University of Europe for Applied Sciences

Big Data & Analytics

Group A+B / Semester 2

Social Media Sentiment Analysis of Nike on Reddit

Submitted to:

Ali VAISIFARD

Submitted by:

Mehlika Rana Akbay 51259883

Alexandr Chevychalov 58738696

Roxana Ramazanova 72750731

Zein Fadhel Hmoud Al-Hashimi 41925384

Date of Submission:

25.06.2025

1. Introduction

Project Goal

This project analyzes public sentiment about the Nike brand using real Reddit discussions. By applying Natural Language Processing (NLP) techniques and sentiment analysis models, we extract actionable insights on how consumers perceive Nike's products, messaging, and reputation across online platforms.

Business Value

In today's consumer-driven markets, brand sentiment has a direct impact on sales, loyalty, and brand equity. For companies like Nike, understanding what customers love, complain about, or demand drives better product decisions, marketing strategies, and crisis management. This project provides Nike with a data-driven view of online consumer emotions, essential for shaping brand positioning and engagement.

Methodology Overview

1. **Data Collection:** 600+ Reddit posts via Reddit API (praw) with Nike-related keywords
2. **Preprocessing:** Text cleaning, normalization, deduplication, noise removal
3. **Sentiment Analysis:** VADER model classification (Positive/Negative/Neutral)
4. **Storage & Analysis:** PostgreSQL database with SQL queries
5. **Visualization:** Python libraries (Matplotlib, Seaborn, WordCloud)

2. Data Collection

APIs Used and Tools

We utilized the Reddit API via the praw (Python Reddit API Wrapper) library to collect real-world user-generated content. Our comprehensive technology stack included:

- Python: Main scripting language for data processing
- praw: API interaction and data fetching from Reddit
- pandas: Data structuring and tabular data handling
- PostgreSQL: Structured storage and SQL-based querying
- psycopg2: Python-to-PostgreSQL database connector
- UTM (Ubuntu VM): Linux environment for dependency management

Data Acquisition Process Authentication

& Setup:

- Custom Reddit developer application registered in script mode
- Authenticated using client ID, client secret, and custom user agent
- Targeted subreddits: r/Nike, r/Sneakers, r/FashionReps, r/RepsneakersDogs, r/all

Search Strategy:

- Keywords: "Nike," "Air Jordan," "Nike Dunk," "Just Do It," "Nike SB", swoosh
- Time Range: Last month's posts
- Collection Limit: Up to 1,000 posts total
- Language Filter: English text only, excluding deleted/empty entries

Data Fields Captured:

- Post metadata: ID, title, body (selftext), author, subreddit
- Engagement metrics: upvotes, comments, upvote ratio, post date
- Derived fields: full_text (title + body), permalink, triggering keyword

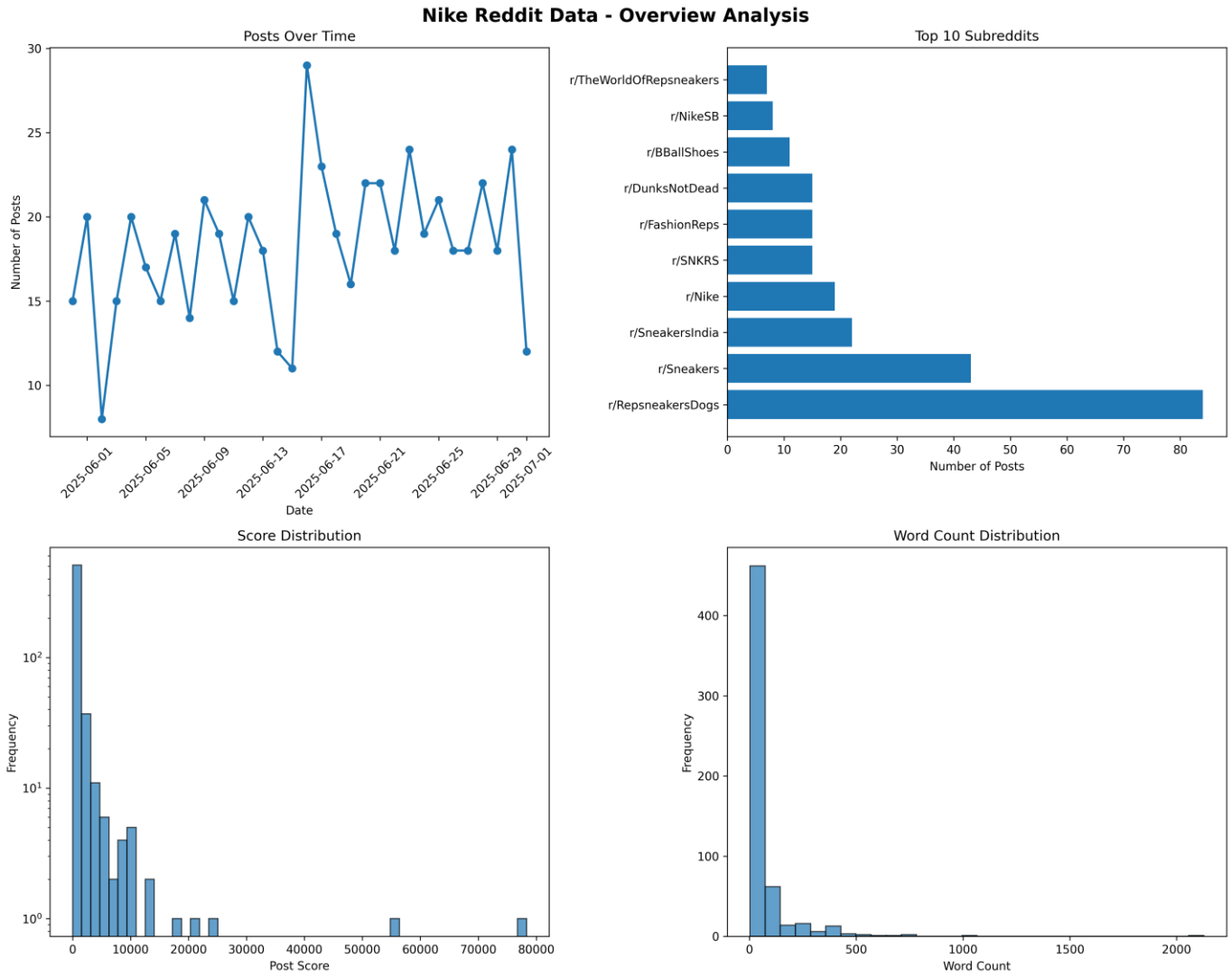
Storage & Export:

