Data Collection and Analysis of Health-Related Communities on Reddit

January 9, 2025

1. Introduction

This report documents the process of collecting and analyzing data from health-related communities on Reddit. The primary objective of this study is to explore user engagement patterns, trends in health discussions, and key insights within these online forums. By examining posts from various subreddits, we aim to understand how users interact and share information related to health topics.

2. Data Collection Process

2.1 Sources of Data

- **Communities Targeted:** The data was collected from 88 health-related subreddits, including but not limited to:
  - r/ADHD
  - r/AlternativeHealth
  - r/VentureBiotech
  - r/Cancer
  - r/infertility
  - ...
- **Justification:** These subreddits were selected for their active user base and relevance to various health topics, ranging from mental health to emerging health technologies.

2.2 Tools and Techniques

- **Web Scraping Techniques:** Web scraping methods were employed for data extraction from Reddit, utilizing Python libraries like `BeautifulSoup` and `Requests` for efficient extraction.
- **Data Handling Libraries:** Libraries such as `Pandas` and `NumPy` were used for preprocessing and cleaning the data.

- **Storage Formats:** The data was stored in CSV format for ease of access and further analysis.

2.3 Challenges Encountered

- Dealing with deleted posts and authors (`[deleted]`) required careful filtering to maintain data accuracy.
- Managing inconsistencies in data formats, such as mixed types in the `createdAt` field.
- API limitations, including rate limits and incomplete data retrieval, necessitated multiple runs to ensure comprehensive collection.

2.4 Data Overview

- **Number of Posts:** A total of 314,538 posts were collected across all subreddits.
- **Time Frame:** Data spans from 2008 to January 2015.
- **Fields Captured:**
  - `authorId`: Unique identifier for the author.
  - `postText`: Content of the post.
  - `createdAt`: Timestamp of when the post was created.
  - `subredditName`: Name of the subreddit.
  - `commentCount`: Number of comments on the post.
  - `postTitle`: Title of the post.
  - `commentsLink`: URL to the comments section.
  - `collectedAt`: Timestamp of when the post was collected.

3. Results

3.1 Key Metrics

- **Total Number of Posts:** 314538 Posts across all topics.

3.2 Post Distribution

- **Time Distribution:**
  - Posts per year and month were analyzed to identify trends over time.
- **Subreddit Analysis:**
  - Number of posts per subreddit.
  - Average word count per post for each subreddit.

4. Data Visualisation

Posts Per Year

This figure illustrates the distribution of posts over the years, showing the trends and changes in the number of posts made on Reddit over time.

