S&P Global
Market Intelligence

Q&A: Leveraging Natural Language Processing to Speed Read Company Filings

Frank Zhao
Senior Director
Quantamental Research
S&P Global Market Intelligence

Company financial filings (Form 10-K) have grown over time and are now equivalent in length to a 240-page novel. Even for analysts with modest coverage universes, thoroughly reading 10-Ks for relevant information can be extremely time consuming. S&P Global Market Intelligence's Machine Readable Filings enables the review process to be done systematically.

Our recent webinar entitled "Speed Reading 10-K & 10-Q Company Filings" walked investors and analysts through the steps required to systematically extract value from the textual content within filings. Joe Gits, President of Social Market Analytics, and Frank Zhao, Senior Director of Quantamental Research, S&P Global Market Intelligence, covered a wide range of topics, including:

- Reviewing how different formatting conventions of 10-Ks and 10-Qs are used to clean, reconcile, and standardize information for a machine-readable state.
- Leveraging a technique from linguistics and NLP called cosine similarity to identify year-over-year textual revisions in filings.
- Comparing the richness and the incremental information among the major sections.
- Describing how Quality and Trend Following (Momentum) relate to 10-K textual changes.
- Evaluating the impact of small versus large textual revisions on stock returns and volatilities.

In this blog, we share questions from the attendees and the responses from Frank Zhao, Senior Director of Quantamental Research, S&P Global Market Intelligence.

Question: If you are doing year-over-year comparisons, how do you deal with quarterly filings that refer to the previous quarter's filing? For example, comparing the 10-Q from Q2 2020 to Q2 2019, and the Q2 2019 says something about Q1 2019 filing and the risk factor. Wouldn't this impact your analysis?

Frank Zhao, S&P Global Market Intelligence: Having more apples-to-apples comparisons is especially important for the Risk Factors section. Given this, we constantly monitor the length of this section in the 10-Qs. When it is below a minimum threshold in terms of the number of words (i.e., 100 words from the Loughran-McDonald[1] Master Dictionary), we then bring in the content from the Risk Factors section in the latest 10-K preceding the 10-Q in question. The dictionary provides a means of determining which collections of characters are actual words, which is important for consistency in word counts. The dictionary also contains sentiment classifications, counts across all filings, and has other useful information about each word.

Question: Once you compute the cosine similarity, how do you build your long and short positions? Do you go long $1 in all the top 10% of non-changes and short $1 on the bottom 10% of changes?

Frank Zhao, S&P Global Market Intelligence: In the analysis that you have seen in our presentation, it is a quintile return spread. So we are betting on 20% of the filers with the greatest year-over-year textual similarity. Then we want to stay away from, or short, the 20% of filers with the least textual similarity. They [the two portfolios] are not dollar-neutral. The stocks in each of the portfolios are equally weighted.

[1] Loughran, T., and B. McDonald. "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks." Journal of Finance 66 (2011): 35-65.

What is also interesting about this category of textual similarity strategy is that the values (i.e., the scores) tend to be very clustered together. But what we have noticed is that even small year-over-year revisions in the text have implications for performance (i.e., returns and risks). And, as one makes bets on filers with more extreme changes, the returns actually go higher without adding more risk.

Question: How many holdings are there on either the long or short side?

Frank Zhao, S&P Global Market Intelligence: The universe we analyzed is pretty broad. It is based on the Russell 3000, and then we have approximately 90+% of historical coverage, so about 2,700 firms. We have about 470 to 480 firms in the portfolio of buys and the portfolio of sells.

Question: How do you handle the incremental words in a given section over time when calculating the similarity scores?

Frank Zhao, S&P Global Market Intelligence: Everything else being the same, cosine similarity actually tends to favor documents that are longer in length. So, if you refer to the slide with the cosine similarity construction, one of the things that we do, besides calculating the term frequency-inverse document frequency, is to normalize by length. We take the Euclidean normal, such that when we are measuring two texts, we normalize and make them of equal length.

Question: Does your analysis have a size bias, going long on the small firms and short on the big ones?

Frank Zhao, S&P Global Market Intelligence: The results that we show in the presentation are based on the Russell 3000. If you decompose the results down to the Russell 1000, a proxy for the larger caps, and the Russell 2000, a proxy for the smaller caps, the inferences do not change. It is true that within the Russell 1000 there is some attenuation in performance.

What is also interesting is that the year-over-year textual similarity signal in the MD&A section tends to work a bit better among the smaller cap stocks whereas in the Risk Factors section tends to work better among the larger cap stocks.

Question: What was the trading frequency of the cosine similarity strategy?

Frank Zhao, S&P Global Market Intelligence: [The trading frequency] of all the analysis that you see is monthly periodicity. We rebalance at every calendar month end with a look-back window of four months.

Question: How do you prevent overfitting from your NLP process?

Frank Zhao, S&P Global Market Intelligence: We don't have a training period. We don't go through an exhaustive amount of analytics and then correlate the scores from each to forward returns to see which one actually shows significance. The issue with this is that a signal does extremely well in periods where it has seen the data, but once new and previously unseen data comes in, the signal decays very quickly to noise.

