Complete Data Analysis Report

DS- 464

Uni Pulse Analytics

Digital Business Analytics

Team Business:

Member 1: Hammad Ahmed

Member 2: Hasham Nadeem

Member 3: Husnain Mushtaq

Table of Contents

Table of Figures

UniPulse Analytics: Sentiment & Discourse Analysis of Pakistani Universities Using Reddit Data

Abstract

UniPulse Analytics is a data-driven system that collects, processes, and analyzes Reddit discussions related to major Pakistani universities—GIKI, LUMS, and NUST. The project applies Natural Language Processing (NLP), sentiment analysis, topic modeling, and Retrieval-Augmented Generation (RAG) to uncover student sentiment, recurring issues, and comparative insights across institutions. A market-ready interactive dashboard built with Streamlit presents actionable insights for students and administrators.

1. Introduction

Student decision-making increasingly relies on peer-generated content on social platforms. Reddit hosts candid discussions on admissions, academics, and campus life, making it a valuable data source for understanding university sentiment. This project aims to transform unstructured Reddit data into structured insights that support informed academic and administrative decisions.

Objectives

- Scrape large-scale Reddit data from university-specific subreddits
- Perform sentiment analysis and thematic categorization
- Compare universities across engagement and sentiment metrics
- Build an AI-powered Q&A system using RAG
- Deliver insights via a professional, interactive dashboard

2. Data Collection

2.1 Data Source

Reddit subreddits:

- r/giki
- r/lums
- r/nust

2.2 Scraping Methodology

Data was collected using the **PRAW (Python Reddit API Wrapper)** library with authenticated credentials loaded via environment variables.

Key scraping characteristics:

- Up to **1000 unique posts per subreddit**
- Multiple listing strategies to bypass Reddit's single-listing cap:
  - Top (All Time)
  - Top (Year)
  - Hot
  - New
- Duplicate prevention using post IDs
- Comment expansion limited to top 15 comments per post
- Rate-limit handling via controlled delays

Each subreddit's data is saved as a separate JSON file.

2.3 Data Schema

Each Reddit post is stored using the following structure:

```
{
  "id": "string",
  "subreddit": "string",
  "title": "string",
  "body": "string",
  "upvotes": "integer",
  "timestamp": "ISO-8601 datetime",
  "num_comments": "integer",
  "comments": ["string"],
  "url": "string"
}
```

3. Data Preprocessing

3.1 Text Preparation

- Combined post title and body into a single text field
- Lowercasing and URL removal
- Removal of punctuation and non-alphabetic characters
- Stopword removal using NLTK

3.2 Feature Engineering

For each post, the following derived attributes were added:

- **Sentiment Score**: Compound score using VADER Sentiment Analyzer
- **Category** (rule-based):
    - Admissions
    - Academics
    - Campus Life
    - General

3.3 Temporal Processing

- Conversion of timestamps to datetime format
- Aggregation by date for trend analysis

4. Exploratory Data Analysis (EDA)

4.1 Engagement Metrics

- Average upvotes per university
- Average comment count (Community Vitality)

4.2 Sentiment Analysis

- Sentiment range: -1 (negative) to +1 (positive)
- Mean sentiment used as **Pulse Index**

4.3 Comparative Analysis

Universities are compared using:

- Pulse Index (average sentiment)
- Community Vitality (average engagement)
- Discussion focus distribution (Academics vs Campus Life vs Admissions)

5. Topic Modeling

Latent Dirichlet Allocation (LDA) was applied to identify dominant discussion themes.

Configuration:

- Vectorizer: CountVectorizer
- N-grams: Trigrams
- Topics per subset: 3

This approach highlights common complaint phrases and recurring discussion points.

6. AI-Driven Insights (RAG System)

6.1 Architecture

The project integrates a **Retrieval-Augmented Generation (RAG)** pipeline:

1. Reddit posts embedded using OpenAI embeddings
2. Stored in a **ChromaDB** persistent vector database
3. User queries embedded and matched via cosine similarity
4. Relevant Reddit context retrieved
5. Final answer generated using an LLM constrained to retrieved context

6.2 Use Case

- Student queries such as:
  - "How is hostel life at GIKI vs NUST?"
  - "Is LUMS academically stressful?"

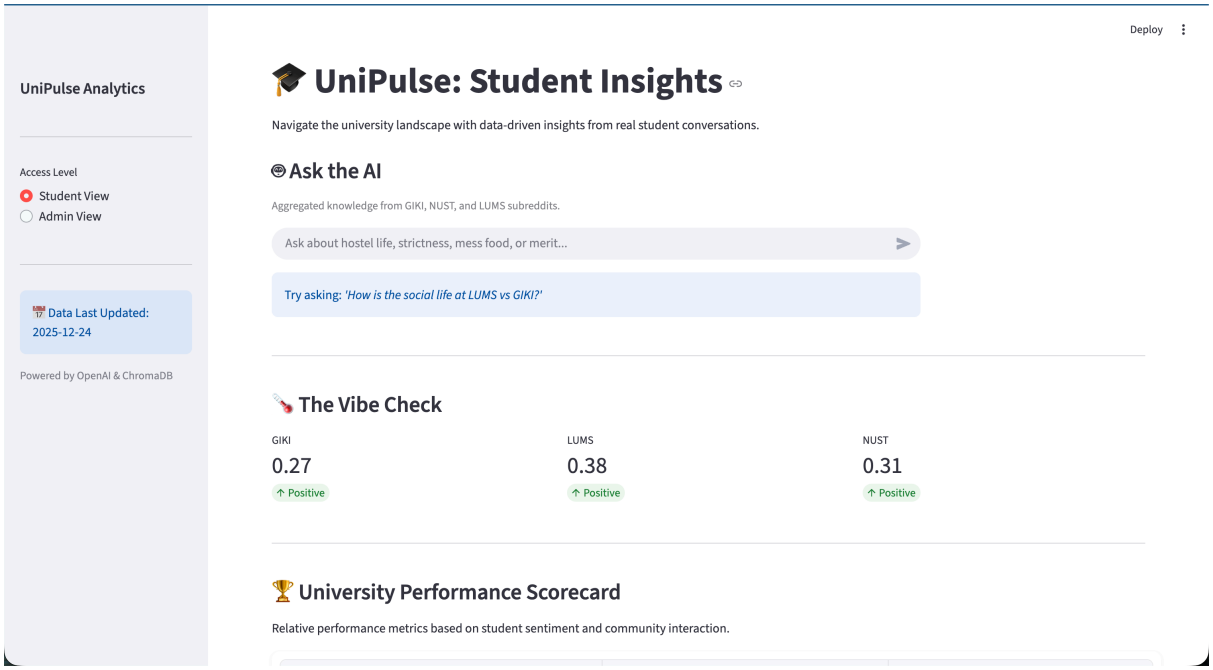
Responses are concise, contextual, and data-grounded.

