



WEB SCRAPING & STOCK MARKET ANALYSIS USING YAHOO FINANCE

A Data-Driven Approach to Extract & Analyze Stock Market Trends

www.finance.yahoo.com



Project Overview

- **Objective:** Extract stock data and news from Yahoo Finance using web scraping.
- **Tools Used:** Python, BeautifulSoup, Pandas, Matplotlib, NLP techniques, etc.
- **Methodology:** Scrape, crawl, clean, analyze, and visualize stock trends & sentiment.





Data Extraction Process



- How Web Scraping Works on Yahoo Finance
- **Chosen Stocks & Sections:**
 - Each group member's section (e.g., Technology, Real-Estate, Mutual Funds, Healthcare, Consumer Cyclical etc.)
 - List of 5-10 stocks analyzed
- **Stock News Extraction:**
 - Extracting all related news articles for each stock



Individual Team Member Analysis

Each person presents their analysis separately based on the section they handled.



Mutual Funds

Process Overview

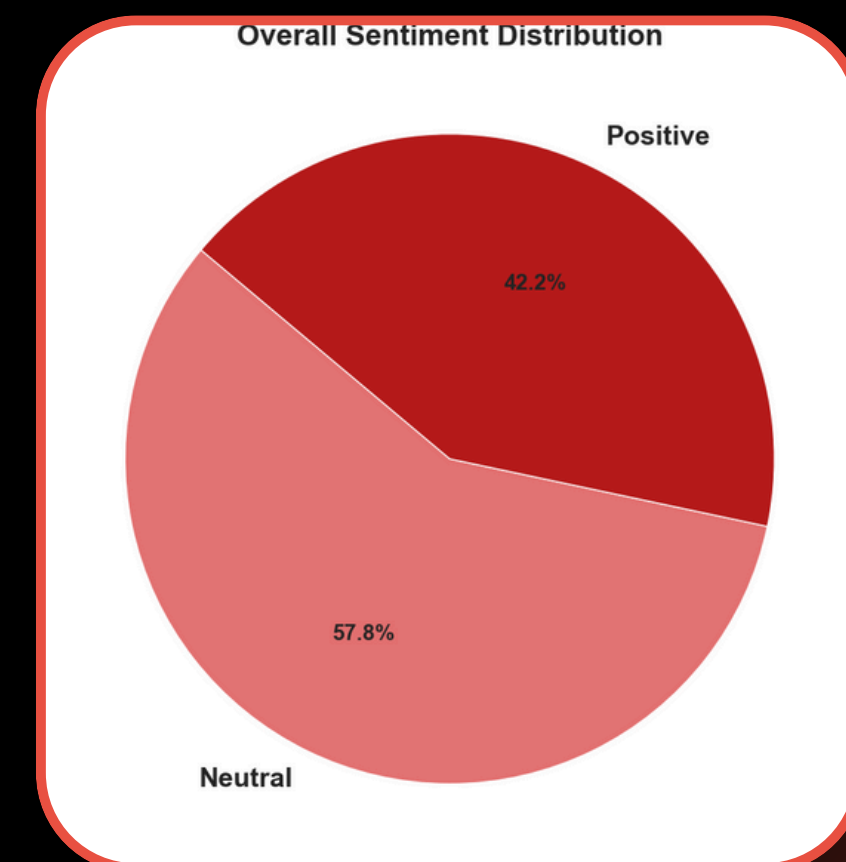
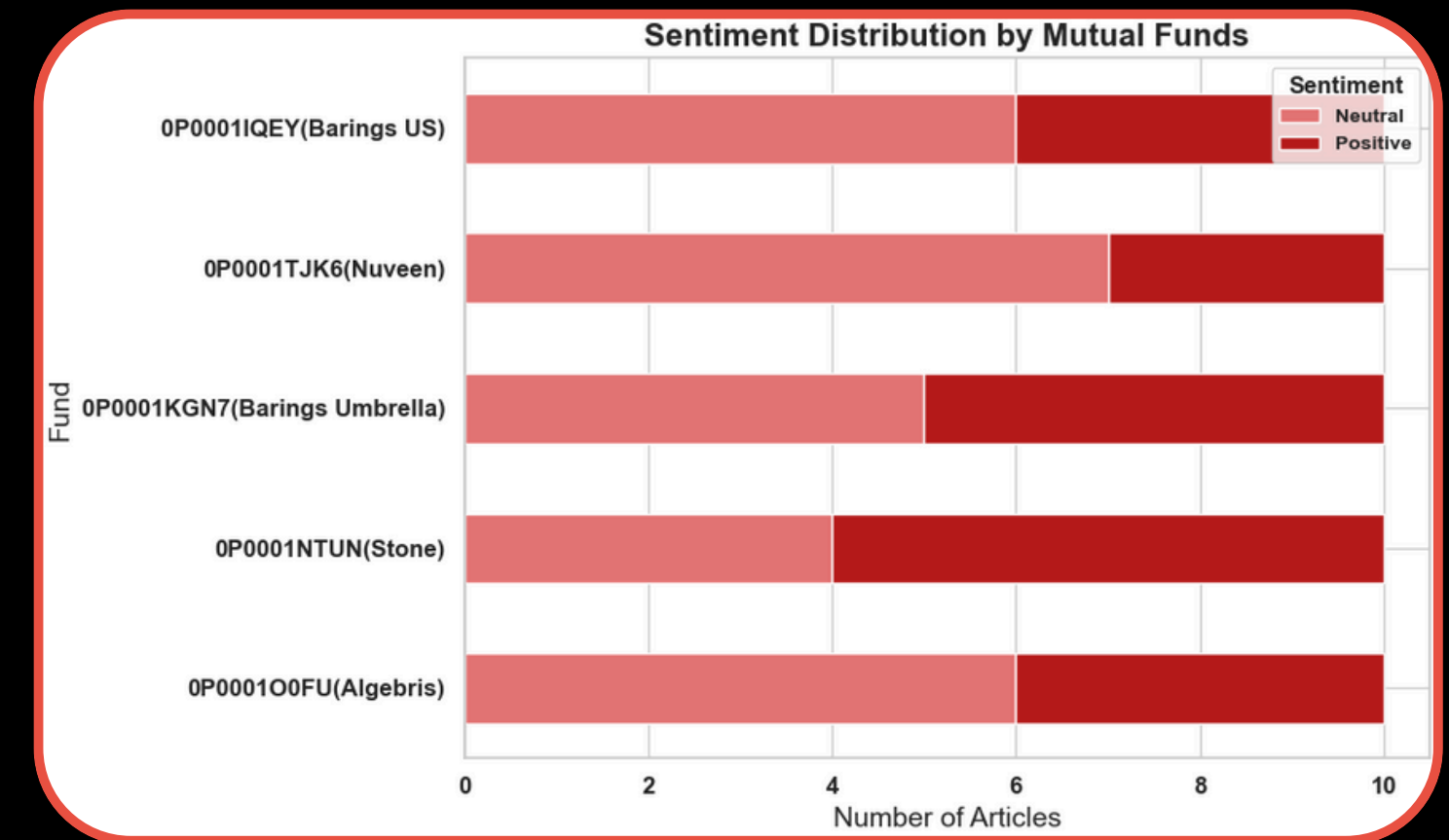
- Used Selenium to extract fund names, symbols, and YTD returns from Yahoo Finance.
- Collected key metrics like YTD Return & 5-Year Return for comparison.
- Used VADER model to classify news articles as Positive, Neutral, or Negative.
- Used Matplotlib & Dash to visualize performance & sentiment trends.

