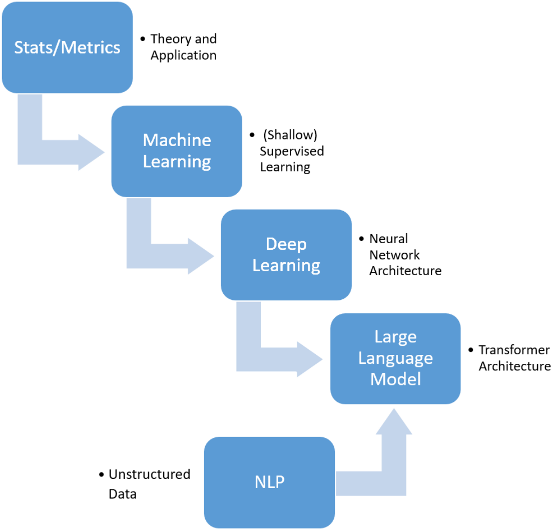
From Machine Learning to Large Language Model

Tian Xie[†]

[†]**Shanghai University of Finance and Economics**

September 20, 2023 @ **Techno-Economics Research Alliance (TERA)**



Stats/Metrics

- Theory and Application

Machine Learning

- (Shallow) Supervised Learning

Deep Learning

- Neural Network Architecture

Large Language Model

- Transformer Architecture

NLP

- Unstructured Data

Formal Introduction

- **Machine Learning Basics**
  - Machine learning is a subset of artificial intelligence (AI) that focuses on developing algorithms and models that can learn patterns from data.
  - Early machine learning techniques include linear regression, decision trees, and support vector machines.
- **The Rise of Deep Learning**
  - Deep learning, a subset of machine learning, gained prominence with the development of neural networks.
  - Neural networks are inspired by the structure of the human brain and consist of interconnected layers of artificial neurons.
- **Natural Language Processing (NLP)**
  - NLP is a branch of AI that deals with the interaction between humans and computers using natural language.
  - Machine learning plays a crucial role in NLP tasks such as language translation, **sentiment analysis**, and speech recognition.

- **Large Language Models**
  - Large language models, like GPT-3, represent a breakthrough in NLP.
  - They are based on deep learning techniques and consist of **millions (or even billions)** of parameters.
- **Significance of Large Language Models**
  - Large language models excel in a wide range of NLP tasks, including text generation, question-answering, and content summarization.
  - They have the potential to revolutionize industries like healthcare, finance, and customer service.
- **Challenges and Ethical Concerns**
  - The development of large language models raises concerns about bias, privacy, and misuse.
  - Researchers and practitioners are working to address these challenges through responsible AI practices.
  - **Not tailor-designed for Economic and Financial analysis.**

- **Future Directions**
  - The field of machine learning continues to evolve, with ongoing research into even larger and more powerful language models.
  - The future of AI and NLP holds exciting possibilities, such as improved language understanding and human-computer interaction.
  - Large language models represent a remarkable journey from traditional machine learning to cutting-edge AI, and they are poised to shape the future of technology.
  - **Field-specific knowledge (training) is essential.**
  - **The possibility of Ensemble estimation (Frequentist or Bayesian).**

Critical Analysis of a Recent Piece

- “Federal Policy Announcements and Capital Reallocation: Insights from Inflow and Outflow Trends in the US” by Yue Qiu, Tian Xie, Wenjing Xie, and Xiangzhong Zheng, *Journal of International Money and Finance*, 2023, vol.139, no.102936, 1-20.
- This paper examines the impact of Federal Open Market Committee (FOMC) policy announcements on international capital reallocation in the U.S. and the implications for financial markets and the economy.
  - We decompose the changes in the fund’s holdings in the US and use the **data imputation** method to identify the channels of change.
  - We quantify FOMC policy announcement text using a recently innovated textual analysis tool, generating a **time-varying dictionaries** for each funds and constructing a diffusion index using principal component analysis.
  - We find that our textual predictors provide significant explanatory power for asset reallocation and showing a great heterogeneity among funds.
- Using the sentiment of the text, we also further explain some facts about international capital flows, e.g. **financial crisis**; **home bias**.

- Our study contributes to the existing literature by expanding upon and complementing previous research on the effects of federal policy announcements on international capital flows.
    - Koepke (2018), Kwabi et al. (2019), Hau and Lai (2016), Albagli et al. (2019), among others
- Our work is also related to a burgeoning literature which examine text-based information of monetary policy statements and its implications for future economic activity.
    - Albagli et al. (2019), Gardner et al. (2022), **Lima et al. (2021, 2023)**, among others.

