



# Textual Analysis in Finance

*FA690 Machine Learning in Finance*

**Dr. Zonghao Yang**

2025 Spring

# Introduction to Textual Analysis in Finance

# Why NLP in Finance?

- Financial markets are driven by information
  - Investors, analysts, and traders rely on unstructured textual data
  - Example sources: news articles, earnings reports, social media, analyst research
- Growing interest in financial NLP
  - Increase in data availability and computational power
  - Rise of deep learning and large language models



**Text**

e.g., news, reports



**NLP Processing**



**Actionable Insights**

e.g., investment, risk assessment

# Challenges in Financial Text Processing

- Domain-Specific Jargon
  - Finance has unique terminology (e.g., "hawkish policy," "liquidity crunch")
  - Pretrained LLMs may require fine-tuning for finance
- High-Stakes Predictions
  - Small errors can lead to significant financial losses
  - Need for interpretability and reliability



# Common Sources of Financial Text Data

- Traditional Financial Reports
  - Earnings call transcripts
  - SEC filings (10-K, 10-Q)
- News & Media: Reuters, Bloomberg, CNBC, Wall Street Journal
- Social Media & Alternative Data
  - X (or Twitter), Reddit (e.g., r/wallstreetbets)
  - Podcast transcripts (e.g., Acquired, In Good Company)
  - Interview transcripts



# Earnings Call Transcripts

- What are they?
  - Verbatim records of company executives discussing quarterly financial results with analysts
  - Often include Q&A sessions where analysts ask about company outlook
- Key Insights from NLP
  - Sentiment analysis to gauge executive confidence
  - Keyword extraction to detect earnings surprises
  - Speech pattern analysis (e.g., hesitation, uncertainty)
- Example:
  - Alphabet Q4 2024 Earnings Call [[Link](#)]
  - SeekingAlpha [[Link](#)]

Transcripts



## The Walt Disney Company (DIS) 2025 Annual Meeting of Shareholders Call Transcript

Mar. 21, 2025 8:52 PM ET | The Walt Disney Company (DIS) Stock, DIS:CA Stock | DIS, DIS:CA | 1 Comment



SA Transcripts  
152.8K Followers

Follow

▶ Play Earnings Call

The Walt Disney Company (NYSE:DIS) 2025 Annual Meeting of Shareholders Conference Call March 20, 2025 1:00 PM ET

### Company Participants

James Gorman - Chairman of the Board  
Robert Iger - Chief Executive Officer  
Horacio Gutierrez - Senior Executive Vice President, Chief Legal and Compliance Officer  
Jason Graham - Inspector of Elections  
Grant Bradski - Shareholder Representative of As You Sow  
Stefan Padfield - Executive Director of the Free Enterprise Project  
Jerry Bowyer - President of Bowyer Research

### Conference Call Participants

#### Operator

Today's presentation may include forward-looking statements that we

Example Earnings Call



# SEC Filings (10-K, 10-Q)

- What are they?
  - Regulatory financial reports submitted to the SEC
  - 10-K: Annual report with comprehensive financial details
  - 10-Q: Quarterly financial updates
- Key Insights from NLP
  - Risk sentiment analysis from "Risk Factors" section
  - Comparison of past filings to detect subtle language shifts
  - Extraction of financial metrics for predictive modeling
- Example: SEC Filings for Tesla, Inc. [[Link](#)]

UNITED STATES  
SECURITIES AND EXCHANGE COMMISSION  
Washington, D.C. 20549  
FORM 10-Q

(Mark One)  
☒ QUARTERLY REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934  
For the quarterly period ended September 30, 2024  
OR  
☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934  
For the transition period from \_\_\_\_\_ to \_\_\_\_\_  
Commission File Number: 001-34756

Tesla, Inc.  
(Exact name of registrant as specified in its charter)

Texas  
(State or other jurisdiction of  
incorporation or organization)  
91-2197729  
(I.R.S. Employer  
Identification No.)  
1 Tesla Road  
Austin, Texas  
(Address of principal executive offices)  
78725  
(Zip Code)  
(512) 516-8177  
(Registrant's telephone number, including area code)  
Securities registered pursuant to Section 12(b) of the Act:


Title of each class	Trading Symbol(s)	Name of each exchange on which registered
Common stock	TSLA	The Nasdaq Global Select Market