7. Visualization & Dashboard(Student)

7.1 Dashboard Overview

The UniPulse Analytics dashboard serves as the primary interface for presenting analytical insights in an intuitive and visually rich format.

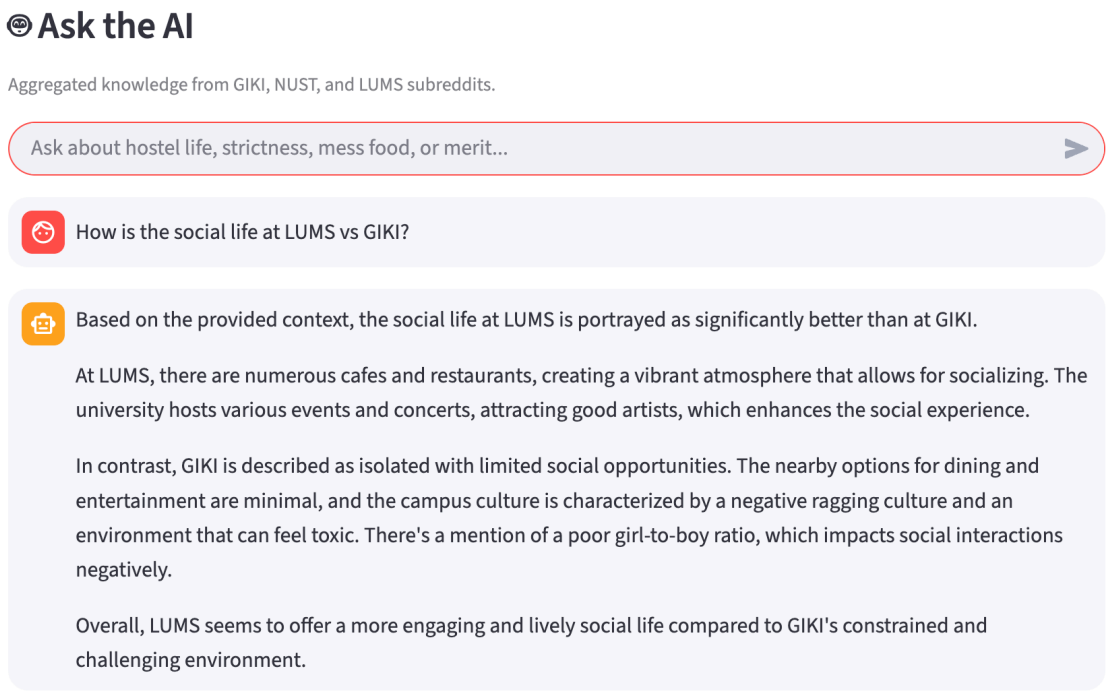
Figure 7.1: UniPulse Analytics – Main Dashboard Interface(User)



7.2 Ask-the-AI (RAG Interface)

This section enables users to query aggregated Reddit knowledge using natural language. The system retrieves semantically relevant posts and generates grounded responses.

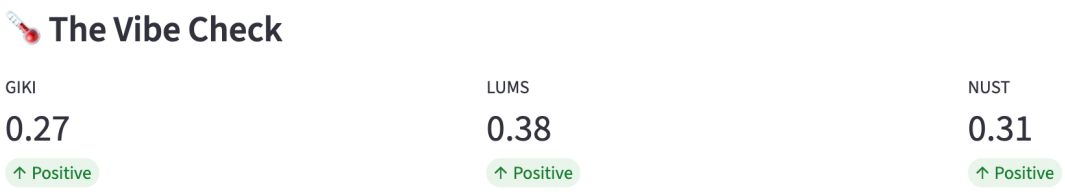
Figure 7.2: Ask-the-AI Interface with Example Query



7.3 Vibe Check – Sentiment Snapshot

The Vibe Check provides an at-a-glance sentiment score (Pulse Index) for each university. Scores are color-coded to indicate sentiment health.

Figure 7.3: Vibe Check Sentiment Indicators for GIKI, LUMS, and NUST



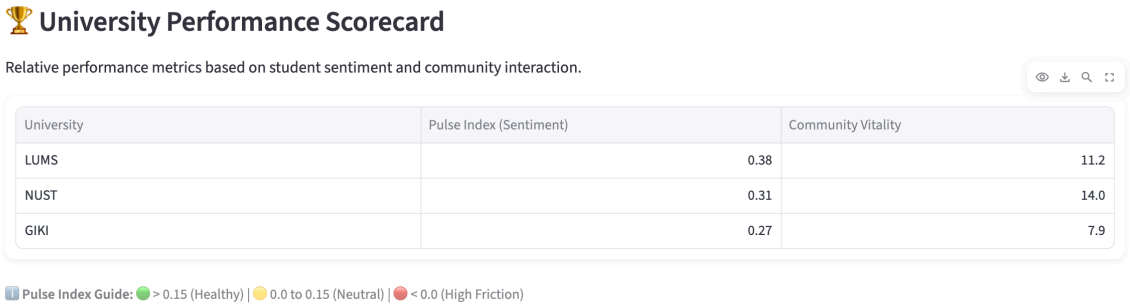
7.4 University Performance Scorecard

This table compares universities using sentiment and engagement-based metrics.

Metrics include:

- **Pulse Index**: Average sentiment score
- **Community Vitality**: Average number of comments per post

Figure 7.4: University Performance Scorecard



7.5 Advanced Visualizations

The dashboard includes multiple analytical visualizations for deeper insights:

- Sentiment distribution box plots
- Radar charts comparing institutional profiles
- Stacked bar charts for discussion categories
- Word clouds for dominant terms
- Time-series sentiment trend plots

