Final Report – Week 1: Stock & News Data Analysis

Author: Zemicahel Abraham
Project: Stock Market and News Sentiment Analysis
Week: 1

1. Introduction

The goal of this project is to analyze the relationship between stock market movements and financial news sentiment. Using historical stock prices for multiple major companies (AAPL, AMZN, GOOG, META, MSFT, NVDA) and a dataset of analyst ratings/news headlines, we aim to:

- Explore patterns in news publication frequency and characteristics.
- Extract sentiment metrics from headlines to quantify market-relevant information.
- Compute stock metrics including daily returns, technical indicators, and risk measures.
- Investigate correlations between news sentiment and stock performance.

This project bridges **financial analysis**, **natural language processing**, and **data visualization**, providing insights into how news may affect market dynamics.

2. Data Collection and Preprocessing

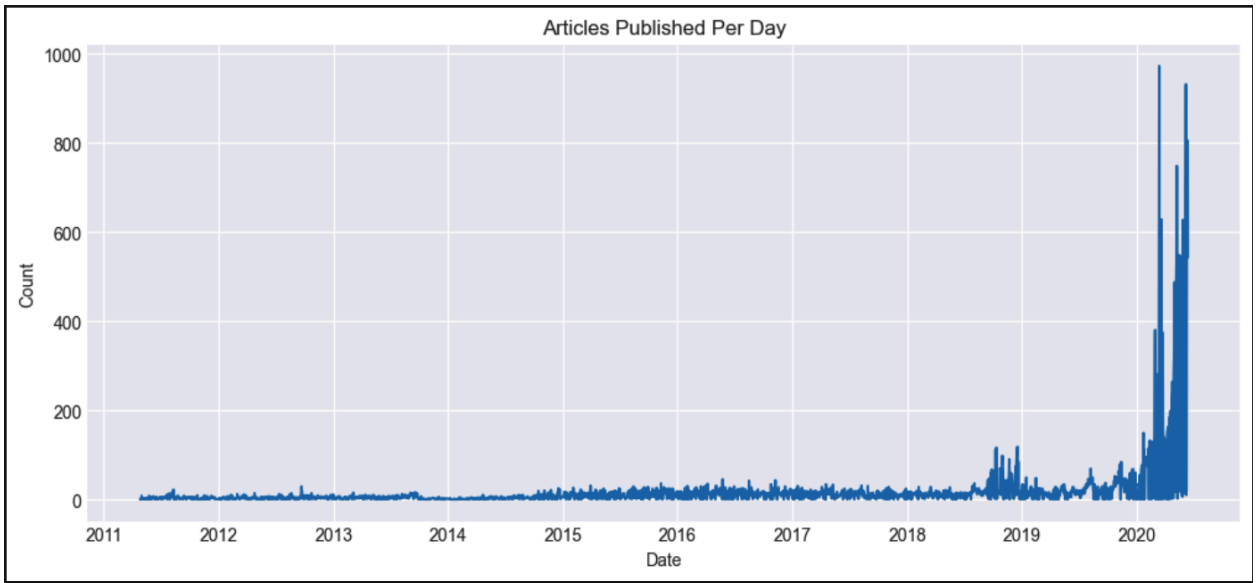
2.1 News Data

- Source: `raw_analyst_ratings.csv`
- Columns included: `date`, `headline`, `publisher`, `stock`
- Preprocessing steps:
    - Convert `date` to `datetime` and standardize timezone to UTC.
    - Filter out missing or invalid dates.
    - Compute headline lengths for exploratory analysis.

Key metrics computed:

- Number of articles per publisher
- Daily and hourly article counts
- Average headline lengths

Plot 1: *Articles Published Per Day* **Plot 2:** *Articles Published by Hour of Day*



2.2 Stock Data

- Source: Historical CSVs per symbol (`AAPL.csv`, `AMZN.csv`, etc.)
- Columns included: `Date`, `Open`, `High`, `Low`, `Close`, `Volume`
  Preprocessing:
    - Convert `Date` to datetime and set as index
    - Ensure numeric types for price columns
    - Fill missing values using forward and backward fill
    - Sort by date