Sentiment by Fund (Bar Chart)

- 0P0001O0FU and 0P0001NTUN received high number of positive sentiment articles, reflecting strong investor confidence.
- 0P0001KGN7 had a balanced mix, indicating moderate market perception.
- 0P0001KEY and 0P0001TJK6 had predominantly neutral coverage, suggesting investors are cautious and waiting for clearer market signals.

Overall Sentiment Distribution (Pie Chart)

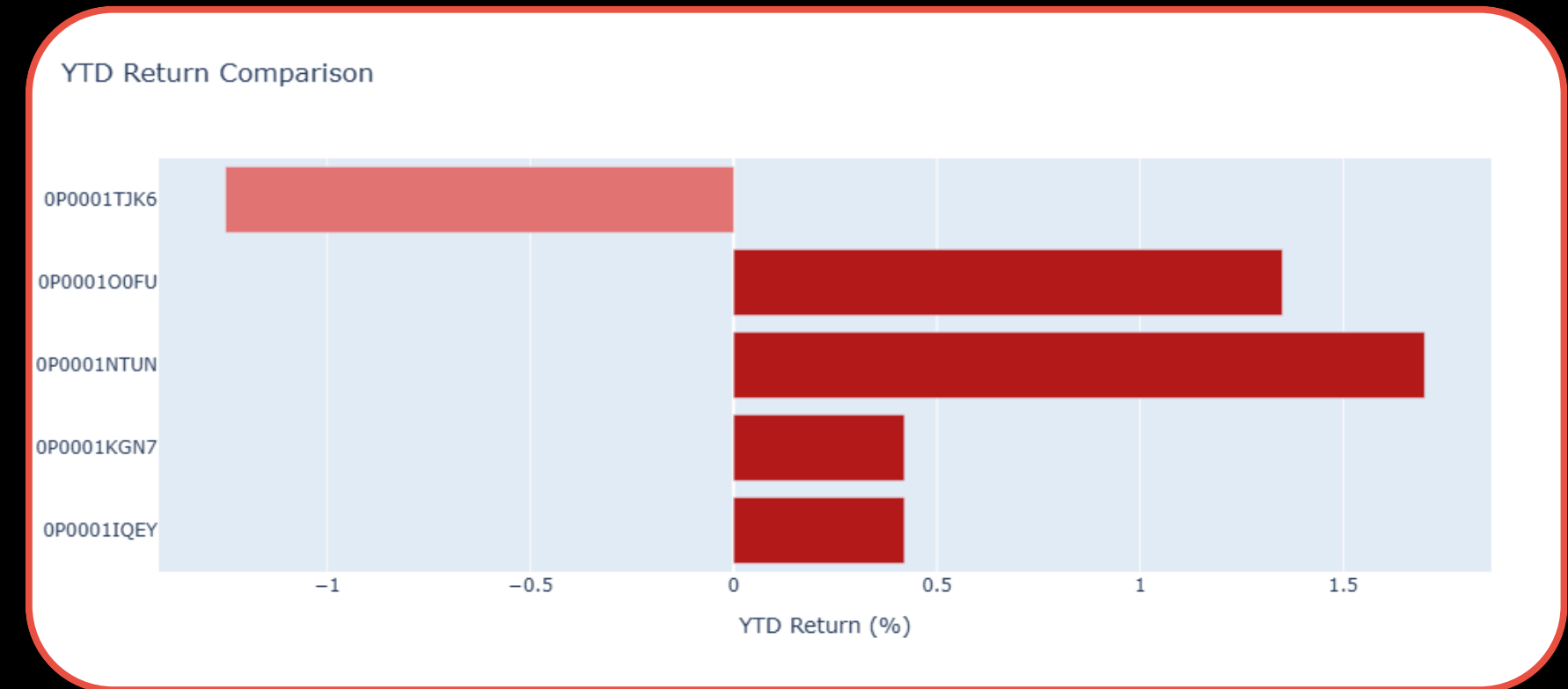
- 57.8% of the articles were neutral (Red), indicating that most reporting is balanced without strong bullish or bearish signals.
- 42.2% were positive (Dark Red), reflecting overall optimism among investors.



Mutual Funds

Mutual Fund YTD Return Comparison

- OP0001NTUN (Stone) is the top performer, showing the highest YTD return among all funds.
- OP0001TJK6 (Nuveen) is the only fund in negative territory, indicating poor performance compared to peers.
- OP0001OOFU (Algebris) and OP0001KGN7 (Barings Umbrella) show moderate growth, sitting between the best and worst performers.
- The performance gap is significant, highlighting a diverse risk-return balance among these mutual funds.



Challenges with Yahoo Finance Scraping

- Dynamic Loading Complexity : Yahoo Finance loads content with JavaScript, making direct scraping tricky. We had to use alternative methods to fetch news links.
- Inconsistent Page Structures : Different mutual funds had varied page layouts, so we had to refine filters and search terms to extract useful articles.
- Blocking and Anti-Scraping : Frequent requests triggered blocks, so we added delays, switched user-agents, and adjusted request headers to avoid detection.
- Limited News Coverage : Some funds had very few news articles, so we had to expand keyword searches and include alternative search terms.
- URL Structure Issues : Yahoo Finance uses different URL formats, so had to manually adjust and test the structure to fetch the correct news articles.

