

# Building a Simple Investment Strategy with Machine Readable Filings



**Contributors:**

Ruchika Mangla  
Guillermo Ruiz-Rico, CFA

# Table of Contents

Highlights .....1

Context of Strategy.....1

Machine Readable Filings.....1

A Simple Investment Strategy..... 2

Portfolio Results ..... 3

Summary..... 4

References..... 5

## Highlights

- An overview of Machine Readable Filings (MRF) and the tools necessary to extract insights are provided through a simple use case aiming to find excess returns.
- 10-K and 10-Q forms from the MRF dataset are leveraged to create a sentiment-based signal emanating from the metric's quarterly changes.
- Using this signal, a strategy is implemented on the S&P 1500 and S&P 500 from 2007 through June 2021. The strategy, which leverages the constituents from the top quintile of each benchmark based on the values of two common filings sections, outperformed relative to each universe: a total return of 461% (Management's Discussion section) and 564% (Risk Factors section) vs 312% for S&P 1500. 314% (Management's Discussion section) and 316% (Risk Factors section) vs 309% for S&P 500.

## Context of Strategy

Following S&P Global Market Intelligence release of the new Machine Readable Filings (MRF) dataset in 2020, several papers<sup>12</sup> were published by the Quantamental Research team. Said papers zoomed in on the strong relationship between historical consistency in commonly shared sections of forms 10-K and 10-Q, and companies' performance as measured by their historical stock prices.

From the perspective of an additional use case, the focus of this exercise is to showcase an example of an investment management workflow wherein sentiment analysis techniques are applied to the 10-K and 10-Q filings in a manner analogous to the already available Textual Data Analytics product. To that purpose, a simple historical strategy simulation is run with the goal of attempting to generate excess returns relative to the S&P 1500 and S&P 500 benchmarks between 2007 and June 2021.

## Machine Readable Filings

The dataset is comprised of textual data sourced from publicly traded companies' filings to the U.S. Securities and Exchange Commission (SEC) with a starting date of 2006. Its value-add proposition is to facilitate the parsing of textual data by providing:

- Metadata specific to each filing.
- Structured raw text as classified in the underlying companies' filings sections.
- Omission of non-textual information, such as tables and images.
- Text-cleansing to enable NLP use cases. Key examples of these steps are maintaining consistency across reporting periods regardless of structure changes, reclassifying of headings sections for standardization purposes, and removing irrelevant elements such as table headers and page numbers.

<sup>1</sup> Yang, Z., and Oyeniyi, T. "Hiding in Plain Sight — Risks That Are Overlooked." S&P Global Market Intelligence Quantamental Research March 2021.

<sup>2</sup> Zhao, F. "U.S. Filings: No News is Good News, Textual Consistency in Corporate Filings Signals Outperformance" S&P Global Market Intelligence Quantamental Research May 2021.

# A Simple Investment Strategy

What does it entail to leverage Machine Readable Filings (MRF) in an investment management workflow?

Natural language Processing<sup>3</sup> (NLP) is leveraged to transform and quantify the MRF dataset, wherein the Loughran-McDonald Master Dictionary is chosen to support this paper's sentiment-based analysis<sup>4</sup>. It should be noted that this is also the dictionary used in the Textual Data Analytics (TDA) product built using the Machine Readable Transcripts dataset.

In this example, the positive-to-negative metric, one of many within the bag-of-words category of sentiment-based signals is selected. Though the suite of tools<sup>5</sup> used to perform these analyses permits to assess the efficacy of several dozens of metrics, for simplification purposes the focus in this document is solely on the positive-to-negative metric<sup>6</sup>.

As the initial step, the metric's coverage is reviewed for the available common document sections as measured by quarterly averages per year using the S&P 1500 as the universe<sup>7</sup>.

While the Management Discussion & Analysis (MD&A) and the Risk Factors (RF) sections are selected, the Quantitative and Qualitative Disclosures about Market Risk (Q&QDMR) section is excluded given the limited section's coverage.

Using the positive-to-negative scores for the aforementioned sections, a logic is built to calculate companies' quarter-over-quarter percentage changes on the positive-to-negative scores. Moreover, to avoid look-ahead bias, we ensure that the associated sentiment scores are as of companies' filing dates.

Next, utilizing the positive-to-negative signals, the following strategy is simulated on both the S&P 1500 and S&P 500 for each section (MD&A and RF) separately:

- Long only
- Quarterly rebalancing i.e. rebalancing at the end of each calendar quarter
- Equally weighted holdings
- Invest on companies in the top quintile based on the quarter-over-quarter positive-to-negative score change
- A company must remain on the top quintile to be held in subsequent quarter(s)

<sup>3</sup> A comprehensive demystifying exercise on what NLP is and how it can be used is carried out at Zhao, F. "Natural Language Processing — Part I: Primer" S&P Global Market Intelligence Quantamental Research, September 2017.