Figure 7.5: Sentiment Distribution Box Plot

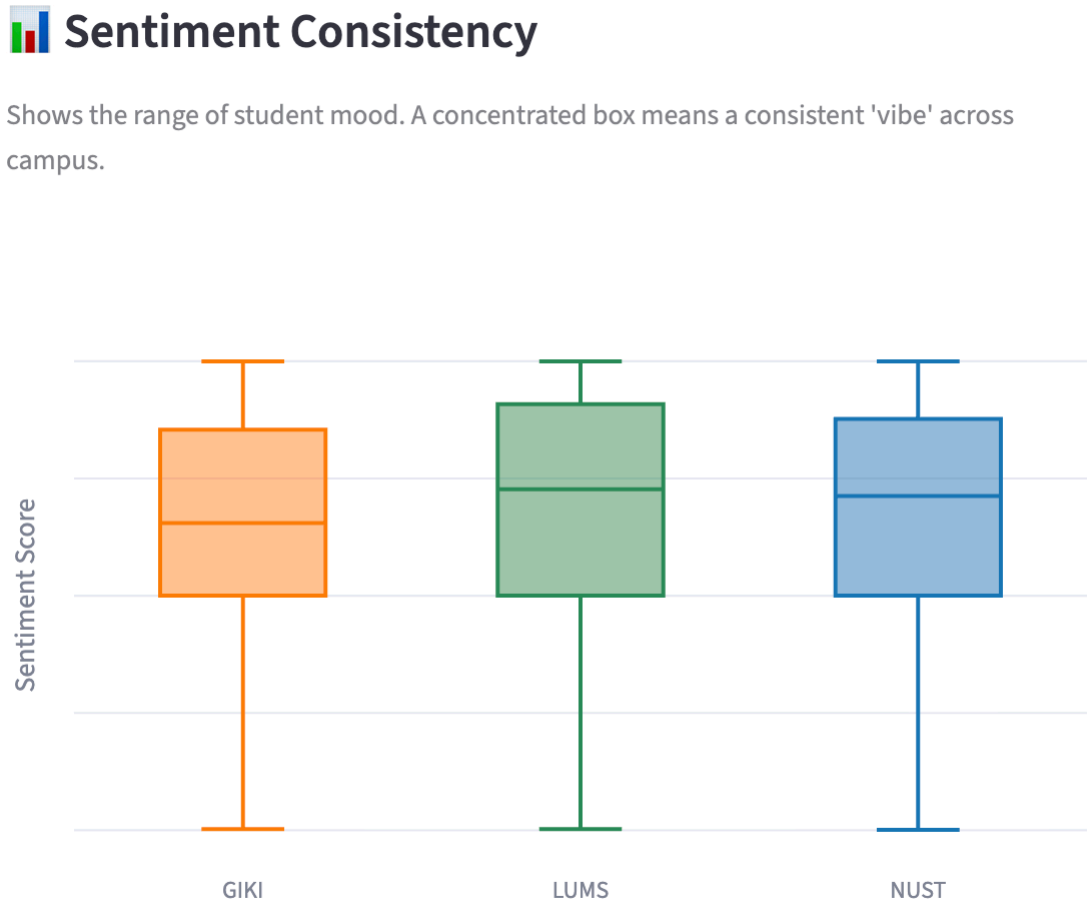


Figure 7.6: University Comparison Radar Chart

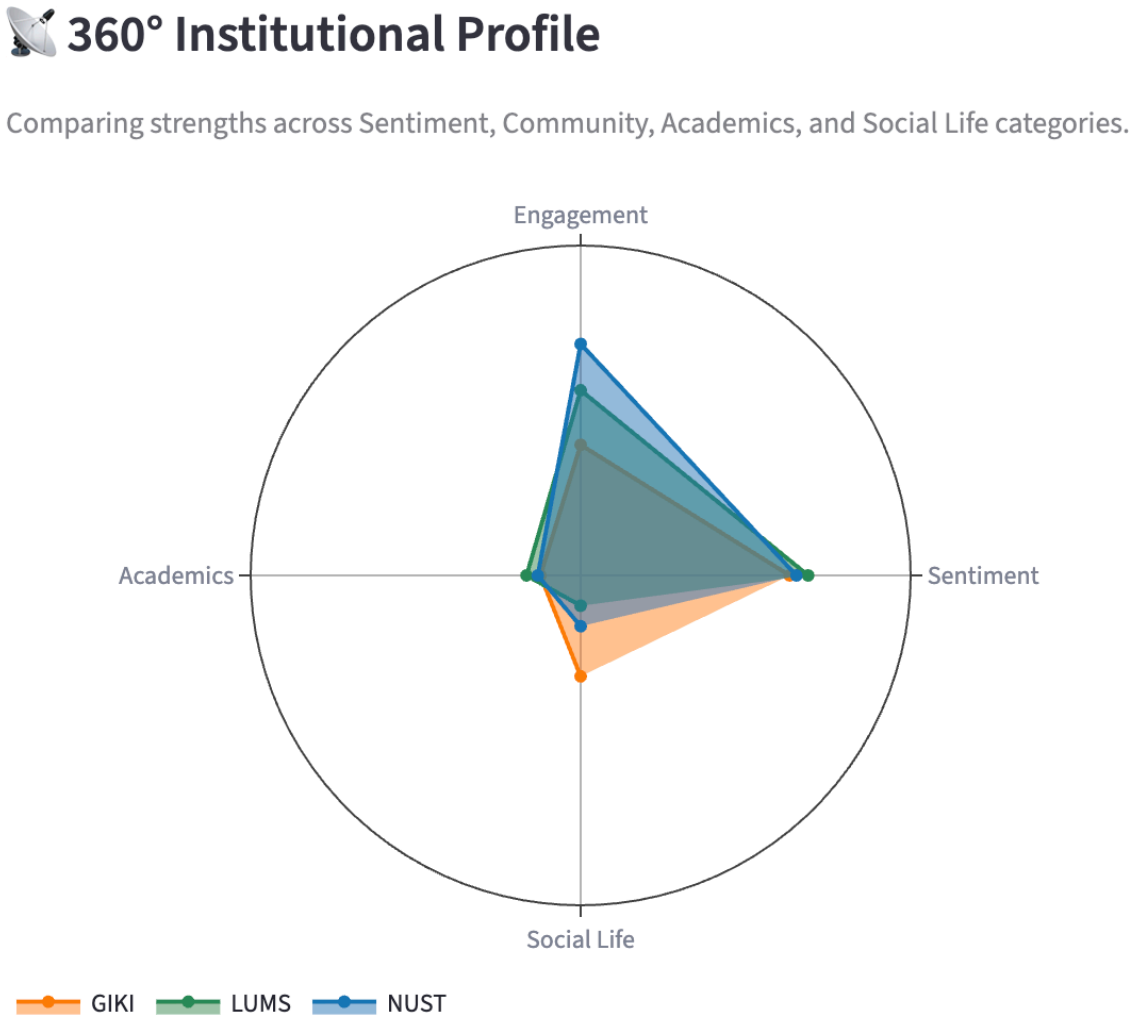


Figure 7.7: Discussion Category Distribution (Stacked Bar Chart)

