Social Media Engagement Patterns in India: A Visual Analytics Study

Group 1
Somya Kumar, Someshwar Singh, Adarsh, Mayank Agrawal, Neela,
Sudeep Chahlia, Prabhakar Raj, Gaurav Bohra, Peeyush Sahu

CS661 - Big Data Visual Analytics
July 2025

Contents

1 Project Repository

The complete codebase, dashboards, and documentation for this project are available on GitHub:
Github Link
Direct Dashboard Link

2 Introduction

The exponential rise of social media platforms has revolutionized how individuals communicate, consume information, and form opinions. In the Indian context, this transformation is especially pronounced due to increasing internet penetration, affordable mobile access, and a digitally active youth population.

This project investigates the behavioral, psychological, and societal implications of social media usage in India, while also drawing comparisons with global trends. Through a combination of interactive dashboards and data visualizations, we explore multiple dimensions of social media engagement — including platform popularity, content preferences, regional usage disparities, toxicity levels, misinformation spread, and signs of digital addiction.

Our aim is to present clear, insightful visual narratives that are accessible to both technical and non-technical stakeholders. By combining publicly available datasets, custom survey inputs, and advanced visualization tools like Plotly and D3.js, this project seeks to inform platform users, educators, and policymakers about emerging trends and their potential consequences.

3 Objective

The primary objective of this project is to analyze and visualize social media usage patterns and their broader impact on Indian society using interactive data-driven dashboards. Specifically, the project aims to:

- Identify the most popular social media platforms across different demographics.
- Examine regional and content-based preferences among users.
- Assess behavioral trends such as engagement frequency, entertainment consumption, and brand interactions.
- Analyze platform-specific issues like toxicity, misinformation, and emotional polarization.
- Investigate indicators of digital addiction and their correlation with health and lifestyle factors.
- Present insights through intuitive, interactive dashboards using HTML, CSS, JavaScript, and visualization libraries such as Plotly and D3.js.

These objectives are designed to facilitate a deeper understanding of the digital behaviors shaping communication, information flow, and well-being in a rapidly digitizing nation.

4 Tools and Technologies Used

To develop and deploy the interactive dashboards and perform data-driven analysis, we utilized a combination of front-end web technologies, data processing tools, and visualization libraries:

- **D3.js**: For building dynamic and responsive data visualizations, such as the bump chart and radar plots.
- **Plotly.js**: Used to create interactive and animated visualizations including violin plots, correlation heatmaps, and bar charts.
- **HTML & CSS**: For structuring and styling the visual analytics dashboards across multiple modules.
- **Google Sheets**: Used for collaborative data entry and simple chart prototyping.
- **Microsoft Excel**: For initial data aggregation, preprocessing, and exploratory summary statistics.
- **Google Colab Notebooks (Python)**: Used for data cleaning, statistical analysis, and preparing JSON/CSV formats for visualization.

This integrated toolchain enabled us to efficiently transform raw datasets into intuitive, interactive, and engaging visual narratives.

5 Visualization Tasks and Insights

5.1 Worldwide Ranking 2025

5.1.1 Task Overview

This visualization displays the global distribution of social media traffic rankings in 2025, using search engine referral tracking data. The platform identifies when users click through from a search engine results page (SERP) to a site that has Statcounter installed. These interactions reveal which platforms dominate in different countries and across device types.

5.1.2 Information Represented

- A world map visualization color-coded by the most visited social platform per country.
- Device-specific filters: All Devices, Mobile Only, Desktop Only, and Tablet Only.
- Ranking breakdowns into Top 1st, 2nd, and 3rd most referred platforms.
- Legend for interpretation.
- Based on search engine referrals, not search query counts — i.e., the action of clicking on a site from SERP.

5.1.3 Observed Trends

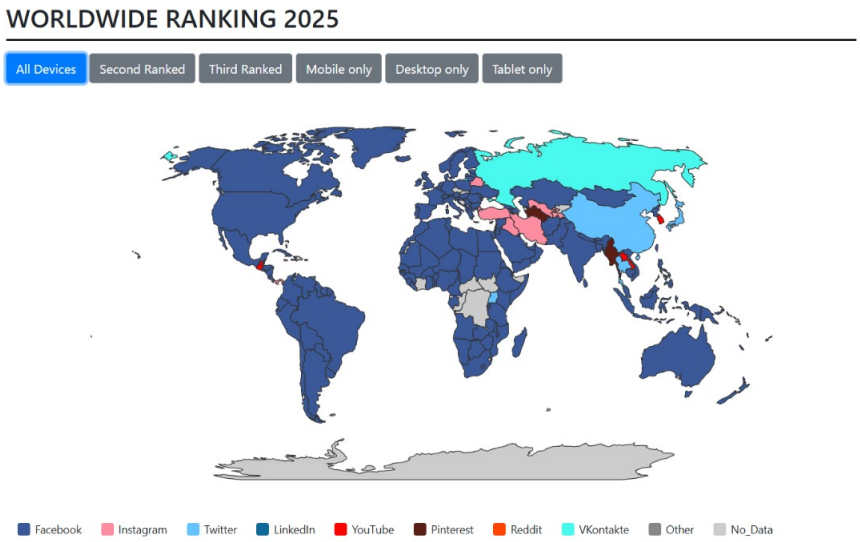
- Facebook and WhatsApp top the global charts in search engine referrals, accounting for over 40% of tracked referrals globally on All Devices.
- WhatsApp emerges as the leading platform in mobile-only environments — especially across India, South America, and parts of Africa.
- LinkedIn and Twitter appear more often in desktop-dominated regions like the U.S., Canada, and the U.K., reflecting professional usage.
- In tablet-based views, YouTube and Facebook share dominance, suggesting use for long-form content and visual browsing.

5.1.4 Inference and Insights

- Mobile-first markets like India show a heavier reliance on messaging apps like WhatsApp for both personal and business communications, based on click-through patterns.
- Search engine referral patterns provide more accurate engagement metrics than pure traffic, as they reveal user intent and action — not just passive exposure.
- Emerging economies contribute a significant portion of social engagement via mobile search, while mature economies have more diversified device usage.
- The Top 3 platform view highlights regional competitors like WeChat in China, showing how walled ecosystems still generate high engagement.

5.1.5 Real-World Implications

- Advertisers should emphasize platforms with higher SERP click-through rates, especially mobile-centric ones in developing countries.
- Tech companies need to ensure their social media landing pages are SEO-optimized to capture search-driven user traffic.
- Policy researchers can gauge digital influence using referral-based rankings, not just app downloads.
- Platform strategies should vary per device type and country, as the referral behavior offers a better proxy of conversion or interest.



Worldwide Rankings 2025

5.2 Social Platforms Traffic Ranking of India

5.2.1 Task Overview

This bump chart visualizes the monthly evolution of social media platform rankings in India, based on search engine referral traffic tracked via Statcounter. Each line traces how a platform has moved up or down in referral rank, providing a historical picture from 2009 to 2025.