<sup>4</sup> As described in S&P Global Market Intelligence TDA User Guide, the Loughran-McDonald Master Dictionary is the de facto financial dictionary for NLP analysis due to its accessibility, its comprehensiveness, its financial-specific context, its lack of dependency on transitory words, and lastly, its less ambiguous and singularly connoted words.

<sup>5</sup> For production purposes, scores for sentiment-based signals were generated using a combination of Alteryx and Python, which is available upon request. Alternatively, sample code using open-source software is also available for execution and manipulation via S&P Global Marketplace Workbench. Sample code addresses the selected metric and should be a close approximation, if not identical. See references for details.

<sup>6</sup> The definition of the Positive-to-Negative is as follows: If NEGATIVE > 0, then [POSITIVE] / [NEGATIVE] else -1.

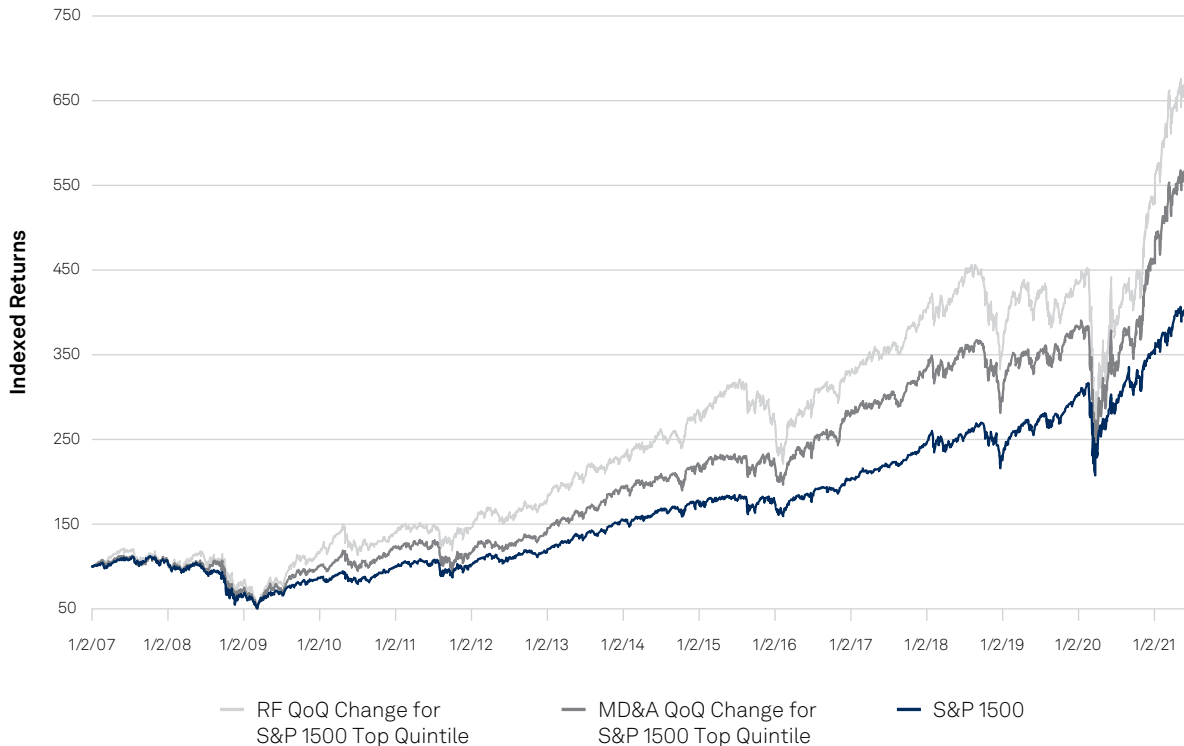
<sup>7</sup> See references for values. Values applicable based on dataset, universe, and embedded process. Analyzed sections were limited to the score availability through the production solution.

# Portfolio Results

The simulated strategies for period 2007 through June 2021 outperform the S&P 1500 as a benchmark:

- The Management Discussion and Assessment section returns 461.57% in comparison to the S&P 1500 total return of 311.79%. Sharpe ratio of this strategy portfolio is superior with a 0.51 compared to the benchmark portfolio 0.49<sup>89</sup>.
- The Risk Factors section returns of 564.41% in comparison to the S&P 1500 311.79% previously mentioned. Sharpe ratio of 0.52 compared to the benchmark's 0.49<sup>101</sup>.

## MRF Pos-To-Neg Strategy vs S&P 1500



Though with smaller excess returns, the proposed strategy also outperforms S&P 500 when executing it using the quintile groups derived from the S&P 500 universe. In this case:

- The Management Discussion and Assessment section returns 314.30% in comparison to the S&P 500 total return of 309.48% and with a Sharpe ratio of 0.43 compared to the benchmark portfolio 0.496<sup>1213</sup>.
- The Risk Factors section returns 316.56% in comparison to the S&P 500 309.48% whereas the Sharpe ratio is 0.40 compared to the benchmark's 0.496<sup>1415</sup>.