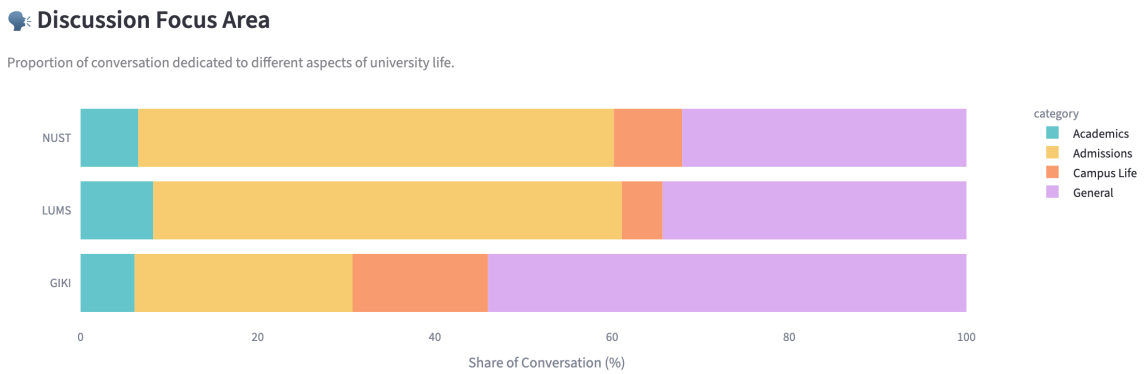


Figure 7.8: Keyword Word Cloud per University

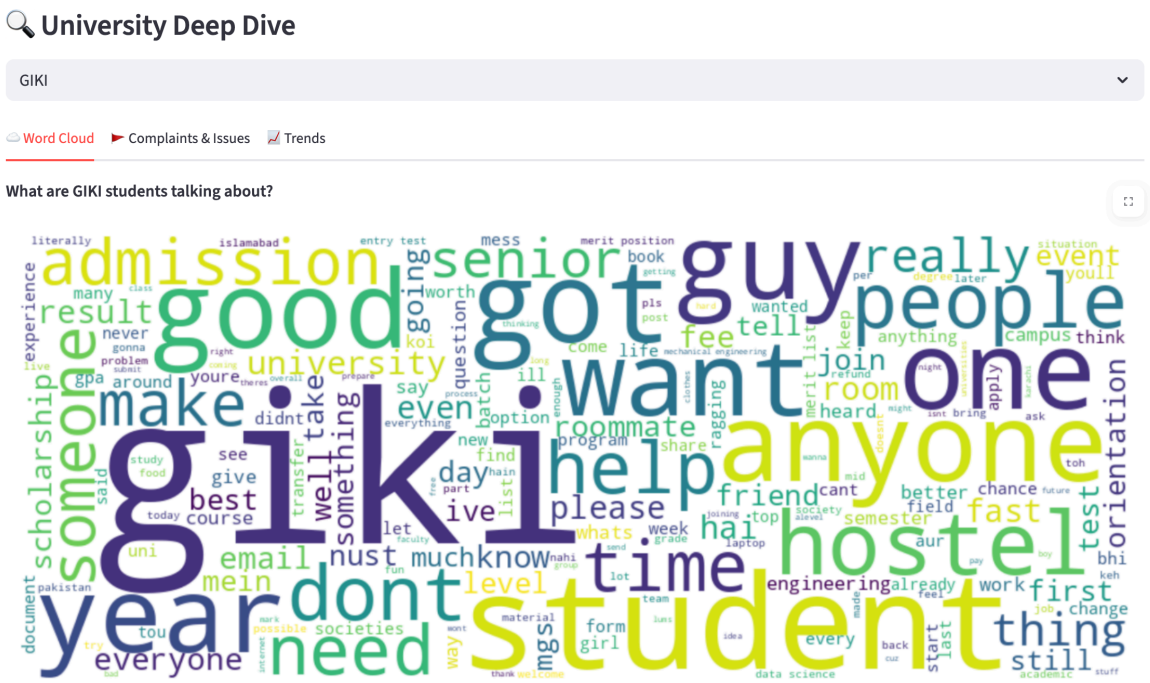
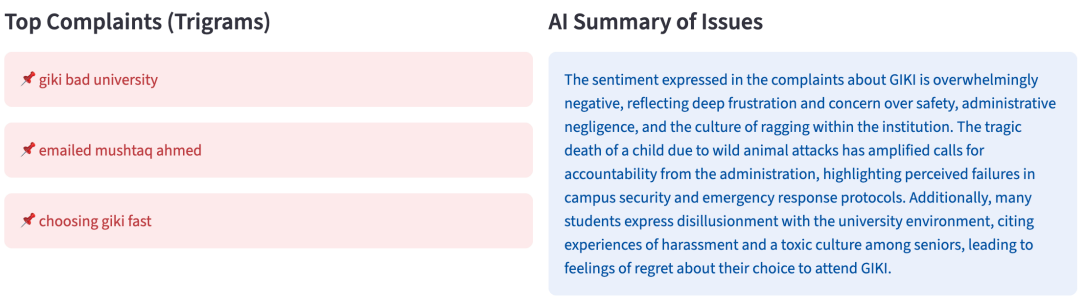


Figure 7.9: Sentiment Trends Over Time

Figure 7.10: Complaints and Isuues (AI Summary of issues)