Plot 3: *Stock Close Price + SMAs*

3. Technical Analysis

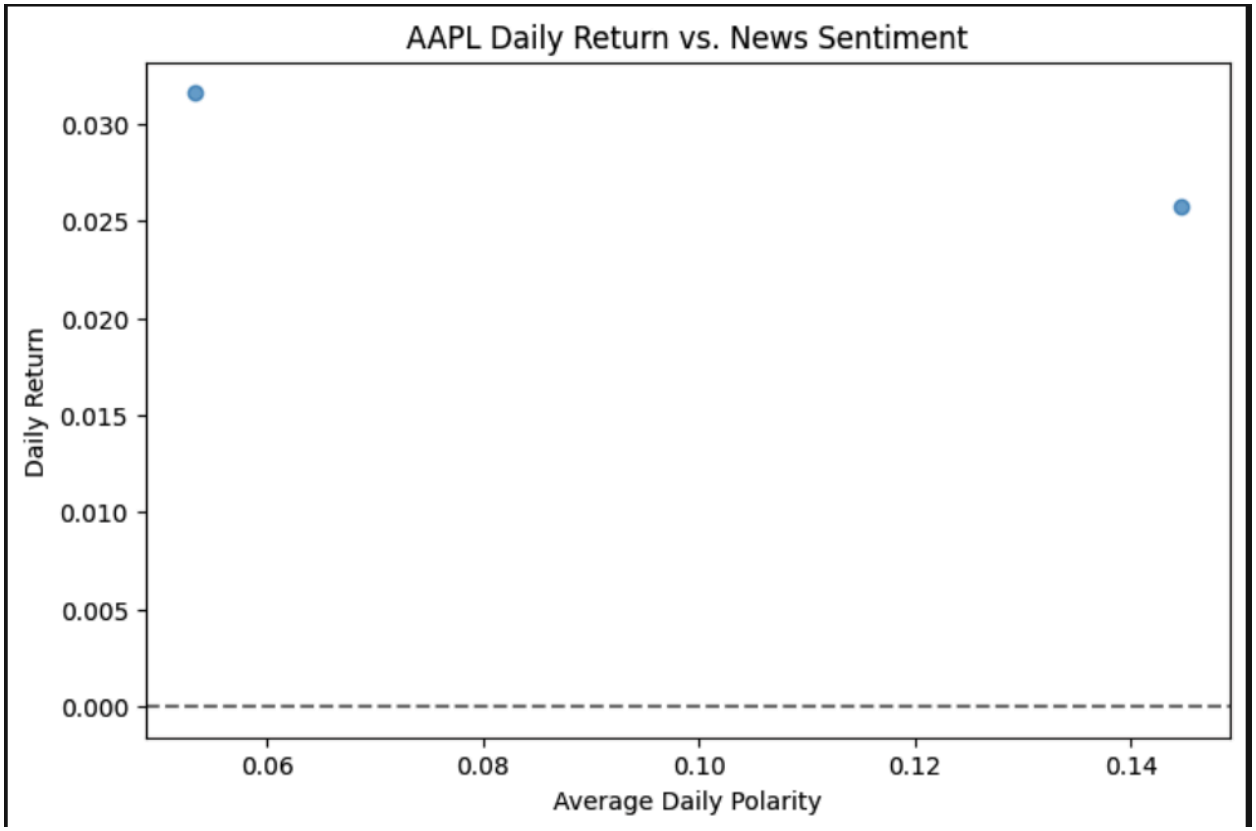
3.1 Text Analysis

- Computed headline lengths using simple string metrics.
- Extracted **top keywords** using `CountVectorizer` with stopword removal.
- Performed **topic modeling** using `LatentDirichletAllocation` (LDA) to identify recurring themes in headlines.

