A Weaverlet-Based Interactive Dashboard for Temporal Analysis of Hate Speech on Reddit

ALEXANDRU LUNGU, University of Twente, The Netherlands

Today, social media platforms have become the dominant medium for information dissemination, with approximately 5.24 billion active users worldwide. This vast reach raises critical concerns about the spread of hate, which now occurs at an unprecedented rate and affects millions of people globally. Hateful messages circulate daily across numerous platforms, impacting a wide range of social groups. Much of this content is closely tied to politics, conflicts, and global events that serve as catalysts for the amplification of hate. Reddit, in particular, stands out as a prominent platform for political discourse and news sharing. This paper introduces an interactive dashboard designed to visualize the spread of hate speech on Reddit. The dashboard enables users to upload datasets, filter them by subreddit, flairs, and time period, and explore the raw data through interactive time-series and influence graphs. Our main contributions are threefold: (i) a novel system for analyzing the temporal dynamics of hate speech, (ii) a module for identifying influential users, and (iii) new insights into the patterns of content propagation through cross-post and domain analysis. The dataset used in this study focuses on political extremism and is drawn from various politically-oriented subreddits, particularly during the period surrounding the 2024 U.S. presidential election.

Additional Key Words and Phrases: Reddit Analysis, hateBERT, Time-Series Visualization, Hate Speech Detection

ACM Reference Format:
Alexandru Lungu. 2025. A Weaverlet-Based Interactive Dashboard for Temporal Analysis of Hate Speech on Reddit. In *Proceedings of 43th Twente Student Conference on IT (TScIT 43).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Social media is a relatively recent phenomenon, first emerging in the late 1990s alongside the infrastructure provided by Web 1.0—the first stage in the evolution of the World Wide Web. By the mid-2000s, giants like Facebook had brought social networking into mainstream culture. Since then, social media has evolved into a massive global industry, encompassing billions of users across dozens of platforms. Reddit is one of the major platforms that emerged around the same time as Facebook. As of 2025, it has over 90 million daily users, making it one of the leading social platforms. Reddit organizes discussions into user-moderated communities called subreddits, each focused on a specific topic. While primarily used for discussion and opinion sharing, it can sometimes become a hub for hateful and extremist content. As a result, some subreddits have become notable sources of hate speech.

A prominent example was the subreddit `r/The_Donald`, a community widely known for its politically charged discourse surrounding Donald Trump. The increasing toxicity and spread of misinformation in that subreddit led to multiple violations of Reddit's rules, ultimately resulting in its ban in June 2020 [17].

Many studies have linked politicians' behavior with shifts in social norms and public prejudices. Research conducted in the United States often centers on Donald Trump's election and provides substantial empirical evidence that his behavior has influenced societal prejudice.

For example, Kim and Ogawa conducted an empirical study using a Twitter network to examine how Governor Yuriko Koike's refusal to attend a memorial for Korean massacre victims in Japan contributed to a rise in hate speech. Their findings demonstrated that the governor's actions led to a significant surge in hateful posts and increased anti-Korean sentiment on social media platforms [7].

Other studies have focused on the United States, particularly on Donald Trump's 2016 election. These works show that Trump's anti-Muslim tweets served as triggers for offline violence and facilitated the spread of xenophobic hashtags [8].

These findings emphasize that politicians have a clear impact on public attitudes and social media now serves as a critical tool for understanding how this influence manifests. Events such as Governor Koike's refusal to participate in a memorial ceremony illustrate how political actions can act as flashpoints for online hate.

Understanding the spread of hate speech is not a new area of research—numerous tools for its detection and analysis have been developed over the years. This field has evolved rapidly, transitioning from traditional machine-learning methods to transformer-based models.[1] [13] Researchers have developed novel visualization systems for tracking the diffusion of hateful content. Many of these tools focus on cross-platform dynamics and aim to map how harmful narratives propagate across different social media spaces.

2 PROBLEM STATEMENT

Prior dashboards have been developed [10], [3],[11], however, Reddit is still understudied and is an emerging platform for data scientists to study. Moreover, very few studies are focusing on day-to-day changes of data that can be surged by political events and spread hate towards many minorities. Based on these issues the following research questions should be answered:

RQ1: How does the frequency of hateful posts on Reddit fluctuate over time, and are there identifiable patterns or trends in these fluctuations?

RQ2: In what ways do the day-to-day fluctuations in the volume and proportion of hate-speech posts vary among politically diverse subreddits, and how do they relate to major offline events?

RQ3: To what extent is the creation of hate-speech content concentrated within a small subset of highly active accounts?

By addressing these research questions, we aim to gain a deeper understanding of how hateful content spreads on Reddit. This involves examining the temporal patterns and fluctuations in hateful posts,

Author's address: Alexandru Lungu, a.lungu-1@student.utwente.nl, University of Twente, P.O. Box 217, Enschede, The Netherlands, 7500AE.