▶▶▶ Webscrapping-Consumer Cyclical

Process Overview:

- **Scraped the Sector Page:** Collected the top 10 stocks from the Consumer Cyclical sector on Yahoo Finance by extracting stock symbols and company names.
- **Navigated to Each Stock's Page:** Accessed individual stock pages to extract relevant financial data and news links.
- **Extracted News Articles:** Focused on articles directly related to each stock, ensuring they were not general market news.
- **Prepared Data for Analysis:** Structured the collected information into a dataset, ready for sentiment analysis and correlation with stock performance.

Challenges Faced:

- **Dynamic Content Loading:** Some elements (e.g., news articles) were loaded dynamically with JavaScript, requiring adjustments in scraping techniques.
- **Avoiding IP Blocks:** Frequent requests triggered anti-scraping mechanisms, so delays and user-agent rotation were implemented to mimic human browsing behavior.

| Largest Companies in This Sector | | | | | View More → |
|---------------------------------------|------------|----------------|---------------|------------|--------------|
| Table View Heatmap View | | | | | |
| Name | Last Price | 1Y Target Est. | Market Weight | Market Cap | Day's Change |
| AMZN Amazon.com, Inc. | 198.02 | 264.71 | 32.26% | 2.082T | |
| TSLA Tesla, Inc. | 249.98 | 343.11 | 12.43% | 802.451B | |
| HD The Home Depot, Inc. | 351.31 | 431.95 | 5.41% | 348.978B | |
| MCD McDonald's Corpor... | 299.83 | 328.36 | 3.33% | 214.864B | |
| BKNG Booking Holdings Inc. | 4,465.48 | 5,636.98 | 2.29% | 147.793B | |
| TJX The TJX Companies,... | 113.26 | 134.53 | 1.97% | 127.322B | |
| LOW Lowe's Companies, I... | 224.44 | 279.81 | 1.96% | 126.73B | |
| SBUX Starbucks Corporati... | 98.11 | 107.57 | 1.72% | 111.237B | |
| NKE NIKE, Inc. | 71.66 | 85.95 | 1.65% | 106.666B | |
| MELI MercadoLibre, Inc. | 2,021.37 | 2,504.75 | 1.59% | 102.478B | |



▶▶▶ Sentiment Analysis(VADER)–Consumer Cyclical

Key Insights:

Apple (AAPL)

- Highly Positive: Likely driven by an AI-powered iPhone or AR/VR headset launch, boosting investor confidence despite competition concerns.

Home Depot (HD) vs. Lowe's (LOW)

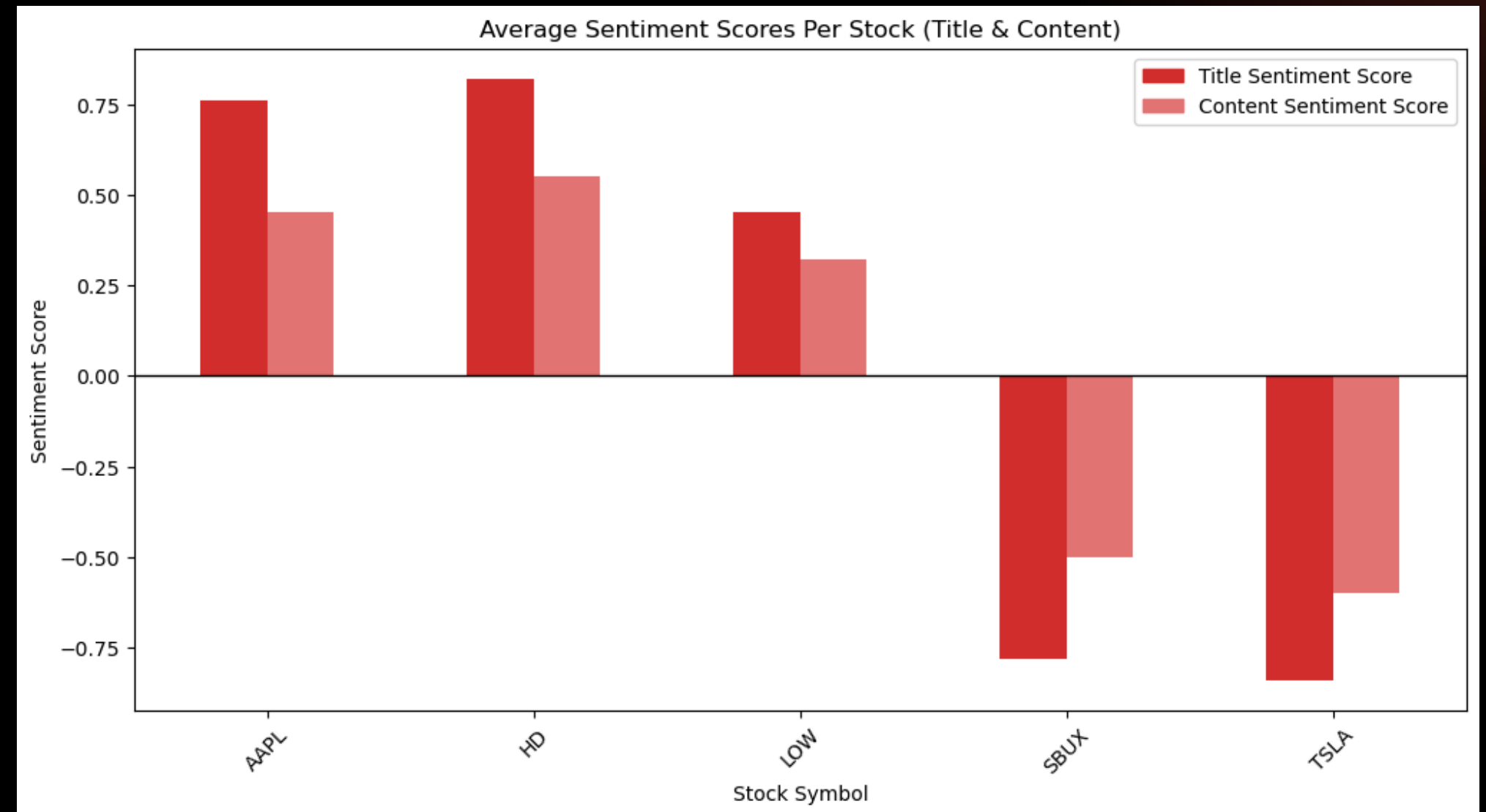
- Varying Positive Sentiment: Home Depot benefits from strong earnings and a housing market boom, while Lowe's lags due to weaker online sales.