5.2.2 Information Represented

- **X-axis:** Time series (from 2009 to 2025).
- **Y-axis:** Traffic rank (1 = highest referrals).
- Each line represents a platform's rank over time based on click-throughs from search engines.
- Reveals user preference shifts through platform ranking dynamics.

5.2.3 Observed Trends

- Facebook led in search engine referrals from 2009 to approximately 2017, but has since seen a steady decline, dropping below 3rd position by 2024.
- YouTube rose sharply post-2018 and by 2023–2025 became the #1 platform, with a significant 25%+ share of search engine referral traffic.
- Instagram surged between 2020–2024, suggesting growing interest in visual and short-form content.
- Telegram and Pinterest showed spikes during specific periods, like the 2021 privacy debate, but later stabilized or dropped.
- LinkedIn maintained steady mid-tier ranks, indicating stable usage in niche professional circles.

5.2.4 Inference and Insights

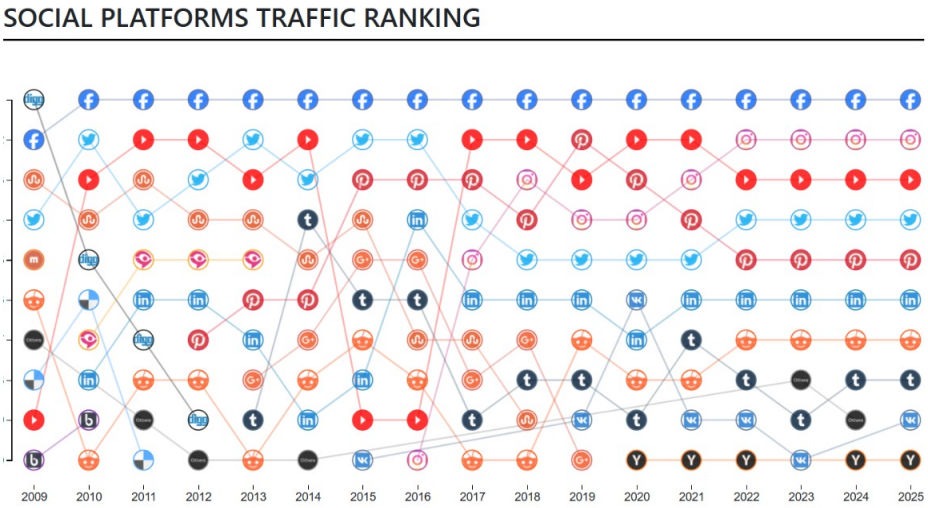
- The data being referral-based means these rankings show which platforms people are actively searching and clicking on, not just using.
- A platform's rank drop (e.g., Facebook) could indicate declining user curiosity or reduced content visibility in SERPs.
- Platforms gaining ground (like YouTube) may have better SEO optimization or content that's actively sought out, including educational, music, and entertainment.
- Peaks in Telegram usage coincide with periods of political activity or platform bans, highlighting referral traffic as a lens for socio-digital behavior.

5.2.5 Real-World Implications

- Brand managers and content strategists should invest more in platforms that are being actively discovered via search (like YouTube or Instagram).
- Referral rankings indicate not just passive user presence, but active information-seeking behavior — important for educational, news, and influencer outreach.
- Search engine optimization (SEO) on platform content (e.g., titles, thumbnails, metadata) can increase search referral rank.
- The correlation between real-world events and platform spikes (e.g., Telegram in 2021) helps media analysts and social researchers track digital sentiment shifts.

5.2.6 Addendum: Search Engine Referral-Based Tracking

The visualizations are based on a specific form of traffic analysis — search engine referrals, not search volume. A referral is logged only when a user searches, views the SERP, and then clicks through to a website that has Statcounter code installed. This methodology prioritizes user intent and action over impression or usage-based metrics. It gives a higher-fidelity picture of user engagement, particularly useful for marketing, product targeting, and digital trend analysis.



Social Platform Traffic Rankings (Bump Chart)