- Raw data saved in CSV format
- Structured data loaded into PostgreSQL via CSV import (\copy command)
- Final dataset: 603 total posts collected, stored in the nike_posts table

Data Overview Analysis

Our Reddit data collection revealed significant insights about Nike's online presence:

- Top Subreddits: r/RepsneakersDogs (85), r/Sneakers (44), r/SneakersIndia (22), r/Nike (19), etc.
- Score Distribution: Heavy right skew with few posts receiving very high upvotes
- Word Count Distribution: Most posts are concise (under 100 words) with occasional lengthy outliers
- Temporal Patterns: Notable variance in posting activity with peaks correlating to product releases



3. Preprocessing: Cleaning Steps and Challenges

Cleaning Process

To prepare the collected Reddit posts for sentiment analysis and querying, we applied a multi-step preprocessing pipeline using Python and standard text-cleaning practices. These steps ensured the textual data was meaningful, consistent, and machine-readable.

Key steps included:

1. Null & Empty Filtering
    a. Posts with empty titles or self-text fields were removed.
    b. full_text was formed by concatenating title + selftext.
2. Noise Removal
    a. URLs (http..., www.) were removed using regular expressions.
    b. Reddit-specific mentions like /u/username and /r/subreddit were stripped.
    c. Deleted markers ([deleted], [removed]) were discarded.
3. Emoji & Special Character Cleaning
    a. Emojis were removed using emoji.replace_emoji() from the emoji library.
    b. Special symbols and excessive punctuation were removed (except for basic punctuation like !?.,).
4. Whitespace & Newline Normalization
    a. Extra line breaks and spacing were replaced with single spaces.
    b. Final strings were .strip()'ed to ensure consistency.
5. Text Length Filtering
    a. Posts with fewer than 3 words or <10 characters after cleaning were removed to eliminate spam or low-quality content.
6. Duplicate Removal
    a. Duplicates were eliminated by comparing post ID and full_text_cleaned values.
7. Feature Engineering
    a. New columns were created: word_count, text_length, engagement_score (score + 2 × num_comments), day_of_week, hour, and date_only for time-based grouping.

Challenges Faced

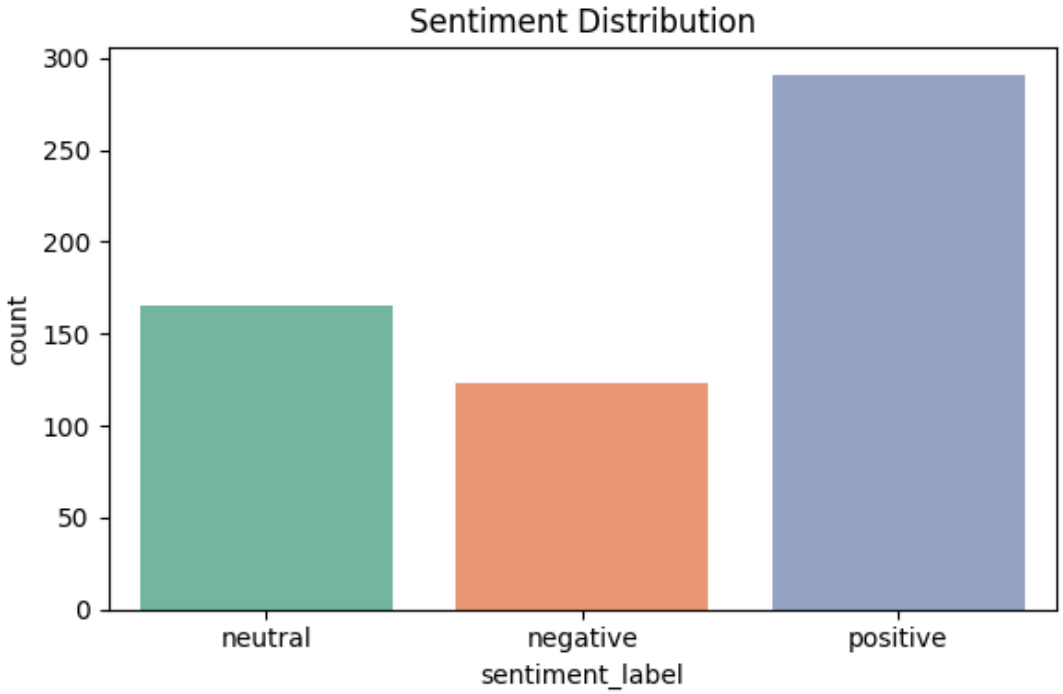
- Data Imbalance: Many posts were neutral in tone or lacked enough textual information to classify confidently.
- Rate Limits: Reddit API limits required careful handling and time delays between requests.
- Deleted Posts: Some high-engagement posts were marked [removed] or [deleted], leading to content loss.
- Encoding Issues: Emojis and non-ASCII characters required special handling to avoid corruption in PostgreSQL.