TScIT 43, July 04, 2025, Enschede, The Netherlands

as well as analyzing user activity through engagement metrics and identifying the key influencers who contribute most significantly to the propagation of hate within subreddits.

3 BACKGROUND

3.1 hateBERT

Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art language model capable of processing sentences bidirectionally, making it a powerful tool for a wide range of Natural Language Processing (NLP) tasks. BERT's bidirectional context modeling allows for a nuanced understanding of language and meaning, which is particularly effective in detecting subtle forms of hate speech. It has been shown to outperform other deep learning approaches in this domain [12].

HateBERT is a retrained version of BERT, specifically fine-tuned for abusive language detection in English. It was trained on 1.5 million offensive Reddit comments sourced from banned subreddits, making it highly suitable for the purposes of this study. In their evaluation, the authors compared HateBERT with the original BERT on three corpora rich in offensive and abusive language. HateBERT outperformed BERT on every dataset and, establishing a new state-of-the-art macro-F1 of 0.765, exceeding BERT's 0.727 and the previous best benchmark of 0.716 on one of the datasets [4].

3.2 Reddit

Reddit is a social media platform where users, known as redditors, submit links, images, videos, or text posts that the community can vote up or down. All content is organized into themed boards called subreddits, making Reddit a highly structured and categorized platform—particularly valuable for data science research. Each subreddit functions as a self-contained community, governed by its own rules regarding content formatting, submission guidelines, and interaction norms making the platform ideal for in-depth analyses of information diffusion, sentiment shifts, and patterns of user activity.

Reddit's API offers researchers free access to data encompassing a wide range of fields. In fact, Reddit content aligns with a majority of topics found in global Google Trends, further demonstrating its relevance as a data source [15].

3.3 Weaverlet

Weaverlet is a component-driven framework built on top of the Plotly Dash framework, offering functionality similar to React but fully implemented in Python [6]. It enables the development of multi-page applications, where each page contains scalable and modular components. These components communicate with one another through a signaling system.

In Weaverlet, every element of the dashboard resides within a component, which is simply a Python class. Each component encapsulates its own layout, unique HTML identifiers, and all associated logic. Weaverlet also includes specialized utilities such as the **SimpleRouterComponent**, which maps URLs to pages using a simple dictionary, and the **RedirectComponent**, which enables client-side navigation through standard Dash callbacks.

A key innovation introduced by Weaverlet is its signal system. Signals allow components to emit payloads that other components can listen to. This supports clean, multi-stage workflows where one signal can cascade into another, with zero risk of circular dependency errors—since signals exist independently of Dash's JSON graph.

4 RELATED WORK

A significant portion of current hate speech detection research relies heavily on Twitter-based datasets. These studies continue to dominate the field due to Twitter's accessibility, volume, and availability of historical data [14] [2]. In contrast, Reddit-specific datasets are comparatively less common, although they are increasingly recognized as crucial for platform-specific analysis. This gap is particularly relevant given Reddit's community-driven structure and diverse topic-focused subreddits, which present unique dynamics not captured by Twitter data.

Some recent studies have adopted a multi-platform approach, combining datasets from Twitter, Reddit, Facebook, and other platforms to build more comprehensive training corpora and enhance model generalization [16]. Despite these efforts, Twitter remains the dominant source, especially in systems where real-time monitoring and linguistic diversity are prioritized.

Three notable platforms illustrate the evolution of hate speech monitoring systems. The MANDOLA platform enables real-time analysis using big data infrastructure and provides filtering capabilities based on time, hate-related topics, and geographical locations (country or city level) [10]. Users can examine specific events across all visualizations to explore potential correlations between spikes in hate speech and offline incidents. MANDOLA also supports interactive exploration and visualization, although it primarily focuses on Twitter data.

The Data Viz Platform for Hate Speech Analysis targets NLP researchers by providing an interactive dashboard that visualizes tweets directed at ethnic minority groups in Italy [3]. It supports multi-dimensional exploration through NLP techniques and manual annotation processes, offering a structured way to investigate how hate speech varies across contexts.

HaterNet stands out for introducing a novel public dataset in Spanish, including two million untagged tweets and 6,000 manually labeled tweets annotated by four expert raters [11]. The platform features two main modules: hate speech detection and social network analysis. Its visual tools include user-mention graphs and word concurrency graphs, enabling analysis of both linguistic patterns and social connections in hate-related discourse.

Despite advances in network analysis, very few systems place emphasis on temporal analysis, and even fewer offer interactive, filterable time-based views for end users. While most dashboards highlight static or aggregated trends, real-time and user-filtered temporal exploration remains underdeveloped in the literature. This presents an important opportunity to enhance interpretability and user engagement through dynamic time-series visualization, particularly in Reddit-focused environments.