Preparing the International Capital Flow Data

- **EPFR**. The raw data set spans from January 2004 to August 2020, covering 3074 funds and 119 countries.
- We categorize the positions of each fund into three groups based on the target countries: US, OECD, and NonOECD. Changes in each group's holding assets from one month to the next indicate the fund's position change.
- Our analysis aims to identify the sources of funds that cause changes in US holdings.
  - For instance, if a fund increases its US holdings, we investigate whether the increased holdings result from reducing OECD or NonOECD positions or from additional capital by the funds.
- We fill in the missing observations by the RF based data imputation technique.

Variable Description

Table: Variable Description

Variable	Description
Panel A: Response Variable	
$y^{(0)}$	Change in US position compared to the previous period.
$y^{(1)}$	Change in US position caused by fund itself.
$y^{(2)}$	Change in US position caused by OECD countries.
$y^{(3)}$	Change in US position caused by Non-OECD countries.
Panel B: Input Variable	
X^a	Net assets of the fund.
X^{da}	The change in net assets compared to the previous period.
X^{div}	The number of countries that the fund has money in.
X^{cash}	Cash held by the fund.
X^{lev}	A dummy variable indicating whether the fund uses leverage.

Quantifying the Textual Data

- We utilize monthly textual data sourced from the U.S. Federal Open Market Committee (FOMC) policy statements and meeting minutes.
- These policy announcements have a significant impact on global financial markets as they provide insights into the Federal Reserve's perspective on the US economy and future policy actions.
  - Several researchers, including **El-Shagi and Jung** (**2015**), **Hansen and McMahon** (**2016**), **Hansen et al.** (**2017**), and **Lima, Godeiro, and Mohsin** (**2021**), among others, have found that these announcements have a substantial impact on financial markets, the economic environment, and future economic growth.

- We transform the raw text into **tidy text**.
  - Tidy data principles are a powerful way to simplify and optimize data handling, and they are no less applicable when dealing with text, as explained in **Wickham** (**2014**).
- To tidy the text data, we follow these steps:
  1. bi-gram tokenization
  2. removal of digits and stopwords
  3. elimination of words that are too long or too short
  4. deletion of specific meaningless words (e.g.names and locations)
- Tokenization is the process of splitting text into meaningful units called tokens, such as words or phrases. To avoid ambiguity, we use bi-grams, which are two-word combinations of tokens.

- Building upon the methodology proposed by **Lima and Goderio (2023)**, we employ a semi-supervised learning approach to quantify the tidy text and combine it with data on international capital flows.
- We start by ranking tokens using their term frequency-inverse document frequency (TF-IDF) and only retain the top $N = 100$ tokens. These N tokens constitute our preprocessed dictionary. The dictionary is constant for all funds.
- Nonetheless, there is notable heterogeneity among the funds as they may favor or prioritize different tokens from the dictionary. To account for this variability in preferences among the funds, we employ a supervised learning approach to select the most predictive words for each fund.
- We calculate the frequency of N tokens appearing in each month's announcement and store this information in an $N \times 1$ vector $\mathbf{X}_t$ for $t = 1, ..., T$. Next, we use the least absolute shrinkage and selection operator (LASSO) to estimate the linear prediction equation and select the most predictive terms for each fund i.

- Specifically, we model the next period response variable $y_{i,t+1}$ for fund $i = 1, ..., n$ at time $t = 1, ..., T-1$ as:

  $$y_{i,t+1} = \mathbf{X}_t^\top \boldsymbol{\beta}_i + \epsilon_{i,t+1}.$$

  Here, $\epsilon_{i,t+1}$ represents the error term, $\mathbf{X}_t$ is the $N \times 1$ predictor defined above at time t, and the coefficient vector $\boldsymbol{\beta}_i$ is estimated by minimizing the following equation:

  $$\hat{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\beta}} \sum_{t=1}^{T-1} (y_{i,t+1} - \mathbf{X}_t^\top \boldsymbol{\beta}_i)^2 + \lambda \|\boldsymbol{\beta}_i\|_1.$$

- In this paper, we use a five-fold OOS evaluation, based on the mean-squared forecast error (MSFE) loss function, to select the optimal value of λ. See **Su, Shi, and Xie (2023)** for the algorithmic details.
- After selecting the optimal value of λ using the OOS evaluation procedure, we apply it to the above equation to identify the most predictive tokens associated with non-zero coefficients. We denote the corresponding predictors as $\mathbf{X}_i^*$.
- In addition, we can separate the selected tokens into two groups. $\mathbf{X}_i^{p*}$ and $\mathbf{X}_i^{n*}$, respectively. Notably, that the combination of $\mathbf{X}_i^{p*}$ and $\mathbf{X}_i^{n*}$ is equivalent to $\mathbf{X}_i^*$.