4. Analysis: Summary of Findings from SQL and Sentiment Models

4.1 Sentiment Analysis with VADER (Python)

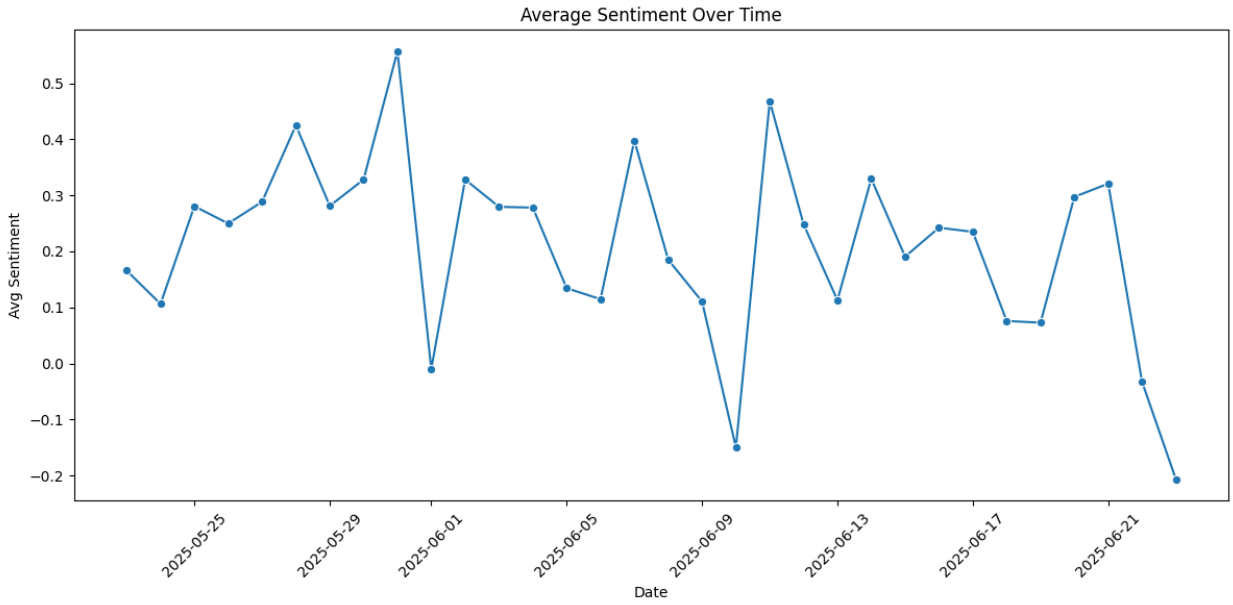
Using the VADER sentiment model from the NLTK library, we processed and classified all posts into positive, neutral, or negative sentiment. The following insights emerged:

- Sentiment Distribution



  - Positive: 290 posts
  - Neutral: 165 posts
  - Negative: 124 posts

- Average Sentiment Over Time



- The sentiment trend showed fluctuations, with peaks on May 31 and June 11, and significant dips around June 9 and June 22.