Indicate by check mark whether the registrant (1) has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 ("Exchange Act") during the preceding 12 months (or for such shorter period that the registrant was required to file such reports), and (2) has been subject to such filing requirements for the past 90 days. Yes ☒ No ☐  
Indicate by check mark whether the registrant has submitted electronically every Interactive Data File required to be submitted pursuant to Rule 405 of Regulation S-T (§232.405 of this chapter) during the preceding 12 months (or for such shorter period that the registrant was required to submit such files). Yes ☒ No ☐  
Indicate by check mark whether the registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, a smaller reporting company, or an emerging growth company. See the definitions of "large accelerated filer," "accelerated filer," "smaller reporting company" and "emerging growth company" in Rule 12b-2 of the Exchange Act:  
Large accelerated filer ☒ Accelerated filer ☐  
Non-accelerated filer ☐ Smaller reporting company ☐  
Emerging growth company ☐

SEC Filings for Tesla, Inc.

# Podcast: Acquired


**ACQUIRED**[HOME](#)[ABOUT](#)[EPISODES](#)[SLACK](#)[CONTACT](#)[MORE ▾](#)[SUBSCRIBE ▾](#)



Indian Premier League Cricket

Spring 2025, Episode 3


March 23, 2025



The Art of Selling Enterprise Software  
(with ServiceNow CEO Bill McDermott)

ACQ2


March 9, 2025



Rolex

Spring 2025, Episode 2


February 23, 2025



Building Web Apps with Just English and AI  
(with Vercel CEO Guillermo Rauch)

ACQ2


February 18, 2025



TSMC Founder Morris Chang

Spring 2025, Episode 1

January 26, 2025



TSMC (Remastered)

Special

January 20, 2025



# Podcast: In Good Company

In Good Company with Nicolai Tangen by Norges Bank Investment Management [[Link](#)]



## 2025

---

HIGHLIGHTS: Jennifer Scanlon - CEO of UL Solutions →

---

UL Solutions CEO: Evolving Safety Testing, AI, and Consumer Protection →

---

Alphabet President and CIO: Advancing AI, Quantum Computing, and Self-Driving Cars →

---

HIGHLIGHTS: Lars Strannegård →

---

Lars Strannegård: Art in Business, Leadership, and the Skills AI Can't Replace →

---

HIGHLIGHTS: Paul Singer →

---

Paul Singer: Activist Investing, Market Risks and Avoiding Losses →

---

HIGHLIGHTS: David Ricks - CEO of Eli Lilly →

---

Eli Lilly CEO: The Weight-Loss Drug Revolution, AI in Pharma, and Innovation →

---

# Overview

## Key NLP Tasks in Finance

- Sentiment Analysis: Predicting market sentiment from text
- Text readability: Measuring the readability of financial document, PR communications, etc.
- Document Representation: Word embeddings for finance
- Challenges & Future Trends: Explainability, multimodal learning



# Sentiment Analysis

# Sentiment Analysis in Finance

- Sentiment: The emotional tone of text (positive, neutral, negative)
- Common techniques: Dictionary-based (or lexicon-based) methods, machine learning, and deep learning
- Why is it important in finance?
  - Market reactions are influenced by sentiment in news, earnings calls, and reports



# Why does Investor Sentiment Matter?

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61(4), 1645-1680.

- Hypotheses: Investor sentiment has a larger impact on stocks whose valuations are
  - Highly subjective: These are stocks where determining the "true" value is challenging due to limited historical data, unclear earnings, or speculative growth potential (e.g., young or unprofitable companies)
  - Difficult to arbitrage: These stocks are hard for rational investors to correct if mispriced, perhaps because they're illiquid, volatile, or costly to trade
- Key Insight: When sentiment swings—whether it's a wave of optimism or pessimism—these stocks are expected to experience bigger price movements compared to stocks that are easier to value or arbitrage





# Empirical Findings

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61(4), 1645-1680.

- When sentiment is low (pessimistic):
  - Stocks with characteristics like small size, young age, high volatility, unprofitability, no dividends, extreme growth potential, or distress tend to have relatively high subsequent returns
  - Interpretation: When investors are pessimistic, these stocks may be undervalued (oversold due to fear or neglect). As sentiment improves or normalizes, their prices rebound, leading to higher returns
- When sentiment is high (optimistic):
  - The same categories of stocks earn relatively low subsequent returns
  - Interpretation: During optimistic periods, these stocks may become overvalued (overbought due to hype or speculation). When sentiment cools, their prices correct downward, resulting in lower returns
- Implication: Investor sentiment plays a significant role in driving stock prices, especially for stocks that are hard to pin down in terms of value and tough for arbitrageurs to trade efficiently



# Investor Implications

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61(4), 1645-1680.

