Data Science Project Report:

Earnings Call Complexity & Stock Performance

Student Name: Suryansh Gupta
Project Duration: 4 weeks

Overview

This data science project examines whether how executives speak on earnings calls conveys information about a firm's near-term performance. Using S&P 500 call transcripts (Hugging Face dataset) and post-event stock returns from Yahoo Finance, I construct a composite Complexity Index that summarizes the clarity versus complexity of management language and test its association with 5-day returns after the call.

The index is built from six interpretable linguistic metrics:

1. **Average sentence length** — mean tokens per sentence, capturing verbosity and syntactic complexity.

2. **Readability (Flesch–Kincaid grade)** — estimates education level needed to understand the text; higher scores imply harder-to-read language.

3. **Lexical diversity (MTLD)** — measures vocabulary variety independent of text length.

4. **Jargon ratio** — share of words appearing in a curated domain/buzzword list (e.g., "synergy," "roadmap," "leverage"), approximating reliance on corporate or technical terms.

5. **Hedging ratio** — share of words/phrases expressing uncertainty (e.g., "might," "could," "believe," "expected"), indicating cautious tone.

6. **Passive-voice ratio** — proportion of sentences containing passive constructions, a proxy for evasiveness or reduced agency.

Objective

1. Measure linguistic complexity in corporate earnings calls.
2. Test if companies that use more complex language experience weaker stock performance.
3. Build a simple predictive model as a stretch goal to explore language-based forecasting.

Methodology

4. Data: Quarterly earnings call transcripts and 1-, 2-, and 5-day forward stock returns (Yahoo Finance).
5. Text Processing: Clean, tokenize, and analyze transcripts using Python libraries — SpaCy, TextStat, NLTK, and LexicalRichness.
6. Metrics: Average sentence length, readability (Flesch–Kincaid), lexical diversity, jargon ratio, hedging ratio, and passive-voice ratio.
7. Complexity Index: Weighted composite of normalized linguistic metrics.

8. Analysis: Correlate each factor — and the combined Complexity Index — with post-call stock returns at multiple time horizons (1 day, 2 days, 5 days).
9. Visualization: Scatter plots and regression trends showing how complexity features relate to short-term stock movements.
10. Stretch Goal: Train simple machine-learning classifiers (logistic regression, random forest) to predict "stock up vs. down."

Key Findings

1. Sentence Length (r = –0.06, p = 0.001):

- **Significant negative relationship.**
- Companies using **longer or more complex sentences** tend to have **slightly lower 5-day returns**.
- Suggests that **verbose or roundabout communication** may signal uncertainty or lack of confidence, leading to mild negative investor reactions.

2. Readability (r = +0.02, p = 0.382)

- No statistically significant effect.
- Simpler or harder-to-read transcripts don't appear to influence short-term stock movements in this dataset.

3. Lexical Diversity (r = +0.06, p = 0.002)

- **Positive and significant correlation.**
- Calls with **more diverse vocabulary** tend to be followed by **slightly better stock performance**.
- Indicates that **linguistic richness** — perhaps reflecting more confident, articulate management — may be positively perceived by investors.

4. Jargon Ratio (r = +0.09, p < 0.001)

- **Strongest significant positive correlation** among all factors.
- Suggests that when companies use **more technical or specialized terms**, stocks perform **better in the following days**.

- This could indicate that **credible or expert language** builds investor confidence, especially when complexity signals competence rather than confusion.