Diffusion Index Model

- The diffusion index (DI) model has a rich history in the field of economics and finance, with its roots traced back to the pioneering work of **Stock and Watson (2002)**.
- We use principal component analysis (PCA) to compute diffusion indices from $\mathbf{X}_i^*$, $\mathbf{X}_i^{p*}$, and $\mathbf{X}_i^{n*}$. The eigendecomposition of the covariance matrix $\mathbf{S}$ of $\mathbf{X}$ is calculated as $\mathbf{S} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$. The principal components of $\mathbf{X}$ are obtained as: $\mathbf{Z} = \mathbf{X}\mathbf{V}$, where $\mathbf{Z}$ is a $n \times p$ matrix whose columns are the principal components.
- We construct our diffusion index by selecting the first column of $\mathbf{Z}$ that captures the most information contained in $\mathbf{X}$ among all the eigenvectors.
- Therefore, we construct the diffusion index using the most powerful eigenvector in $\mathbf{X}$ to deliver a straightforward and intuitive explanation.

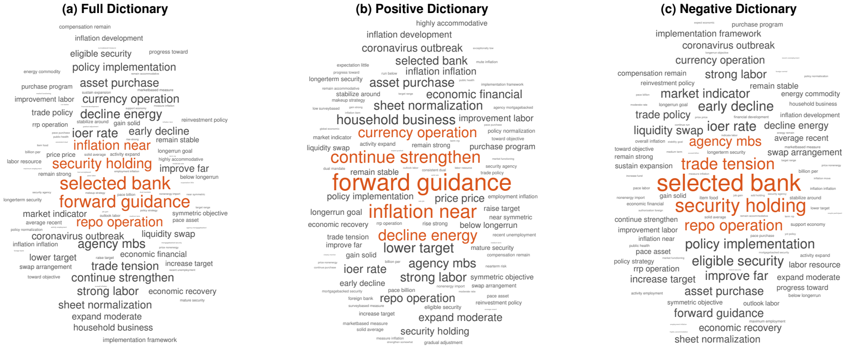
Data Description

- Combining the all the data, resulting in 65,360 observations covering 706 funds from January 2004 to August 2020

Statistic	X^{da}	X^{a}	X^{div}	X^{cash}	X^{lev}
Mean	15.6169	2154.6231	20.0776	2.2653	0.6140
Median	0.8857	417.0383	19.0000	1.4200	1.0000
Maximum	25574.3293	397113.2253	67.0000	76.6156	1.0000
Minimum	-45501.4995	0.9176	3.0000	-31.4244	0.0000
Std. Dev.	506.5555	10843.5866	10.2799	3.2956	0.4868
Skewness	-6.1218	22.4912	1.2184	3.3057	-0.4683
Kurtosis	1722.3238	653.4936	5.1627	37.8488	1.2193
JB Test	0.0010	0.0010	0.0010	0.0010	0.0010
ADF Test	0.0010	0.0010	0.0010	0.0010	0.0010

Statistic	$y^{(0)}$	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	-
Mean	1.8890	1.7574	0.1219	-0.0463	-
Median	0.0000	0.0000	0.0000	0.0000	-
Maximum	3617.2200	3453.3897	1689.1847	730.4047	-
Minimum	-3540.0562	-3037.0130	-1716.9785	-834.1834	-
Std. Dev.	60.2344	54.8042	15.6954	15.8125	-
Skewness	-0.0596	-0.0251	11.5165	-2.1646	-
Kurtosis	599.3891	580.6846	5183.4998	936.7828	-
JB Test	0.0010	0.0010	0.0010	0.0010	-
ADF Test	0.0010	0.0010	0.0010	0.0010	-

Wordclouds



(a) Full Dictionary

(b) Positive Dictionary

(c) Negative Dictionary

Figure: Wordclouds of the Full, Positive, and Negative Dictionaries under $y^{(0)}$