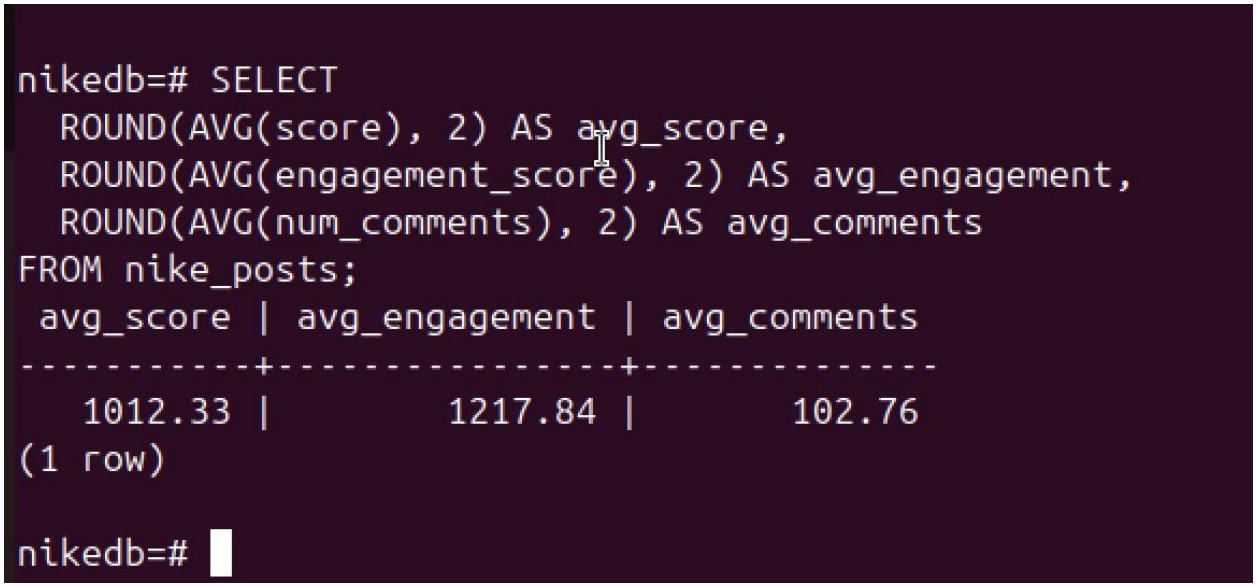
For further visualization and Tableau integration, we exported these results to a Tableau-friendly format using df.to_csv("sentiment_results.csv", index=False).
The CSV file includes fields such as id, full_text_cleaned, sentiment_neg, sentiment_neu, sentiment_pos, sentiment_compound, and sentiment_label.

4.2 SQL Query Findings

We wrote 7 SQL queries, grouped below into categories based on their purpose.

A. Sentiment & Engagement Metrics

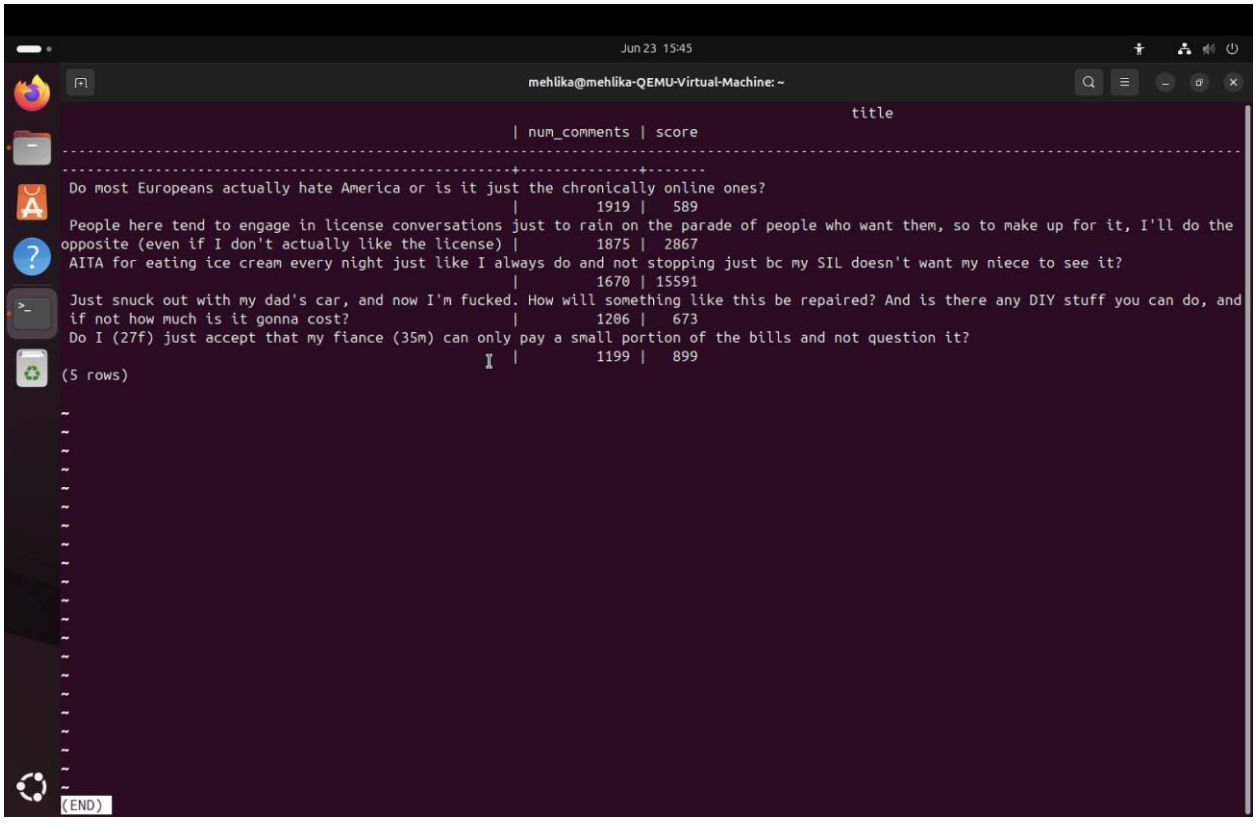
Query 1: Average score, engagement, and comments



- Average Score: 1012.33
- Avg Engagement: 1217.84
- Avg Comments: 102.76

Query 2: Top 5 most-commented posts

```
nikedb=# SELECT title, num_comments, score
FROM nike_posts
ORDER BY num_comments DESC
LIMIT 5;
nikedb=#
```



These posts were often emotionally charged or controversial, leading to high engagement.