Starbucks (SBUX)

- Moderately Negative: Labor strikes and consumer backlash over price hikes drive negative sentiment, reflecting ongoing unionization challenges.

Tesla (TSLA)

- Highly Negative: Likely impacted by a Cybertruck recall or regulatory scrutiny on self-driving tech, reinforcing Tesla's volatile public perception.



Broader Market Context: Tech sentiment is split between Apple's innovation and Tesla's setbacks, while strong home improvement demand supports retail stocks.



▶▶▶ Webscrapping-Healthcare Sector

Process Overview:

- **Scraped the Sector Page:** Scraped the Sector Page: Collected top healthcare stocks from Yahoo Finance by extracting stock symbols and company names.
- **Navigated to Each Stock's Page:** Accessed individual stock pages to extract relevant financial data and news links.
- **Extracted News Articles:** Focused on stock-specific news, avoiding general market trends.
- **Prepared Data for Analysis:** Structured data into a dataset for sentiment analysis and market trend evaluation.

Challenges Faced:

- **Dynamic Content Loading:** News articles and stock data required JavaScript handling, adjusting scraping techniques.
- **Avoiding IP Blocks:** Implemented request delays and user-agent rotation to mimic human browsing behavior and prevent blocks.

| Sentiment Analysis of Stock News Articles | | | | | |
|---|--------------|---|-----------------------|--------------------------|-------------------------|
| | Stock Symbol | News Title | Title Sentiment Score | Title Sentiment Category | Content Sentiment Score |
| 0 | LLY | Tesla's problems, Eli Lilly's weight loss pill... | -0.6124 | Negative | -0.8393 |
| 1 | LLY | Is Eli Lilly and Company (LLY) the Best Qualit... | 0.6369 | Positive | 0.9924 |
| 2 | LLY | AstraZeneca Stock Scores RS Rating Upgrade | 0.0000 | Neutral | 0.4404 |
| 3 | LLY | Eli Lilly (NYSE:LLY) Sees 1.59% Rise As Jaypir... | 0.7096 | Positive | 0.9953 |
| 4 | LLY | Eli Lilly Reverses Its Recent Upswing On Broad... | -0.4404 | Negative | -0.4767 |
| ... | ... | ... | ... | ... | ... |
| 193 | DHR | Danaher Full Year 2024 Earnings: EPS Misses Ex... | -0.2263 | Negative | 0.4329 |
| 194 | DHR | Is Danaher Corporation (DHR) the Best Scienti... | 0.6369 | Positive | 0.9950 |
| 195 | DHR | Danaher Announces Appointment of Charles Laman... | 0.0000 | Neutral | 0.9966 |
| 196 | DHR | Is Danaher Corporation (DHR) The Best Manufact... | 0.6369 | Positive | 0.9964 |
| 197 | DHR | 3 Healthcare Stocks Skating on Thin Ice | 0.0000 | Neutral | 0.9981 |



▶▶▶ Sentiment Analysis(VADER)–Healthcare

Key Insights:

Danaher (DHR)

- Positive Sentiment due to being acknowledged for making tuberculosis (TB) testing more accessible in low-income countries, boosting public perception.

Johnson & Johnson (JNJ)

- Positive Sentiment: Boosted by strong sales of Darzalex and its \$14.6B acquisition of Intra-Cellular Therapies, strengthening its pharmaceutical portfolio.

Eli Lilly (LLY)

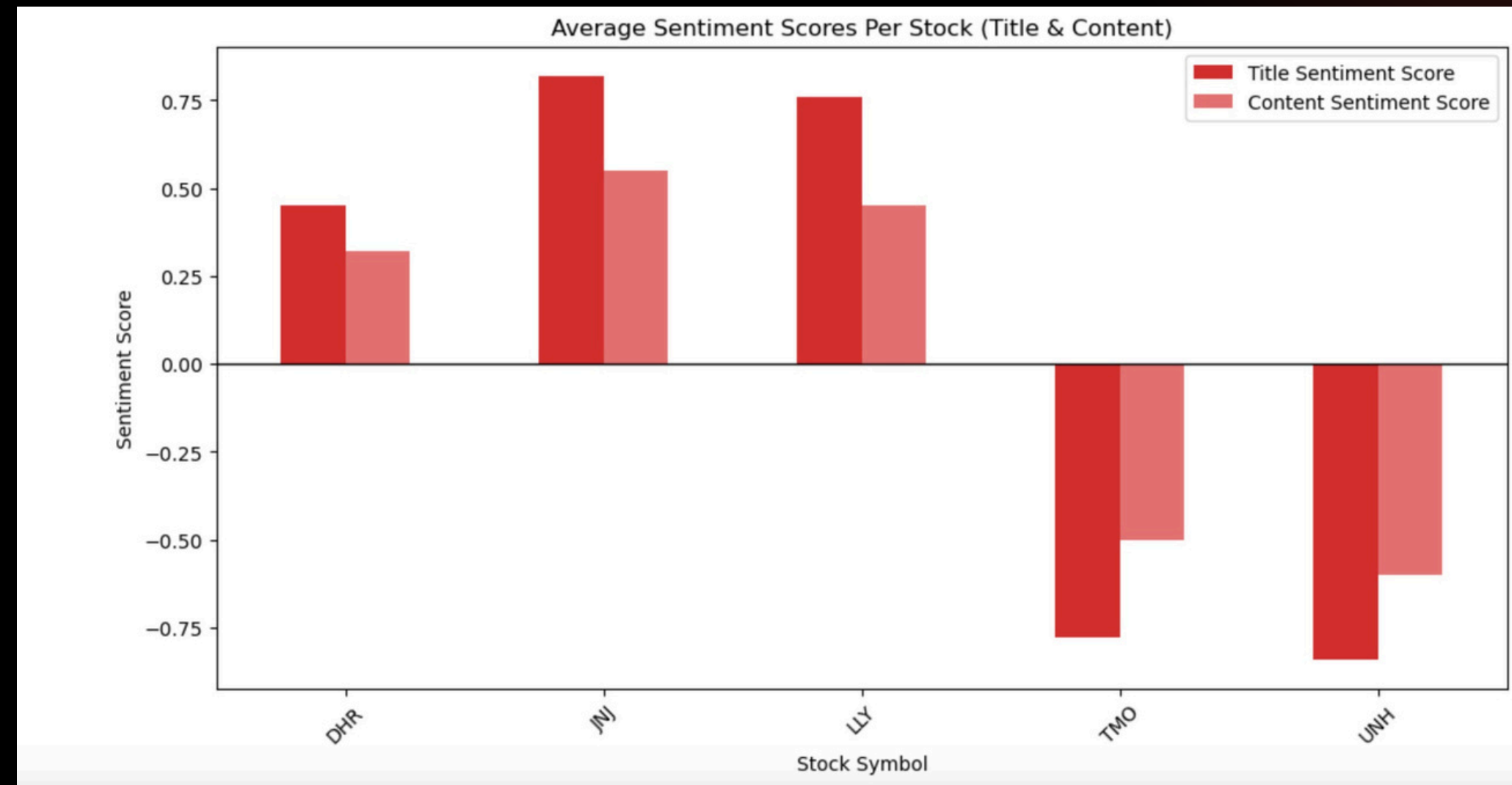
- Positive Sentiment: Investor optimism following \$27B investment in U.S. manufacturing expansion and commitment to affordable TB treatment..