5.3 Internet Usage Statistics Dashboard — Demographic Access Across India (2025)

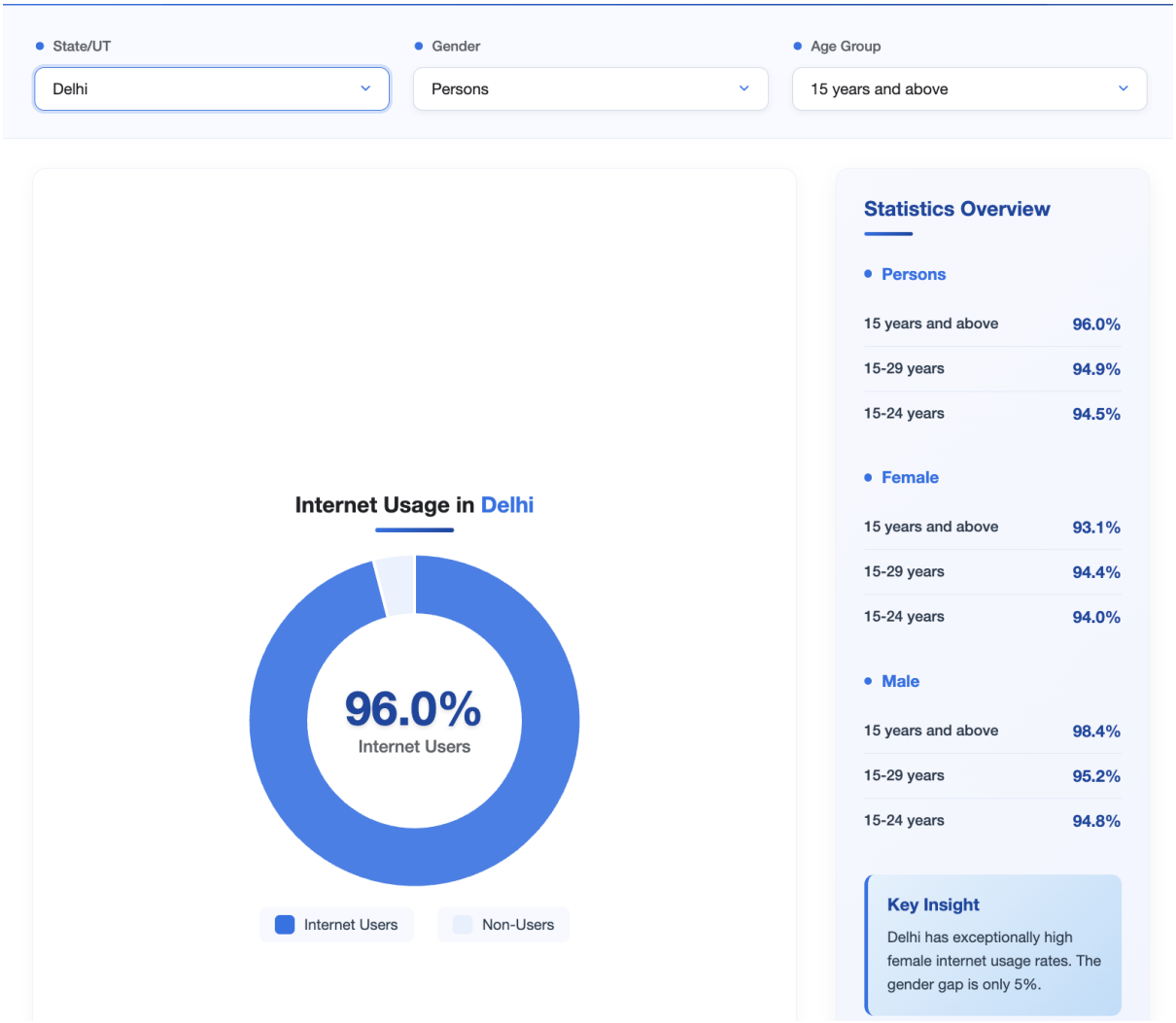


Figure 1: Demographic Internet Usage

5.3.1 Task Overview

This task involves an interactive state-wise dashboard (`try3.html`) that visualizes internet accessibility across various demographic segments in India as of July 2025. The dashboard enables users to explore the percentage of the population with internet access, segmented by geographic location, gender, and age. Key features include filters for demographic subgroups and highlights for states with extreme values in usage and gaps, thereby aiding the study of digital inclusion and accessibility disparities.

5.3.2 Information Represented

The dashboard represents a variety of demographic and geographical data in a visually interactive format. Key data elements include:

- **Geographical Scope:** All 36 Indian states and union territories.

- **Demographic Filters:**
  - **Age Groups:** 15+, 15–24, 15–29.
  - **Gender Categories:** Male, Female, Persons (Total).
- **Visual Elements:**
  - Pie Charts: Show proportions of internet users vs. non-users.
  - State-wise Bar Graphs: Internet usage by gender and age group.
  - Summary Panels: Display states with highest/lowest usage and largest gender gaps.
- **Data Source:** Simulated national digital access surveys (2025).

5.3.3 Observed Trends

The dashboard reveals several notable demographic and geographic trends:

- **Youth (15–24 years)** demonstrate extremely high internet usage rates, often exceeding 90% in most regions.
- **Females** consistently show lower usage rates than males, although this gap narrows in more urbanized and digitally developed states.
- **Delhi** reports the highest access rate (96%), whereas **Chhattisgarh** shows the lowest at 71.8%.
- **Rajasthan** exhibits the largest gender gap in internet access, approximately 19.5%, reflecting socio-cultural disparities.

5.3.4 Inference and Insights

The visual analytics system allows us to derive the following insights:

- **Gender-based digital inequality** remains a prominent concern in several parts of India, especially in the northern and central states.
- High internet adoption among the youth highlights the importance of **digitally enabled education and employment opportunities**.
- **Urban regions** exhibit near-universal digital access, suggesting that infrastructure and policy gaps are primarily in rural or underdeveloped areas.

5.3.5 Real-World Implications

This dashboard has several applications for policy, education, and business:

- **Policy Makers:** Can design digital literacy programs and broadband expansion strategies based on demographic gaps.
- **Non-Governmental Organizations (NGOs):** Can identify areas requiring training and digital outreach initiatives.

- **EdTech and Digital Startups:** May target states with lower access but large untapped youth populations for digital services.
- **Gender-focused Campaigns:** Should prioritize bridging the gender digital divide in low-access states.

5.4 India's Content Preferences Heatmap — Platform-Wise Digital Consumption (2025)

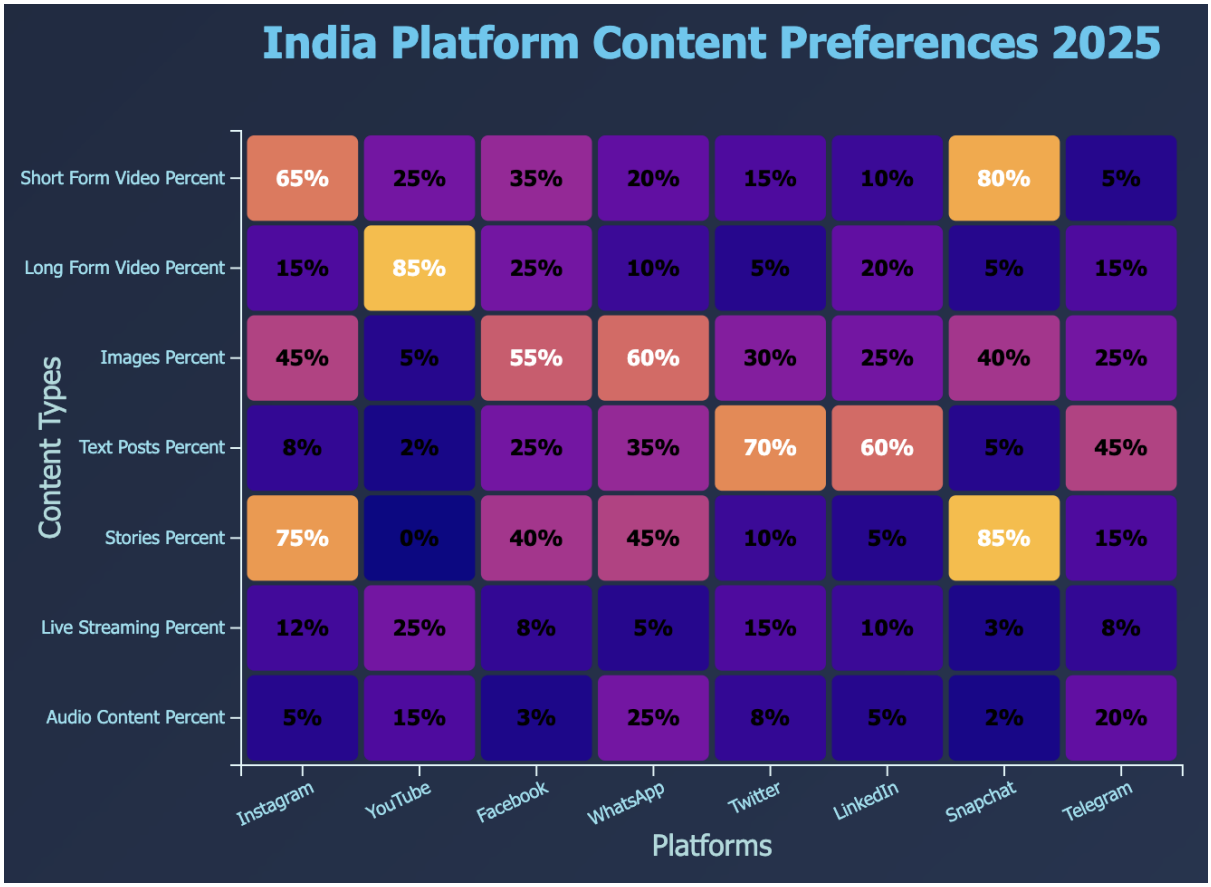


Figure 2: Content Preference Heatmap Across Indian Digital Platforms

5.4.1 Task Overview

This task features a heatmap-based visual analytics dashboard (`contentHmap.html` / `try2.html`) to illustrate the content type preferences of Indian users across popular digital platforms in 2025. The system seeks to decode user engagement styles — whether visual, textual, short-form, or long-form — and analyze how these preferences vary by platform. The heatmap facilitates a platform-wise comparison, highlighting evolving digital content consumption patterns in India.