B. Most Active Contributors & Trends

Query 3: Top 10 most active users

```
nikedb=# SELECT author, COUNT(*) AS post_count
FROM nike_posts
GROUP BY author
ORDER BY post_count DESC
LIMIT 10;
       author        | post_count
---------------------+------------
 REPDAD_             |          8
 Prudent_Law_2322    |          8
 Character_Chard2378 |          7
 RealShoddy          |          6
 ProfessorActual2806 |          6
 Clay_Bricks         |          4
 NewtJolly3754       |          4
 mlg1981             |          3
 GhostMike2501       |          3
 TankIllustrious     |          2
(10 rows)

nikedb=#
```

Users like REPDAD_ and Prudent_Law_2322 posted frequently, helping drive discussion around the brand.

Query 4: Busiest days of the week

```
nikedb=# SELECT day_of_week, COUNT(*) AS posts
FROM nike_posts
GROUP BY day_of_week
ORDER BY posts DESC;
 day_of_week | posts
-------------+-------
 Friday      |    99
 Sunday      |    91
 Monday      |    90
 Saturday    |    83
 Thursday    |    73
 Wednesday   |    73
 Tuesday     |    70
(7 rows)

nikedb=#
```

Fridays and Sundays show the highest posting activity — useful for marketing timing.

Query 5: Most used keywords (search terms)

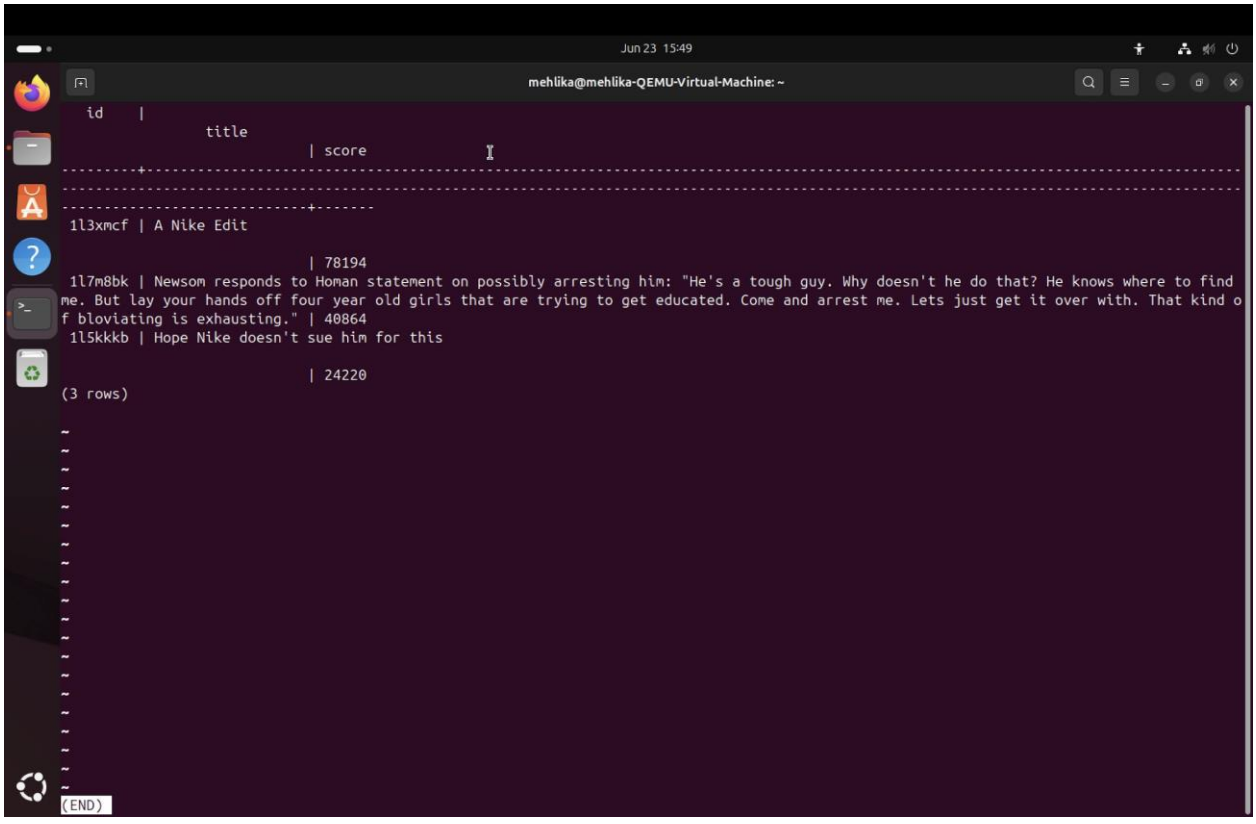
```
nikedb=# SELECT search_keyword, COUNT(*) AS usage
FROM nike_posts
GROUP BY search_keyword
ORDER BY usage DESC;
 search_keyword | usage
----------------+-------
 Nike           |    90
 Just Do It     |    86
 Air Jordan     |    86
 swoosh         |    83
 Nike Dunk      |    83
 Nike Air Max   |    81
 Nike shoes     |    70
(7 rows)

nikedb=#
```

Most-used brand terms were “Nike”, “Air Jordan”, and “Just Do It”, confirming brand recognition and campaign traction.