- Enhance investment strategy to capitalize on sentiment-driven opportunities
  1. Target sentiment-sensitive stocks like small, volatile, or growth-oriented firms
  2. Track sentiment indicators to assess market mood
  3. Take a contrarian stance—buying when sentiment is low and selling or avoiding when it's high
  4. Time your trades, managing risks, and grounding your decisions in fundamentals



# Measuring Sentiment: Dictionary

- Dictionary-based methods

- Tetlock, Paul C. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market.” *Journal of Finance* 62, no. 3 (2007): 1139–68.
- Loughran, Tim, and Bill McDonald. “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks.” *Journal of Finance* 66, no. 1 (2011): 35–65.

- Dictionary-based sentiment measure

$$\text{Positive Sentiment} = \frac{\text{Number of Positive Words}}{\text{Total Words in Document}}$$

Negative and uncertainty measures are similarly defined

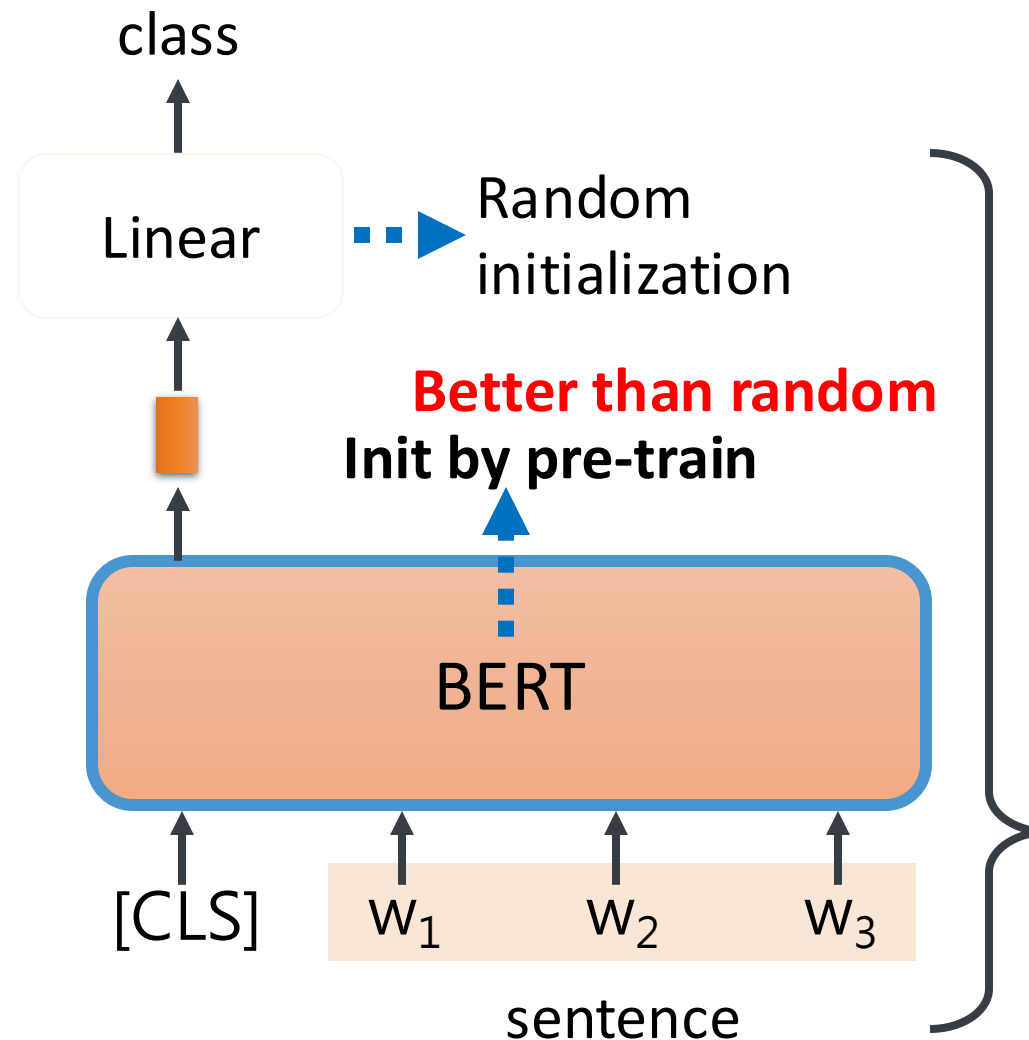
- Comments

- The finance sentiment dictionary, such as Loughran and McDonald (2011), is finance specific
- Simple but lacks context awareness