5.4.2 Information Represented

The heatmap visualizes the engagement percentages (0–100%) for a range of content types across major platforms. The color intensity indicates the relative popularity of each format.

- **Content Types:**
  - Short-form Videos
  - Long-form Videos

  - Text Posts
  - Stories
  - Images
  - Live Streaming
  - Audio Content

- **Platforms Analyzed:** Instagram, YouTube, Facebook, WhatsApp, Twitter (X), LinkedIn, Snapchat, Telegram

5.4.3 Observed Trends

Analysis of the heatmap reveals several consumption preferences among users:

- **Snapchat (80%)** and **Instagram (65%)** lead in short-form video engagement.
- **YouTube (85%)** continues to dominate long-form video content.
- **WhatsApp** exhibits a balanced usage pattern across images (60%), stories (45%), and text (35%) — indicating a strong multimodal communication trend.
- **Twitter (70%)** remains the primary platform for text-heavy content, aligning with its identity as a real-time information and commentary source.

5.4.4 Inference and Insights

From the visual analysis, the following conclusions can be drawn:

- Platforms are increasingly **specializing in distinct content formats:**
  - Snapchat and Instagram for short-form visuals.
  - Twitter and LinkedIn for text-based updates.
  - Facebook and WhatsApp as mixed-media platforms.
- **Stories and live streaming** are becoming prevalent, signaling a shift towards real-time, ephemeral content.
- **Audio content** remains underutilized, gaining mild traction only on Telegram and YouTube.

5.4.5 Real-World Implications

The insights derived from this dashboard offer value to multiple stakeholders:

- **Content Creators:** Enables alignment of content format with user preferences native to each platform.
- **Marketers:** Facilitates data-driven decisions for campaign strategy — such as using reels on Instagram or long tutorials on YouTube.
- **Policy Makers and Educators:** Provides guidance on selecting suitable content modes for outreach, awareness, and e-learning campaigns tailored to platform strengths.

5.5 Social Media Platform Usage Radar Chart — Comparative Activity Profiling

Figure 3: Content Preference Heatmap Across Indian Digital Platforms

5.5.1 Task Overview

This task introduces an interactive radar chart-based visualization (`spider.html`) to compare how users engage in various social media activities across different platforms. The system allows users to either explore each activity type individually or analyze overall usage patterns in a unified radar/spider view. This comparative profiling helps identify the primary function each platform serves in the broader digital ecosystem.

5.5.2 Information Represented

The radar chart visualizes user engagement data across multiple dimensions:

- **Platforms Analyzed:** Facebook, Instagram, TikTok, LinkedIn, Snapchat, X (Twitter), Reddit, Pinterest
- **User Activities:**
  - Entertainment
  - Brand Research
  - News & Current Events
  - Messaging
  - Posting Photos/Videos
- **Visualization Features:**
  - Color-coded polygons for each activity type.
  - Hover-based tooltips displaying usage percentages and platform ranks.
  - Multi-layered chart enabling comparative profiling of each platform's strength.

5.5.3 Observed Trends

Several clear patterns emerge from the radar chart:

- **TikTok** dominates the entertainment category with a usage rate of 79.6%, followed by Instagram at 67.1%.
- **Facebook and Instagram** perform well across both messaging and brand research activities.
- **X (formerly Twitter)** leads in the news and current events category with a notable 60.5% user engagement.
- **Snapchat** excels in messaging and image-based content sharing.
- **LinkedIn** ranks low in entertainment but shows relatively better engagement in brand-related searches and professional networking.

5.5.4 Inference and Insights

Key insights derived from this task include:

- Platforms show distinct activity-based specialization: TikTok for entertainment, LinkedIn for brand research, and X for news.
- Brands, educators, and agencies can optimize content strategies by aligning with platform-specific user behavior.
- The radar visualization effectively communicates how digital users differentiate platforms based on intent, utility, and engagement style.

5.5.5 Real-World Implications

This multi-dimensional view of user engagement has strategic relevance for multiple stakeholders:

- **Marketers:** Helps fine-tune campaigns by choosing platforms best suited for each message or product type.
- **Platform Developers:** Can identify underrepresented features on their platforms and strategize enhancements.
- **Digital Behavior Researchers:** Offers an empirical basis to study cross-platform user preferences and specialization in social media behavior.