Thermo Fisher (TMO)

- Negative sentiment as they acquired Solventum's filtration unit \$4.1B, but investor caution persists due to market competition and integration risks.

UnitedHealth (UNH)

- Negative sentiment due to being weighed down by security concerns related to an executive incident and regulatory scrutiny.



Broader Market Context: Healthcare sentiment is split, with pharma leaders (JNJ, LLY) thriving on expansion & acquisitions, while healthcare services (TMO, UNH) face investor skepticism.



▶▶▶ Stock Market Analysis – Technology

Key Stocks Scraped: Key Stocks Scraped: AAPL (Apple), MSFT (Microsoft), NVDA (NVIDIA), AMZN (Amazon), AVGO (Broadcom), ORCL (Oracle), CRM (Salesforce), CSCO (Cisco), IBM, PLTR (Palantir)

Objective: Analyze stock news sentiment, market trends, and key topics.

Data Collected

- Extracted real-time stock market data & news articles from Yahoo Finance.
- Collected headline news, sentiment, and stock trends for 10 key tech stocks.

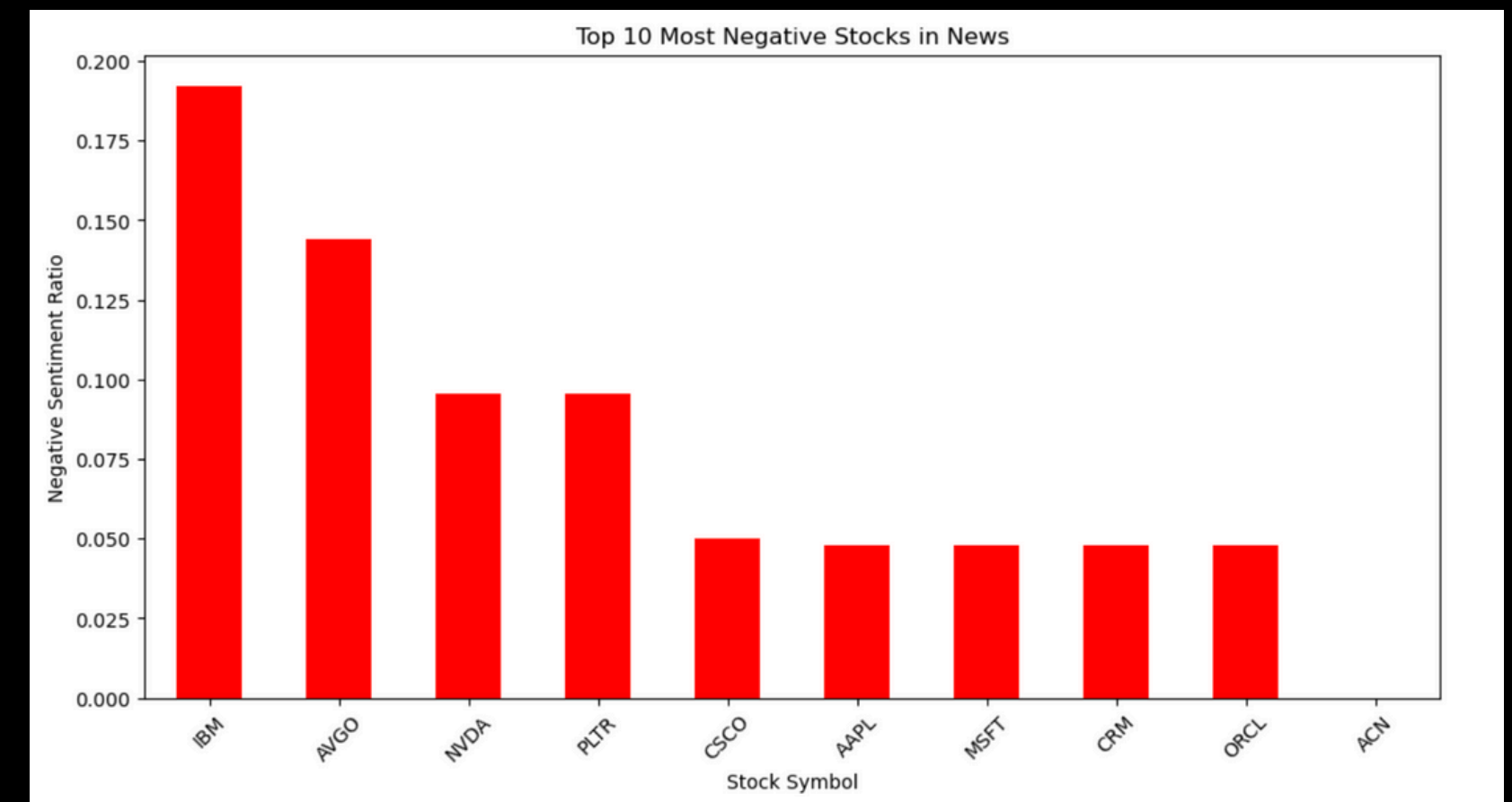
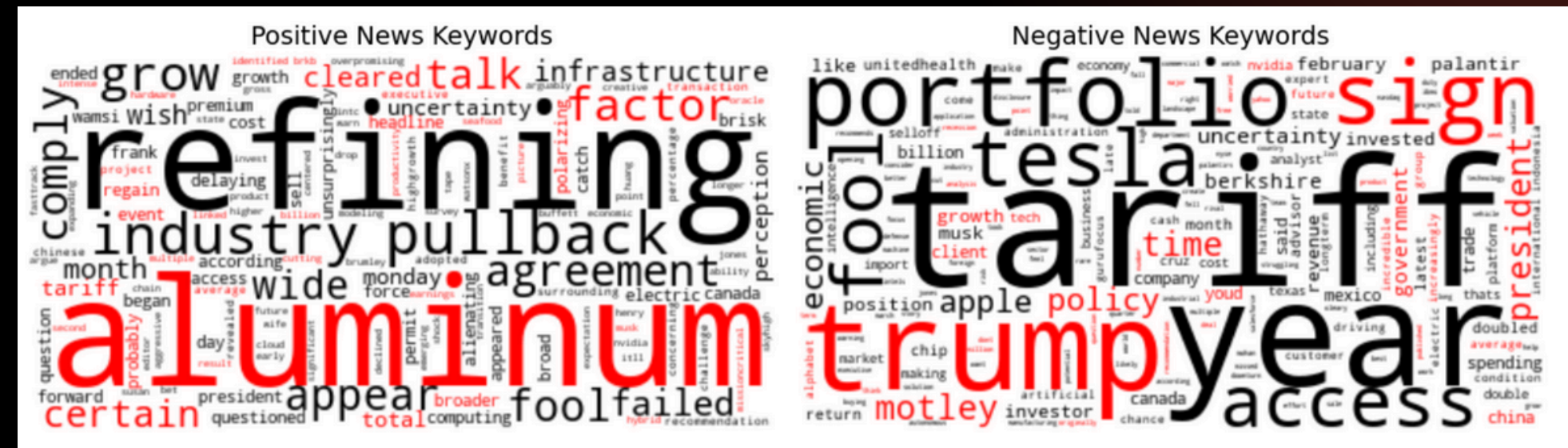
Insights from Word Cloud

