Final Project Progress Report

Project Title: MASLD AWARENESS TRACKER

Project scope update

The project scope was expanded from three to five data sources to provide more comprehensive MASLD awareness tracking. Media Cloud was added for news media analysis, and stock market analysis was enhanced beyond the original plan. Kaggle datasets were deprioritized due to limited relevance. The project remains consistent with the original proposal and now integrates all five data sources to assess the impact of key FDA approvals on public awareness.

Data sources

Data Source	Status	API/Method Used	Data Obtained & Analysis
Google Trends	COMPLETE	pytrends API	Search interest trends with FDA event integration. Analysis includes timeline visualization and fallback to pre-collected data via Google Drive. 148 data points (2023-2025) for MASLD, NAFLD, Rezdiffra, Wegovy, Ozempic. Search interest trends validated in tests.py.
Reddit	COMPLETE	PRAW (Python Reddit API Wrapper)	9,255 posts/comments from 12 subreddits. VADER sentiment analysis implemented with development fallback to pre-processed data for consistent results.
PubMed	COMPLETE	NCBI E-utils API (Biopython)	1,344 scientific publications on MASLD/NAFLD and treatments. Timeline analysis with disease+drug mention combinations operational.
Stock Data	COMPLETE	Yahoo Finance API (yfinance)	707 trading days for Novo Nordisk (NVO) and Madrigal Pharmaceuticals (MDGL). Price trend visualization with FDA event markers implemented.
Media Cloud	COMPLETE	Google Drive (gdown package) + File Processing	39,376 news articles across disease-focused, Resmetirom-focused, and GLP-1-focused datasets. Comparative timeline and source analysis functional.

- Media Cloud is a new addition beyond original scope. Stock Data focus shifted from drug pricing to market impact analysis. Kaggle was intentionally excluded after evaluation.
- Project structure follows requirements with modular design: data collection (load.py), processing (process.py), analysis (analyze.py), configuration (config.py), and testing (tests.py). Comprehensive test suite (tests.py) validates API connectivity and data quality across multiple sources. Core pipeline provides basic visualizations while some advanced analyses are available in interactive Jupyter notebook (results.ipynb).
- GDrive Links Justification: Pre-collected datasets are hosted on Google Drive to ensure project reproducibility during evaluation and address potential API rate limiting with Google Trends and Reddit APIs. This provides immediate data access for testing the complete analysis pipeline without dependency on external API availability or web scraping constraints.
- API Implementation:
  - Reddit PRAW: API authentication with client credentials with rate limit handling and fallback to pre-collected data.
  - PubMed E-utils: Biopython Entrez with query optimization for MASLD/NAFLD and drug terms.
  - Google Trends: pytrends with search query setup for 5 search terms across 2023-2025 timeframe.
  - Yahoo Finance: yfinance with multi-level header processing for comparative stock analysis.
  - Error Handling: Comprehensive try-except blocks with graceful degradation to Google Drive fallback.

Issues / difficulties

Issues Faced & Resolved:

- Google Trends API Rate Limiting: Successfully handled HTTP 429 errors with retry logic and fallback to pre-collected data.
- Reddit Data Volume Management: Optimized processing for 9,255 records with efficient VADER sentiment analysis implementation.
- Stock Data Structure Complexity: Resolved multi-level header processing from yfinance API responses.
- Development Strategy: Implemented dual analysis approach with basic visualizations in main pipeline and advanced exploratory analyses in Jupyter notebook for flexibility.
- During final testing, two minor data validation issues were identified and addressed: Reddit data naturally extends slightly beyond the study period (expected behavior with live data), and stock data structure validation required adjustment for yfinance's multi-level headers. All core functionality is confirmed operational.
- Package Dependencies: All required packages (praw, yfinance, pytrends, biopython, vaderSentiment) installed and configured.
- Project structure reorganization completed to match final project requirements.
- Statistical Analysis Scope: The current implementation focuses on descriptive analytics and data pipeline development. Next, I will conduct advanced statistics (including inferential statistics and correlation testing) to gain deeper analytical insights.

Potential Issues Expected:

- API Quota Limitations during Testing: Testing may encounter rate limits with Google Trends and Reddit APIs, though fallback mechanisms are in place.

- Data Freshness Discrepancies: Real-time API calls may return slightly different results than the pre-collected datasets used for development.
- Environment Configuration: You may need to install multiple specialized dependencies (praw, yfinance, pytrends, biopython) for full functionality.
- Large Dataset Processing: The 9,255 Reddit records and 39,376 Media Cloud articles require substantial memory and processing time.
- Academic Network Access: PubMed API access may behave differently on university networks versus development environment.

Mitigation Strategies Implemented: These include comprehensive test validation (tests.py), Google Drive fallback systems, detailed documentation (README.md, requirements.txt), dual analysis environment, and modular architecture design to ensure project reliability and reproducibility.