7.6 Deployment & Accessibility

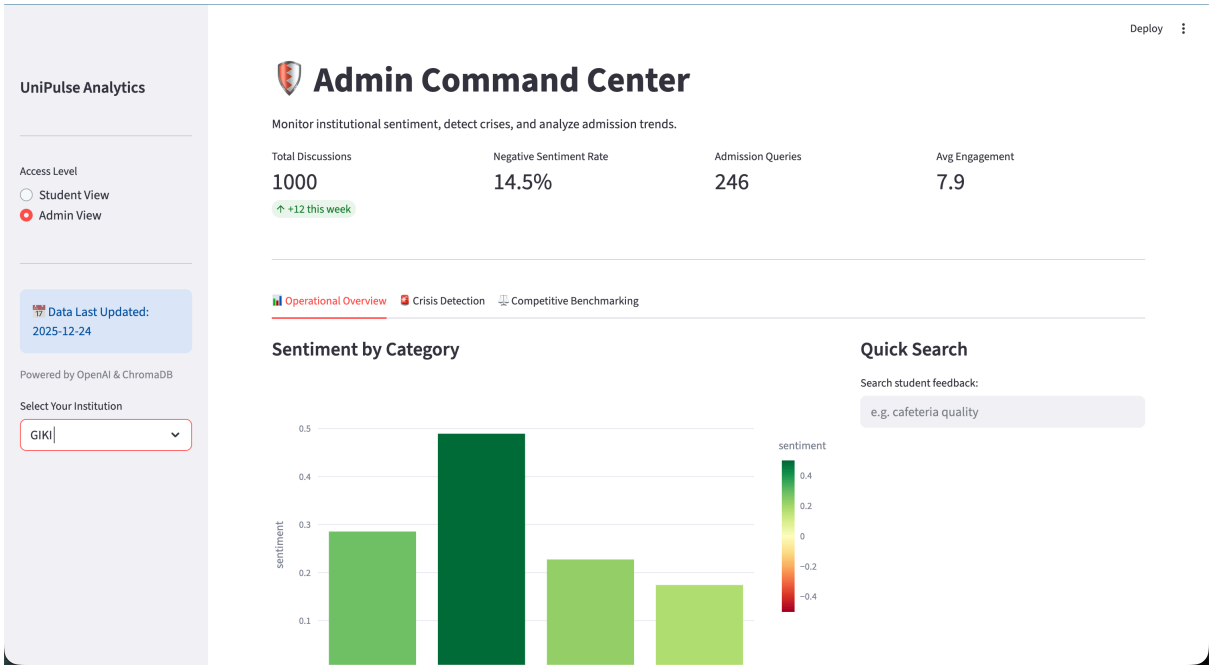
The dashboard is deployed using Streamlit and can be hosted locally or on cloud platforms. It supports modular extension for additional universities and analytics.

8. Visualization & Dashboard(Admin)

8.1 Dashboard Overview

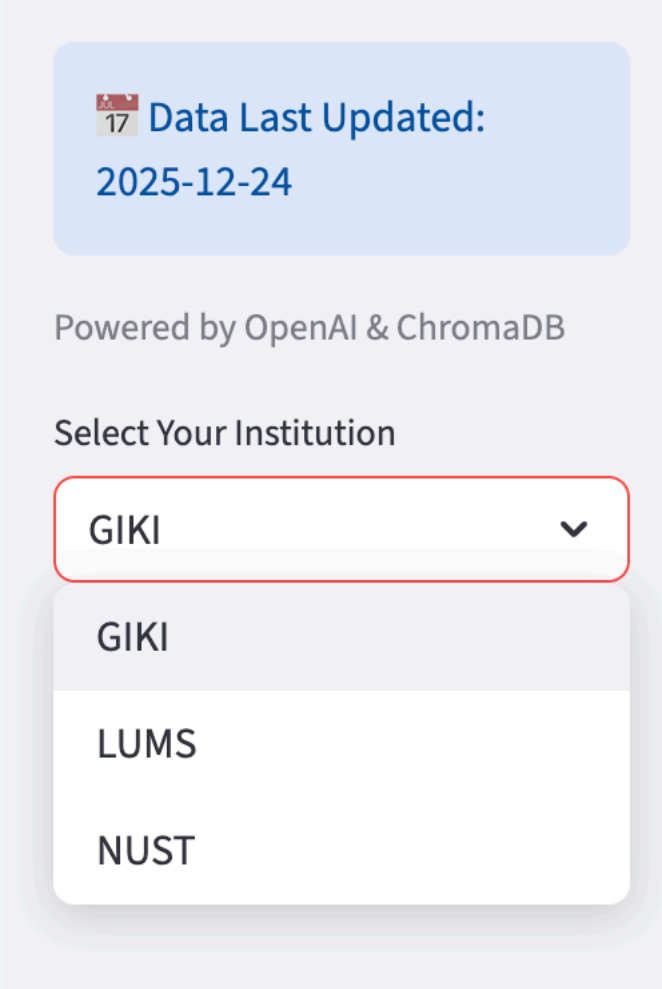
The UniPulse Analytics dashboard serves as the primary interface for presenting analytical insights in an intuitive and visually rich format.

Figure 8.1: UniPulse Analytics – Main Dashboard Interface(Admin)



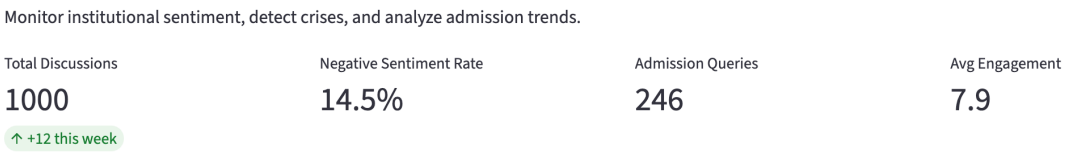
8.2 University Selection for Admin.

Figure 8.2: Select univeristy(Admin)



8.3 Monitor institutional sentiment, detect crises, and analyze admission trends.

Figure 8.3: Monitor institutional sentiment, detect crises, and analyze admission trends.(Admin)



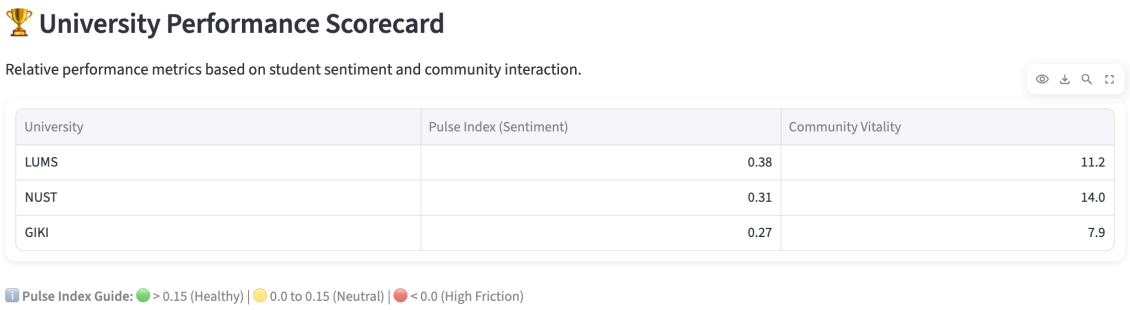
8.4 University Performance Scorecard

This table compares universities using sentiment and engagement-based metrics.