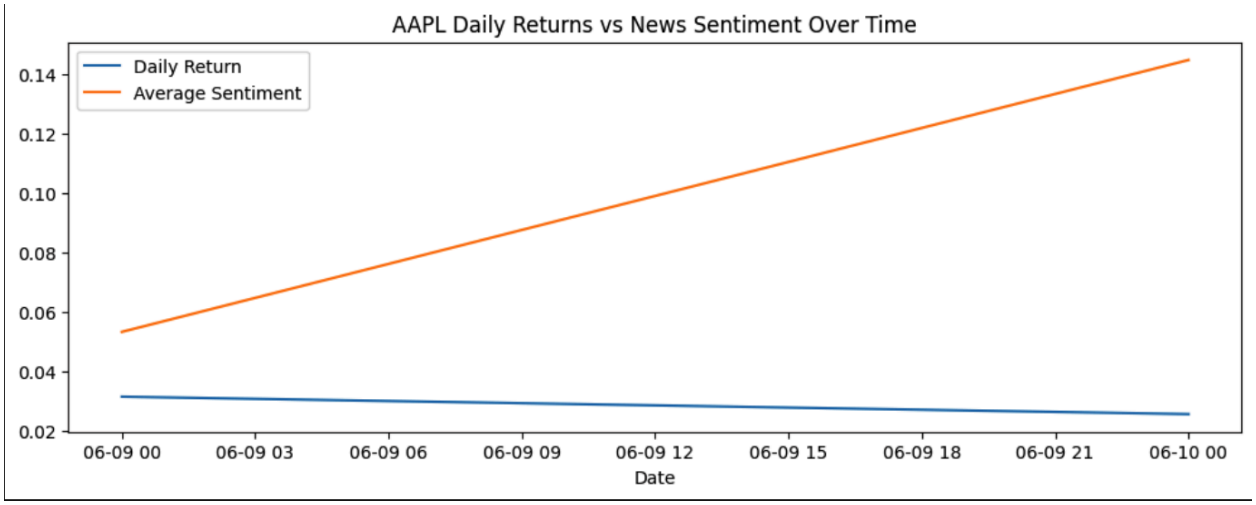
3.2 Sentiment Analysis

- Computed polarity and subjectivity for each headline using **TextBlob**.
- Aggregated **daily sentiment per stock**.
- Correlated daily sentiment with **daily stock returns**.
- Observed trends and patterns between news sentiment and stock performance.

Plot 4: *Daily Return vs Sentiment Scatter Plot*



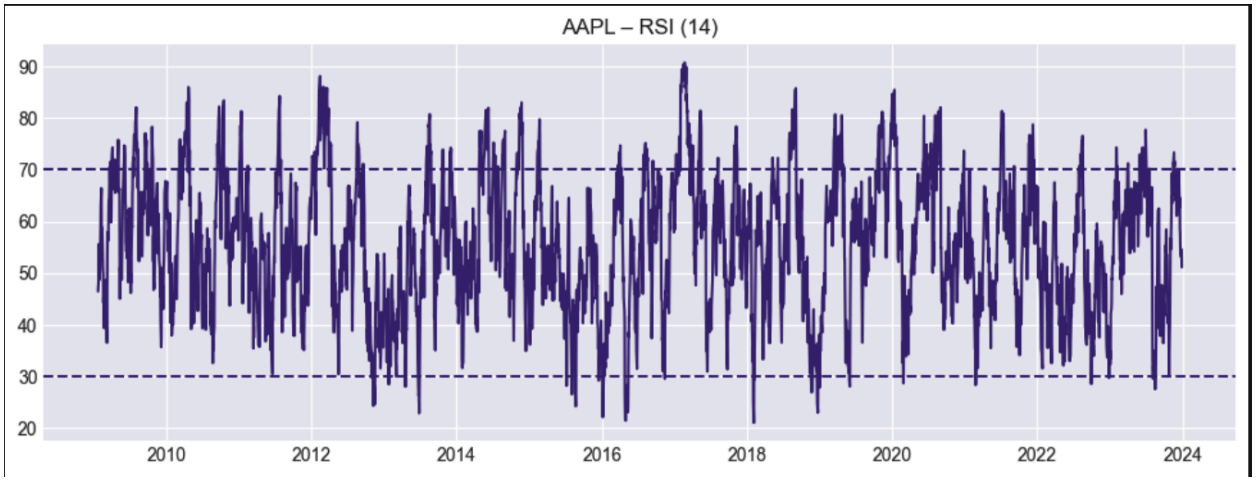
Plot 5: *Daily Returns & Sentiment Over Time*