5 METHODOLOGY

5.1 Fine-tuned hateBERT Model

To further enhance performance, I used a fine-tuned HateBERT using two complementary datasets [9] [5]. The Davidson dataset contains 24,000 tweets annotated as hate speech, offensive language, or neither. Its three-class labeling forces the model to learn the subtle yet crucial distinctions between overt hate speech and general toxicity—a distinction that is often blurred in politically charged Reddit threads. ETHOS complements this with a dual annotation scheme: both binary and multi-label classifications across eight categories such as race, religion, and gender. Collected from Reddit and YouTube using active learning strategies, ETHOS prioritizes edge cases and minority-class instances, deepening the model's grasp of intersectional hate and reducing bias toward more explicit or common slurs. As a result of this fine-tuning, the model achieved an accuracy of 0.8628 with a validation loss of 0.3388.

5.2 Reddit Data Collection

To develop the dashboard central to this research, I focused on extracting and analyzing several key categories of data. Subreddit data includes variables such as subscriber count, visibility status, and NSFW designation. These communities vary in accessibility: subreddits may be public, restricted, private, or archived. Some are flagged as "over 18," indicating that they contain NSFW (Not Safe for Work) material. Post data encompasses various types of submissions, including text posts (common in discussion-oriented subreddits), images, videos, external links, and polls.

Posts typically feature a title and body, along with engagement metrics such as upvotes, downvotes, a comments section, and optional flairs. Flairs are labels assigned either to posts or users, offering an additional layer of categorization within a subreddit. Posts may also be awarded virtual badges—known as awards—purchased with Reddit Coins (a form of virtual currency bought with real money) and used by users to express appreciation for high-quality or entertaining content. The metadata for each post includes the associated URL, which may be external, internal, or empty in the case of self-posts. Additionally, some posts are cross-posts, meaning they reshare content from one subreddit to another.

The comments system on Reddit enables users to engage in nested, threaded discussions, forming a tree-like structure where replies appear beneath the comment they are responding to. Comments, like posts, can be upvoted, downvoted, and awarded by other users. Finally, user data includes karma, a reputation metric that quantifies the popularity of a user's posts and comments. Karma is split into two categories: post karma and comment karma. Users may also have paid memberships that grant access to premium features or act as moderators—volunteers who oversee subreddit activity, enforce rules, remove content, and ban users when necessary.

This overview provides essential context for understanding the structure of Reddit and the specific data types utilized in the development of the dashboard, which forms the foundation of the analysis presented in this research.

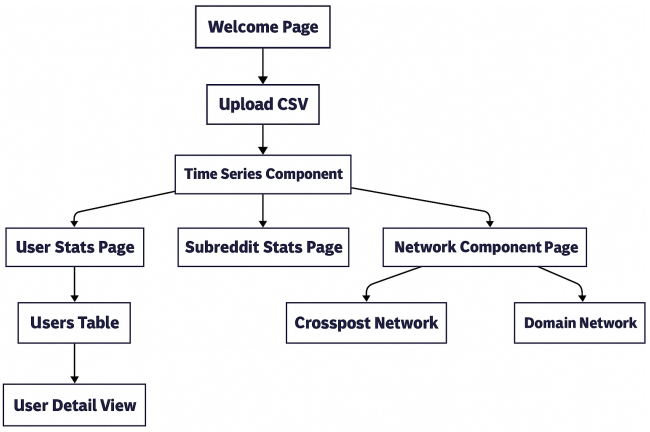

Fig. 1. Flow chart Weaverlet Dashboard.

5.3 Weaverlet Flow of the Dashboard

In our dashboard, Weaverlet is used throughout the entire implementation. The application flow is illustrated in Figure 1. All navigation—whether to User Stats, Subreddit Stats, or the Network View—is handled via the **RedirectComponent**. This makes adding a new page like "Top Domains" simple: write a new component class, add an entry to the router dictionary, wire a button to a redirect, and the integration is complete.

Time-Series page is implemented as a single component that orchestrates subcomponents: a Filter panel, a Graph panel, and navigation buttons. Each subcomponent manages its own IDs and callbacks, while the parent page handles their coordination.

For example, the Filter panel communicates with the Graph panel via a standard Dash callback, since they operate closely together. However, when a CSV file finishes loading on the Welcome page, a signal is emitted to notify all components that depend on fresh data.

Weaverlet enables developers to stay entirely within Python while simulating advanced front-end development patterns commonly associated with frameworks like React.

5.4 Dataset and pre-processing

The dataset was collected using Reddit's free public API. The retrieved data was categorized into submissions, comments, users, and the number of crossposts (see table 1). Both the submissions and comments were analyzed using the fine-tuned BERT model described in Section 5.1.

