Final Project Progress Report

Project Title: MASLD AWARENESS TRACKER

Project scope update

The project scope was strategically expanded from three to five data sources to provide more comprehensive MASLD awareness tracking. I added Media Cloud for news media analysis and enhanced stock market analysis beyond the originally planned sources. Kaggle datasets were deprioritized as they proved less relevant than initially anticipated. The project scope remains fully consistent with the original proposal. All five data sources are operational: Google Trends (148 data points), Reddit (9,255 posts with sentiment analysis), PubMed (1,344 publications), Stock Data (707 trading days), and Media Cloud (39,376 articles). Complete data collection, processing, and analysis pipelines are implemented and generating results for measuring FDA approval impacts (Resmetirom - March 2024, GLP-1 agonists - August 2025) on MASLD public awareness.

Data sources

Data Source	Status	API/Method Used	Data Obtained
Google Trends	COMPLETE	pytrends API	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 and functional in process.py
PubMed	COMPLETE	NCBI E-utils API (Biopython)	1,344 scientific publications on MASLD/NAFLD and treatments. Full terminology and publication rate analysis operational
Stock Data	COMPLETE	Yahoo Finance API (yfinance)	707 trading days for Novo Nordisk (NVO) and Madrigal Pharmaceuticals (MDGL). FDA impact analysis implemented
Media Cloud	COMPLETE	Google Drive (gdown package) + File Processing	39,376 news articles across disease-focused, Resmetirom-focused, and GLP-1-focused datasets. Timeline 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. All 5 current sources are actively contributing to analysis. All data acquisition pipelines are operational, with successful API connectivity demonstrated for Reddit, Google Trends, Yahoo Finance, and PubMed. Data integrity validated through comprehensive test suite (tests.py).
- Project structure follows requirements with modular design: data collection (load.py), processing (process.py), analysis (analyze.py), configuration (config.py), and testing (tests.py). Core data collection infrastructure operational. Comprehensive test suite (tests.py) validates API connectivity and data quality across multiple sources. All five data sources are fully operational with automated data collection and analysis pipelines generating comparative visualizations and insights.
- GDrive Links Justification: Pre-collected datasets are hosted on Google Drive to ensure project reproducibility during evaluation. This approach addresses potential API rate limiting with Google Trends and Reddit APIs, and provides immediate data access for testing the complete analysis pipeline without dependency on external API availability or web scraping constraints.

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.
- 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 matching final project requirements.

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), and modular architecture design to ensure project reliability and reproducibility.