Ask EDGAR: Insights into Asset Management from Big Data and Natural Language Processing

Simona Abis and Anton Lines[*]

Abstract

Understanding capital allocation is key to understanding the functioning of market economies. We propose to use textual data from mutual fund mandatory disclosures stored in the SEC's EDGAR database to examine how investors parse and evaluate soft information when making investment decisions. Prior work has found low explanatory power when relating capital flows to hard information such as past returns and risk, but such information comprises only a small fraction of the information available to investors. Recent advances in machine learning and natural language processing should allow us to quantify some of this soft information, enabling a more detailed examination of investor decision-making. In addition, we can use this setting to analyze information processing subject to limited attention, as the amount of available text vastly exceeds the information processing capacity of most investors.

1 Motivation

The mutual fund industry has frequently been used as a laboratory to study capital allocation in financial markets due to its relative transparency and its central place in the financial decision-making of ordinary investors (e.g. Sirri and Tufano (1998, JF); Del Guercio and Tkac (2002, JFQA), Berk and van Binsbergen (2016, JFE); Berk, van Binsbergen, and Liu (2017, JF)). These papers typically regress capital flows on past returns (raw and risk-adjusted), performance ranks, fund return volatility, tracking error volatility, and past fund assets, reporting adjusted R-squares in the range of only 10%-20%. Funds with similar track records and starting assets can grow at significantly different rates after opening to investors.

However, potential fund investors (either directly or via financial advisors) also have access to a wealth of soft information, including marketing materials, regulatory disclosures, and even in-person meetings with fund managers. While we still cannot observe this full information set, textual information from prospectuses and other mandatory disclosures seems likely to be correlated with the ways in which fund managers talk about themselves in general. The importance of prospectuses is highlighted by the SEC, which urges investors to be weary of past performance and to read the prospectuses carefully (Rule 156). In 1998 the regulator published "A Plain English Handbook" urging managers to produce readable reports. Looking at a sample of fund prospectuses, Johnson (2004, JFRC) finds that their readability significantly improved after an SEC ruling on the topic (Rule 421) became effective. Philpot and Johnson (2006, JFSM) further document that mutual funds write the risk section of their prospectuses more clearly than their objectives section, and this is particularly true for poorly performing funds.

Given the costs associated with mandatory disclosure and its increasing emphasis by the SEC, it is crucial to understand whether disclosures contain meaningful information, and particularly whether their usage by investors leads to a more efficient or inefficient allocation of resources. Can we extract informative signals from the text of mutual fund prospectuses (and more generally from their communication with investors) above and beyond what we could learn from an analysis of hard information about these funds? Are investors following SEC guidelines and paying attention to this information in their decision-making process? Are they doing so in an efficient manner? Investors are constrained in their information processing capacity but are overloaded with textual information about different funds. If this text does indeed contain meaningful signals about fund quality or risk-taking behavior, investors would need to parse it efficiently before being able to use it in their decision-making process.

Hillert, Niessen-Ruenzi and Ruenz (2016) provide some evidence that mutual fund investors pay attention to the tone of shareholder letters, directing capital to those with a more personal writing style and a less negative tone.

[*] Columbia Business School, 3022 Broadway, New York, NY 10027. E-mail: simona.abis@columbia.edu, anton.lines@columbia.edu

They find managers' writing style to be correlated with performance and future risk-taking. By contrast, Spickers and Petersen (2016) analyze manager's descriptions of themselves on a social trading platform and show that managers who provide a more positive description of themselves attract less capital.

However, the existing literature on this topic is limited and constrained by the difficulties of working with a large and unstructured dataset of textual information. Existing work also does not make use of more advanced machine learning techniques that would enable us to answer more general questions about capital allocation and investor information processing. For example, one influential recent theory suggests that investors may choose fund managers on the basis of trust (Gennaioli, Shleifer, and Vishny (2015, JF)). This is a difficult theory to test empirically because "trust" is not easily measurable using quantitative data. However, using natural language processing, it should be possible to extract measures of trust from the language used in prospectuses. We also aim to uncover the extent to which fund managers are able to signal their type through textual disclosures. On the one hand, effective signaling could lead to an efficient allocation of resources. On the other hand, fund managers with poor investment skill might invest more effort in crafting the text of their disclosures in other to attract flows, creating an inefficient allocation of resources if investors respond favorably.