5.6 Facebook Religious Sentiment Analysis (2010–2025)

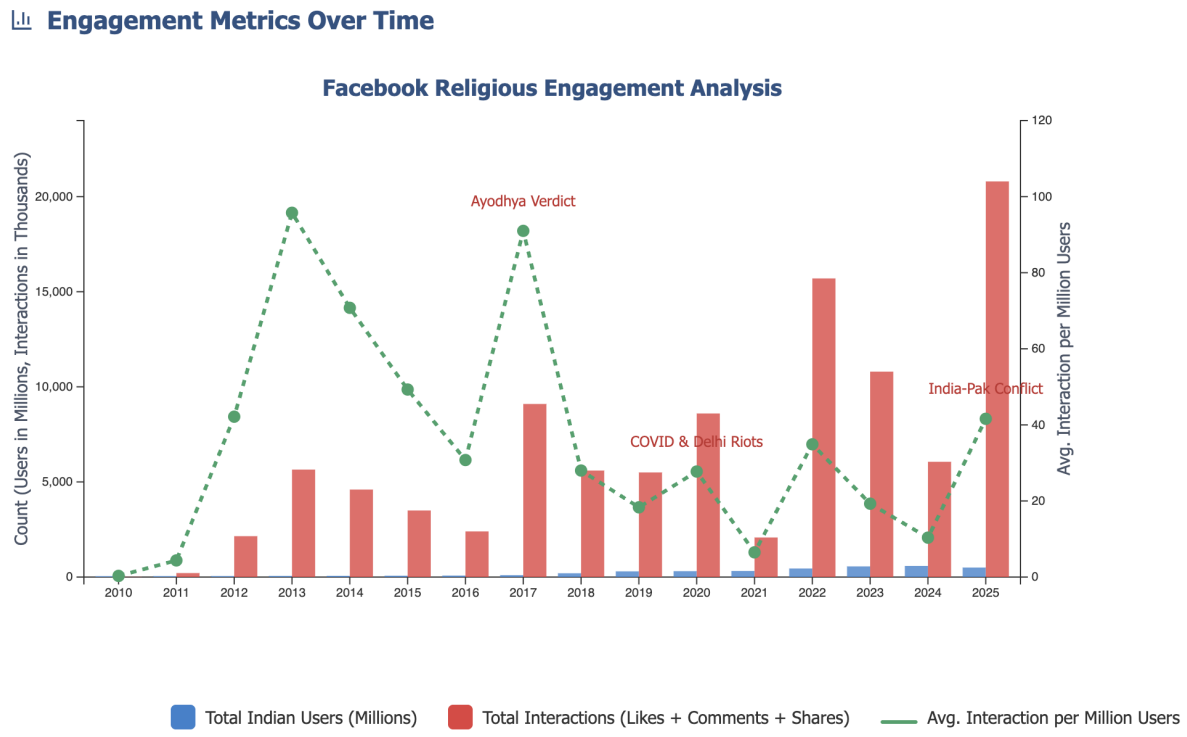


Figure 4: Religious Sentiment Trend Analysis Dashboard (Placeholder)

5.6.1 Task Description

This task focuses on analyzing the evolution of religious sentiment expression on Facebook in India by studying user engagement with posts containing the phrase *"Jai Shri Ram"* over a 15-year period (2010–2025). The objective is to understand how public interaction with religious content on social media correlates with real-world events, platform growth, and regulatory interventions.

The visualization system (`religious.html`) is a web-based dashboard built using D3.js and features:

- Total Indian Facebook users (in millions)
- Total interactions (likes, comments, shares) on *"Jai Shri Ram"* posts
- Average interactions per million users
- Annotated timeline with political, religious, and societal events per year

5.6.2 What Information It Represents

The dashboard captures three major quantitative dimensions:

1. **Total Users:** Reflects platform growth and penetration across India.
2. **Total Interactions:** Measures raw engagement with religion-related content.

3. **Average Interactions per Million Users:** A normalized metric revealing engagement intensity independent of user base size.

Each year is also annotated with a description of major real-world events to provide context for notable changes in engagement patterns.

5.6.3 Observable Trends

Several distinct phases can be identified from the dashboard:

- **2010–2013:** Marked by a rapid increase in both users and engagement as Facebook expanded in India.
- **2014–2017:** Engagement patterns became more volatile due to rising political-religious discourse. A major spike in 2017 aligns with the Ayodhya verdict.
- **2018–2020:** Although overall engagement declined slightly, 2020 saw renewed interest—likely due to the COVID-19 lockdown shifting religious expression online.
- **2021–2025:** Despite increasing user base, engagement became inconsistent. Government crackdowns, algorithmic moderation, and user fatigue contributed to lower intensity. However, 2025 saw another spike, potentially due to India–Pakistan conflict.

5.6.4 Inferences and Insights

From the trend analysis, the following inferences can be made:

- Online religious sentiment is **event-driven**, with spikes corresponding to real-world unrest or socio-political triggers.
- Post-2020 moderation and saturation have **dampened engagement**, even as user numbers continued to grow.
- Normalized data reveal a **decline in sentiment intensity**, pointing to changing behavior, content fatigue, or better platform governance.
- Government regulations and moderation policies have a **direct impact on sentiment expression** and public engagement.

5.6.5 Broader Implications

The findings from this task have practical implications across sectors:

- **For Governments:** Enables proactive policy response to predict and manage digital sentiment during sensitive periods.
- **For Platforms:** Highlights the need for timely and transparent content moderation during elections or religious events.
- **For Researchers:** Offers a longitudinal case study of how religion, technology, and society interact in the digital space.
- **For Society:** Encourages awareness of how online discourse can influence or amplify collective sentiment and behavior.

5.7 Social Media Toxicity Dashboard — Platform-Wise Toxic Behavior Analysis

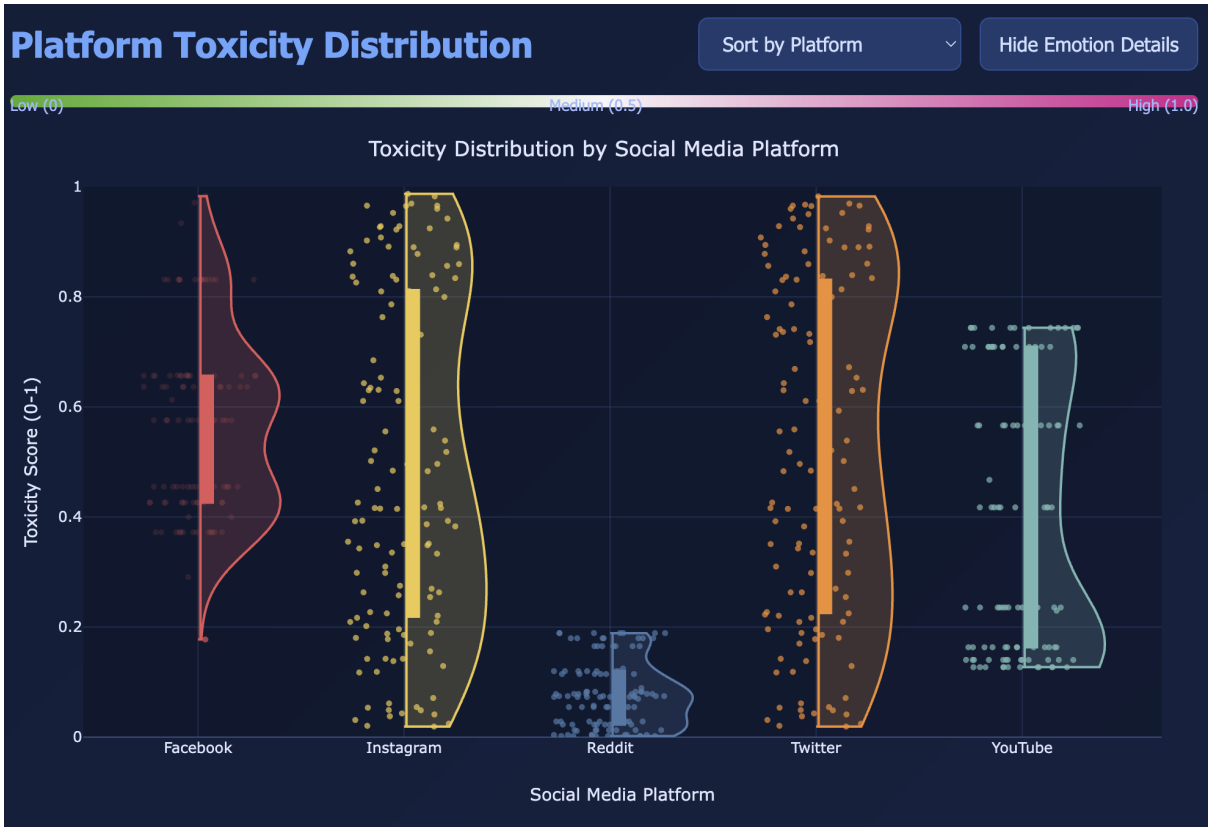


Figure 5: Interactive Toxicity Analysis Dashboard (Placeholder)

5.7.1 Task Overview

This task investigates toxic content prevalence across major social media platforms using an interactive web dashboard `toxicity_rate.html`. The dashboard provides comprehensive visual summaries such as violin plots, box plots, and emotion-based distributions. These visualizations enable users to identify which platforms exhibit higher toxicity and how toxicity levels vary across different emotional sentiments.