# Measuring Sentiment: Deep Learning Approach

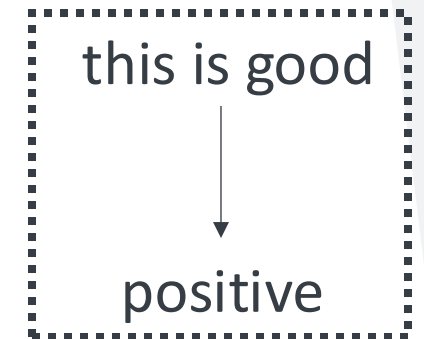
## BERT or FinBERT

- Sentiment analysis using BERT can be framed as a prediction problem where the target is the stock return or binary stock price movement
- “Sentiment” measures how much does the underlying text predict the price movement of the corresponding stock



Input: sequence  
output: class

Example:  
Sentiment analysis



This is the model  
to be learned.

# Sentiment to Investment

- Investment: Given a universe of assets (e.g., stocks), which assets to invest in, and how much to invest in each asset?
- Mean-variance portfolio optimization framework

$$\begin{aligned} r &= (r_1, r_2, \dots, r_N) \\ E &= \begin{bmatrix} \sigma_{11} & & & \\ \sigma_{21} & \sigma_{22} & & \\ \cdot & \cdot & \cdot & \\ \sigma_{N1} & \sigma_{N2} & \cdot & \sigma_{NN} \end{bmatrix} \end{aligned} \longrightarrow \begin{aligned} &\max_w w^T \tilde{r} \\ &s.t. \\ &\sum_{i=1 \dots T} w_i = 1 \\ &w^T \tilde{E} w \leq \sigma \end{aligned}$$

- The expected returns ( $r$ ) are usually estimated based on historical returns, which may contain estimation errors for future returns
- The Black-Litterman model is a portfolio optimization framework that combines an investor's subjective views on asset returns to update expected return estimation

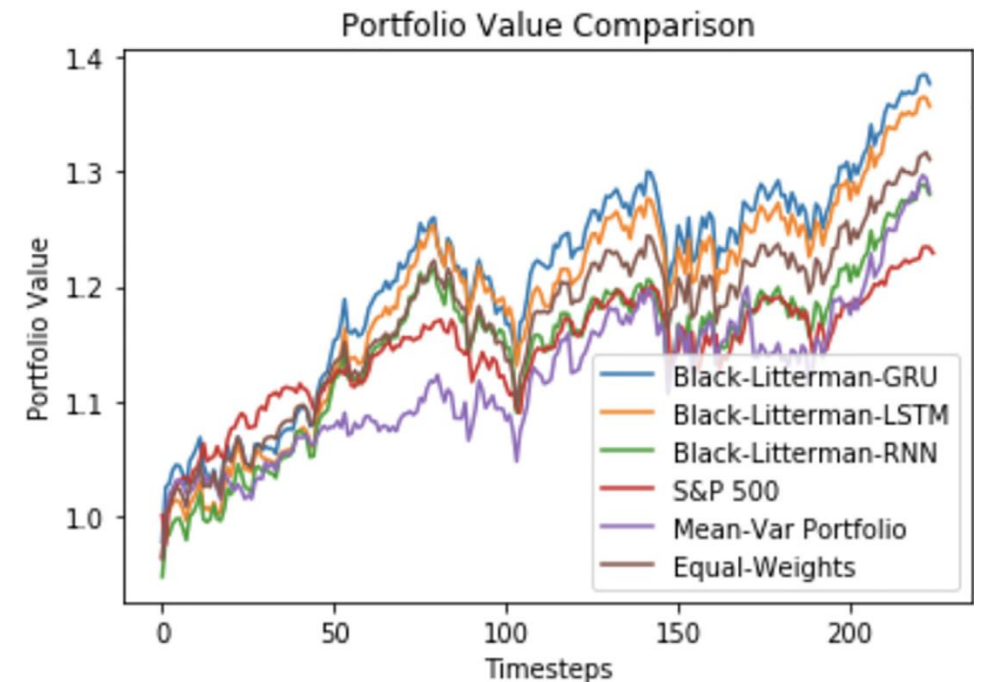


# Sentiment to Investment

Hung, M. C., Hsia, P. H., Kuang, X. J., & Lin, S. K. (2024). Intelligent portfolio construction via news sentiment analysis. *International Review of Economics & Finance*, 89, 605-617.