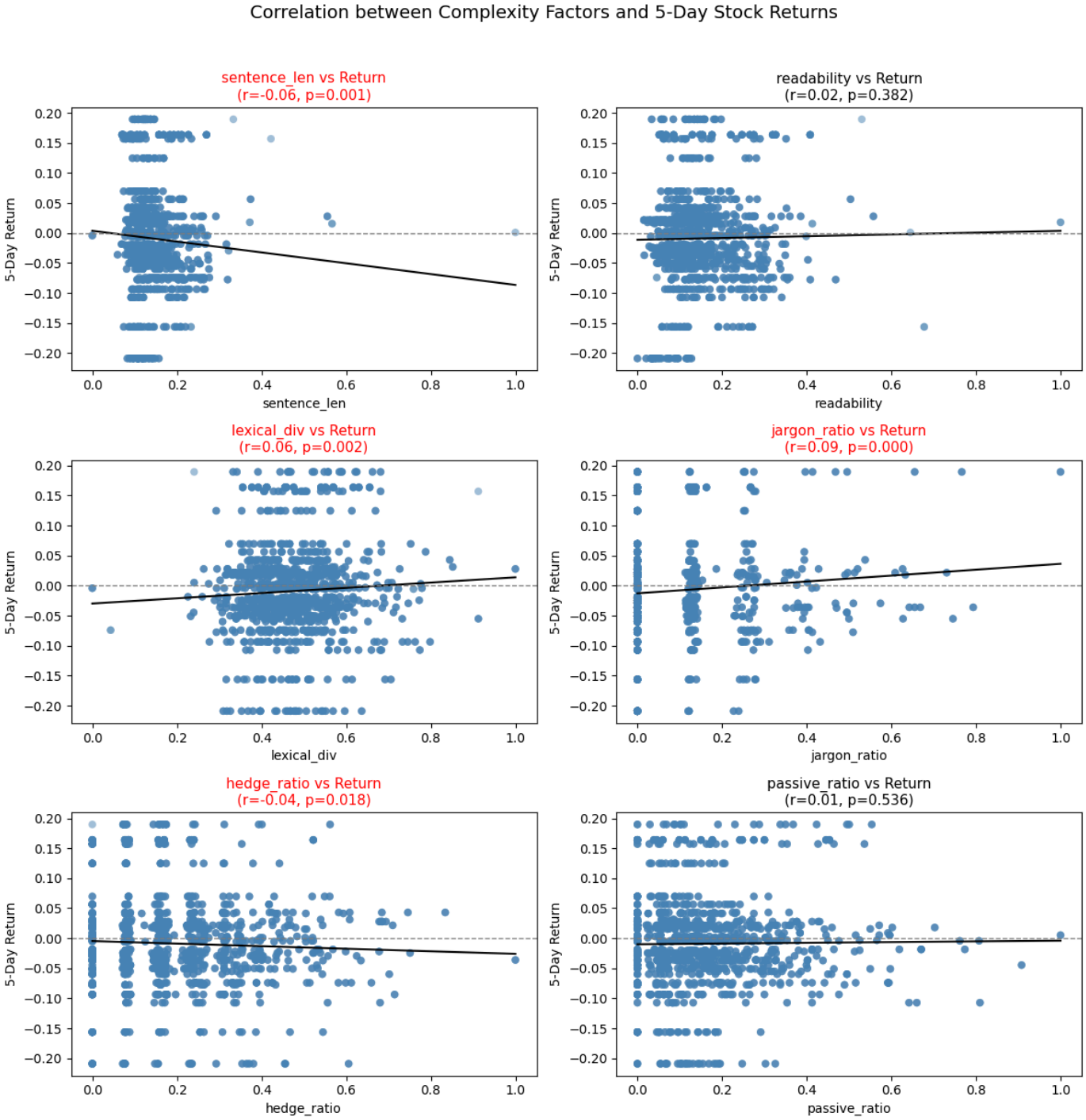
5. Hedging Ratio (r = –0.04, p = 0.018)

