Market Intelligence from Twitter Data

Developer: Neha Garai

Project Details:

To extract real-time Indian market sentiment signals from Twitter using a

logged-in browser automation system and generate actionable trading intelligence.

Key Market Focus:

• NIFTY50

• Sensex

• BankNifty

• Intraday trading sentiment

Architecture:

Main Components

- Selenium-based scraper using Chrome User Profile
- Data cleaning and deduplication
- Sentiment analysis + TF-IDF text intelligence
- Composite market signal calculation
- Storage in Parquet (efficient for analytics)

Data Flow:
Twitter -> Selenium Scraper ->JSON -> Parquet ->Analysis ->Signals -> Insights

Tech Stack:

- Scrapping and Automation -> Selenium, chrome driver
- Data Processing -> Pandas, NumPy
- NLP & signal Analysis -> Text Blob, Scikit-Learn, (TF-IDF+SVD)
- Data Engineering: Parquet/ Pyarrow
- Logging: Python Logging Framework
- Visualization: Matplotlib

Data Extraction:

1. 2400 tweets collected in one session
2. Live feed from “Top market hashtags”
3. Automatic infinite scrolling + rate limit safety
4. Hash-based duplicate removal

Key Features:

- Real-time scraping from logged-in profile
- Automatic scroll-based tweet loading
- Duplicate removal using hash-set
- Sentiment scoring (polarity)
- TF-IDF signal for linguistic market strength
- Composite signal = sentiment + information score
- Efficient output for ML modelling
- Output in CSV + Parquet + visual insights

Dataset field:

1. Username -> Twitter handle
2. Content -> Tweet text
3. Timestamp-> Original tweet timestamp
4. Sentiment -> Polarity: [-1, 1]
5. tfidf_strength -> Information density score
6. signal -> Composite normalized trading signal
7. hashtags -> Extracted tags
8. mentions -> Extracted user mentions

Signal and Insights:

- Higher signal = more bullish market sentiment
- Bank Nifty & Intraday show highest sentiment volatility
- NIFTY50 sentiment reacts strongly near macro events

Error Handling & Anti-Bot Measures:

- Delays between scrolls
- Randomized human behaviour timing
- Logged-in Chrome profile
- Disabled automation flags
- Try/except based safe parsing
- Logging for debugging and monitoring

How to Run:

1. Install dependencies:
2. All bash runnable code - pip install -r requirements.txt
3. Run scraper: python main_market_24h_2000_scraper.py
4. Run analysis: python analyze_tweets.py

Outputs stored inside:

output/

data.parquet

signals.csv

hashtag_summary.csv

*.png

Use Cases:

• Trend detection of Indian indices

• Macro event reaction monitoring

• Short-term volatility prediction

Future Improvements:

1. Engagement-weighted signals (likes, retweets)
2. Transformer-based sentiment (FinBERT)
3. Symbol mapping directly to stocks
4. Deployment as scheduled cloud job

Project Outcome:

A production-style, modular, and scalable real-time

market intelligence pipeline using social sentiment

Final Notes:

This system demonstrates skill in:

- Python automation
- Real-time data engineering
- NLP signal processing
- Market intelligence generation