Metrics include:

- **Pulse Index**: Average sentiment score
- **Community Vitality**: Average number of comments per post

Figure 8.4: University Performance Scorecard(Admin)



8.5 Advanced Visualizations

The dashboard includes multiple analytical visualizations for deeper insights:

- Sentiment distribution box plots
- Radar charts comparing institutional profiles
- Stacked bar charts for discussion categories
- Word clouds for dominant terms
- Time-series sentiment trend plots
- Sentiment By Category
- Crisis Detection

Figure 8.5: Sentiment Distribution Box Plot(Admin)

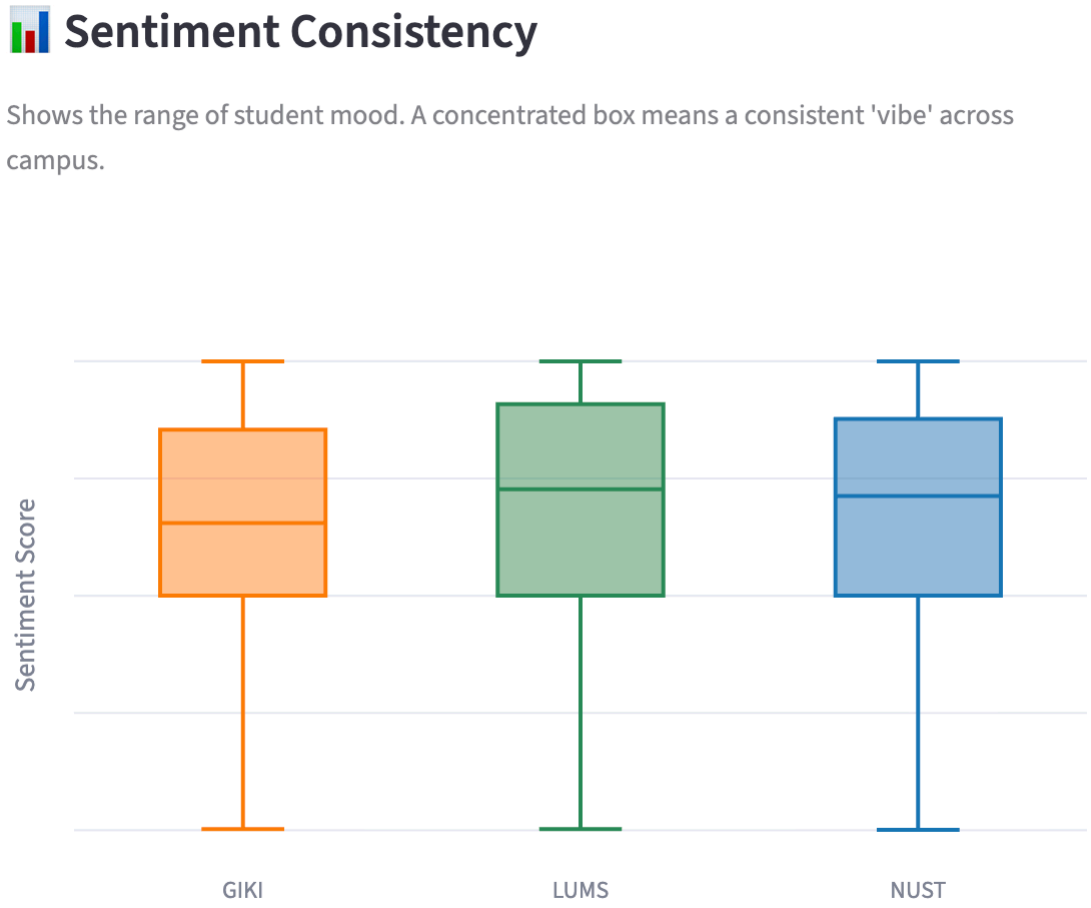


Figure 8.6: University Comparison Radar Chart(Admin)

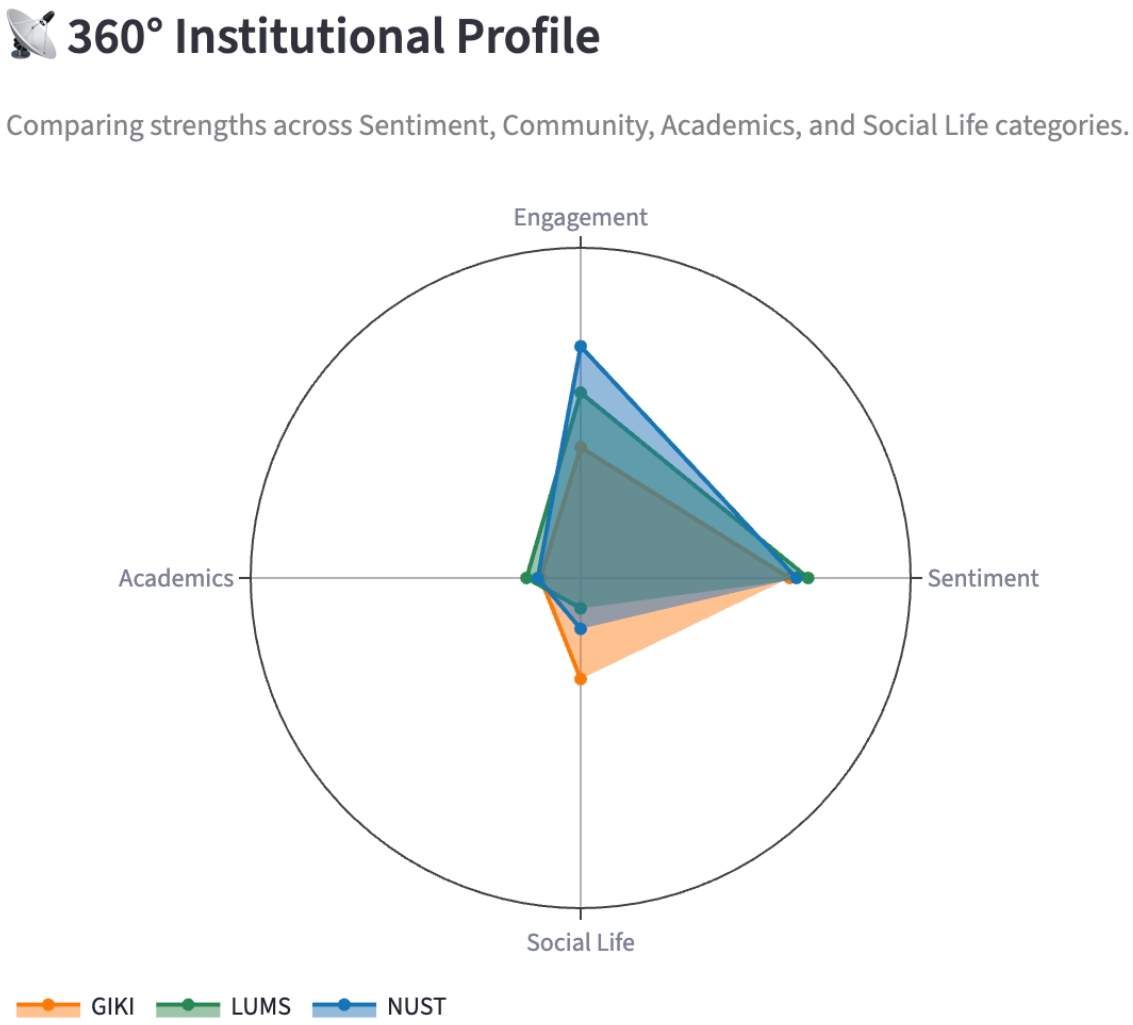


Figure 8.7: Discussion Category Distribution (Stacked Bar Chart) (Admin)