So, what we do with our work is that we try to come up with economic hypotheses beforehand, whether from literature or from our domain knowledge in equity investing, and try to think about why a particular signal may be priced by market participants through two lenses. One lens is risk-based where investors are getting compensated with extra returns for taking on additional risk. The other, which is the crux of this presentation, is rooted in behavioral finance. Behavioral biases that we humans have and the signals try to exploit those biases.

In this case, it is the sheer size of the filings that investors have trouble reading and processing. And then the changes are subtle, which may be hard to pick up without the help of computers.

Explore the research and datasets behind the webinar

U.S. Filings: No News is Good News

Investors have historically overlooked the implications of year-over-year textual revisions in the newest corporate filings due to the voluminous amount of text and the small amount of changes. This Quantamental Research publication using U.S. corporate filings data finds that filers with the greatest year-over-year textual similarity outperformed those with the least similarity up to 5% per year after considering commonly used stock selection and risk strategies. **Read Research >**

Hiding in Plain Sight — Risks That Are Overlooked

10-Q & 10-K filings are a vital source of information, but their length and complexity may result in investors overlooking important details. In this report, the Quantamental Research group uses Natural Language Processing to identify companies that have made significant changes to the "Risk Factors" section of their corporate filings. **Read Research >**

U.S. Machine Readable Filings

The Machine Readable Filings feed provides all of the textual portions of public filings, broken down into the various sections identified by the company, with extraneous information such as page numbers, images, and tables removed.

The data is delivered in a structured format through Xpressfeed™ and Snowflake. This allows customers to perform Natural Language Processing (NLP) against this set of data without having to do the document cleanup and structuring themselves. **Learn More >**

Global Machine Readable Filings

The Global Machine Readable Filings feed expands on the U.S. Machine Readable Filings feed by providing parsed text for non-US annual and interim reports. **Learn More >**

Subscribe to Quantamental Research

The S&P Global Quantamental Research group leverages the uniqueness and depth of S&P Global's data and analytics offerings to uncover **actionable insights** for investors and decision makers across a variety of business types. **Subscribe Now >**

Explore all our datasets and solutions on the S&P Global Marketplace.

S&P Global
Market Intelligence

Contact Us

Asia-Pacific
+852-2533-3565

Europe, Middle East & Africa
+44-207-176-1234

The Americas
+1 877 863 1306

These materials have been prepared solely for information purposes based upon information generally available to the public and from sources believed to be reliable. No content (including index data, ratings, credit-related analyses and data, research, model, software or other application or output therefrom) or any part thereof (Content) may be modified, reverse engineered, reproduced or distributed in any form by any means, or stored in a database or retrieval system, without the prior written permission of S&P Global Market Intelligence or its affiliates (collectively, S&P Global). The Content shall not be used for any unlawful or unauthorized purposes. S&P Global and any third-party providers, (collectively S&P Global Parties) do not guarantee the accuracy, completeness, timeliness or availability of the Content. S&P Global Parties are not responsible for any errors or omissions, regardless of the cause, for the results obtained from the use of the Content. THE CONTENT IS PROVIDED ON “AS IS” BASIS. S&P GLOBAL PARTIES DISCLAIM ANY AND ALL EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, ANY WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE OR USE, FREEDOM FROM BUGS, SOFTWARE ERRORS OR DEFECTS, THAT THE CONTENT’S FUNCTIONING WILL BE UNINTERRUPTED OR THAT THE CONTENT WILL OPERATE WITH ANY SOFTWARE OR HARDWARE CONFIGURATION. In no event shall S&P Global Parties be liable to any party for any direct, indirect, incidental, exemplary, compensatory, punitive, special or consequential damages, costs, expenses, legal fees, or losses (including, without limitation, lost income or lost profits and opportunity costs or losses caused by negligence) in connection with any use of the Content even if advised of the possibility of such damages.

S&P Global Market Intelligence’s opinions, quotes and credit-related and other analyses are statements of opinion as of the date they are expressed and not statements of fact or recommendations to purchase, hold, or sell any securities or to make any investment decisions, and do not address the suitability of any security. S&P Global Market Intelligence may provide index data. Direct investment in an index is not possible. Exposure to an asset class represented by an index is available through investable instruments based on that index. S&P Global Market Intelligence assumes no obligation to update the Content following publication in any form or format. The Content should not be relied on and is not a substitute for the skill, judgment and experience of the user, its management, employees, advisors and/or clients when making investment and other business decisions. S&P Global keeps certain activities of its divisions separate from each other in order to preserve the independence and objectivity of their respective activities. As a result, certain divisions of S&P Global may have information that is not available to other S&P Global divisions. S&P Global has established policies and procedures to maintain the confidentiality of certain non-public information received in connection with each analytical process.

S&P Global may receive compensation for its ratings and certain analyses, normally from issuers or underwriters of securities or from obligors. S&P Global reserves the right to disseminate its opinions and analyses. S&P Global's public ratings and analyses are made available on its Web sites, www.standardandpoors.com (free of charge) and www.ratingsdirect.com (subscription), and may be distributed through other means, including via S&P Global publications and third-party redistributors. Additional information about our ratings fees is available at www.standardandpoors.com/usratingsfees.