Table 1. Dataset Statistics

Category	Count	Notes
Subreddits	12	Politics-related subreddits
Posts	9,580	Total number of posts
Hate Speech (Posts)	1,013	Posts labeled as hate
Comments	51,231	Total number of comments
Hate Speech (Comments)	5,399	Comments labeled as hate
Users	20,381	Unique user IDs
Cross-posts	355	Total number of cross-posts

The classified Reddit data was divided into two categories: submissions and comments. For comments, the body text was directly used for classification. In the case of submissions, the title was concatenated with the body text, if present. Most posts are self-text, meaning they consist only of text and are suitable for direct analysis. However, a significant number of posts include only videos, images, or external links, which results in the body being empty. In such cases, only the title of the submission is passed through the classification process, as it is the only available textual content.

After this initial structuring, the text undergoes a preprocessing pipeline. This process ensures the content is properly encoded and has a well-formed Unicode representation. The next step involves cleaning the text by removing markdown syntax, bullet points that may prefix lines, and all forms of hyperlinks. Finally, emojis are converted into descriptive text representations, enabling the model to interpret their semantic and emotional content more effectively. This step helps the model infer tone and sentiment from emojis, which are often strong indicators of user intent(see table 2).

Table 2. Text Before and After Processing

Stage	Text
Before processing	`> * Wow!! Elon MuskÃćâĆňâĂİcritics say ï£ï£ï£ï check this out [here](https://example.com) <https://another.com> () []`
After processing	`Wow!! Elon Musk–critics say :cat: check this out here`

6 SYSTEM DESIGN AND IMPLEMENTATION

6.1 Time Series Analysis

The time-series analysis is the core and most significant feature of the dashboard. It is built upon the dataset that has been gathered and classified, with a primary focus on content labeled as hate. The objective is to provide an interactive graph that allows users to explore this data across various dimensions. Users can filter the graph (see figure 2) by subreddit, and then choose whether to analyze submissions (posts) or comments. Some subreddits apply custom flair to their content, for example, the subreddit AskTrumpSupporters assigns flairs to users such as "Supporter" or "NonSupporter." Posts can also be flaired according to the subreddit's categories, in this case, the topic is labeled "Elections 2024".

Additionally, users can filter data by specific date ranges. Once the filters are applied, the graph dynamically updates to reflect the selected criteria (see figure 3). Each point on the graph represents one day. Clicking a point reveals aggregated engagement metrics for all posts from that day: total number of posts, total number of comments under those posts, total upvotes, and average upvote ratio.

Figure 3 shows that posts filtered by the "Election 2024" flair experienced noticeable surges during the peak of the campaign period. Following the announcement of Trump's election victory, the graph

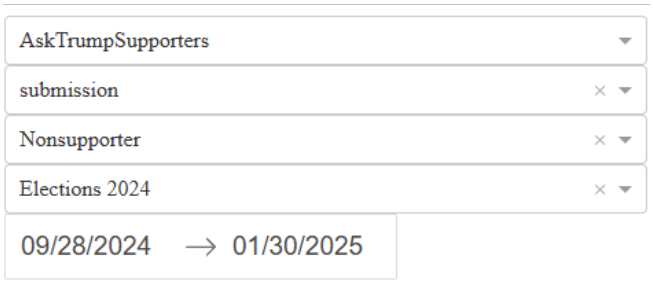

Fig. 2. Time Series Filters.

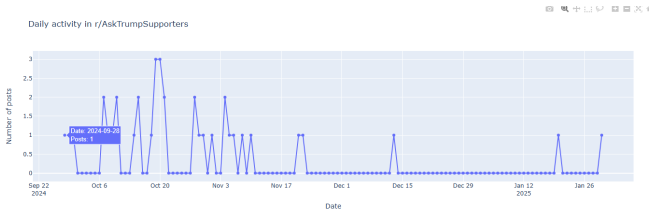

Fig. 3. Daily activity in r/AskTrumpSupporters based on filters

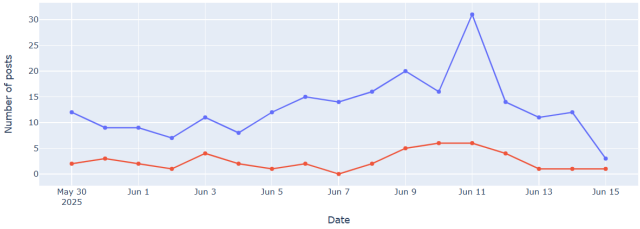

Fig. 4. Daily activity for hateful(red line) and non-hateful(blue line) posts

stabilizes, with significantly fewer mentions of the election in subsequent months.

Another insightful visualization is presented in Figure 4, which displays the red line representing hateful posts and the blue line representing non-hateful posts. The figure is based on data from the r/trump subreddit. Additionally, the graph focuses exclusively on posts authored by users labeled as "ULTRA MAGA" as this group was found to contribute the highest volume of hateful content across the dataset. Both trajectories are mostly stable throughout the observation window but rise sharply on the 11th of June, with hateful and non-hateful counts nearly doubling. This synchronous spike suggests that an external catalyst, such as a political announcement or an international event, triggered an intense surge of activity within the subreddit.