5.7.2 Information Represented

The following metrics and dimensions are visualized in the dashboard:

- **Platforms Compared:** Facebook, Twitter (X), Instagram, YouTube, Reddit.
- **Metrics Visualized:**
  - Toxicity Score Distribution via Violin and Box Plots.
  - Median and Variance of Toxicity Scores.
  - Emotion Categories: Angry, Happy, Sad, Neutral (with proportions visualized per platform).
- **Ranked Summary Card:** Lists platforms in descending order of median toxicity.

5.7.3 Observed Trends

Based on the dashboard visualizations, the following patterns are evident:

- **Reddit** shows the widest range and highest median toxicity, commonly associated with controversial or loosely moderated content.
- **Twitter (X)** demonstrates a high concentration of toxic scores, often stemming from politically charged or polarizing threads.
- **Instagram and YouTube** display comparatively lower toxicity, likely influenced by their visual-centric content and stronger moderation policies.
- **Anger** emerges as the dominant emotion in toxic posts, whereas **Neutral** and **Happy** emotions are more prevalent on Instagram and YouTube.

5.7.4 Inference and Insights

Key insights derived from the analysis include:

- Platform culture and moderation strategies play a pivotal role in shaping toxicity levels.
- Text-heavy platforms tend to be more prone to toxic interactions due to ease of argumentation and user anonymity.
- Emotion-linked toxicity opens avenues for AI-based moderation tools to predict and preemptively flag potentially harmful content based on sentiment detection.

5.7.5 Real-World Implications

This analysis has several practical applications:

- **Content Moderation Teams** can utilize platform-specific insights to tailor and strengthen their moderation frameworks.
- **Developers of AI Filters** may integrate emotion-toxicity correlations to enhance automated detection systems.
- **Sociologists and Behavior Analysts** can leverage these findings to better understand digital aggression and design healthier, safer online spaces.

5.8 Misinformation Analysis Dashboard — Understanding Digital Influence (2020–2021)

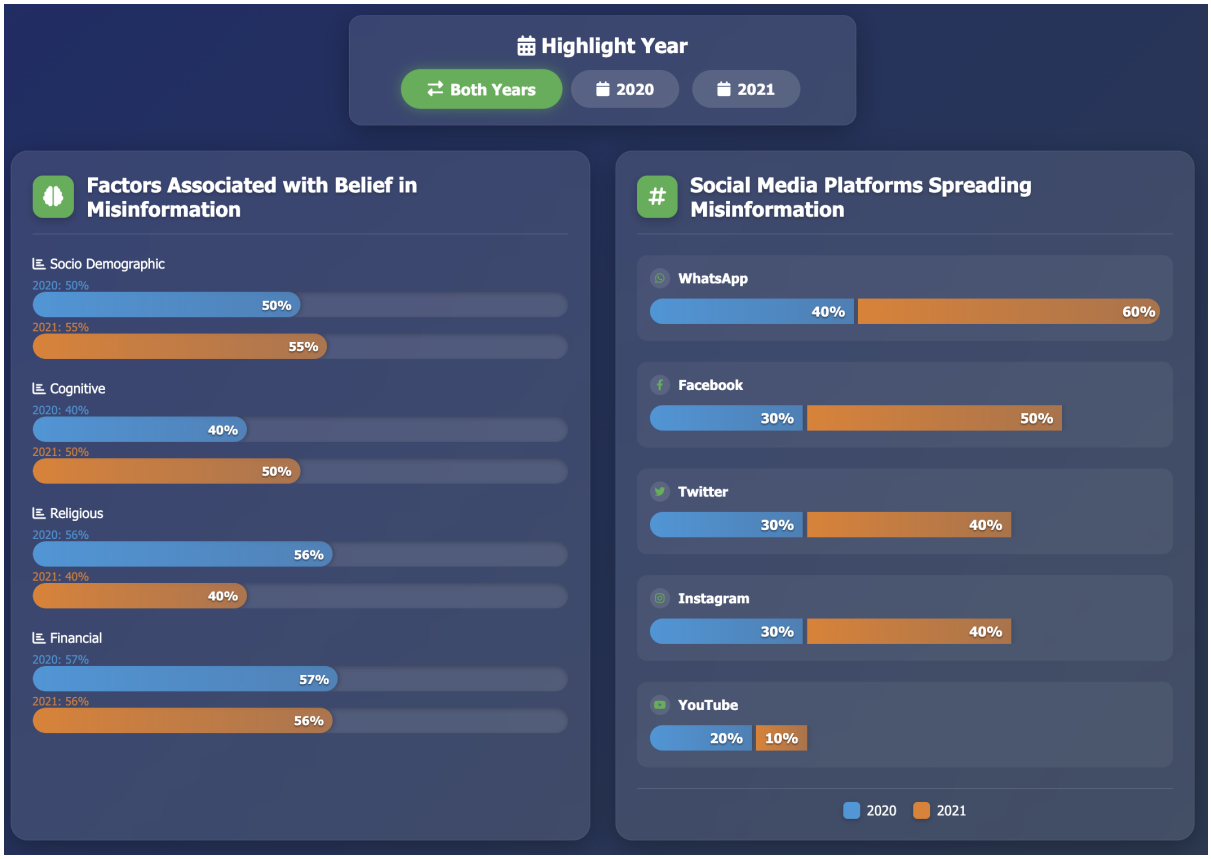


Figure 6: Interactive Misinformation Influence Dashboard (2020–2021) — Placeholder

5.8.1 Task Overview

This task explores the spread of misinformation across digital platforms during the pandemic years of 2020 and 2021. The interactive dashboard (`missleadinginfo.html`) presents dual bar charts for influential user factors and misinformation-prone platforms, along with year-wise toggles to visualize changes over time. This analysis is crucial to understanding how belief in misinformation evolved and which platforms were most responsible for its dissemination.

5.8.2 Information Represented

The dashboard incorporates the following elements:

- **Key Factors Associated with Misinformation Belief:**
  - Socio-demographic
  - Cognitive
  - Religious
  - Financial

- **Primary Platforms Analyzed:** WhatsApp, Facebook, Twitter (X), Instagram, YouTube
- **Temporal Comparison:** Interactive toggles allow comparison across 2020, 2021, and combined data. Color-coded bars and dynamic animations visually communicate the proportion and intensity of misinformation impact.

