Research Engineer Intern Assignment for SimPPL

Social Media Analysis Dashboard

Hosted Platform: http://20.244.49.140/
GitHub Link: https://github.com/ddihora1604/research-engineering-intern-assignment_1
YouTube Video: https://youtu.be/t_dPmjmsaB0

1. Project Overview

The Social Media Analysis Dashboard is a comprehensive and interactive platform built to analyze, visualize, and interpret social media activity—specifically using data from Reddit. By leveraging natural language processing (NLP), machine learning models, and advanced visualizations, this tool allows researchers and analysts to explore patterns in conversations, detect coordinated behaviors, and extract trends across communities.

2. Objective and Motivation

The primary objective of this project is to transform raw social media data into actionable insights through interactive and intelligent interfaces. Social platforms generate enormous volumes of unstructured data, and understanding user behavior, sentiment shifts, and community dynamics requires sophisticated tools.

Key motivations include:

- Enabling researchers to discover emerging discussion topics
- Tracking sentiment evolution over time
- Detecting coordinated posting behavior which may indicate manipulation or spam
- Providing semantic understanding of conversations using embeddings and clustering
- Supporting natural language querying through integration with generative AI

3. Features of the Dashboard

Each component of the dashboard is crafted to serve a specific analytic need. Below is a breakdown of the major features included:

3.1 Social Media Insights – Trends and Discussions

- Uses NLP techniques to identify patterns in discussions
- Extracts topics, clusters conversations, and detects sentiment polarity
- Summarizes dominant themes and trending words from social activity

3.2 Key Metrics Analysis

- Displays quantitative summaries such as:
  - Total post volume
  - Engagement metrics (comments, reactions)
  - Number of active users
- Useful for high-level monitoring and detecting spikes in activity

3.3 Word Cloud Generator

- Dynamic word clouds built using D3.js
- Visual representation of term frequencies in user conversations
- Provides a quick overview of commonly discussed words or hashtags

3.4 Top Contributors Analytics

- Ranks and profiles the most active users
- Metrics include:
  - Frequency of posts
  - Number of comments
  - Engagement rates (interactions per post)
- Enables understanding of influential voices within communities

3.5 Comprehensive Narrative Data Story

- Generates an automated summary story using the LLaMA model via Groq API
- The AI constructs a readable narrative based on:
  - Detected trends
  - Topic distributions
  - Sentiment summaries

- Bridges the gap between raw data and human-readable insights

3.6 Time Series Trend Visualization

- Tracks the volume of posts and engagement over time
- Provides options for custom date range filtering
- Visualizes patterns such as seasonal spikes, event-driven activity, or long-term trends

3.7 Topic Evolution Graph

- Implements Latent Dirichlet Allocation (LDA) to extract dominant topics
- Visualizes the change in topic prominence over time
- Interactive graph powered by D3.js, allowing exploration of topic shifts and continuity

3.8 Community Distribution Insights

- Pie chart visualization of:
    - Distribution of users across different subreddits or groups
    - Volume of posts per community
- Helps identify core communities and compare relative activity

3.9 Topic Modeling and Analysis

- Implements customizable LDA topic modeling
- Adjustable parameters include:
    - Number of topics
    - Word weights
- Shows top words per topic and example documents for clarity
- Aids in exploring hidden themes within the data

3.10 Semantic Map

- Dimensionality reduction using UMAP with SentenceTransformer embeddings
- Generates a 2D semantic space where similar posts cluster together
- Reveals natural groupings of conversation topics, even across different communities
- Topics are color-coded and labeled for easier exploration

3.11 Coordinated Posting Detection

- Identifies suspicious behavior patterns through:
    - Temporal similarity of posts
    - Content overlap and similarity scoring
- Highlights groups of users that may be participating in coordinated campaigns
- Useful for uncovering spam, bot networks, or orchestrated disinformation

3.12 AI-Powered Chatbot

- Integration of Google's Gemini 2.0 Flash model
- Users can type natural language queries like:
    - "What are the major themes discussed in May?"
    - "Who are the most active users in r/technology?"
- The chatbot provides insightful, human-like answers based on the analyzed data
- Enhances the accessibility of insights for non-technical users

4. Technical Architecture and Design

The system follows a modular and scalable architecture, combining backend processing, frontend interactivity, and AI integrations.

4.1 Backend (Flask Application)

- Serves API endpoints for:
    - Data loading and preprocessing
    - LDA topic modeling
    - Sentiment analysis and coordinated group detection
- Integrates with:
    - Groq API for LLaMA-based narrative generation
    - Gemini API for chatbot functionality

4.2 Frontend (HTML + D3.js)

- Custom dashboards created with HTML templates and JavaScript
- D3.js handles most of the interactive visualizations, including:

    - Word clouds
    - Time series graphs
    - Pie charts
    - Topic evolution maps

4.3 AI and ML Models

- LDA: For extracting latent topics
- UMAP + Sentence Transformers: For semantic similarity visualization
- Temporal Pattern Detection: For identifying coordinated behavior
- LLaMA via Groq: For story generation
- Gemini via Google: For natural language interaction

5. Data Requirements

The dashboard is built to process data in the JSONL (JSON Lines) format, with each line containing a Reddit post object.

Expected JSON Format:

```
{
  "data": {
    "id": "post_id",
    "author": "username",
    "created_utc": 1627484400,
    "title": "Post title",
    "selftext": "Post content",
    "subreddit": "subreddit_name",
    "num_comments": 42,
    "permalink": "/r/subreddit/comments/post_id/title/"
  }
}
```

Required Fields:

- id: Unique identifier for each post
- author: Username of the poster
- created_utc: Unix timestamp for the post
- title: Title of the Reddit post
- selftext: The main text content
- subreddit: Subreddit name where it was posted
- num_comments: Count of comments received
- permalink: URL to the post
- parent_id (optional): Used to trace comment hierarchy

This structure ensures consistent parsing and supports all analytic modules in the dashboard.

6. Impact and Applications

The Social Media Analysis Dashboard is not only a powerful research tool, but also serves potential use cases in:

- Misinformation detection
- Community management
- Brand monitoring and reputation analysis
- Campaign performance tracking
- Academic research in sociology, linguistics, or digital humanities

Its interactive nature and AI integration make it accessible to both technical and non-technical stakeholders, encouraging wide adoption in journalism, academia, and corporate intelligence.

7. Conclusion

This project exemplifies the integration of advanced AI models, statistical analysis, and web development to create a holistic and interactive dashboard. It empowers users to dig deep into the nuances of social media activity and derive rich insights from raw, noisy data. With its modular design, extensible architecture, and focus on user interactivity, the Social Media Analysis Dashboard sets a benchmark for future analytic platforms in the digital age.