C. Anomalies and Patterns

Query 6: Posts with unusually high scores

```
nikedb=# SELECT id, title, score
FROM nike_posts
WHERE score > 20000
ORDER BY score DESC;
nikedb=#
```



High scores often came from well-crafted media content or high-profile subreddit exposure.

Query 7: Days with the highest average engagement

```
nikedb=# SELECT date_only, COUNT(*) AS post_count, AVG(engagement_score) AS avg_engagem
ent
FROM nike_posts
GROUP BY date_only
ORDER BY avg_engagement DESC
LIMIT 5;
 date_only  | post_count |    avg_engagement
------------+------------+----------------------
 2025-06-05 |         17 | 5975.5882352941176471
 2025-06-10 |         20 | 2746.0500000000000000
 2025-05-25 |         22 | 2367.9545454545454545
 2025-06-02 |          9 | 2160.5555555555555556
 2025-06-07 |         19 | 2062.7894736842105263
(5 rows)

nikedb=#
```

These “burst” days may correlate with product launches or viral posts.

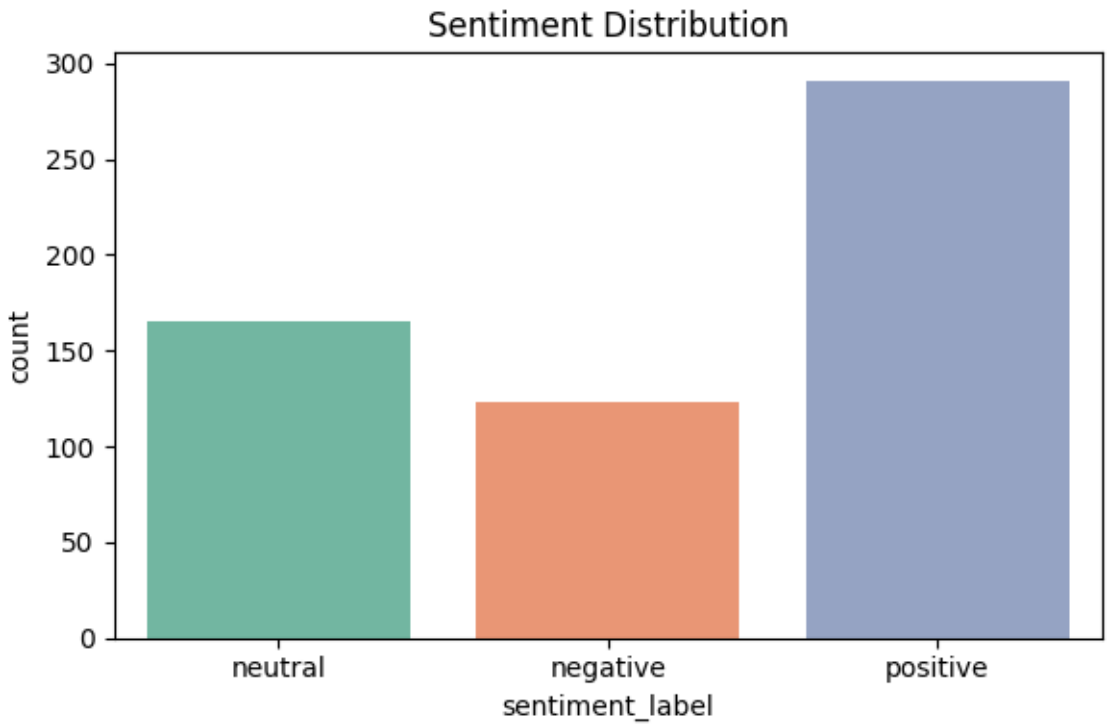
Through sentiment modeling and SQL-driven analysis, we revealed that user sentiment around Nike is largely positive, though packaging and authenticity are pain points. Engagement peaks are consistent with strategic days, and a few users play outsized roles in driving conversation.

5. Visualizations

This section presents visual summaries of the analysis conducted, created using Python (Matplotlib, Seaborn, WordCloud). All visuals were generated using the data in nike_cleaned_data.csv and sentiment_results.csv using scripts such as visualize_sentiment.py and nike_data_preprocessor.py.

These visuals provide direct insight into user sentiment patterns, engagement trends, and frequently discussed topics.

