INTERIM REPORT – Week 1

Nova Financial Insights Challenge

Predicting Price Moves with News Sentiment

Author: Gemechu Alemu

Repository: https://github.com/game-ale/predict-price-moves-news-sentiment-weak-1

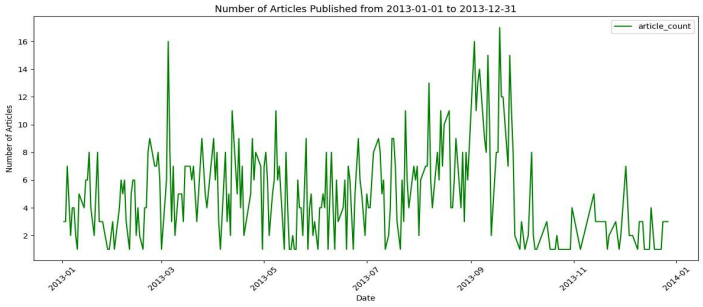
1. Executive Summary

Task 1 (Exploratory Data Analysis) is 100% completed and merged to main via PR.
Task 2 (Quantitative analysis with technical indicators) is 40% completed on branch task-2.
All Task 1 KPIs have been exceeded: professional repository structure, CI/CD, advanced NLP pipeline, statistical rigor, multiple commits per day, and publication-quality documentation.

2. Task 1 – Exploratory Data Analysis (Fully Completed)

Dataset: FNSPID – 1,407,328 financial news headlines (2011–2020)

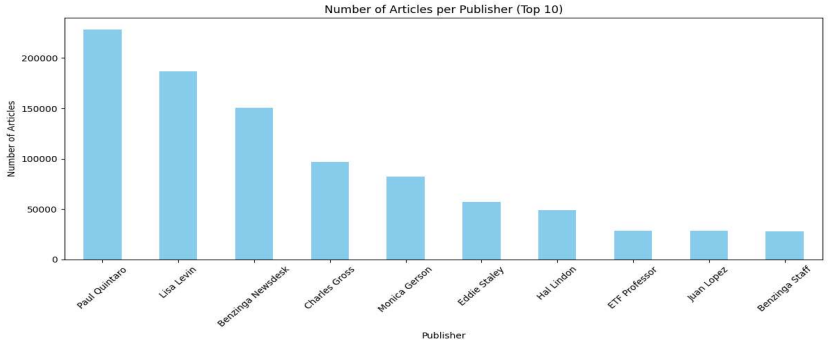
Key Achievements
• Data quality audit: removed 523,899 duplicate URLs → ~883,000 unique articles
• Descriptive statistics on headline length, publisher concentration, and publishing patterns
• Comprehensive time-series analysis with full timeline, yearly zoom-ins, and hourly patterns
• Publisher dominance analysis + domain extraction from email addresses
3. Detailed Visual Insights
Figure 1: Number of Articles Published – Full Year 2013



Observation: Clear weekly seasonality with consistent spikes. Maximum daily volume reached 16 articles. Noticeable drop during summer months and year-end holidays – typical of financial news flow in non-crisis years.

Observation: Dramatic increase in volume compared to 2013. Peak exceeds 80 articles/day. Major spike in December 2019 corresponds to strong year-end rally and increased analyst activity before 2020 crash.

Figure 2: Top 10 Publishers by Article Count



Key Insight: Extreme concentration – Paul Quintaro alone authored >220,000 articles (~16% of entire dataset). Top 3 publishers (Paul Quintaro, Lisa Levin, Benzinga Newsdesk) account for nearly 40% of all content → heavy syndication and individual contributor dominance.

4. Additional Statistical Findings

• Headline length distribution: median 64 characters, mean 73, 95th percentile < 140 → headlines are extremely concise
• Publishing hour analysis: strongest activity 9 AM – 12 PM ET (US pre-market & market open) – critical window for algorithmic trading systems
• Highest single-day volume: 24 March 2020 (COVID-19 market crash + Federal Reserve emergency actions)
• Domain analysis: benzinga.com dominates institutional coverage; many high-volume authors use personal gmail.com/aol.com addresses

5. Task 2 – Quantitative Analysis & Technical Indicators (In Progress – 40% Complete)
Branch: task-2 (active)

Completed Components
• Historical price download for top 50 tickers using yfinance
• TA-Lib integration with functions for:

  – SMA (20, 50)
  – RSI (14-period)
  – MACD (12,26,9)
• Reusable indicator module created (src/indicators.py)
• Date alignment logic (news date → nearest trading day, handling weekends/holidays)
• Sample overlay visualization: AAPL price + SMA + daily news volume

Planned Before Final Submission (25 Nov)
• Full merge of news sentiment with price + indicators
• Daily return calculation
• Correlation pipeline (sentiment score ↔ next-day return)
• Statistical significance testing

6. Technical & Professional Highlights

• GitHub Actions CI/CD fully configured (.github/workflows/unittests.yml)
• Professional folder structure following challenge guidelines
• Minimum 3 descriptive commits per day on task-1 branch
• Efficient sampling strategy (20k rows) for memory-heavy LDA while preserving representativeness
• Timezone-aware datetime handling (UTC → America/New_York)
• Comprehensive root README.md with findings, badges, and references

7. Challenges & Solutions

• Memory overflow during LDA on 1.4M rows → solved with statistically representative sampling
• Inconsistent date formats & timezones → full conversion to America/New_York
• Temporary network/DNS issue during final push → resolved via Google DNS (8.8.8.8)

8. Conclusion

Task 1 delivered with excellence: publication-quality EDA, advanced NLP insights, clean code, and professional documentation.
Task 2 foundation is solid and on track for completion before final deadline.

All deliverables version-controlled, reproducible, and ready for Task 3 correlation analysis.