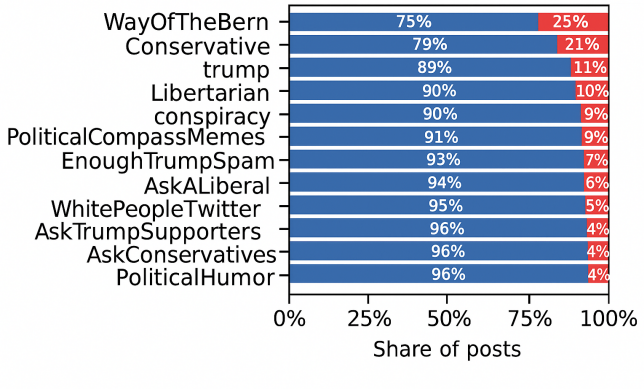

Fig. 5. Hateful and non-hateful percentage of posts by subreddit

Fig. 6. Hateful and non-hateful percentage of posts by subreddit

These patterns underscore the close link between offline events and online discourse and illustrate how flair-based filtering exposes meaningful behavioral signals.

6.2 Subreddits Behaviour

Our dataset covers twelve politically themed subreddits that span the U.S. ideological spectrum and showcase several distinct conversational styles. Although each community centers on politics, their posting conventions differ substantially.

r/AskALiberal, r/AskConservatives, and r/AskTrumpSupporters function as moderated, text-only Q&A forums whose threads are heavily flaired and categorised.

r/PoliticalCompassMemes, r/WhitePeopleTwitter, and r/PoliticalHumor rely on humor delivered through images, GIFs, and short videos.

r/WayOfTheBern, r/Conservative, and r/EnoughTrumpSpam revolve around discussion threads, most of which link to external news articles. The remaining communities adopt a discussion-centered format but offer a broader mix of content. This deliberately diverse collection enables more nuanced insights into how different political communities on Reddit express ideological discourse.

Figure 5 indicates that the vast majority of subreddits in our sample contain fewer than 10 % hateful posts. Three communities break this pattern—r/WayOfTheBern, r/Conservative, and r/trump—but the first two are especially notable: hate-speech prevalence reaches 25 % in r/WayOfTheBern and 21 % in r/Conservative, whereas r/trump remains just above the 10 % threshold.

Figure 6 examines the timeline of the r/Conservative subreddit in greater detail. This timeline spans 3 June – 14 June and comprises 969 posts, of which 21 % are classified as hateful. For most of the

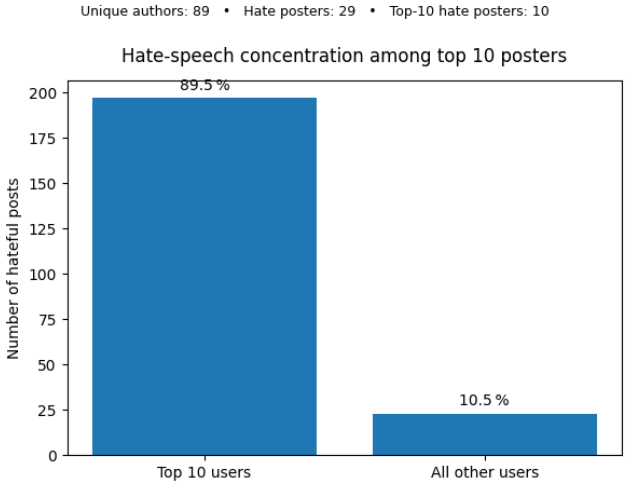

Fig. 7. Hate-speech concentration among top 10 posters in WayOfTheBern subreddit

interval, there is little correlation between hateful and non-hateful activity, yet both series surge on 13 June. Comparable spikes appear in other subreddits, r/WayOfTheBern and r/Libertarian, suggesting a network-wide response to the same external trigger. Two earlier dates show similar behavior: 6 June and 5 March. Each peak aligns with real-world events, Israel's strike on Iran (13 June), mass protests in Los Angeles (6 June), and the announcement of new U.S. visa-restriction policies (5 March). These correlations reinforce the dashboard's capacity to reveal how offline events can rapidly amplify online hate speech across Reddit's political communities.

6.3 Influencer Analysis

The next key component of the dashboard is user analysis. This module identifies the top 10 influential users within a given subreddit, ranking them based on the number of posts they have contributed. This ranking is visualized in Figure 10 for the r/WayOfTheBern and serves as a starting point for deeper insights into user behavior and engagement patterns.

Using the same subreddit, additional visualizations were generated that clarify how hate speech is distributed among redditors. Figure 7 shows that the ten most active users account for 89.5 % of all hateful content, leaving only 10.5 % to the remaining contributors. These percentages are computed from a pool of 89 distinct authors, 29 of whom used hateful language at least once, underscoring the influence a small group can exert over the tone of the community.