Positive News Keywords

- **Growth & Industry Expansion:** Keywords like refining, aluminum, infrastructure highlight development and market optimism.
- **Compliance & Stability:** Terms such as pullback, agreement, comply suggest businesses adapting to regulations while maintaining stability.
- **Technology & Forward Momentum:** Mentions of electric, computing, factor indicate innovation and investor confidence in emerging tech sectors.

Negative News Keywords

- **Trade & Economic Instability:** Tariff, uncertainty, economic suggest concerns over global trade policies and financial markets.
- **Politics & Regulations:** Trump, policy, government highlight the influence of political decisions on the stock market.
- **Big Tech & Market Risks:** Tesla, Apple, portfolio appearing in a negative context may signal volatility in major tech stocks.



▶▶▶ Stock Market Analysis – Technology

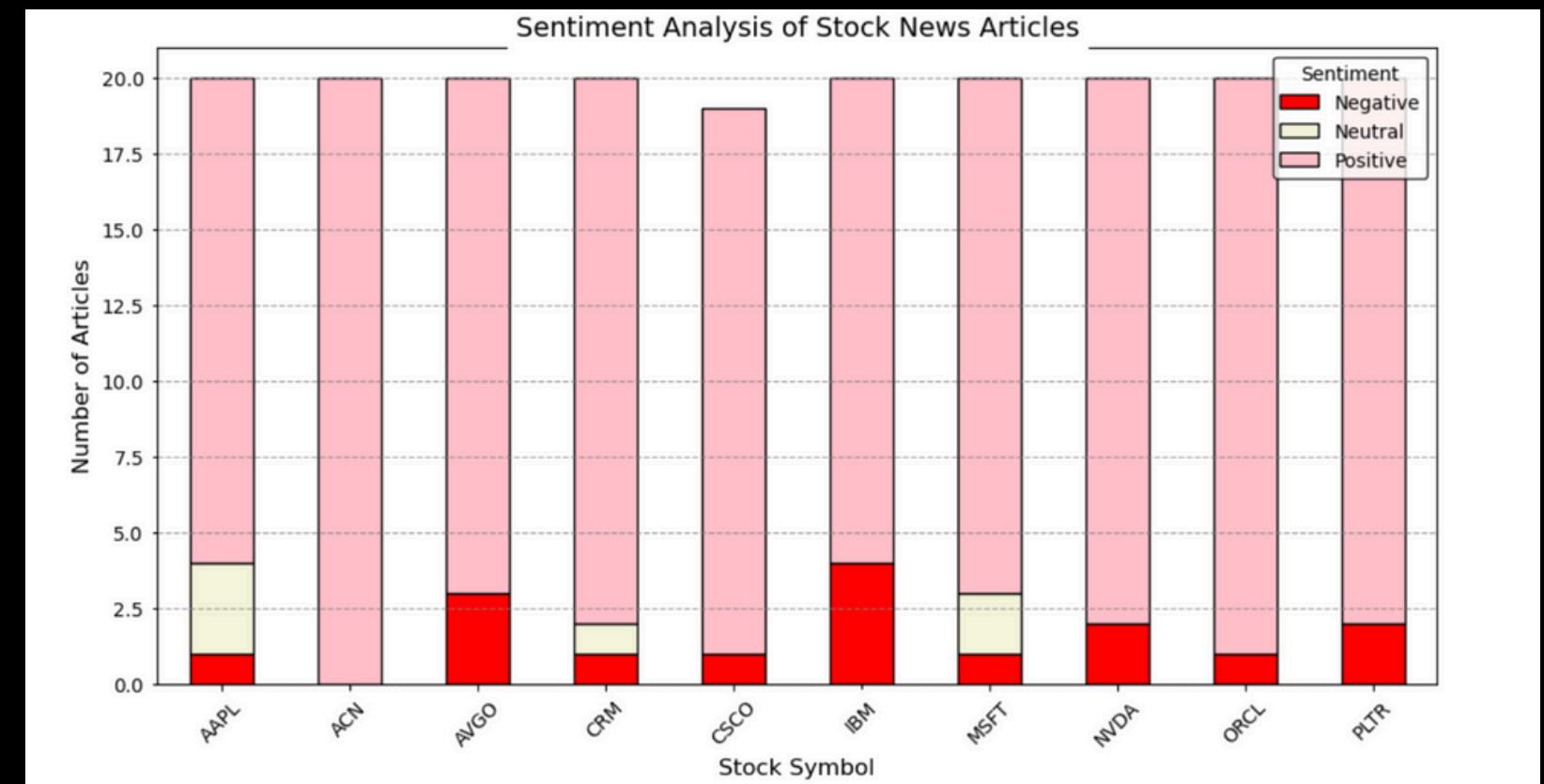
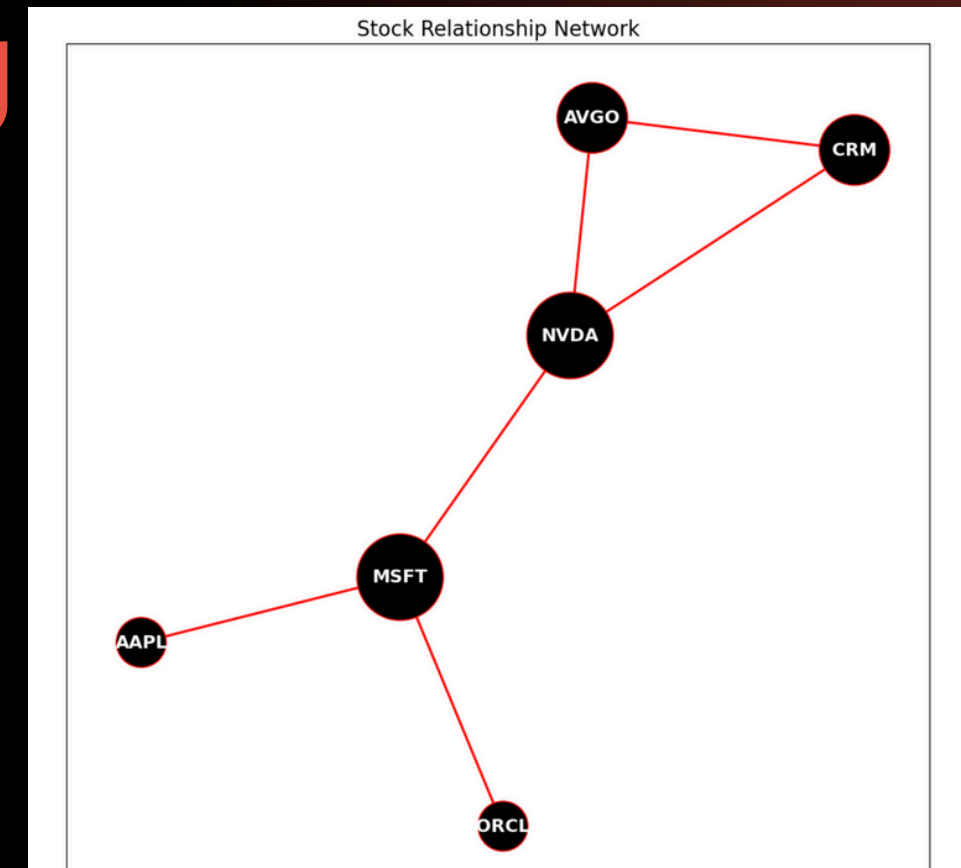
Key Stocks Scraped: Key Stocks Scraped: AAPL (Apple), MSFT (Microsoft), NVDA (NVIDIA), AMZN (Amazon), AVGO (Broadcom), ORCL (Oracle), CRM (Salesforce), CSCO (Cisco), IBM, PLTR (Palantir)

Insights for Stock Relationship network

- Microsoft (MSFT) & Nvidia (NVDA) are central – frequently mentioned with other stocks, highlighting their strong industry influence in AI and cloud computing.
- CRM & AVGO form a close cluster with NVDA, suggesting connections in AI, enterprise software, or cloud computing.
- MSFT acts as a bridge between multiple stocks, linking AI, cloud, and software sectors.
- Apple (AAPL) & Oracle (ORCL) are less connected, indicating more sector-specific news coverage rather than broad industry discussions.

Insights for Sentiment Analysis


- Tech stocks generally receive positive coverage, but IBM, NVDA, and AVGO have considerable negative sentiment.
- Neutral sentiment is rare, indicating that most financial news tends to lean towards an opinion rather than just reporting facts.
- IBM appears to be the most controversial stock, as it has the highest proportion of negative news.



►►► Stock Market Analysis – **Real Estate**

Key Stocks Scraped: PLD Prologis, Inc., AMT American Tower Corporation, WELL Welltower Inc., EQIX Equinix, Inc., SPG Simon Property Group, Inc., PSA Public Storage, DLR Digital Realty Trust, Inc., O Realty Income Corporation, CCI Crown Castle Inc., CBRE CBRE Group, Inc..

