**Forecasting IPO Underpricing Using Multimodal Sentiment, Prospectus Analysis, and Macro-Financial Signals**

**Muhammad Farhan**

Courant Institute of Mathematical Sciences

Department of Computer Science

New York University

New York City, USA

m.farhan@nyu.edu

**Abstract.** TBD

**Keywords:** IPO underpricing, sentiment analysis, prospectus, machine learning, financial forecasting, macro-financial.

1. **Introduction**

TBD

1. **Literature Review**

Recent advances in financial machine learning and natural language processing (NLP) have enriched the prediction of IPO underpricing by introducing features derived from textual disclosures and investor sentiment. Traditionally, IPO research focused on firm fundamentals and market timing; however, these newer approaches highlight behavioral and information asymmetry components, offering a multidimensional view of underpricing drivers.

Bodnaruk et al. (2020) integrated S-1 filing sentiment and structured data using machine learning methods, finding that tone from specific sections like “Risk Factors” significantly enhances predictive power. Their Random Forest model outperformed baselines, with sentiment polarity and uncertainty measures proving particularly useful. This confirmed the utility of NLP techniques in capturing the informational content of filings.

Von Bodman (2024) examined European IPO prospectuses, employing sentence embeddings and similarity scores. The study found that uncertain and litigious language was consistently associated with higher underpricing, underscoring how regulatory disclosures convey latent signals to investors.

Hashem et al. (2025) used multinomial logistic regression on structured textual features, showing that prospectus composition without absent sentiment was a strong predictor of IPO outcomes. This demonstrates the significance of document structure, including section headers, length, and disclosure balance.

Deb et al. (2020) explored prospectus tone using Loughran-McDonald dictionaries and fuzzy logic models like GARSINFIS. The study showed that optimistic tone correlates with stronger 30-day returns, while models with interpretability (e.g., neuro-fuzzy) performed on par or better than complex black-box methods.

Liu et al. (2021) demonstrated that Twitter and Reddit sentiment scores in the 10 days preceding IPOs had predictive power for underpricing, acting as a proxy for retail investor enthusiasm. Similarly, Liang and Chen (2024) extended this by incorporating Google Trends, Twitter, and StockTwits data to show how online attention correlates with IPO demand and pricing.

Aziz and Bari (2025) focused on the emotional spectrum of social sentiment. Their NLP-based classification revealed that emotions such as joy and trust were positively correlated with IPO underpricing, while fear and anger were associated with conservative pricing. These findings emphasize that emotional granularity enhances traditional polarity-based models.

Together, this literature supports a multimodal approach by integrating text analysis, online behavior, and macro signals as the most promising strategy for forecasting IPO underpricing.

1. **Data Sources**

To develop a robust and interpretable model of IPO underpricing from 2010 to 2024, I integrate diverse datasets that capture structural, behavioral, and macroeconomic signals. Each data source is selected based on prior literature and aligned with predictive features found to be significant.

1. **IPO Metadata**

The foundation of our dataset is the IPO calendar from NASDAQ and deal-level data extracted from SEC EDGAR. Key fields include IPO date, ticker symbol, offering price, number of shares offered, shares outstanding, and company name. These allow us to compute the insider holdings ratio:

This ratio reflects the extent of ownership retained by insiders post-IPO, a proxy for insider confidence and long-term alignment. I also generate a binary flag for venture capital (VC) participation by identifying whether pre-IPO shareholders (typically early investors) sold shares during the offering. These ownership variables are supported by studies such as Liu et al. (2021), which link insider behavior to post-IPO returns.

1. **Prospectus Textual Data**

Using EDGAR’s S-1 filings, I extract structured sections such as “Risk Factors,” “Business Description,” and “MD&A.” These sections are parsed using NLP pipelines to compute several textual features:

* Sentiment polarity and subjectivity via FinBERT
* Lexical complexity using the Fog Index

Section-specific tone and complexity metrics have been shown to improve prediction, especially in Bodnaruk et al. (2020) and Deb et al. (2020). I also include structural markers such as word count and section length as in Hashem et al. (2025).

1. **Social Media and Retail Sentiment**

I gather sentiment from Reddit and StockTwits via Playwright-based scraping tools, focusing on the 90-day window prior to each IPO. I extract features such as:

* Post volume (attention proxy)
* Sentiment polarity and emotional categories (e.g., joy, fear)
* Sentiment volatility (standard deviation)

This retail-driven sentiment data mirrors the work of Liu et al. (2021) and Aziz and Bari (2025), who showed that real-time investor discourse significantly predicts short-term IPO returns.

1. **Google Search Trends**

I access Google Trends via PyTrends to obtain interest levels in IPO company names and tickers 90 days before the IPO date. I derive the slope and peak level of search intensity, aligned with findings by Liang and Chen (2024) that suggest retail curiosity often precedes demand surges.

1. **News Sentiment**

I use NewsAPI and FinBERT to analyze financial news articles mentioning IPO-bound companies. I compute daily average sentiment scores and volatility in the 30 days preceding the IPO. These metrics serve as a professional sentiment counterpart to retail emotion, complementing social media data.

1. **Macroeconomic Signals (VIX and Fed Rate)**

Two key macro-financial indicators are incorporated. First, the CBOE Volatility Index (VIX) is obtained from Yahoo Finance (^VIX) to reflect market uncertainty on the IPO date. Second, the effective federal funds rate is collected from the Federal Reserve Economic Database (FRED) as of the month prior to each IPO. Prior literature (e.g., von Bodman, 2024) notes that high volatility and tight monetary policy correlate with lower IPO pricing and weaker short-term performance.

1. **Return Metrics**

Our target variables are derived from Yahoo Finance:

* **1-Day Return**:
* **1-Month Return**: based on adjusted close prices

These are standard measures of IPO underpricing and short-term aftermarket performance, used across all referenced studies.

By combining prospectus language, online sentiment, ownership structure, and macro-financial indicators, I aim to forecast IPO pricing outcomes with both statistical rigor and economic interpretability.

1. **Methodology**

TBD

1. **Results**

TBD

1. **Conclusions and Discussions**

TBD

**References**

Aziz, A., & Bari, A. (2025). *Social Media Emotions and IPO Returns*.

Bodnaruk, A., Loughran, T., & Ritter, J. R. (2020). *Textual Information and IPO Underpricing: A Machine Learning Approach*.

Deb, N., & Lahiri, S. (2020). *Do Words Matter? Predicting IPO Performance from Prospectus Sentiment*.

Hashem, M., & Yilmaz, M. T. (2025). *Prediction of IPO Performance from Prospectus Using Multinomial Logistic Regression: A Machine Learning Model*. Data Science in Finance and Economics, 5(1), 105–135.

Liang, S., & Chen, Y. (2024). *The Power of Prediction With Google Searches and Social Media Posts: Retail Investor Interest and IPO Pricing*.

Liu, S., Zhang, Y., & Wang, H. (2021). *Social Media Sentiment and IPO Pricing*.

Von Bodman, C. (2024). *The Impact of Prospectus Language on IPO Underpricing: A Textual Analysis of European IPOs*.