Figure 8 examines individual behavior more closely. Its horizontal axis records each user's total post count, while the vertical axis shows how many of those posts were hateful; every bubble represents a redditor, and the bubble's area increases with the author's overall activity. The chart reveals that most users posted fewer than ten times and produced either zero or one hateful message. By contrast, one author published roughly 170 posts, more than 80

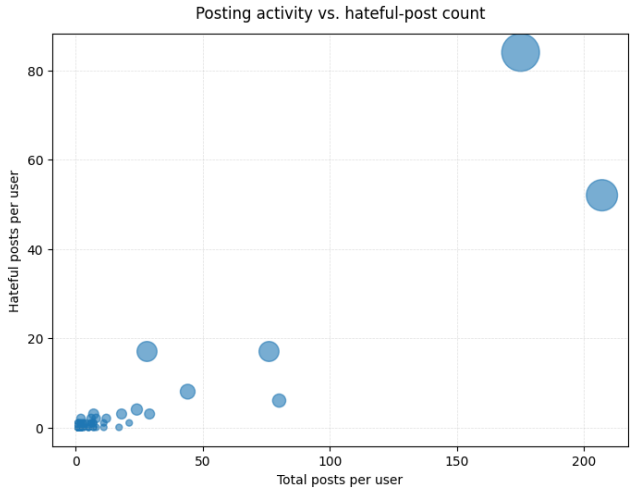

Fig. 8. Posting activity vs. hateful-post count in WayOfTheBern subreddit

of which were hateful, and another produced about 210 posts with approximately 50 hateful messages. These two outliers demonstrate how a handful of participants can drive a disproportionate share of hateful discourse.

As shown in the rank, there are typically two or three users who contribute a disproportionately high number of posts compared to the rest of the community.

For each identified user, several engagement metrics are calculated to help quantify their influence and reputation. These include the total number of active days, the average number of posts per day, the average score or awards received per post, and the average karma per comment. These metrics collectively offer a more nuanced view of each user's role and presence within the community (see figure 10).

The dashboard enables users to select a specific influencer and examine detailed profile metrics. Below, we analyze User 1, the most prolific contributor in r/WayOfTheBern.

User 1

- First day of activity: 2025-05-27
- Last day of activity: 2025-06-14
- Account created: 2015-06-21
- Total karma: 1,011,437
- Link karma: 796,871
- Comment karma: 214,566
- Is moderator? True
- Has Reddit Gold? False
- Verified e-mail? True
- Reddit employee? False

The details for User 1 indicate that they have accumulated a high amount of link karma, which typically corresponds to a large number of submitted posts. This suggests that the user has contributed content frequently and that their posts have been well-received by the community, implying a strong reputation. Additionally, the

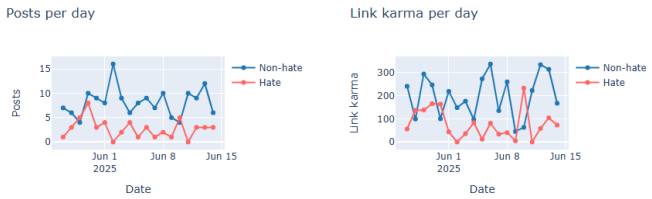

Fig. 9. Graphs showing daily activity for user 1

user holds moderator status and has a verified email address, which reduces the likelihood that the account is operated by a bot.

Moreover, the dashboard includes a visualization of the user's daily activity alongside the karma they received over time shown in figure 9. The first chart traces two lines: a blue line for non-hate posts and a red line for hate posts. The user-contributed with both types of content throughout the observation window, although non-hateful posts were more common overall. Minor upticks in hate-speech output appeared around 30 May, when the user produced roughly eight hateful posts, and 10 June, with 5 posts.

The other chart records the karma accumulated each day, calculated as the total score of that day's submissions. Karma is an indicator of how positively the subreddit responds to the user's contributions and its curve rises sharply on 30 May and 10 June, mirroring the spikes in hate-post counts. By contrast, karma tied to non-hateful content fluctuates far more widely because individual posts sometimes resonate with the community. This dynamic was observed on 6 June, when possibly some popular submission drove the day's karma sharply upward even though the user posted no more than usual.

The two graphs reveal that the user posted consistently, and there is a clear correlation between posting frequency and the amount of karma obtained.

Together, these visualizations provide a comprehensive picture of each user's activity, consistency, and social impact within the subreddit. This level of analysis supports a deeper understanding of how individual users contribute to the propagation and engagement of hateful content.

7 CONTRIBUTIONS

The dashboard introduced in this paper fills a crucial gap in our understanding of how harmful content spreads on Reddit. Whereas earlier tools such as MANDOLA [10] and HaterNet [11] concentrated on Twitter data, our system focuses solely on Reddit.