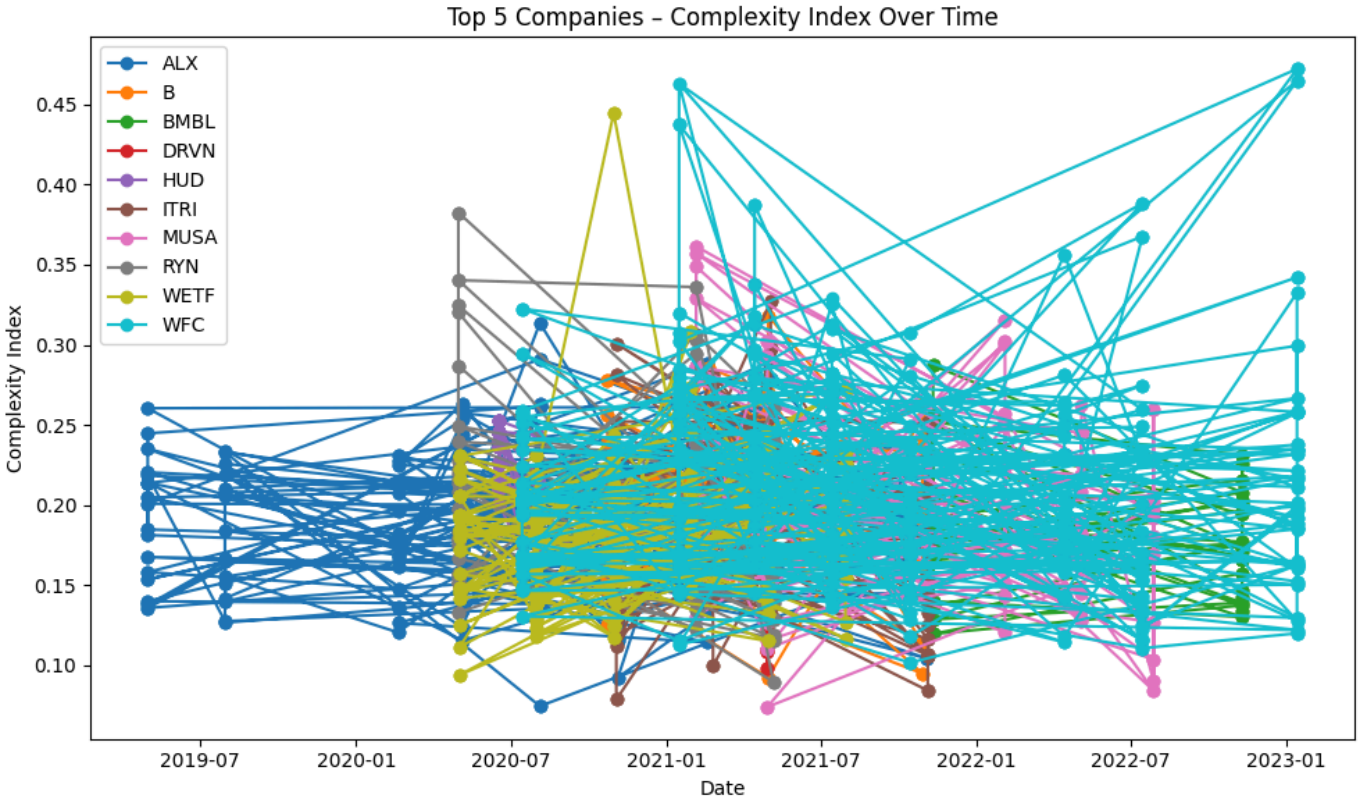
- **Significant but weak negative correlation.**
- Firms that use **more cautious or uncertain phrasing** ("may," "might," "could") tend to see **slightly weaker returns**.
- Implies that investors may interpret **hedging language as lack of confidence or risk aversion**.