5.8.3 Observed Trends

The year-wise analysis reveals several important shifts and patterns:

- **WhatsApp** showed a dramatic rise in misinformation involvement, increasing from 40% in 2020 to 60% in 2021.
- **Religious and financial factors** were dominant drivers in 2020, possibly linked to pandemic-related fear and uncertainty.
- **Cognitive and socio-demographic factors** gained more prominence in 2021, suggesting a shift toward more psychological and group-based targeting.
- **Facebook and Twitter** also exhibited significant increases in misinformation spread over the two years.

5.8.4 Inference and Insights

The data leads to several key insights:

- Private **messaging platforms like WhatsApp** are the most effective vectors for misinformation due to end-to-end encryption and minimal content moderation.
- The shift in influential factors from religious/financial to cognitive/socio-demographic reflects how misinformation strategies adapted to evolving public vulnerabilities.
- No platform is immune; **diverse content types and algorithmic recommendation systems** contribute to misinformation proliferation across all platforms.

5.8.5 Real-World Implications

The study provides actionable insights for various stakeholders:

- **Governments:** Can use the data to shape platform regulation policies, especially around encrypted messaging and crisis communication.
- **Social Media Platforms:** Should enhance transparency in content distribution algorithms and invest in real-time misinformation detection systems.
- **Public Educators and NGOs:** Can develop targeted digital literacy campaigns and fact-checking initiatives focused on susceptible user groups.

5.9 Social Media Addiction Dashboard — A Visual Analytics Study

Figure 7: Social Media Addiction and Psychological Effects

5.9.1 Task Overview

This task centers around the development and analysis of an interactive web-based dashboard (`effects.html`) that investigates the psychological and behavioral impact of social media usage among students. The dashboard leverages data from 728 students across 65 countries, aiming to uncover correlations between digital habits and well-being indicators such as addiction scores, mental health, sleep duration, and social dynamics. It serves as an exploratory visual tool for analyzing digital overuse and its implications on daily life.

5.9.2 Information Represented

The system presents a variety of key variables related to social media usage and its consequences:

- **Metrics Visualized:**
  - Addiction Score (0–10 scale)
  - Mental Health Score
  - Average Sleep Duration per Night (in hours)
  - Daily Social Media Usage (in hours)
  - Frequency of Social Media-Induced Conflicts
  - Age
- **Visualization Components:**
  - Correlation Heatmap — displays Pearson correlation coefficients between variables.

  - Interactive Scatter Plots — generated dynamically based on user selection from the heatmap.
  - Metric Cards — present high-level statistics like average sleep, maximum addiction score, and daily screen time averages.

5.9.3 Trends and Observations

The dashboard reveals several statistically and behaviorally significant trends:

- A strong negative correlation of **-0.95** exists between addiction scores and mental health, suggesting that increased social media usage correlates with worse mental health outcomes.
- Students with addiction scores above 7 average only **5.6 hours of sleep** per night — well below the recommended 8 hours.
- **Instagram and TikTok** users report the highest addiction levels (mean score 7.8), while **LinkedIn** users show the lowest (3.9).
- A strong positive correlation of **0.93** between addiction and frequency of conflicts suggests excessive use may result in increased interpersonal tension.

5.9.4 Inference and Insights

From the trends and visualized correlations, the following insights can be drawn:

- Social media addiction is a multi-faceted issue affecting not only digital engagement but also sleep patterns, emotional health, and social harmony.
- Platforms promoting high-frequency interaction, such as short-form video platforms, may inherently foster addictive behaviors.
- Indicators such as reduced sleep and increased conflicts can serve as early warning signs of problematic usage.
- The close tie between addiction and mental health underscores the need for digital wellness interventions, especially in educational environments.

5.9.5 Real-World Implications

This analysis offers meaningful applications across different sectors:

- **Healthcare Professionals:** Can use these findings to tailor mental health support for digitally overexposed youth.
- **Educational Institutions:** May embed digital wellness into student health programs.
- **Technology Developers:** Should consider ethical design features that reduce compulsive interaction loops.
- **Policy Makers:** Can use this empirical basis to frame digital well-being regulations and promote healthy usage habits.

5.10 Sentiment Analysis of Twitter Data

This project involves analyzing tweets using the *Sentiment140* dataset provided by Stanford. The objective is to classify tweets as either positive or negative using machine learning algorithms and various feature extraction techniques.

This is the sentiment140 dataset. It contains 1,600,000 tweets extracted using the twitter api . The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment .

Content

It contains the following 6 fields:

target: the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
ids: The id of the tweet (2087)
date: the date of the tweet (Sat May 16 23:58:44 UTC 2009)
flag: The query (lyx). If there is no query, then this value is NO_QUERY.
user: the user that tweeted (robotickilldozr)
text: the text of the tweet (Lyx is cool)
The linked Jupyter notebooks used in this project are summarized as follows:

- **First.ipynb** – Preparation and cleaning of the dataset.
- **Second.ipynb** – Analysis, visualization, and preparation of the data for further processing.
- **Third.ipynb** – Implementation of Zipf's Law and visualization of tweet tokens.
- **Fourth.ipynb** – Dataset splitting and application of the TextBlob sentiment analyzer as a baseline. Feature extraction using `CountVectorizer` and classification using Logistic Regression on unigrams, bigrams, and trigrams.
- **Fifth.ipynb** – Feature extraction using TF-IDF and Logistic Regression applied to unigrams, bigrams, and trigrams. Performance comparison with other classification algorithms such as Ridge Classifier, Perceptron, Passive-Aggressive Classifier, Stochastic Gradient Descent, LinearSVC, L1-based LinearSVC, KNN, Nearest Centroid, Multinomial Naive Bayes, Bernoulli Naive Bayes, and AdaBoost.
- **Sixth.ipynb** – Implementation of the `Doc2Vec` model using Gensim for feature extraction. Experiments included DBOW (Distributed Bag of Words), DMC (Distributed Memory Concatenated), DMM (Distributed Memory Mean), and combinations such as DBOW + DMC and DBOW + DMM on unigrams.
- **Seventh.ipynb** – Implementation of phrase modeling using Gensim, followed by DBOW, DMC, DMM, DBOW + DMC, and DBOW + DMM applied to bigrams and trigrams. Additional classification algorithms were also applied to the dataset.