Our first contribution is the development of a novel time-series analysis module that allows day-to-day temporal analysis. Users can filter datasets by subreddit, flairs, and date ranges to visualize day-level fluctuations in hateful content. Unlike existing dashboards that typically provide only weekly or monthly aggregation, this platform reveals nuanced daily patterns and correlations with external events.

Second, we offer a detailed influencer analysis that identifies prolific users and evaluates their impact through engagement metrics such as karma scores and posting frequency. Our case study of the subreddit r/WayOfTheBern highlights how a small group of active

contributors disproportionately influences the propagation of hate speech, an insight that previous tools have rarely addressed in depth. Lastly, the dashboard showcases a pioneering application of Weaverlet, Python-native framework. This research marks the first known instance of deploying Weaverlet for large-scale social media analytics, proving its effectiveness in managing complex visualization workflows.

8 LIMITATIONS

The system is novel and introduces new concepts; however, it also has drawbacks and areas for improvement. The dataset currently used is obtained via Reddit's free API, which limits data retrieval to a maximum of 1,000 submissions per subreddit. This is a relatively small sample size—especially for large subreddits with high activity levels—where 1,000 submissions may represent only four to five days' worth of data. This constraint makes it unsuitable for meaningful temporal analysis.

Furthermore, while the system effectively detects hate speech in text, Reddit hosts a diverse range of content formats, including images, videos, memes and external links. These media types are currently outside the scope of the system's detection capabilities. As a result, hate speech embedded in non-textual content cannot be analyzed at this stage.

Another limitation is related to scalability. Although the current implementation handles tens of thousands of entries within a few seconds, processing significantly larger datasets—containing millions of records—may lead to considerable delays or even system crashes, posing a major reliability issue.

9 FUTURE WORK

For future work, the system should be expanded to handle larger datasets and better adapted to analyze a wider variety of content types. Reddit includes not only text but also images, videos, memes and external links, all of which can contain hateful content that the current system cannot detect. Incorporating computer vision models would enable the analysis of such multimedia content, significantly broadening the system's detection capabilities.

In the context of Reddit, temporal analysis conducted before and after subreddit bans could be used to evaluate the effectiveness of these moderation actions. Such analysis would help determine whether banning a subreddit leads to a measurable reduction in hate speech over time. To achieve this, historical data should be obtained using tools like Pushshift, which allows access to older Reddit posts and comments.

Another valuable area for future development is transitioning the system to real-time processing. This would allow for the continuous monitoring of hate speech as it emerges, offering more immediate insights into the dynamics of online discourse. Real-time analysis would also enhance the system's practical utility by enabling quicker intervention and more timely responses from moderators or platform administrators.

10 CONCLUSION

This paper presented a Weaverlet-based interactive dashboard designed for fine-grained, temporal exploration of hate speech on Reddit. By integrating a newly fine-tuned HateBERT classifier with Reddit's rich platform metadata, the system provides filterable time-series visualizations, influencer analysis, and cross-post insights—entirely within a Python-native, modular framework. Applied to a politically oriented dataset centered around the 2024 U.S. presidential election, the dashboard revealed four key findings. (RQ1) Hate speech volume tends to rise in sharp bursts, often coinciding with major offline events. (RQ2) Temporal patterns emerge across ideologically similar subreddits, which often exhibit synchronous spikes in hate-speech activity in response to the same external events. (RQ3) Content production is highly concentrated within a few prolific users.

Beyond answering these core research questions, the platform offers two key benefits for Reddit moderators and the broader public. First, by synthesizing engagement trends, temporal dynamics, and user-level concentration into a unified interface, it delivers actionable insights that streamline moderation workflows and prioritize interventions where they are most needed. Second, by overlaying hate-speech trends onto timelines of political events, the dashboard fosters a deeper public understanding of how sociopolitical developments influence on-platform discourse, thereby informing both community governance and scholarly analyses of online political dynamics.

REFERENCES

[1] Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access* 13 (2025), 20871–20892.

[2] Saad Almohaimeed, Saleh Almohaimeed, Ashfaq Ali Shafin, Bogdan Carbunar, and Ladislau Bölöni. 2023. THOS: A benchmark dataset for targeted hate and offensive speech. (Nov. 2023). arXiv:2311.06446 [cs.CL]

[3] Arthur T. E. Capozzi, Viviana Patti, Giancarlo Ruffo, and Cristina Bosco. 2018. A Data Viz Platform as a Support to Study, Analyze and Understand the Hate Speech Phenomenon. In *Proceedings of the 2nd International Conference on Web Studies (WS.2 2018)*. ACM, 28–35. https://doi.org/10.1145/3240431.3240437

[4] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. HateBERT: Retraining BERT for Abusive Language Detection in English. https://doi.org/10.48550/ARXIV.2010.12472

[5] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. https://doi.org/10.48550/ARXIV.1703.04009

[6] Alberto García-Robledo. [n. d.]. *Weaverlet*. https://github.com/observatoriogeo/weaverlet

[7] Taehee Kim and Yuki Ogawa. 2024. The impact of politicians' behaviors on hate speech spread: hate speech adoption threshold on Twitter in Japan. *J. Comput. Soc. Sci.* (April 2024).