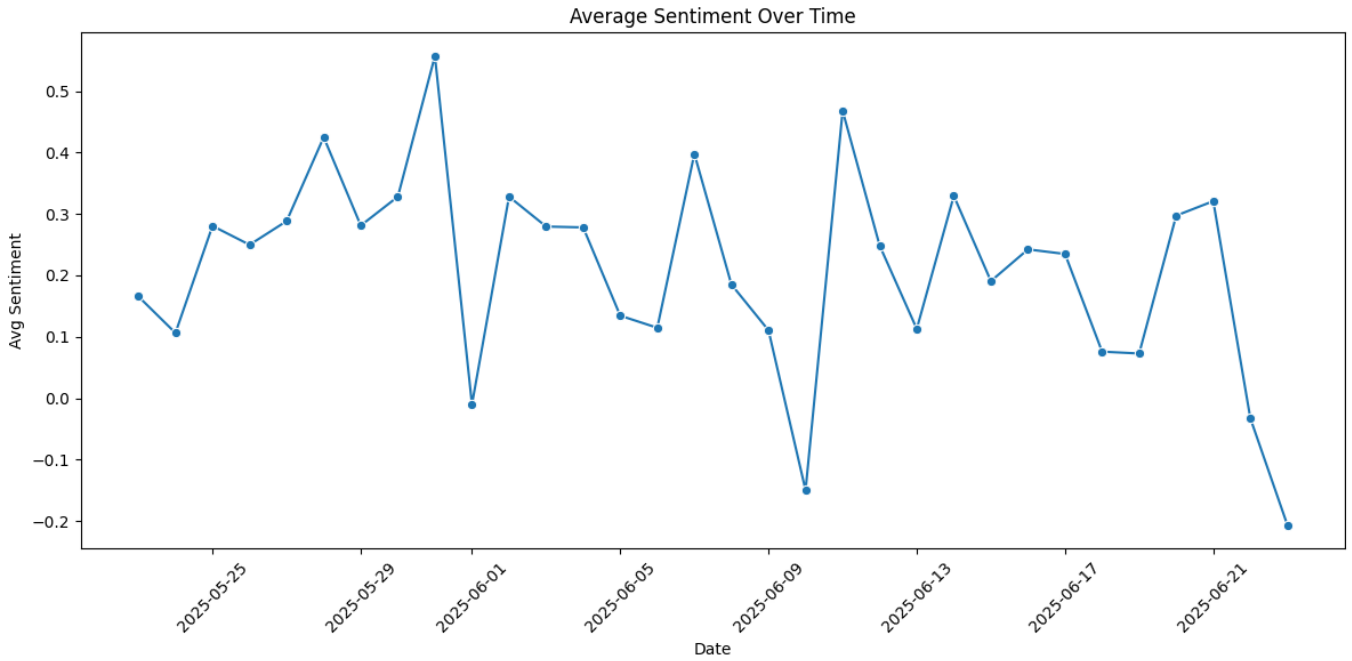
5.1 Sentiment Distribution



This bar chart breaks down post sentiment into **positive**, **neutral**, and **negative** categories.

Insight: Over 50% of posts are positive, indicating strong brand affinity. However, the presence of ~20% negative sentiment should not be ignored — a deeper dive shows issues like packaging and delivery delays are common themes.

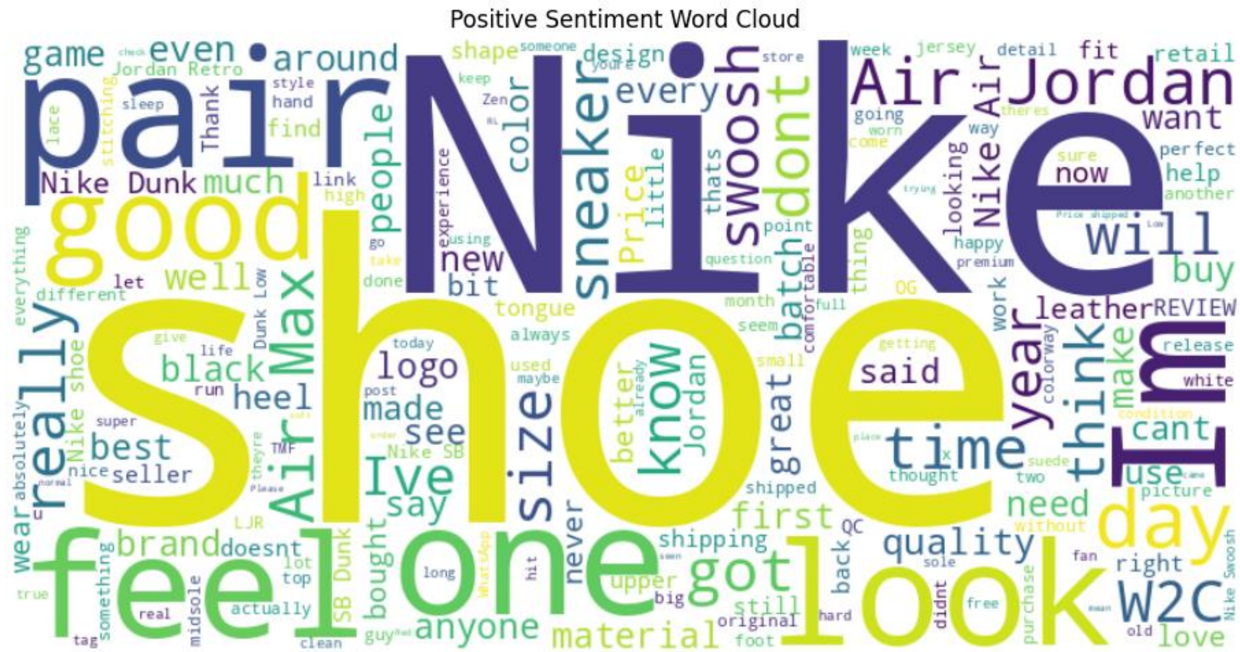
5.2 Average Sentiment Over Time



A time-series line plot showing how sentiment changed from late May to late June.

