Project Proposal

Daniel Buckman, Santiago Etchepare, Michael Logan, Livia Mucciolo

Research Question

Can sentiment extracted from unstructured company-generated documents (e.g., SEC filings, press releases, earnings calls) be used to predict future excess stock returns? Specifically, we ask: *Do changes or sequential patterns in document sentiment contain predictive power beyond what is captured by standard asset pricing models?*

Motivation

A core tenet of modern finance is that publicly available information should be reflected in asset prices. However, increasing evidence suggests that markets underreact to complex and unstructured information, such as the language used in corporate disclosures. Prior work by Loughran and McDonald [2011] has shown that sentiment in SEC filings can predict returns. More recently, Kelly et al. [2021] demonstrate that sophisticated textual analysis captures risk factors and investor expectations more effectively than traditional metrics.

Our hypothesis is that these documents contain forward-looking signals implicitly communicated by management—either through tone, choice of language, or the structure of updates—that are not immediately priced in by the market. This is particularly likely in industries like biotech, where long-run value is tied to opaque, uncertain R&D pipelines and management may hint at future developments before they are formally announced.

We also believe that sentiment is not static—it evolves over time and may follow sequences that are meaningful. For example, a steady deterioration in sentiment across multiple quarters may indicate worsening fundamentals that are not yet reflected in earnings forecasts or analyst reports. By explicitly modeling sentiment dynamics, we aim to capture this temporal structure.

While our baseline predictive framework will use OLS regressions for interpretability and benchmarking against established asset pricing models, we will also apply machine

learning models—including random forests, gradient boosting, and transformer-based sequence models—to explore whether added model complexity yields stronger predictive performance. Comparing results across models will allow us to assess the tradeoff between accuracy and interpretability, and evaluate whether there is true virtue in complexity for this domain.

Data

- **Text Data:** SEC filings (10-K, 10-Q, 8-K) from EDGAR via WRDS. Later stages may include press releases, earnings call transcripts, and audio.
- **Return Data:** Monthly/daily returns from CRSP.
- **Factor Data:** Fama-French 5 factors + Momentum, from Ken French's data library.
- **Metadata:** Document types, timestamps, industry classifications (NAICS), and firm identifiers.

Methods

We propose a tiered modeling approach:

Level 1: Apply open-source sentiment models (e.g., FinBERT) to full 10-K and 10-Q. Predict excess returns at 1, 3, 6, and 12 months using OLS with FF5 + Momentum factors. Evaluate signal significance and out-of-sample performance via a rolling 80/20 training/test window.

Level 2: Use RAG (retrieval-augmented generation) to extract targeted sections (e.g., MD&A, Risk Factors). Add intra-document similarity scores and lagged sentiment as features.

Level 3: Incorporate additional document types (e.g., press releases, earnings call transcripts). Evaluate whether added sources enhance predictive power.

Level 4: Focus on biotech firms. Use specialized lexicons to analyze sentiment about R&D pipelines, trial outcomes, and regulatory milestones.

Level 5: Extend analysis to global firms. Examine cross-country variation in sentiment predictability based on regulatory regimes and market sophistication.

Across all levels, we will compare OLS to machine learning models to assess marginal value of complexity. We will report performance metrics (e.g., R^2, Sharpe ratio) and use portfolio sorting to assess economic relevance.

Preliminary Bibliography

References

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35–65.

Kelly, B., Kostyshak, S., & Manela, A. (2021). Textual Analysis for Economics and Finance. *Journal of Economic Literature*, 59(1), 1–80.

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3), 712–729.

Ke, B., & Zhang, S. (2021). The Information Content of Earnings Call Audio: Evidence from Vocal Cues. *Journal of Accounting Research*, 59(2), 497–548.