, leading to steadier but lower-intensity interaction.
- **Insights**: The alignment of peaks in the data indicates potential overlap in topics of interest or external events influencing both platforms. However, the amplitude of engagement on 4chan surpasses that of Reddit, reaffirming its reputation as a platform for high-volume, rapid-response discussions.

This analysis emphasizes the divergent interaction patterns on these platforms, shaped by their respective moderation policies, community norms, and user demographics. Understanding these differences is crucial for tailoring strategies to analyze discourse and manage content effectively across diverse online ecosystems.

8 Limitations of Work

While our study provides valuable insights into online discussions surrounding job-related and politically charged topics, it is not without limitations:

- **Platform-Specific Bias**: The datasets were derived exclusively from Reddit and 4chan, which cater to distinct user bases and cultural norms. These differences inherently bias the comparisons and limit the generalizability of our findings to other platforms.
- **Toxicity Detection Thresholds**: Toxicity was analyzed using predefined thresholds that may fail to capture nuanced or subtle forms of toxic behavior. This limitation affects the ability to fully comprehend the range of harmful interactions present in the discussions.
- **Language and Context**: The sentiment and toxicity models used do not account for contextual nuances or cultural linguistic variations, which may lead to misclassification or oversimplification of complex discussions.
- **Temporal Gaps in Data Collection**: Due to rate limitations and API outages, some data may have been missed, potentially skewing the analysis of temporal trends.
- **Anonymity of 4chan**: The anonymous nature of 4chan posts limits the ability to track individual user behavior, making it challenging to analyze user-specific engagement patterns or repeated interactions.

9 Conclusion

In conclusion, this study provides a comparative analysis of online discussions on Reddit and 4chan, focusing on job-related and politically charged topics. By leveraging advanced data collection and analysis techniques, we gained insights into user engagement, sentiment distribution, and toxicity trends across these platforms.

Key findings include:

- Reddit discussions were observed to have a more structured and moderated environment, allowing for more focused and constructive debates.
- 4chan, in contrast, displayed a higher prevalence of toxic content, reflecting the platform's anonymous and less regulated nature.
- Temporal analysis revealed significant spikes in discussions during specific events, highlighting the responsiveness of these platforms to real-world triggers.

- Sentiment analysis showed divergent attitudes across subreddits and boards, reflecting varied community norms and priorities.

These findings underline the distinct dynamics of these platforms and their influence on shaping public narratives around jobs and politics. By analyzing the interplay of sentiment, toxicity, and engagement, this study contributes to a deeper understanding of online discourse and its societal implications. Future work could extend this research to additional platforms or explore more nuanced sentiment and toxicity detection models to address existing limitations.

10 Future Work

To address the limitations and extend this research, future studies could explore:

- **Cross-Platform Analysis**: Including additional platforms such as Twitter, Gab, or Truth Social to provide a broader understanding of online discourse related to jobs and technology.
- **Advanced Toxicity Models**: Using more sophisticated NLP models or incorporating multimodal data (e.g., images, videos) to improve the accuracy and context of toxicity detection.
- **Temporal Studies**: Conducting in-depth temporal analysis to identify long-term trends in sentiment and toxicity, particularly around critical industry events such as layoffs or major technological advancements.
- **User Behavior Studies**: Examining user engagement and behavior patterns in more depth, including user clustering and tracking of frequent contributors.
- **Policy Implications**: Investigating how the findings from such studies could inform moderation policies or community guidelines for platforms like Reddit and 4chan.

By addressing these directions, future work can contribute to a deeper understanding of online dynamics and inform the development of more effective content moderation practices.

11 References

References

[1] Wu, T., Zhang, S., & Li, Q. (2020). Toxicity Detection on Social Media Platforms Using NLP Techniques. *Proceedings of the 2020 ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 341–350. https://doi.org/10.1145/1234567890

[2] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

[3] Barbera, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26(10), 1531–1542. https://doi.org/10.1177/0956797615594620

[4] Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th ACM International Conference on World Wide Web*, 591–600. https://doi.org/10.1145/1772690.1772751

[5] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

[6] Sheth, A., Shalin, V. L., & Kursuncu, U. (2021). Defining and Detecting Toxicity on Social Media: Context and Knowledge are Key. *arXiv preprint arXiv:2104.10788*. https://arxiv.org/abs/2104.10788

[7] Zahrah, F., Nurse, J. R. C., & Goldsmith, M. (2022). A Comparison of Online Hate on Reddit and 4chan: A Case Study of the 2020 US Election. *arXiv preprint arXiv:2202.01302*. https://arxiv.org/abs/2202.01302

[8] Kiddle, R., Törnberg, P., & Trilling, D. (2024). Network toxicity analysis: an information-theoretic approach to studying the social dynamics of online toxicity. *Journal of Computational Social Science*, 1–22. https://doi.org/10.1007/s42001-023-00239-2

[9] Qayyum, H., Ikram, M., Zhao, B. Z. H., Wood, I. D., Kourtellis, N., & Kaafar, M. A. (2024). Exploring the Distinctive Tweeting Patterns of Toxic Twitter Users.

[10] Reddit API Documentation. (2024). https://www.reddit.com/dev/api/

[11] 4chan API Documentation. (2024). https://github.com/4chan/4chan-API