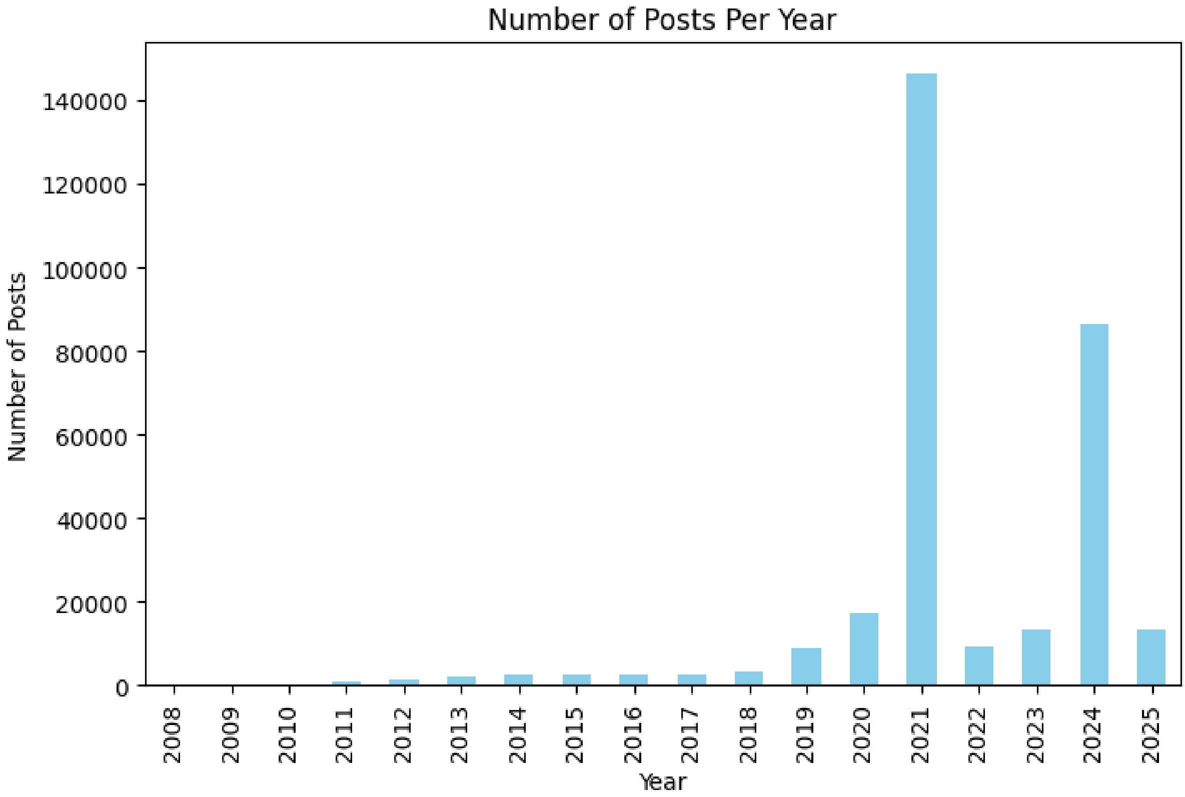

Figure 1: Distribution of Posts Per Year

Posts Per Month

This figure displays the distribution of posts across different months, highlighting any seasonal trends or monthly patterns in posting behavior.

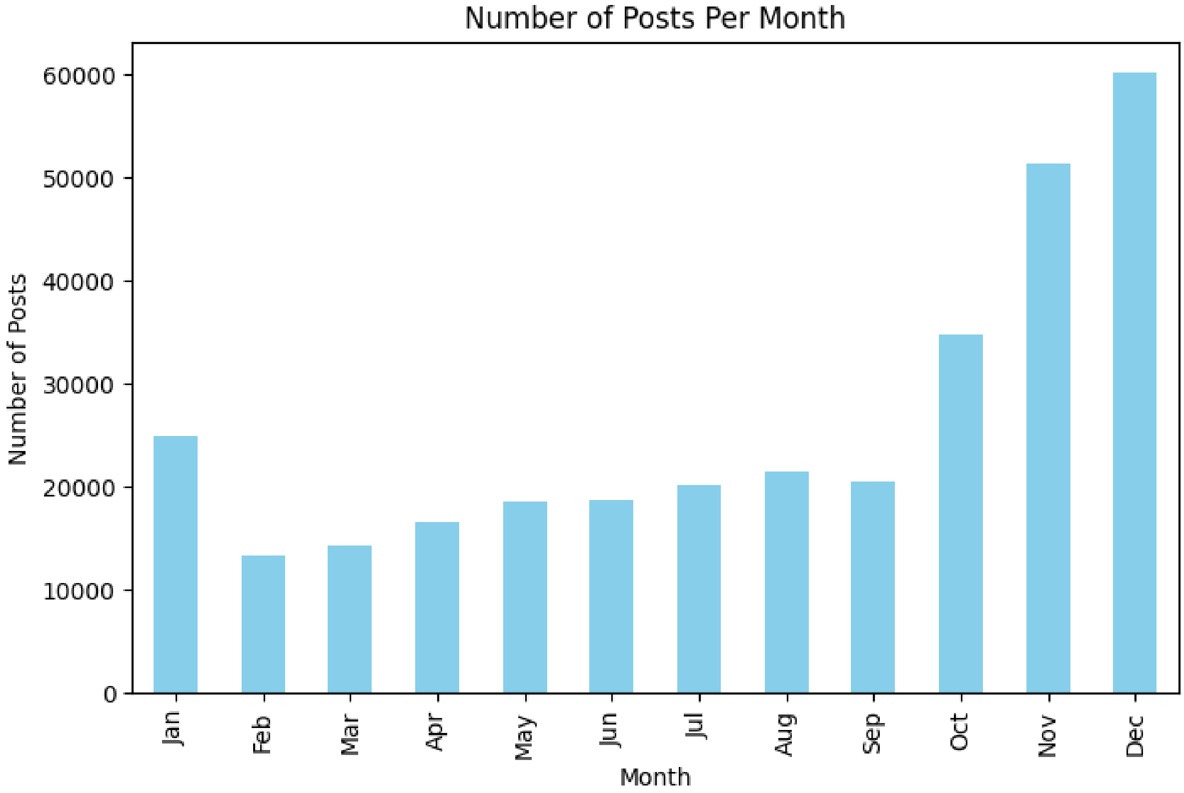

Figure 2: Distribution of Posts Per Month

Posts Per Author

This figure represents the distribution of posts made by each author, offering insights into the activity level and contributions of individual users in the subreddit communities.