## ■ Questions:

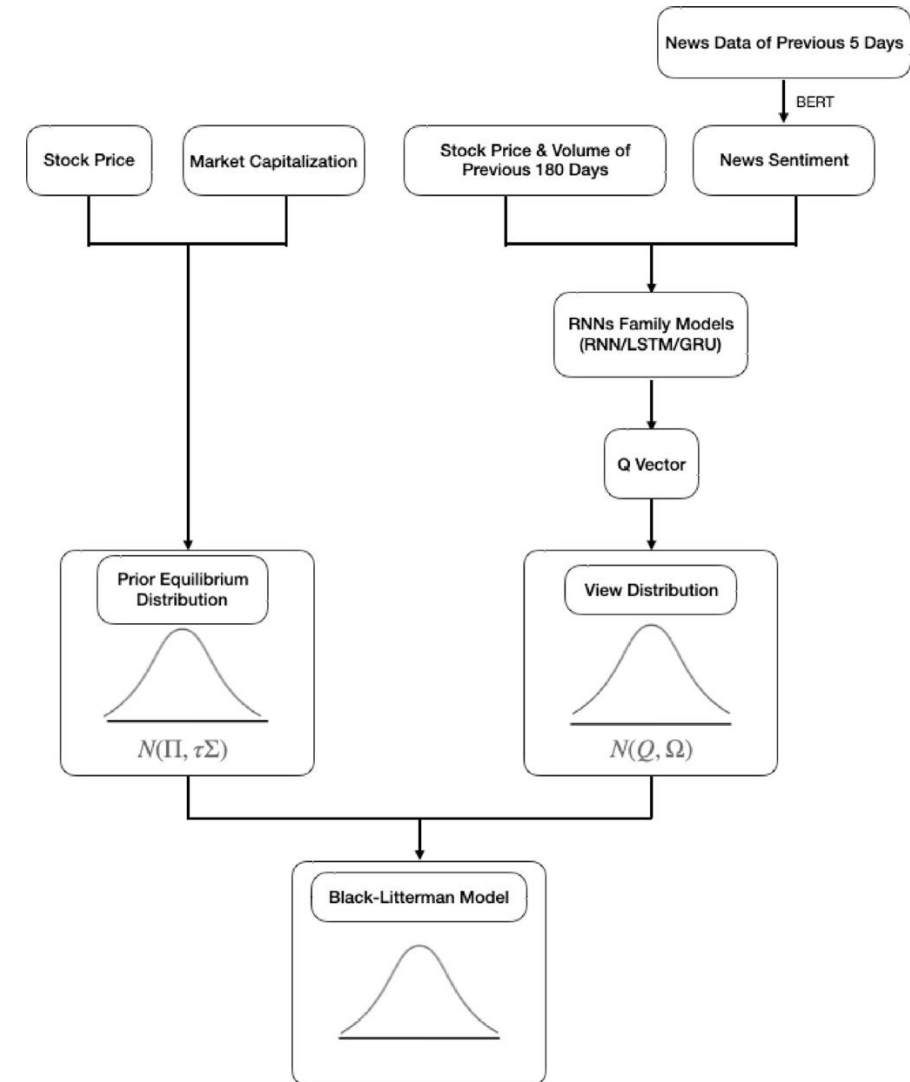
- How can news sentiment analysis translate into forming investment portfolios?
  - Black-Litterman model (Hung et al., 2024)
- Does sentiment analysis integrated into the Black-Litterman model enhance portfolio performance?



# Sentiment to Investment

Hung, M. C., Hsia, P. H., Kuang, X. J., & Lin, S. K. (2024). Intelligent portfolio construction via news sentiment analysis. *International Review of Economics & Finance*, 89, 605-617.