Empirical Results

Panel Regression Results

Table: Panel Regression with Textual Predictor

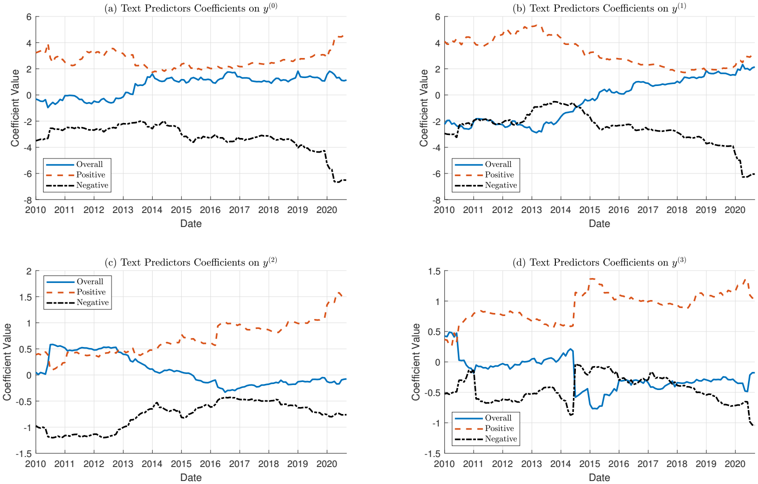
Variable	With Textual Predictor			
	$y^{(0)}$	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$
Panel A: Textual Predictor				
X^*	0.7856	1.0522**	-0.0604	-0.2928
	(0.6023)	(0.5153)	(0.1380)	(0.2352)
X^{p*}	3.6386***	2.8408***	1.0520***	0.9564***
	(0.4781)	(0.4283)	(0.2098)	(0.1818)
X^{n*}	-4.7656***	-4.2632***	-0.6307***	-0.5809**
	(0.5822)	(0.5383)	(0.0980)	(0.2379)
Panel B: Control Variable				
X^{da}	0.0218***	0.0228***	0.0012***	-0.0022***
	(0.0034)	(0.0034)	(0.0003)	(0.0006)
X^a	0.0002	0.0002	0.0000	-0.0000
	(0.0002)	(0.0002)	(0.0000)	(0.0001)
X^{div}	-0.0321	-0.0389	-0.0464*	0.0085
	(0.0909)	(0.0815)	(0.0259)	(0.0314)
X^{cash}	0.5448***	0.4990***	0.0562*	0.0521*
	(0.1365)	(0.1290)	(0.0305)	(0.0296)
X^{lev}	0.3777	0.1897	0.0731	0.1233
	((0.4547)	(0.4263)	(0.1187)	(0.1559)
Panel C: Model Description				
Individual F.E	Yes	Yes	Yes	Yes
Time F.E	Yes	Yes	Yes	Yes
Centered R^2	0.0737	0.0900	0.0073	0.0134

The Heterogeneous Effects

Table: Regression for Different Quartiles of the Sample

Variable	$y^{(0)}$	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$
Panel A: First Quartile				
X^*	0.0393	0.0749	0.0131	0.0928
	(0.0798)	(0.0636)	(0.0159)	(0.0703)
X^{p*}	0.2108***	0.1531***	0.0191**	-0.0376***
	(0.0501)	(0.0493)	(0.0093)	(0.0520)
X^{n*}	-0.2345***	-0.2271***	-0.0607***	-0.1096**
	(0.0606)	(0.0645)	(0.0224)	(0.0513)
Panel B: Second and Third Quartiles				
X^*	0.1049	0.4147*	0.0959	-0.0291
	(0.3154)	(0.2477)	(0.0839)	(0.1364)
X^{p*}	1.8672***	1.5602***	0.2344***	0.3696***
	(0.2471)	(0.1986)	(0.0619)	(0.1213)
X^{n*}	-1.9018***	-1.7215***	-0.3557***	-0.3659***
	(0.2329)	(0.2097)	(0.0572)	(0.0986)
Panel C: Fourth Quartile				
X^*	1.8943	0.6224	-0.2234	-0.5580
	(1.7558)	(1.6289)	(0.4369)	(0.6148)
X^{p*}	9.1058***	7.6636***	3.2061***	2.1594***
	(1.3849)	(1.3735)	(0.6893)	(0.4746)
X^{n*}	-12.8104***	-9.6859***	-1.8026***	-1.4686**
	(1.7338)	(1.5659)	(0.3373)	(0.7133)
Individual F.E	Yes	Yes	Yes	Yes
Time F.E	Yes	Yes	Yes	Yes

Dynamic Features



(a) Text Predictors Coefficients on $y^{(0)}$

(b) Text Predictors Coefficients on $y^{(1)}$

(c) Text Predictors Coefficients on $y^{(2)}$

(d) Text Predictors Coefficients on $y^{(3)}$

Figure: Reallocation Channels of Capital Inflow and Outflow in the US Market.

