10 Academy: Artificial Intelligence Mastery

Predicting Price Moves with News Sentiment: Week 1 Challenge

Interim Report

by: Yonas Zelalem

May 30, 2025

Introduction

This is an interim report on the Financial News and Stock Price Integration Dataset (FNSPID) as part of the 10 Academy AIM Week 1 Challenge, working to uncover insights that might connect financial news to stock market trends. This interim report shares my initial findings, outlines my methodology, and highlights challenges faced so far.

Methodology

I started by loading the raw_analyst_ratings.csv dataset, which have over 1.4 million entries with columns like headline, url, publisher, date, and stock. Using Python tools—Pandas, NLTK, and Matplotlib/Seaborn.

- **Data Loading and Cleaning**: converted the date column to a datetime format (UTC-04:00) and ran basic stats to grasp the dataset's structure.
- **Descriptive Statistics**: calculated headline lengths and counted articles per publisher to spot key contributors.
- **Text Analysis**: With NLTK, I tokenized headlines to extract the top 10 keywords, stripping out stopwords to reveal dominant themes.
- **Time Series and Visualization**: plotted publication frequency over time and created two bar charts—one for all 1,034 publishers and another for the top 10.
- **Domain Extraction**: pulled email domains from publisher names to explore organizational diversity.

This hands-on approach is laying the foundation for Task 2, where I'll build predictive models to link news to stock prices.

Initial Findings

- **Headline Insights**: The average headline length is 73 characters, ranging from 3 to 512. This mix of short updates and detailed reports could affect how readers engage with the news.
- **Publisher Powerhouses**: Among 1,034 publishers, I found Paul Quintaro leading with 228,373 articles, followed by Lisa Levin (186,979) and Benzinga Newsdesk (150,484). The long tail—down to Matthew Ely with just 1—shows a few dominate the scene.
- **Time Trends**: Publications span 2009 to 2020, with a peak of 806 articles on June 10, 2020. This spike might tie to market events worth exploring further.

- **Keyword Highlights**: Top words like stocks (161,702), vs (138,835), eps (128,801), and est (122,289) suggest a focus on earnings and comparisons—key for investors.
- **Publisher Domains**: Benzinga.com tops the list with 7,937 articles, while Gmail.com adds 139, indicating a blend of professional and individual sources.

Challenges Encountered

- **Visualization**: The "All Publishers" plot got messy with 1,034 publishers, even at 12x20 inches.
- **Domain Extraction Limits**: Pulling email domains worked for some (8,082 entries), but many publishers lack emails, shrinking the sample. This hampers diversity analysis.

Next Steps

With Task 1 in the bag, I'm gearing up for Task 2 (Quantitative Analysis) and Task 3 (News-Stock Correlation). Here's my streamlined plan to crush it before the 3:00 AM EAT deadline:

- **Task 2: Quantitative Analysis with PyNance and TA-Lib**
    - Load stock price data into a Pandas DataFrame with Open, High, Low, Close, Volume using yfinance (2009-2020).
    - Calculate TA-Lib indicators: SMA, RSI, MACD, and PyNance metrics (e.g., volatility).
    - Visualize with Matplotlib/Seaborn.
- **Task 3: Linking News Sentiment to Stock Moves**
    - Align news and stock data by normalizing dates (match trading days).
    - Use TextBlob for sentiment analysis on headlines (positive/negative/neutral), aggregating daily scores.
    - Compute daily stock returns (percentage change in closing prices).
    - Calculate Pearson correlation between sentiment and returns; visualize with scatter/time series plots.

Conclusion

This interim report marks my halfway point in the Week 1 Challenge, showcasing a strong start with rich insights and a clear methodology. Challenges like cluttered plots and limited domains are driving my problem-solving, and I'm optimistic about linking news to stock prices with this dataset's depth. Check my progress on GitHub, and feel free to share your thoughts—happy analyzing, everyone!