- Sentiment analysis: Binary classification based on stock prices using BERT
- Machine learning prediction: Combining sentiment, historical prices and volumes to predict the stock return
  - RNN/LSTM/GRU models
- Apply the Black-Litterman model to update the return estimation, and thus, the optimal portfolio, based on the return prediction



# Does Finance Benefit Society?

Reference: Jha, M., Liu, H., & Manela, A. (2025). Does finance benefit society? A language embedding approach. *Review of Financial Studies*.

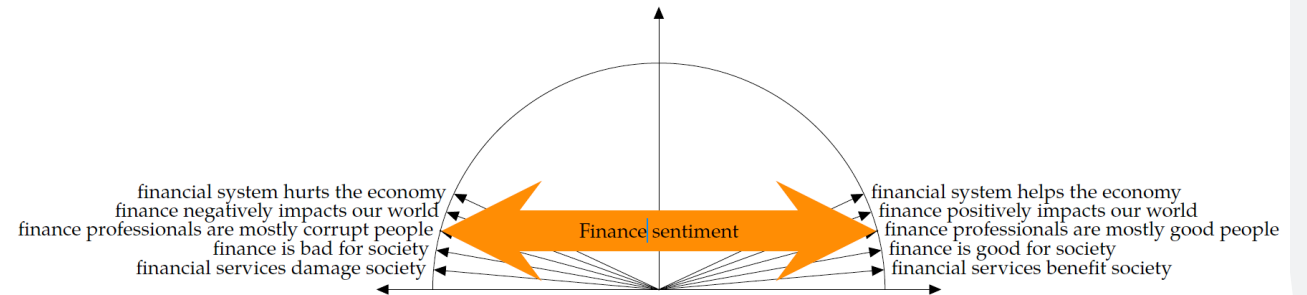
- Research question: How does public sentiment toward finance evolve across countries (1870–2019), and how does it correlate with economic outcomes?
- Data:
  - Books in 8 languages and 140+ years (Google Books)
  - Focus: Sentences with "finance"/"financial" (5-gram snippets)
- Key findings
  - Persistent Sentiment Gaps: Higher in capitalist economies (US/UK) vs. lower in China/Russia
  - Predictive Power: Sentiment growth predicts GDP/credit expansion (1 SD  $\uparrow$   $\rightarrow$  0.1–0.3% GDP boost)
  - Crisis Signals: Sentiment drops before financial crises (e.g., 2008, 1929)



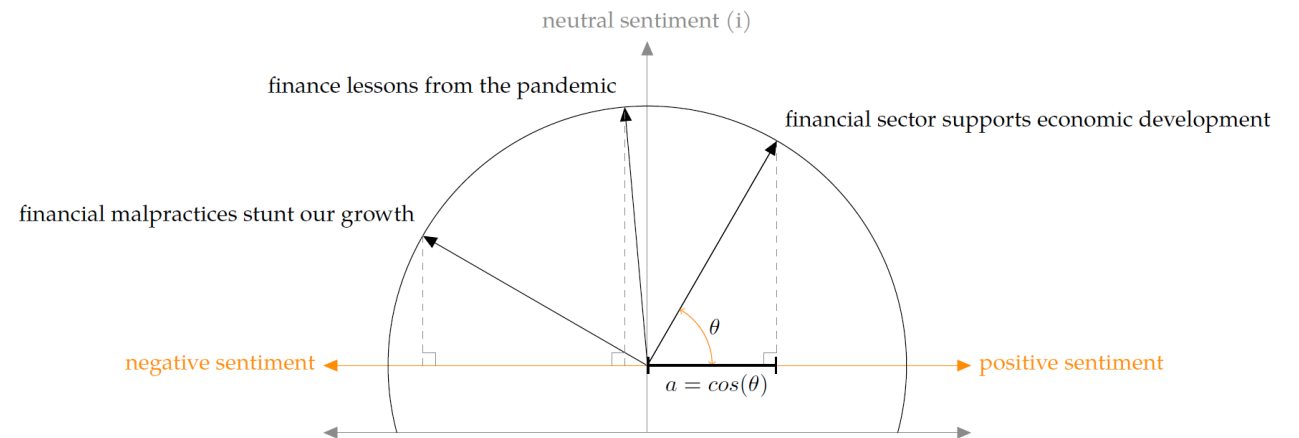
# Does Finance Benefit Society?