Regression: During, and After the 2008 Financial Crisis

Table: Fixed-Effect Regression Before, During, and After the 2008 Financial Crisis

Variable	$y^{(0)}$	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$
Panel A: Before Crisis				
X^*	-1.2151	-1.7773	0.2990	0.4981
	(1.8209)	(1.3477)	(0.3065)	(1.3495)
X_p^*	1.5271	2.0783	0.6093**	0.1069
	(2.5413)	(1.3287)	(0.2963)	(1.4525)
X_n^*	-2.3622**	-1.5450	-0.9339**	-1.2581***
	(1.1249)	(1.3470)	(0.4290)	(0.3988)
Panel B: During Crisis				
X^*	0.6532	-2.1301	-0.2213	1.0111
	(2.6211)	(2.9244)	(0.6585)	(1.2931)
X_p^*	7.6774**	8.4556**	0.1563	0.4252
	(3.7533)	(4.2385)	(0.4935)	(1.3047)
X_n^*	-4.6552***	-4.2008**	-1.3748**	-0.0298
	(1.7432)	(1.7926)	(0.7047)	(0.5336)
Panel C: After Crisis				
X^*	0.9766	1.2644**	-0.0960	-0.2620
	(0.6556)	(0.5498)	(0.1521)	(0.2465)
X_p^*	3.6325***	2.8176***	1.1228***	0.9479***
	(0.5116)	(0.4574)	(0.2305)	(0.1967)
X_n^*	-5.0603***	-4.4875***	-0.5982***	-0.6331***
	(0.6379)	(0.5879)	(0.1078)	(0.2440)

Portfolio Bias

Table: Fixed-effect within OECD and non-OECD

Variable	$y^{(0)}$	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$
Panel A: OECD				
X^*	0.8044	1.0796*	-0.1205	-0.3265
	(0.6028)	(0.5518)	(0.1460)	(0.2533)
X_p^*	3.7908***	2.9448***	1.1326***	1.0375***
	(0.4794)	(0.4517)	(0.2261)	(0.1942)
X_n^*	-4.8506***	-4.3318***	-0.5996***	-0.5894**
	(0.6247)	(0.5885)	(0.1020)	(0.2574)
Panel B: OECD without US				
X^*	0.2043	1.0418	0.0845	-0.1650
	(0.6744)	(0.6797)	(0.1908)	(0.1695)
X_p^*	2.9811***	1.9204***	0.6015***	0.8069***
	(0.4682)	(0.4832)	(0.1795)	(0.1764)
X_n^*	-3.5622***	-3.8233***	-0.5953***	-0.6437***
	(0.5482)	(0.5515)	(0.1390)	(0.1282)
Panel C: Non-OECD				
X^*	0.4027	-0.5806	0.7640**	0.2136
	(0.9860)	(0.7006)	(0.3178)	(0.2965)
X_p^*	-0.4086	0.5372	0.0687	-0.0592
	(0.7663)	(0.5617)	(0.1262)	(0.2834)
X_n^*	-1.9994***	-0.4663	-1.0830***	-0.6387***
	(0.6225)	(0.4976)	(0.3107)	(0.1972)

Out of **706** funds, **643**, or 91%, are domiciled in OECD countries. Excluding funds domiciled in the U.S., there are **363** remaining, or 51.4%. The rest **62** funds are domiciled in non-OECD countries, accounting for 8.5% of the total.

Conclusion

- In this study, we combine machine learning techniques with textual analysis to gain insights into the complex dynamics of international capital allocation and the role of government policy in shaping global financial markets.
- Our findings reveal that FOMC textual sentiments have a significant effect on cross-border capital flows, with larger funds being more sensitive to these announcements than smaller funds.
- We also found that attention to FOMC announcements changed dramatically during and after the 2008 financial crisis, suggesting that the surge of international capital flows after the crisis can be attributed to the increasing attention paid to FOMC announcements.
- Furthermore, our study contributes to the literature by examining the impact of federal policy announcements on investor diversification. Our results indicate that U.S. and non-OECD funds exhibit a strong home bias, while OECD funds, excluding the U.S., do not display such a bias in either direction.

Future Extension!

- Incorporate additional generative and interactive policy elements.
  - For example, the press conference! [switch to webpage]
  - Interaction
  - Verbal tone
  - Stronger sentiment
  - ... and more
- Note that expanding our content in these directions will necessitate the use of more sophisticated data analysis tools, such as the Language Model (LLM).

More Future Extension!

- Lima et al.'s LASSO-based framework may pose challenges.
- Analyzing each fund individually is suboptimal and relies on strong assumptions.
- To address this, consider exploring **simultaneous** textual analysis across all funds.
  - This approach introduces higher dimensionality to the response variable.
  - It also entails a more intricate LASSO-selection process.
- We have **identified a potential solution** to this challenge, which I will defer to our future presenter to elaborate on.