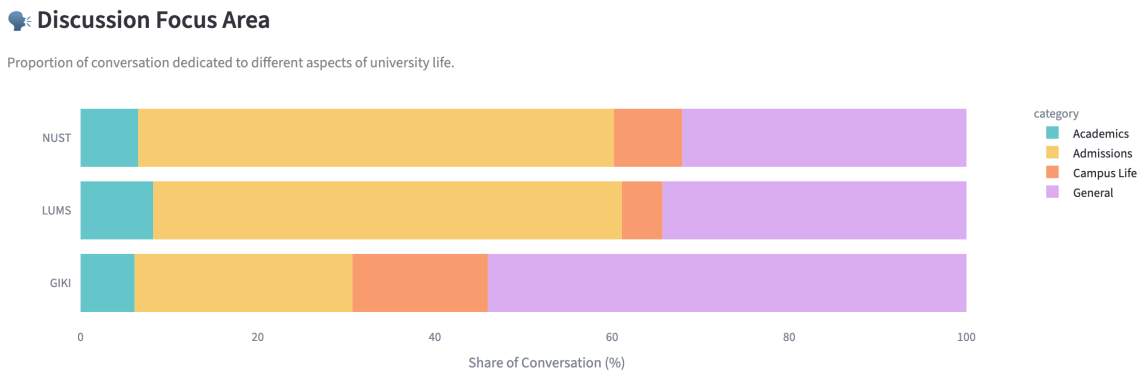


Figure 8.8: Keyword Word Cloud per University(Admin)

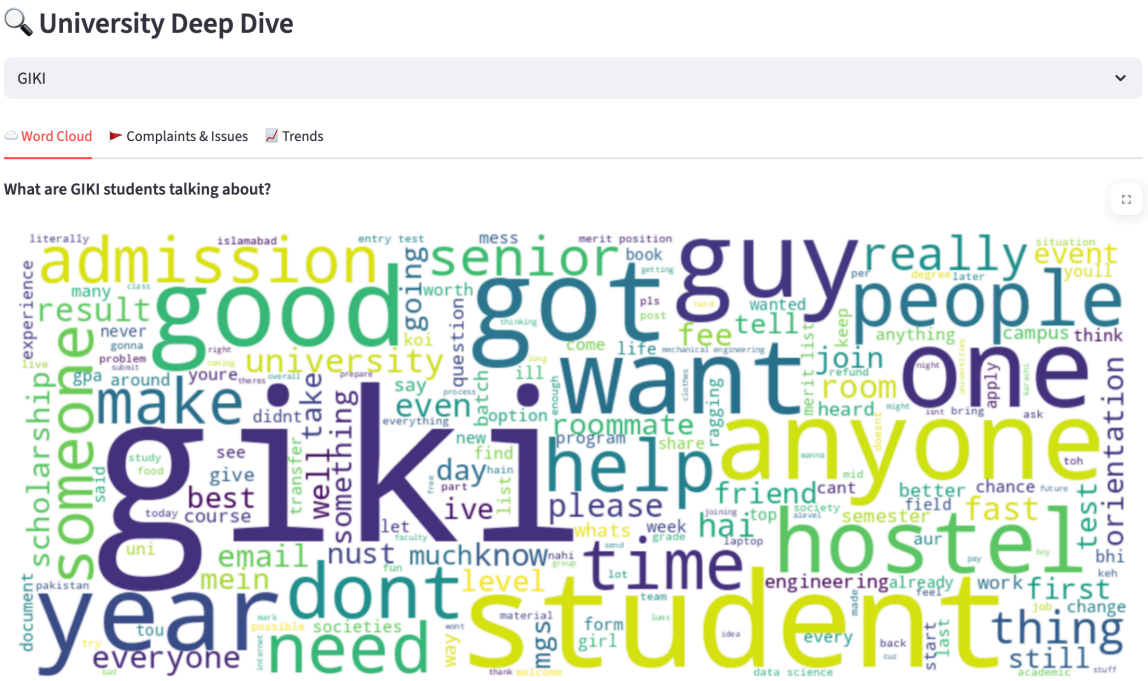


Figure 8.9: Sentiment Trends Over Time(Admin)

Figure 8.10: Complaints and Isuues (AI Summary of issues) (Admin)

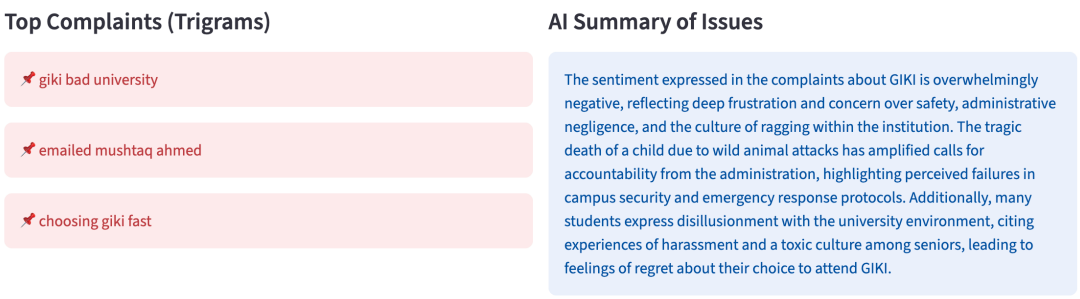


Figure 8.11 : Sentiment by Category(Admin)