## Methodology

- BERT Model : Pre-trained on 3.3B words (context-aware vectors)
- Measuring sentiment: Cosine similarity with sentiment axis
  - Seed phrases: "finance fosters growth" vs. "finance causes crises"



(a) Defining the positive minus negative finance sentiment dimension

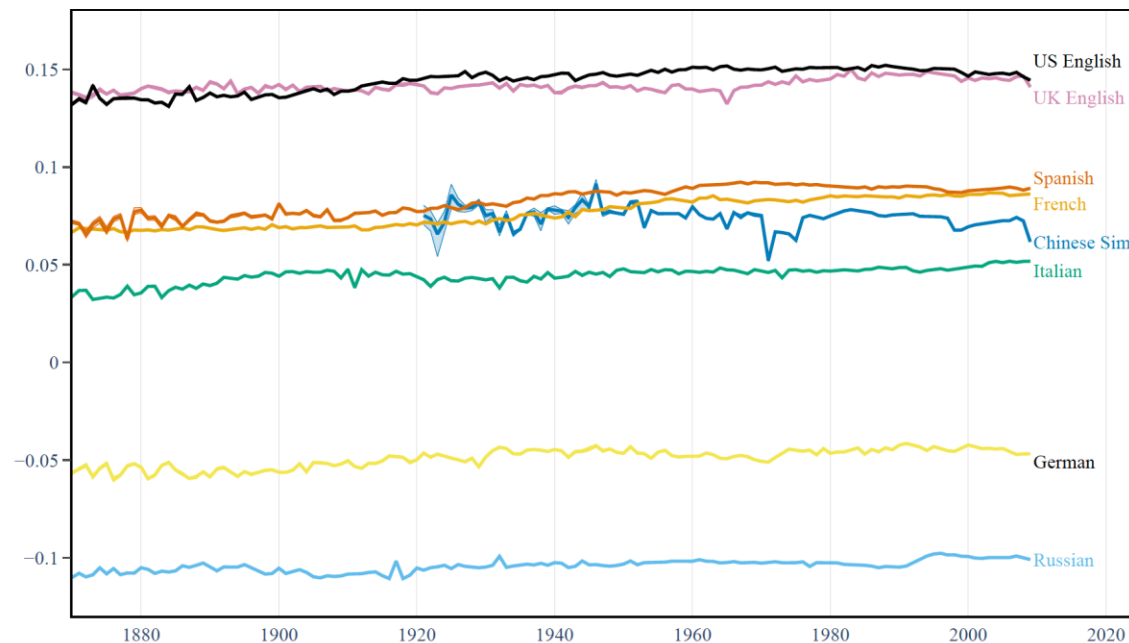


(b) Projection of sentences onto positive minus negative sentiment dimension

# Does Finance Benefit Society?

## Results and Validation

- Validation
  - Placebo Tests: No correlation with coal/paper/tobacco sentiment
  - Robustness: Results hold with broader definitions of "finance"





# Does Finance Benefit Society?

- Implications
  - Policy: Leading indicator for credit booms/crises
  - History: Tracks trust shifts during events like the Great Depression
  - Research: Extend method to tech/energy sectors
- Limitations
  - Endogeneity: Reverse causality (economic downturns → sentiment)
  - Data Lag: Books published years after events (vs. real-time news)
- Consider text that is available in real time: e.g., news article, social media posts



# Readability in Financial Texts

# Readability in Financial Texts

- Why does readability matter?
  - Investors and analysts process large amounts of text
  - Difficult-to-read disclosures may increase information processing costs
  - Readability affects market reactions, stock volatility, and analyst forecasts
- Traditional Readability Measures
  - Fog Index: Equal to  $0.4 * (\text{average number of words per sentence} + \text{fraction of complex words})$ , where complex words are those with more than two syllables
  - File Size: Assumes longer documents are harder to read
  - Limitations: Ignore word context and target audience comprehension
- Example of complex words that will count towards Fog index: *financial, company, interest, agreement, including, operating, period, and related*
  - Most frequently appearing complex words in annual report



# Measuring Readability with Language Predictability

Zang, Amy, Jiexin Zheng, and Rong Zheng. “Measuring Readability with Language Predictability: A Large Language Model Approach.” SSRN (2024)

- Large language models: Next token word predictions
  - Predict  $p(w_{t+1}|w_1, \dots, w_t)$ ; then generate the next word by sampling from the conditional probability distribution
- Zang et al. (2024) propose to use predictability as a measure for readability
  - Language Predictability Score (PLS)
$$\log p(w_i | w_{i-256}, \dots, w_{i-1})$$
  - Large language models: BERT and GPT2
  - Intuition: For a given sentence, the more closely the words are in line with the LLM predictions, and more readable is the sentence
- Compared with traditional measures
  - Capture context-dependent readability
  - Consider the target audience’s (investors, analysts) expectations by fine-tuning the LLM



# Measuring Readability with Language Predictability

Zang, Amy, Jiexin Zheng, and Rong Zheng. “Measuring Readability with Language Predictability: A Large Language Model Approach.” SSRN (2024)

- Example: “Our company faced challenges in integrating the acquired businesses last year.”