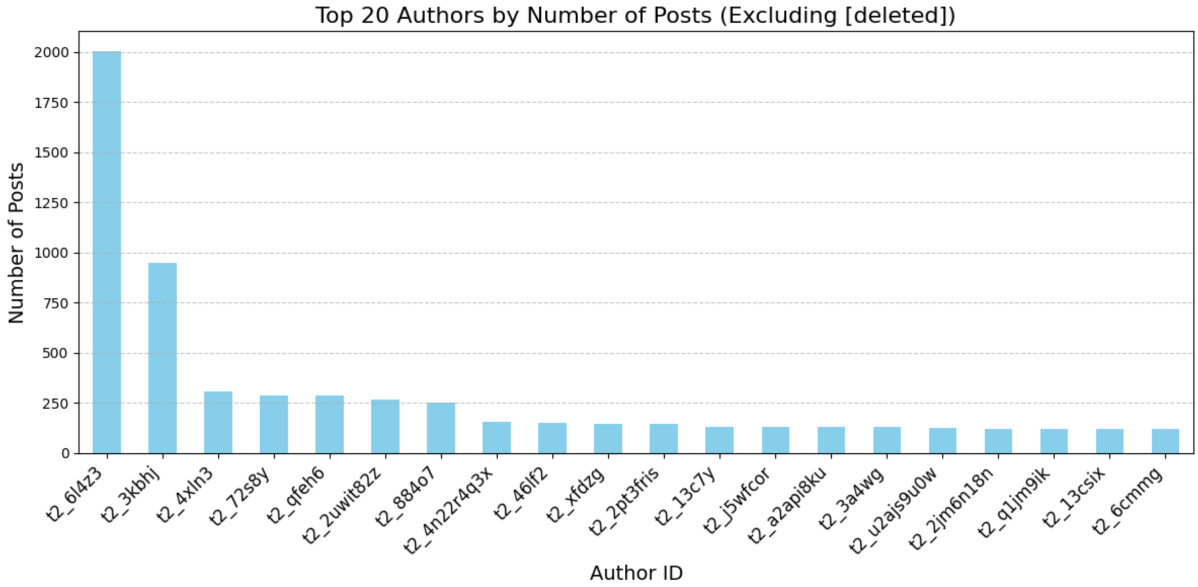

Figure 3: Distribution of Posts Per Author

Posts Per Topic

This figure displays the distribution of posts across different topics (or communities), highlighting the varying levels of activity and engagement in each community over time.

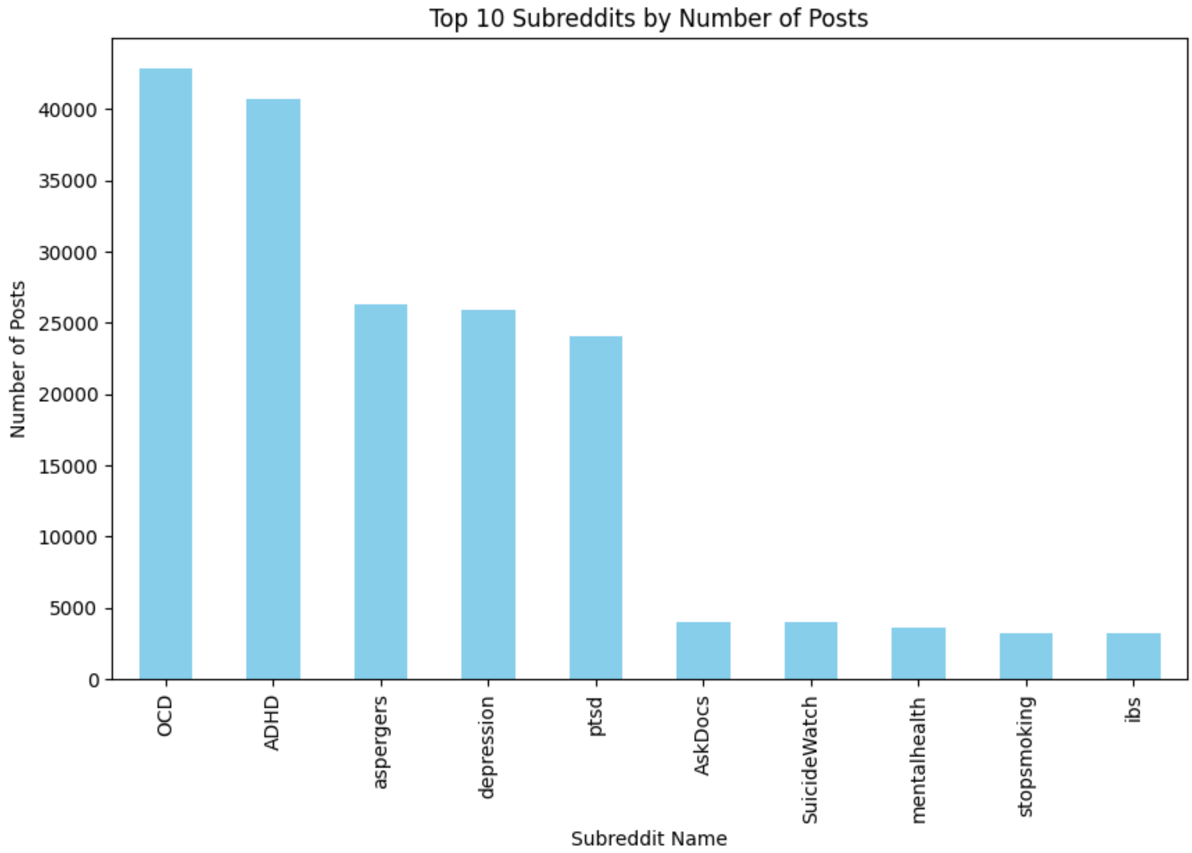

Figure 4: Distribution of Posts Per Topic

5. Conclusion

The data collected from Reddit health-related communities provides valuable insights into user engagement and health topic discussions. By analyzing trends, user contributions, and community activity, we can better understand the role of online forums in health information dissemination.