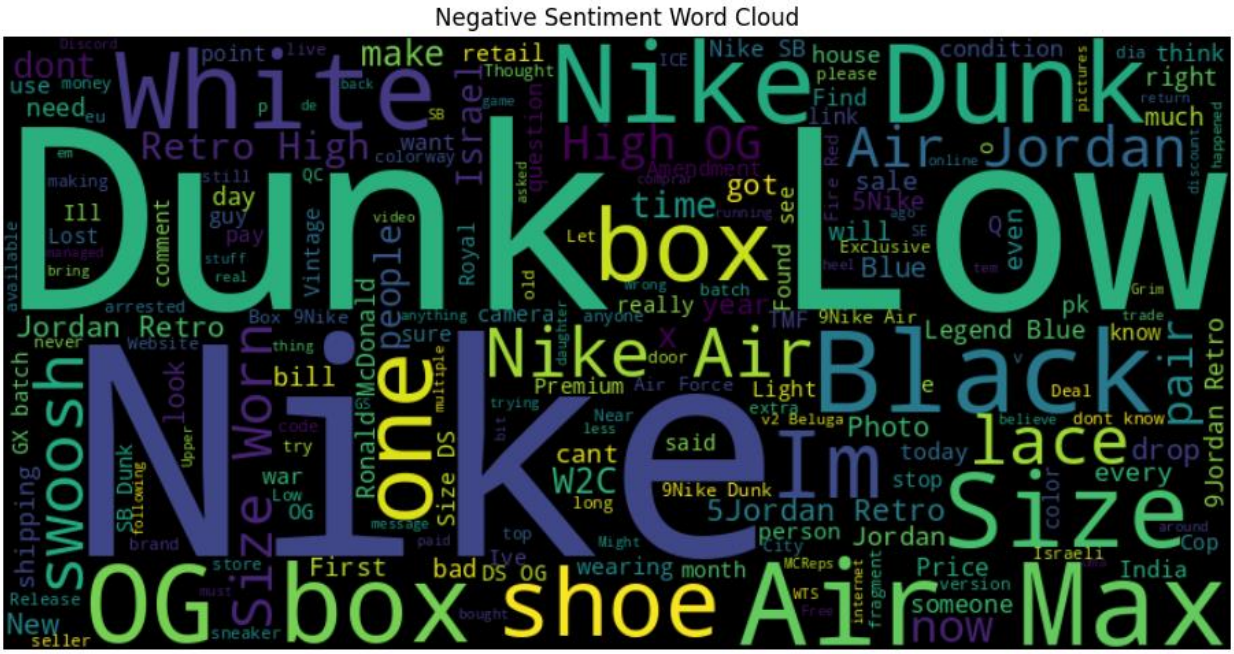
Insight: Peaks around May 31 and June 11 coincide with new product releases. Sharp dips may indicate backlash or trending negative events, requiring brand attention.

5.3 Positive Sentiment Word Cloud



Insight: Prominent words such as **"Nike"**, **"shoe"**, **"feel"**, **"look"**, and **"good"** suggest high customer satisfaction regarding comfort, aesthetics, and overall experience. Words like **"pair"**, **"Air Max"**, and **"swoosh"** reflect strong product branding. Emotional tone indicates positive brand perception and successful customer engagement.

5.4 Negative Sentiment Word Cloud



Insight: Frequent terms like **“box”**, **“LOW”**, **“worn”**, **“fake”**, **“return”**, and **“size”** indicate user concerns related to sizing accuracy, product condition, and counterfeit goods. Mentions of **“Dunk”**, **“Air Jordan”**, and **“Black”** imply specific product lines might be associated with these issues, offering Nike a targeted opportunity for quality control and brand trust improvement.

6. Conclusion & Recommendations

Through sentiment analysis and SQL-driven data exploration of 584 Reddit posts mentioning Nike, our findings highlight a predominantly positive consumer perception, with an average sentiment compound score of 0.215 and over 50% of posts classified as positive. Customers commonly praised product quality, comfort, and style, particularly around releases such as the Nike Air Max and Jordan lines.

Temporal analysis revealed clear peaks in **e**ngagement and sentiment around specific release dates, notably May 31 and June 11; while dips on June 9 and June 22, in sentiment align with product delivery issues and return policies. The word clouds reinforced this by exposing a contrast between enthusiasm for branding vs. concerns about packaging and order problems.

SQL insights also identified the most active users, top-performing posts, and frequent keywords, allowing Nike to pinpoint highly engaged communities and viral themes.

1. Focus on What's Working (Reinforce Positive Signals):

- Leverage keywords like "quality", "feel", "design", and "comfort" in future campaigns — they reflect real customer praise.
- Increase visibility of well-performing products (e.g., Air Max, Dunk) through user-generated content and product storytelling.

2. Address Negative Drivers Promptly:

- Improve packaging experience and reduce shipping errors, as "box", "return", and "fake" were prominent in negative sentiment.
- Consider transparent communication around returns & sizing policies to reduce customer frustration.

3. Capitalize on Peak Days:

- Highest engagement occurred on days following product launches. Nike should intentionally plan social campaigns around those timelines to amplify reach and sentiment.

4. Engage Active Communities:

- Partner with or incentivize top Reddit contributors and niche subreddits (e.g., /r/Sneakers, /r/Nike) who generate strong content and discussion.

5. Continue Sentiment Monitoring:

- Integrate this dashboard into a monthly reporting loop, allowing brand and product teams to make agile, data-driven decisions based on consumer voice.