Information Production

Chris Hansman, Marcin Kacperczyk, Savitar Sundaresan, and Antoine Uettwiller

Proposal for the NBER Big Data Session

Abstract

Research typically measures the informational content of analyst reports on the basis of widely available ``hard'' information such as earnings forecasts and investment recommendations. We exploit a new, large, text-based data set of reports and show that the traditional measures miss out on important aspects of the information produced by analysts. We consider additional dimensions of informational content based on the complexity, sentiment, and size of reports as well as the composition, ordering, and location of the analysts authoring the reports. We examine the recent implementation of MiFID II in the EU as a structural change to the pricing of information, and show that considering only ``hard'' information provides an incomplete picture of the resources devoted to information production, and the quality of information produced.

Research Strategy and Discussion

Data: The novel element in our data is the complete text of approximately 1.7 million analyst reports from 2014 to 2018. These reports represent a wide range of U.S. and international brokerages and cover a broad set of equities. We have merged or are in the process of merging these data with a series of additional sources. This includes roughly half a million regulatory filings from the SEC covering the same time span, as well as more standard data sets including I/B/E/S and Compustat, which provide quantitative information on analyst forecasts and firms, respectively.

Access to the complete text of the reports allows us to widen the scope of hard information available to researchers. Scraping the reports provides a more thorough picture of complete *teams* of analysts collaborating on a report, as opposed to the single lead analyst available in typical data sets. We are also able to both precisely date reports and validate the forecasts provided by other data sources. Finally, we are able to uncover previously unavailable forecasts.

Perhaps more importantly, the text itself provides a valuable and underexplored window into the informational content of analyst reports. With

this text, we are able to examine the complexity, sentiment, length, and similarity of reports. Such measures may reflect both the effort necessary to produce a given report, and provide implicit information regarding the precision or potential bias of any quantitative predictions.

Strategy: A key challenge in our analysis comes in converting unstructured text data into useful measures of a given report's informational content. We have or are currently in the process of doing so in several ways. First, we have scraped, or are in the process of scraping the reports to extract all quantitative forecasts or predictions, and developed algorithms to determine the location and identity of all analysts involved in producing a report. Second, we have used standard dictionaries and indexes to convert unstructured text into structured data. These include the Loughran-McDonald dictionary—which yields the number of words classified as 'positive', 'negative', 'uncertain', 'litigious', 'constraining', 'superfluous', and 'interesting' in each report—as well as readability measures like the gunning fog index and the SMOG index, which estimate the years of formal education necessary to understand a text. Third, we have explicitly counted the number of characters, words, and sentences in each document, and are developing methods to count and quantify graphs and other non-text objects. Finally, we have generated a term frequency matrix, as well as term frequency-inverse document frequency to represent the corpus of words/n-grams present in the reports, which will allow us to develop and implement topic models.

To demonstrate the value of this information, we examine the change in informational content of reports—measured in the dimensions described above—surrounding the implementation of MiFID II, the EU regulation which requires sell-side firms to unbundle research (such as analyst reports) from services relating to the trade execution. We use event-study and difference-in-difference strategies to examine the impact of this structural change to the pricing of information on the hard and soft information contained within analyst reports.

Computing Languages Used:

The majority of our analysis is conducted in Python.