Word	Preceding Context (Simplified)	$p(w_i \text{context})$
company	"During the fiscal year"	0.60
faced	"...year our"	0.45
challenges	"...our company faced"	0.70
integrating	"...faced challenges in"	0.30
acquired	"...challenges in integrating"	0.65
businesses	"...in integrating the acquired"	0.80
last	"...the acquired businesses"	0.25
year	"...acquired businesses last"	0.90

$$LPS = \frac{1}{N} \sum_{i=1}^9 \log p(w_i|\text{context}) = -0.0626875$$





# Measuring Readability with Language Predictability

Zang, Amy, Jiexin Zheng, and Rong Zheng. “Measuring Readability with Language Predictability: A Large Language Model Approach.” SSRN (2024)

- Key findings based on Language Predictability Score
  - Stock Return Volatility: Lower LPS (less predictable language) is significantly associated with higher post-filing stock return volatility, indicating greater uncertainty among market participants
  - Analyst Forecasts: Lower LPS correlates with higher analyst forecast dispersion (less agreement among analysts) and decreased forecast accuracy, suggesting increased processing difficulty for sophisticated users
  - Comparison to Traditional Measures:
    - LPS outperforms the Fog index, Bog index, and file size in explaining these outcomes, even after controlling for firm fixed effects and other variables
    - Traditional measures often show insignificant or inconsistent relationships with these market outcomes
  - Boilerplate vs. Non-Boilerplate: The readability of boilerplate language (high LPS) does not affect analyst forecasts, while non-boilerplate language (lower LPS) significantly impacts dispersion and accuracy, indicating analysts can discount predictable, less informative content



# “Who” is Reading Financial Documents?

- Think about readability from another perspective: “Who” is actually reading financial documents like 10-K?
- Loughran & McDonald (2017) show that the average publicly traded firm’s 10-K is downloaded only approximately 28 times immediately after the filing
- Human vs Machines?

Reference: Loughran T, McDonald B. 2017. The use of EDGAR filings by investors. J. Behav. Finance 18:231–48

# Document Representation and Financial Text Embeddings

# Document Representation

- Traditional methods
  - Document-term matrix: Bag-of-words, TF-IDF
  - Topic modeling (Example: [The Structure of Economic News](#))
- Limitation: The lack of context; not able to capture the semantic meaning of words
- Modern methods
  - Word2Vec: Li, Kai, Feng Mai, Rui Shen, and Xinyan Yan. “Measuring Corporate Culture Using Machine Learning.” *Review of Financial Studies* 34, no. 7 (2021): 3265–3315.
  - FinBERT: Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2), 806-841.



# FinBERT

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2), 806-841.

- Available on HuggingFace [[Link to FinBERT](#)]
- FinBERT: Fine-tune the pretrained BERT model on financial text (4.9B tokens)
  - Corporate Reports 10-K & 10-Q (2.5B tokens)
  - Earnings Call Transcripts: 1.3B tokens (1.3B tokens)
  - Analyst Reports: 1.1B tokens (1.1B tokens)
- Limitations
  - FinBERT only releases the sentiment analysis function
  - One cannot use FinBERT to create embeddings for document representation and downstream training



# Conclusion and Discussion

# Conclusion

- New information source: Many financial data is available in textual format
- Dimensionality reduction: The textual data is unstructured and inherently ultra-high dimensional
  - Text representation: Use vector embedding to represent text that is aware of context
  - Consistent measure of text: For example, sentiment and readability
  - Information and event extraction: Named Entity Recognition (NER), relation extraction, fine-tuned LLMs
- Downstream analysis
  - Relationship with stock prices considering the idiosyncrasies of firms
  - Relationship with macro economic conditions
  - Investment
- With data and technology, there are many research opportunities in the finance context





# THANK YOU

**Stevens Institute of Technology**  
1 Castle Point Terrace, Hoboken, NJ 07030