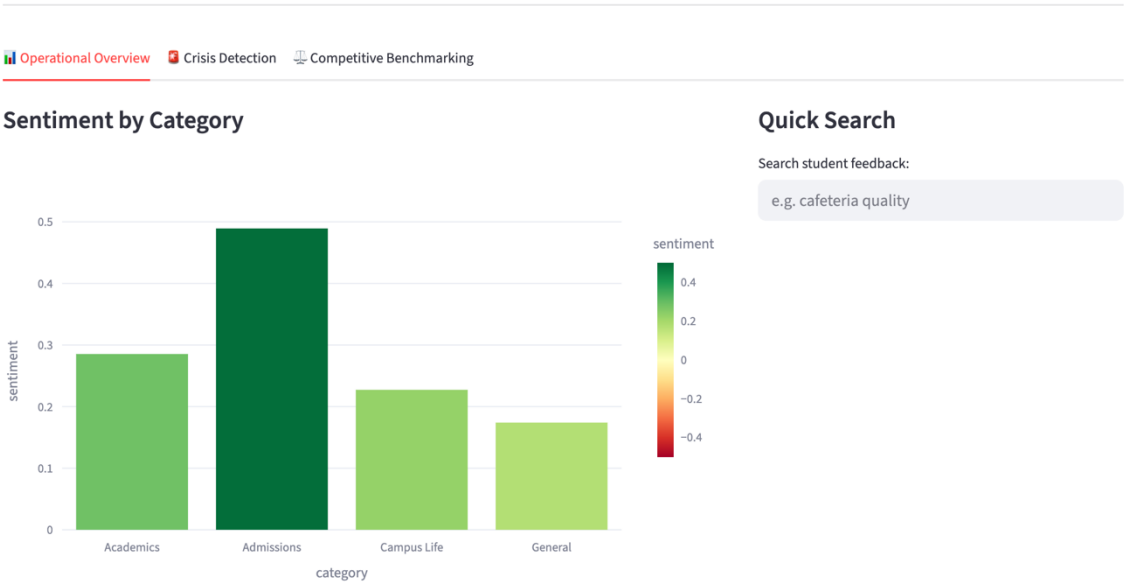
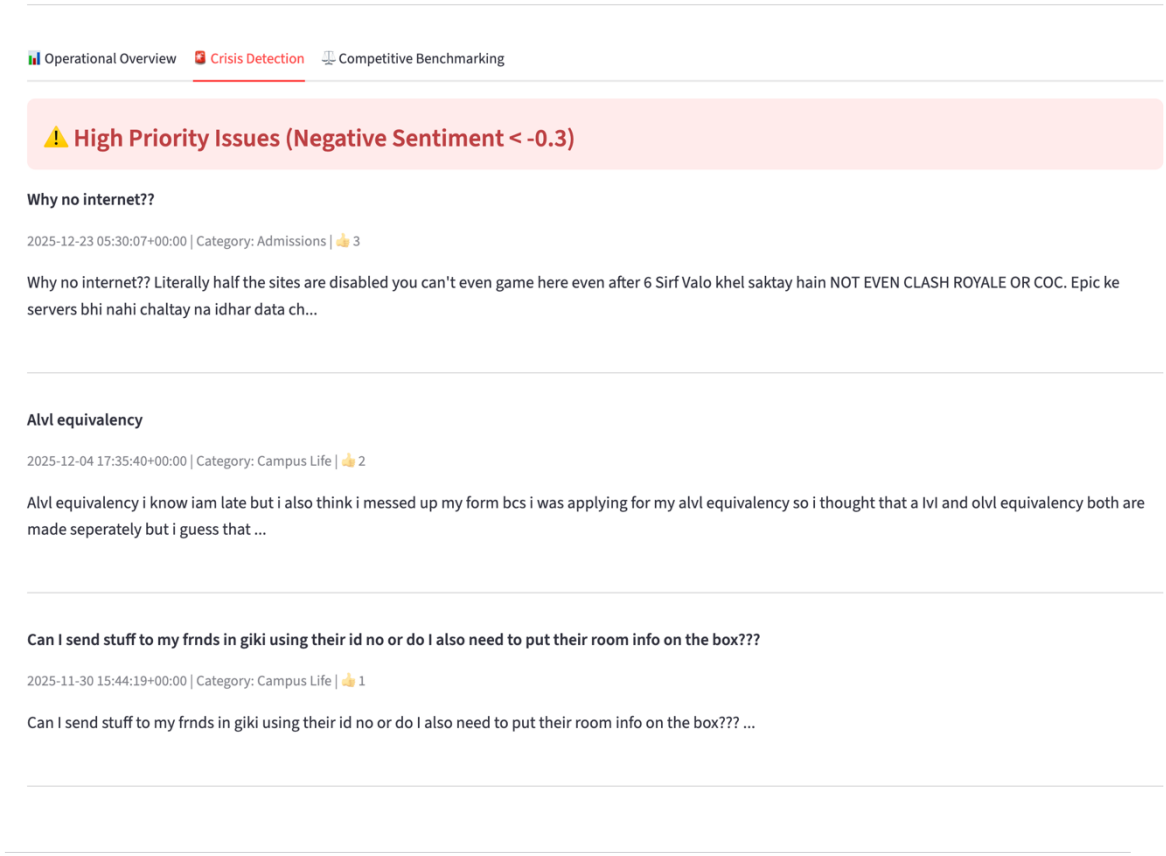


Figure 8.12: *Crisis Detection(Admin)*



8.6 Deployment & Accessibility

The dashboard is deployed using Streamlit and can be hosted locally or on cloud platforms. It supports modular extension for additional universities and analytics.

9. Results & Findings

Key insights observed:

- Universities exhibit distinct sentiment “signatures”
- High engagement does not always correlate with positive sentiment
- Academics-related discussions show higher sentiment volatility
- Campus life discussions tend to be more positive and expressive

10. Limitations

- Reddit users may not represent the entire student population
- Sentiment analysis may misinterpret sarcasm
- Topic modeling depends on sufficient text volume

11. Future Work

- Expand to additional universities (FAST, PIEAS, IBA)
- Integrate multilingual (Urdu/Roman Urdu) sentiment models
- Add anomaly detection for crisis monitoring
- Deploy dashboard on cloud infrastructure

12. Conclusion

UniPulse Analytics demonstrates how social media data can be transformed into meaningful institutional intelligence. By combining large-scale data scraping, NLP, and AI-powered retrieval, the system delivers transparent, student-centric insights in a professional, market-ready format.

13. GitHub Repository

GitHub Link: https://github.com/HAK-04/NLP-Analytics-Dashboard

Note: The repository includes scraping scripts, processing pipelines, json data files and this report.

References

- Reddit API Documentation
- NLTK & VADER Sentiment Analysis
- scikit-learn LDA Documentation
- ChromaDB Vector Store
- Streamlit Framework