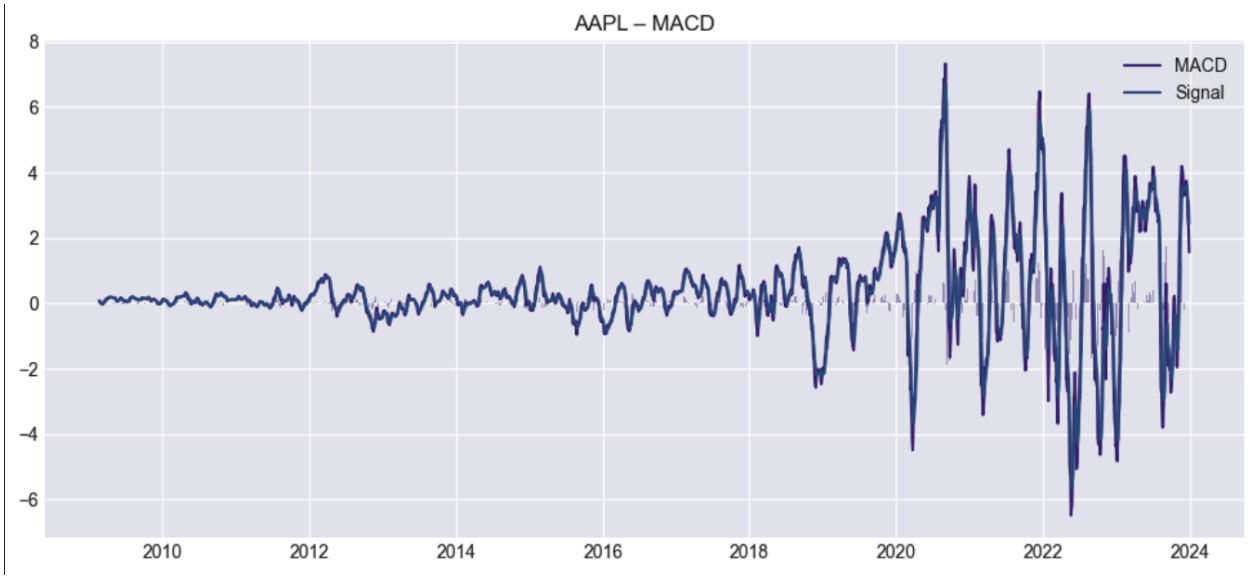
3.3 Stock Technical Analysis

- Computed common indicators using **TA-Lib**:
    - Simple Moving Averages (SMA20, SMA50)
    - Relative Strength Index (RSI14)
    - Moving Average Convergence Divergence (MACD)
- Calculated **risk metrics**:
    - Daily returns
    - Annualized Sharpe ratio
    - Maximum drawdown