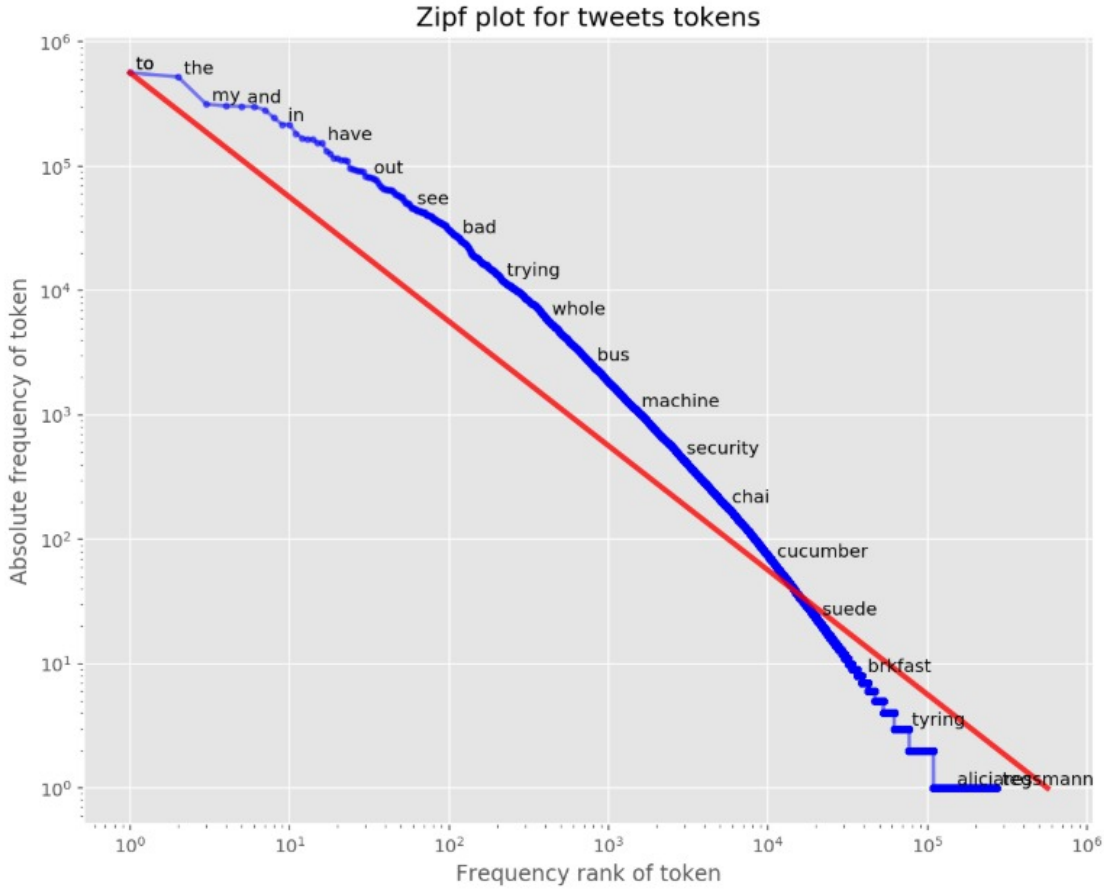


Figure 8: Deviating above the expected line on higher ranked words and deviating below the expected line on lower ranked words.

6 Challenges and Limitations

One of the primary challenges encountered during the course of this project was the unavailability of a single comprehensive dataset that could fulfill all our analytical requirements. Given the diverse nature of the visualizations we aimed to develop—spanning platform usage statistics, user behavior trends, and sentiment analysis—we found it necessary to gather data from multiple heterogeneous sources. This involved extensive effort in identifying credible and relevant datasets, ensuring compatibility across formats, and performing necessary preprocessing steps such as normalization and cleaning. Integrating data from various platforms and time periods also posed challenges in maintaining consistency and accuracy, but careful curation allowed us to construct a cohesive foundation for our visual analytics system.

Figure 9: Words which have the highest frequency among all the tweets, positive and negative both.

7 Conclusion

This project presents a comprehensive visual analytics exploration of social media trends with a focus on Indian and global usage patterns. Through interactive dashboards, we analyzed platform rankings, content preferences, user behavior, toxicity levels, misinformation prevalence, and addiction symptoms.

Key findings include:

- Instagram and YouTube lead in engagement for entertainment and short-form content.
- Facebook remains dominant in messaging but shows higher toxicity and misinformation.
- Youth are the most active demographic, yet also most vulnerable to addiction and sleep deprivation.
- Emotional polarization and regional content variation are significant on platforms like Twitter and Facebook.

The visual interfaces allowed for intuitive interpretation and revealed deeper patterns across demographic and temporal dimensions. This work demonstrates the power of combining web-based data visualization with behavioral analytics to uncover meaningful societal insights.

8 Future Work

While the current project focuses on descriptive and comparative visualizations, several directions can enhance its depth and scope:

- **Time-Series Tracking:** Adding timeline-based changes to engagement, toxicity, and content consumption would help identify evolving trends.
- **Sentiment Analysis:** Applying NLP to extract user sentiments on posts can enrich emotional and opinion-based understanding.
- **Predictive Modeling:** Building models to forecast misinformation spread, user addiction likelihood, or platform growth could assist in early interventions.
- **Platform Policy Comparison:** Analyzing moderation policies across platforms can explain variance in toxicity or fake news.
- **More Granular Demographics:** Incorporating profession, education level, or income groups can help build more targeted insights.

By extending these analyses, future iterations can offer actionable recommendations for users, policymakers, and platform designers to promote healthier digital habits and reduce online harm.

9 Team Contributions

Name	**Contribution**
Somya Kumar	Task 1: Worldwide Ranking Visualization
Neela	Task 2: Social Platform Traffic Analysis
Someshwar Singh	Task 3: Internet Usage Dashboard
Adarsh	Task 4: India Platform Content Preferences
Peeyush Sahu	Task 5: Usage Pattern Spider Map
Gaurav Bohra	Task 6: Facebook Religious Sentiment Analysis
Sudeep Chahlia	Task 7: Social Media Toxicity Analysis
Prabhakar & Mayank Agrawal	Task 8: Misinformation Visualization
Somya & Adarsh	Task 9: Addiction Dashboard
Mayank Agrawal & Somya	Task 10: Twitter Sentiment Analysis

Table 1: Individual Contributions by Team Members

10 References

[1] Statcounter Global Stats. *Social Media Traffic Rankings in India (2009-2025)*. Accessed July 2025. Available at: https://gs.statcounter.com/social-media-stats/all/india/#monthly-200903-202506

[2] Datareportal. *Digital 2025 India Report*. Accessed July 2025. Available at: https://datareportal.com/reports/digital-2025-india

[3] Kaggle Datasets. *Census 2016 Population by County*. Accessed July 2025. Available at: https://www.kaggle.com/

[4] ResearchGate. *Popular social media platforms spreading misinformation (2020–2021)*. Accessed July 2025. Available at: https://www.researchgate.net/figure/Popular-social-media-platforms-spreading-misinformation_fig2_367228139

[5] Datareportal. *Social Media Users Statistics*. Accessed July 2025. Available at: https://datareportal.com/social-media-users

[6] *16 yrs–Random Sample–"JaiShriRam" Tag–FB_India*. Kaggle, 2022. Available at: https://www.kaggle.com/datasets/gauravb37/16-yrs-random-sample-jaishriram-tag-fb-india

[7] CS661 Big Data Visual Analytics Lecture Slides, IIT Kanpur, 2025.