Key Insights from the Word Cloud

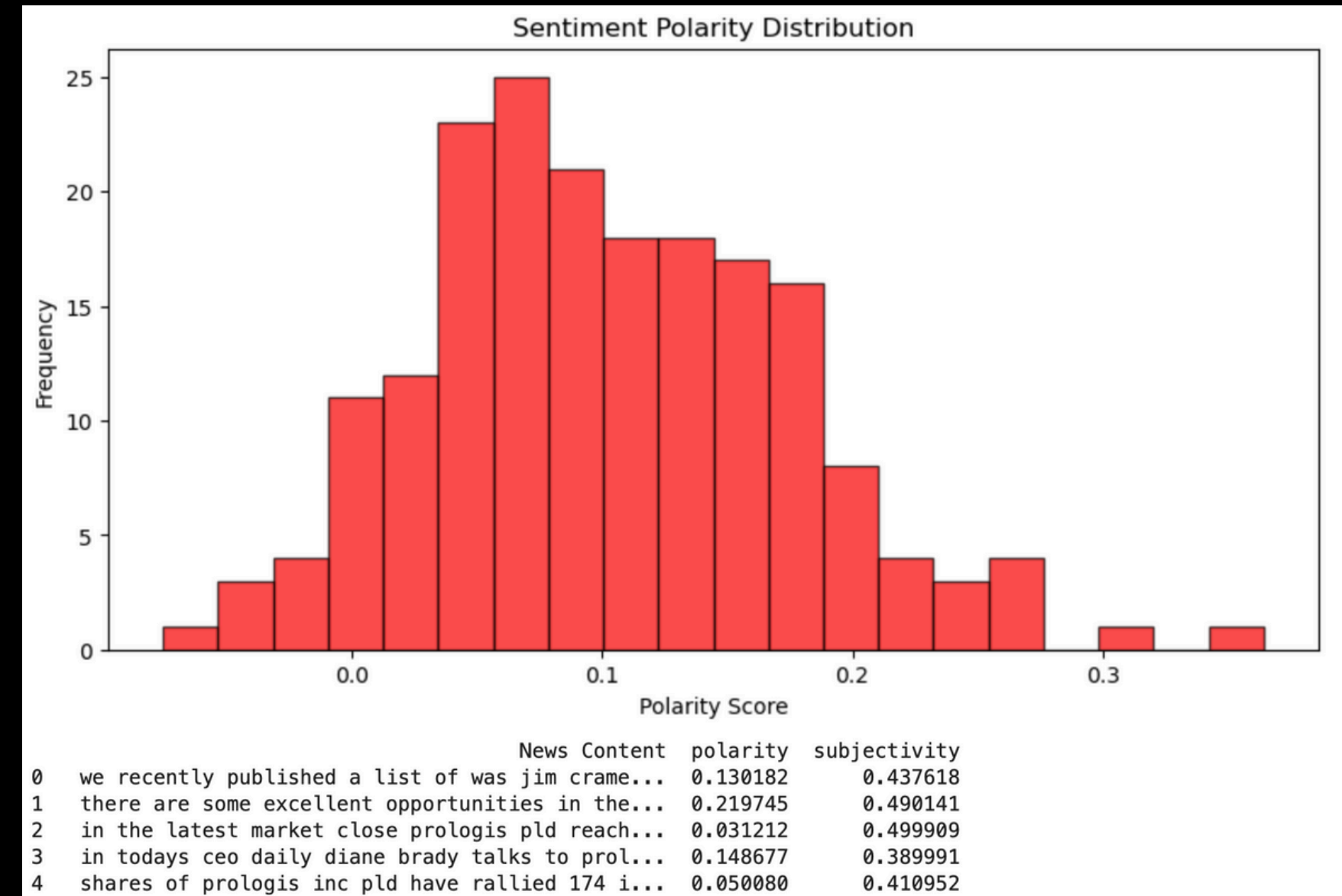
- Investment and Market Focus:** frequently mentioned terms- "share," "real estate," "business," "market," "billion," and "revenue"-highlight a strong emphasis on financial performance and stock-related discussions.
 - Real Estate and Digital Infrastructure:** "Data center" and "digital reality" indicate discussions around technology-driven real estate investments.
- 
- A word cloud visualization of terms related to real estate and digital infrastructure. The words are arranged in a horizontal, overlapping manner. The most prominent words are 'real', 'estate', 'billion', 'portfolio', 'digital', 'reality', 'investor', 'expect', 'consensus', 'analysis', 'period', 'months', 'ended', 'last', 'year', 'looking', 'reit', 'crown', 'castle', 'quarter', 'access', 'portfolio', 'zacks', 'really', 'net', 'addition', 'benefit', 'impact', 'growth', 'transaction', 'ffo', 'significant', 'increase', 'demand', 'forecast', 'predicted', 'analyst', 'ratings', 'stock', 'valuation', 'market', 'trends', 'revenue', 'performance', 'discussed', 'future', 'market', 'predictions', 'growth', 'potential', 'opportunities', 'significant', 'increase', 'impact', 'demand', 'forecast', 'predicted', 'analyst', 'ratings', 'stock', 'valuation', 'market', 'trends', 'revenue', 'performance', 'discussed', 'future', 'market', 'predictions', 'growth', 'potential', 'opportunities'.
- Financial & Performance Indicators:** Words like "billion," "transaction," "ffo (funds from operations)," "growth," and "revenue" highlight financial performance, stock valuation, and market trends. In addition to mentions of "Zacks rank" and "consensus estimate" suggest analyst ratings and stock evaluation metrics are frequently discussed.
 - Market Trends & Sentiment:** "Expected," "estimate," "consensus," and "forecast" indicate that many articles discuss future market predictions and analyst expectations. And the presence of "significant," "increase," "impact," and "demand" suggests an emphasis on growth potential and market opportunities.



▶▶▶ Stock Market Analysis – Real Estate

Key Insights : Polarity vs Subjectivity

- **Overall Sentiment Trend:** The distribution is skewed toward neutral to slightly positive polarity (0.0 to 0.2). Implying that most articles have **mildly positive sentiment**, indicating neutral to **optimistic** reporting rather than extreme positive or negative tones.
- **Absence of Extreme Sentiment:** Very few articles have strongly positive or negative polarity (close to 1 or -1) suggesting the news content is fact-based rather than highly opinionated.
- **Implications for Market Sentiment :** A prevalence of neutral to slightly positive sentiment may indicate steady market conditions with no extreme volatility.
- **Moderate Objectivity in News Content:** The subjectivity scores in the table range from 0.38 to 0.49, indicating that the articles contain a mix of objective facts and subjective opinions. A subjectivity score closer to 0 suggests purely factual content, while scores near 1 indicate highly opinionated writing.
- **Implications for Market Interpretation:** Higher objectivity means news articles are less likely to be speculative or biased, making them more reliable for investors and analysts.





Conclusion

For this project, we explored different web scraping tools, including Scrapy, Selenium, and BeautifulSoup, to collect stock data and news articles from Yahoo Finance. Scrapy and Selenium were difficult to use due to dynamic content loading and anti-bot protections, making them less efficient for our needs. BeautifulSoup worked best as it allowed us to extract data quickly and with fewer technical issues.

Our findings show that news sentiment plays a significant role in market trends. Positive sentiment often aligns with strong investor confidence, while negative sentiment reflects potential risks or uncertainties. Since many investors rely on news when deciding whether to buy, hold, or sell stocks, sentiment analysis provides an objective way to assess market conditions. This approach can be valuable for predicting stock movements and understanding how public perception impacts financial markets.



Thank you!

For any questions, please email us at:

sjdid1@babson.edu

asinha1@babson.edu

nthamma1@babson.edu

asharma16@babson.edu

Ododo1@babson.edu