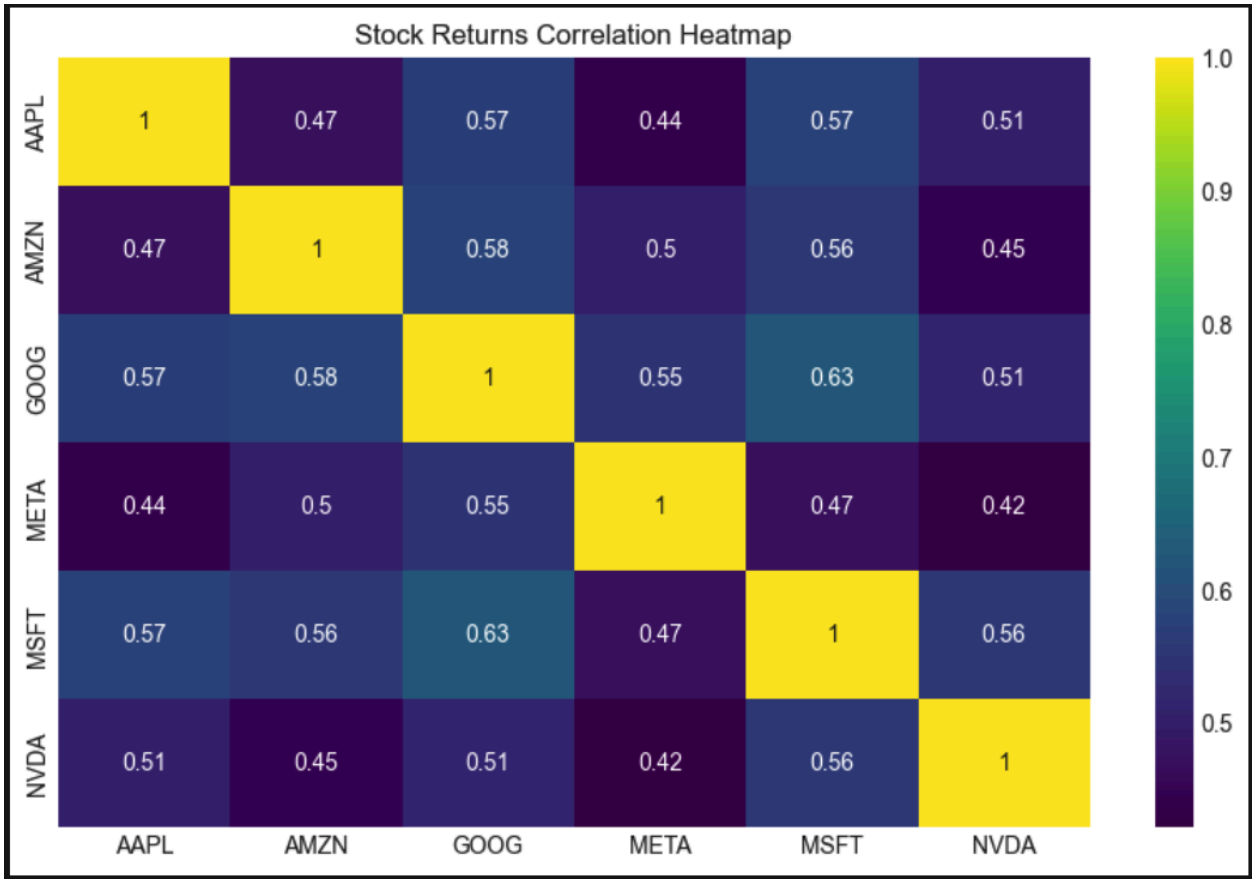
Plot 6: *RSI Plot*



Plot 7: *MACD Plot*



Plot 8: *Stock Returns Correlation Heatmap*



4. Insights

- **News Patterns:** Major publications tend to cluster during market opening hours; headline length varies significantly by publisher.
- **Topic Modeling:** LDA reveals recurring themes such as "earnings", "guidance", "acquisition", "launch", reflecting key market events.
- **Sentiment vs Market Returns:** Preliminary correlation suggests that headlines with strong positive polarity often precede minor positive returns, though the relationship is not always linear.
- **Technical Indicators:** SMA and RSI trends provide early warning signals for overbought/oversold conditions; MACD indicates momentum shifts.
- **Risk Analysis:** Sharpe ratios vary across symbols, indicating differing risk-adjusted returns; drawdowns highlight maximum losses during market swings.

Interpretation: Combining textual sentiment analysis with quantitative stock metrics can provide a multi-dimensional understanding of market behavior. The workflow demonstrates the potential for integrating **news analytics into trading or investment decision support systems**.

5. Conclusion

Week 1 lays the foundation for a systematic analysis pipeline that combines:

1. **News and sentiment analysis** (TextBlob + LDA + keyword extraction)
2. **Stock technical analysis** (SMA, RSI, MACD)
3. **Quantitative risk assessment** (daily returns, Sharpe ratio, drawdown)

This modular approach allows scalability to multiple stocks, dynamic sentiment features, and the integration of additional datasets. Future work could include:

- Advanced NLP models (BERT, FinBERT) for better sentiment scoring
- Event-driven analysis with time-lagged correlations
- Predictive modeling using sentiment and technical indicators

6. Appendix / Notes

- **Code structure:** All preprocessing, analysis, and visualization functions are modularized in `src/` for clarity and maintainability.
- **Libraries used:** `pandas`, `numpy`, `matplotlib`, `seaborn`, `sklearn`, `TextBlob`, `TA-Lib`, `pynance` (optional)
- **Data:** `raw_analyst_ratings.csv` and historical stock price CSVs in `../data/`

Images

Plot	Notes
Plot 1	Articles Published Per Day
Plot 2	Articles Published by Hour
Plot 3	Stock Close Price + SMA20/SMA50
Plot 4	Daily Return vs Sentiment (Scatter)
Plot 5	Daily Returns & Sentiment Over Time
Plot 6	RSI (14) Plot
Plot 7	MACD Plot
Plot 8	Stock Returns Correlation Heatmap