More generally, artificial intelligence can be used to predict the future performance of mutual funds using features extracted from the available textual information, creating a benchmark against which to evaluate whether investors make efficient use of the available information. This will also allow us to examine potential discrepancies between what investors consider important and what actually predicts fund performance.

2 Empirical Challenges

The first technical aim of this project is to compile a structured, categorized, and usable version of *all* textual information contained in mutual fund disclosures (available through the SEC online database, EDGAR) and use it to better understand investor choices. Ultimately, we plan to extend this effort to all regulatory disclosures (including corporate disclosures). The second technical aim is to apply state-of-the-art artificial intelligence techniques, which are only just beginning to see use in academic finance and economics, in order to understand the informational content of this large body of textual data.

Being able to draw meaningful conclusions from the SEC data presents a number of challenges.

First, we must map information in each mandatory disclosure to the particular regulatory question being addressed. For instance, different sections of prospectuses may be written by different individuals (legal firm, fund manager, fund family) and might serve different purposes. So any meaningful comparison must be conditioned on the section being analyzed. This is not a trivial problem given the size of the dataset and the lack of structure in historical disclosures. This phase of the project will involve automated web download, mapping of disclosures to existing fund identifiers, text cleaning and parsing followed by categorization and database storage. All of our analysis will be done in Python, which has become the standard language for machine learning applications. To parse and categorize the data, we will be using packages such as beautiful soup and pyparsing (one of us—see Abis (2018)—has already begun working on this problem). To perform this categorization for the entire EDGAR database, we will require large storage capacity and fast computing.

Once the data is categorized and cleaned, we plan to use both simple textual measures, such as the length and readability of different sections, as well as supervised and unsupervised learning in order to extract meaningful signals from the data. Some of the signals we would be able to extract are: complexity, transparency, novelty and tone of the text as well as trust-building or deceiving language. Supervised learning would allow us to categorize funds along a number of pre-specified dimensions such as whether they follow quantitative or discretionary strategies (see Abis (2018)). We could also use unsupervised learning techniques such as clustering to group funds according to similarity in strategy or risk descriptions. For this part of the project, we would use Python packages such as scikit-learn and nltk.

The final stage of the project would require relating these signals to fund performance and risk taking behavior first and to fund flows next. It has traditionally been difficult to measure fund flows as no structured dataset of inflows and outflows existed; for this reason most previous research has used changes in AUM. Detailed information about inflows and outflows, though, is available in the N-SAR filing (also available through EDGAR), which we plan to use. Further, to alliviate endogeneity concerns, we plan to link asset growth in the first year after inception to differences in the inital prospectus/marketing material and to study inflows and outflows following changes in prospectuses due to modifications in mandatory disclosure rules by the SEC.

Bibliography

- Abis, S. (2017). Man vs. Machine: Quantitative and Discretionary Equity Management. Working Paper.
- Berk, J. B., & Van Binsbergen, J. H. (2016). Assessing asset pricing models using revealed preference. Journal of Financial Economics, 119(1), 1-23.
- Berk, J. B., Van Binsbergen, J. H., & Liu, B. (2017). Matching capital and labor. The Journal of Finance, 72(6), 2467-2504.
- Del Guercio, D., & Tkac, P. A. (2002). The determinants of the flow of funds of managed portfolios: Mutual funds vs. pension funds. Journal of Financial and Quantitative Analysis, 37(4), 523-557.
- Gennaioli, N., Shleifer, A., & Vishny, R. (2015). Money doctors. The Journal of Finance, 70(1), 91- 114.
- Hillert, A., Niessen-Ruenzi, A., & Ruenzi, S. (2016). Mutual fund shareholder letters: flows, performance, and managerial behavior.
- Johnson, D. T. (2004). Has the Security and Exchange Commission's Rule 421 made mutual fund prospectuses more accessible?. Journal of Financial Regulation and Compliance, 12(1), 51-63.
- Philpot, J., & Johnson, D. T. (2007). Mutual fund performance and fund prospectus clarity. Journal of Financial Services Marketing, 11(3), 211-216.
- Sirri, E. R., & Tufano, P. (1998). Costly search and mutual fund flows. Journal of Finance, 53(5), 1589- 1622.
- Spickers, T., & Petersen, G. K. (2016). Too good to be true?–The influence of manager self-descriptions on investor behavior.