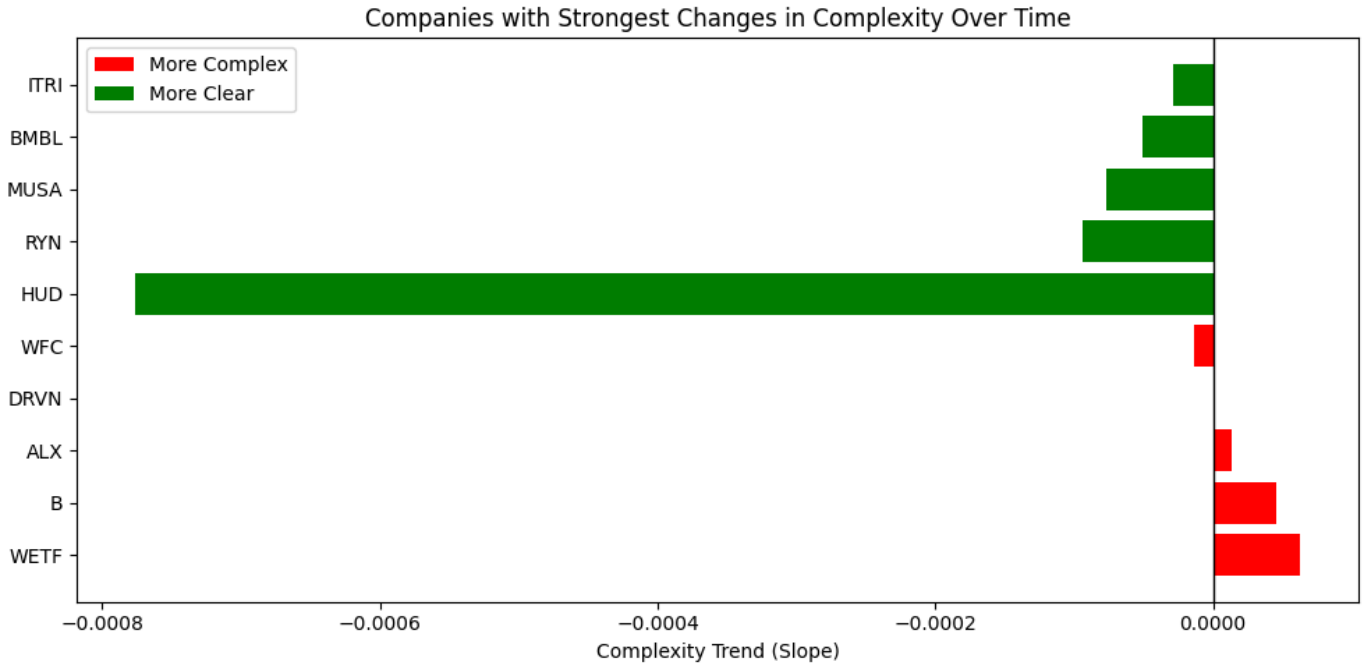
6. Passive Ratio (r = +0.01, p = 0.536)