[8] Karsten MMller and Carlo Schwarz. 2018. Making America hate again? Twitter and hate crime under trump. *SSRN Electron. J.* (2018).

[9] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems* 8, 6 (Jan. 2022), 4663–4678. https://doi.org/10.1007/s40747-021-00608-2

[10] Demetris Paschalides, Dimosthenis Stephanidis, Andreas Andreou, Kalia Orphanou, George Pallis, Marios D. Dikaiakos, and Evangelos Markatos. 2020. MANDOLA: A Big-Data Processing and Visualization Platform for Monitoring and Detecting Online Hate Speech. *ACM Transactions on Internet Technology* 20, 2 (March 2020), 1–21. https://doi.org/10.1145/3371276

[11] Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and Monitoring Hate Speech in Twitter. *Sensors* 19, 21 (Oct. 2019), 4654. https://doi.org/10.3390/s19214654

[12] Gilang Rabbani and Lili Ayu Wulandhari. 2024. Hate Speech Classification Using Mixed Language Feature Modification and Machine Learning Approach. In *2024 11th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. IEEE, 735–741. https://doi.org/10.1109/eecsi63442.2024.10776321

[13] Gil Ramos, Fernando Batista, Ricardo Ribeiro, Pedro Fialho, Sérgio Moro, António Fonseca, Rita Guerra, Paula Carvalho, Catarina Marques, and Cláudia Silva. 2024. A comprehensive review on automatic hate speech detection in the age of the

transformer. *Social Network Analysis and Mining* 14, 1 (Oct. 2024). https://doi.org/10.1007/s13278-024-01361-3

[14] Kamal Safdar, Shibli Nisar, Waseem Iqbal, Awais Ahmad, and Yawar Abbas Bangash. 2023. Demographical Based Sentiment Analysis for Detection of Hate Speech Tweets for Low Resource Language. *ACM Transactions on Asian and Low-Resource Language Information Processing* (Aug. 2023). https://doi.org/10.1145/3616867

[15] Jan Sawicki, Maria Ganzha, Marcin Paprzycki, and Amelia Badica. 2022. Exploring Usability of Reddit in Data Science and Knowledge Processing. *Scalable Computing: Practice and Experience* 23, 1 (April 2022), 9–22. https://doi.org/10.12694/scpe.v23i1.1957

[16] Paras Sheth, Tharindu Kumarage, Raha Moraffah, Aman Chadha, and Huan Liu. 2023. PEACE: Cross-platform hate speech detection- A causality-guided framework. (June 2023). arXiv:2306.08804 [cs.CL]

[17] Amaury Trujillo and Stefano Cresci. 2022. Make Reddit great again: Assessing community effects of moderation interventions on r/the_Donald. *Proc. ACM Hum. Comput. Interact.* 6, CSCW2 (Nov. 2022), 1–28. https://doi.org/10.1145/3555639

A AI STATEMENT

During the preparation of this work the I used *OpenAI ChatGPT exclusively* to refine wording,improving clarity, coherence, and academic tone, without generating any new content. The tool was also used to generate code fragments and to debug existing code. After each use, the I carefully reviewed and edited all outputs and accept full responsibility for the final text, code, and overall content of this work.

B ADDITIONAL FIGURES

User Statistics Dashboard

Select a Subreddit:

WayOfTheBern

	User	Posts (NH)	Posts (H)	% Hate	Posts/day (NH)	Posts/day (H)	Avg score (NH)	Avg score (H)	Awards/post (NH)	Awards/post (H)	Karma/comment (NH)	Karma/comment (H)
○	User 1	155	52	25.1%	8	3	24	28	0	0	1384	4126
○	User 2	91	84	48.0%	5	5	12	10	0	0	6	6
○	User 3	74	6	7.5%	6	2	32	22	0	0	104	1289
○	User 4	59	17	22.4%	4	2	17	31	0	0	503	1745
○	User 5	36	8	18.2%	2	1	14	6	0	0	274	1233
○	User 6	26	3	10.3%	2	1	11	31	0	0	505	4375
○	User 7	11	17	60.7%	1	2	18	23	0	0	213	138
○	User 8	20	4	16.7%	2	1	11	8	0	0	9923	49616
○	User 9	20	1	4.8%	2	1	22	5	0	0	6449	128982
○	User 10	15	3	16.7%	2	1	65	26	0	0	957	4784

Fig. 10. Ranking of the top 10 most prolific users in r/WayOfTheBern subreddit