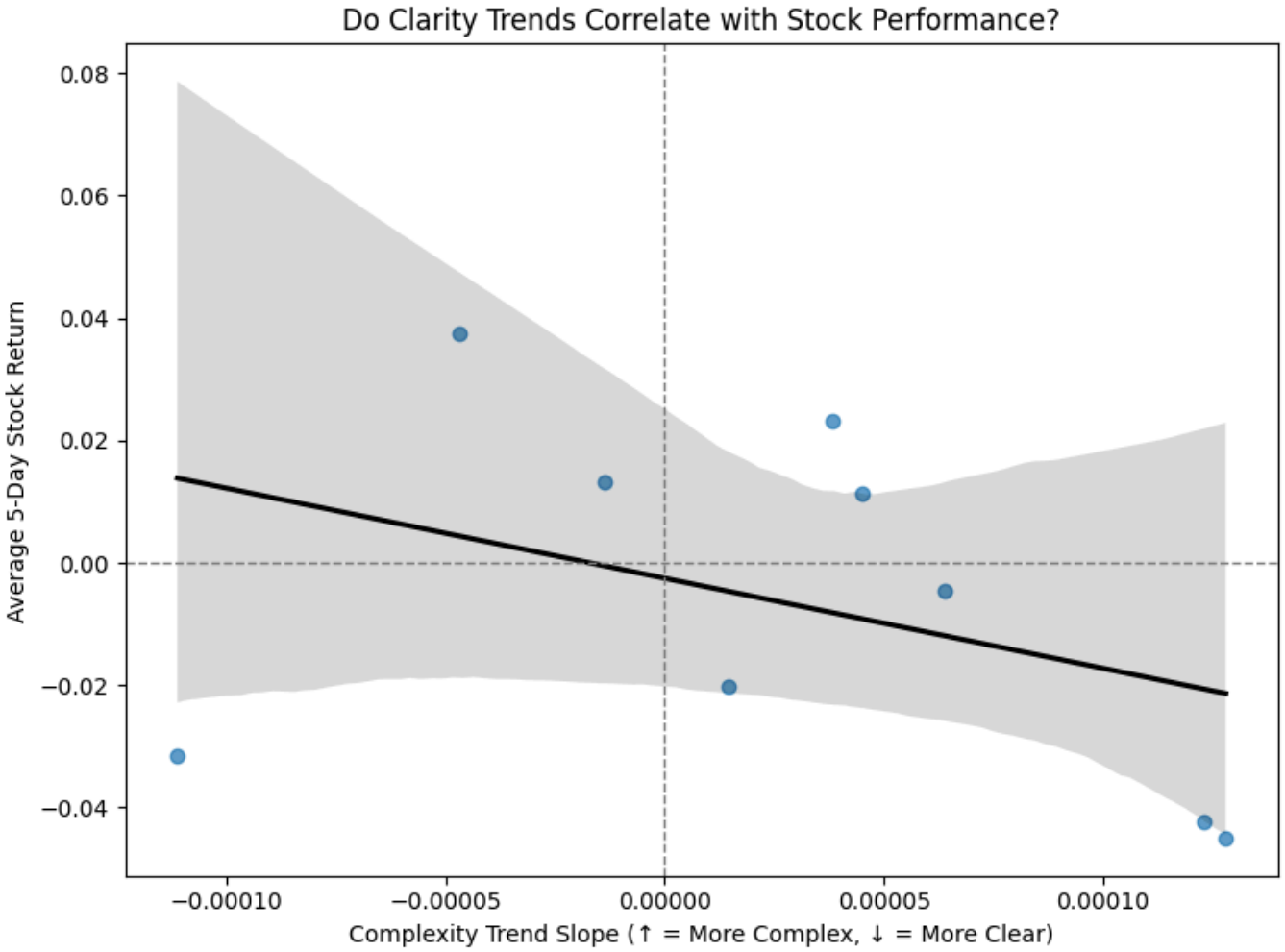
- No meaningful relationship.
- The frequency of **passive-voice usage** does not affect stock outcomes — investors may not react to this stylistic element.

Visual Findings




Top 5 Companies – Complexity Index Over Time
ALX
B
BMBL
DRVN
HUD
ITRI
MUSA
RYN
WETF
WFC
Complexity Index
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
2019-07
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
Date


Companies with Strongest Changes in Complexity Over Time
More Complex
More Clear
ITRI
BMBL
MUSA
RYN
HUD
WFC
DRVN
ALX
B
WETF
−0.0008
−0.0006
−0.0004
−0.0002
0.0000
Complexity Trend (Slope)



The visual analysis confirms that clearer communication often aligns with stronger market reactions. The downward slope in the clarity-trend regression line suggests that firms improving clarity over time tend to achieve better 5-day post-earnings returns.

Conclusion

Language complexity reflects more than style. It signals confidence, clarity, and potentially underlying business strength. The Earnings Call Complexity Index integrates natural language processing and financial analytics to quantify executive tone and its market impact. Key conclusions from this analysis are:

1. Investors appear to reward confident, precise, and articulate communication (higher lexical diversity, appropriate jargon use).
2. Conversely, overly complex or hedged language correlates with weaker post-call performance.
3. While these effects are small, they suggest managerial tone and communication clarity carry informational value beyond financial metrics.