<sup>8</sup> MD&A of S&P 1500: Annual rate of return is 12.64%, vs S&P 1500 10.25%.

<sup>9</sup> MD&A of S&P 1500: Realized standard deviation is 24.80%, vs S&P 1500 20.78%.

<sup>10</sup> RF of S&P 1500: Annual rate of return is 13.95%, vs S&P 1500 10.25%.

<sup>11</sup> RF of S&P 1500: Realized standard deviation is 26.75%, vs S&P 1500 20.78%.

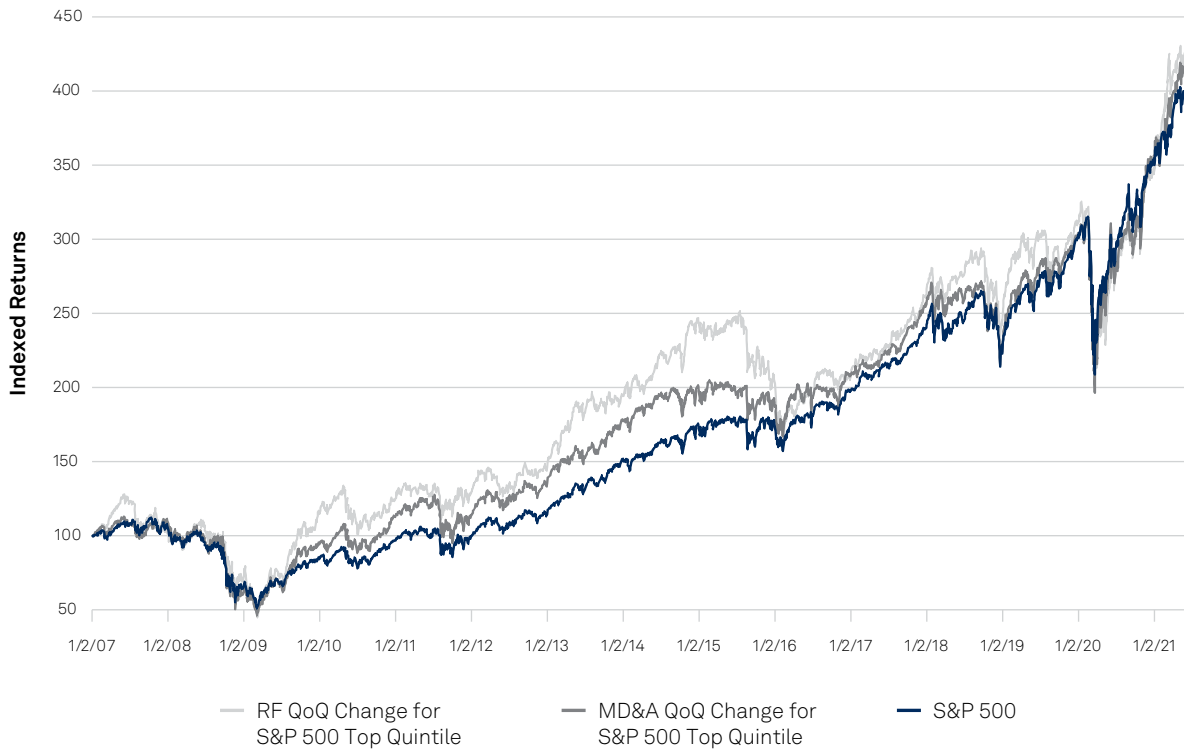
<sup>12</sup> MD&A of S&P 500: Annual rate of return is 10.30%, vs S&P 500 10.21%.

<sup>13</sup> MD&A of S&P 500: Realized standard deviation is 23.93%, vs S&P 500 20.59%.

<sup>14</sup> RF of S&P 500: Annual rate of return is 10.34%, vs S&P 500 10.21%.

<sup>15</sup> RF of S&P 500: Realized standard deviation is 25.75%, vs S&P 500 20.59%.

## MRF Pos-To-Neg Strategy vs S&P 500



## Summary

This document described the steps to leverage the Machine Readable Filings dataset for creating a signal to guide a simple investment strategy.

The exposed use case was purposefully simple and of narrow focus but revealed the ability of generating excess returns.

Readers may attempt replicating this strategy. The implementation addressing the NLP requirements to produce the selected metric used in this document is available upon request. Alternatively, a sample and customizable open-source code may be explored as a starting point. Said code is available via the [S&P Global Marketplace Workbench](#).

Overall, the ever-expanding NLP dictionaries and techniques, coupled with MRF's large number of investment management applications, makes this dataset a valuable proposition to practitioners, either as the sole driver or to complement existing strategies.

# References

- Coverage for U.S. MRF using S&P 1500 yearly averages of quarterly filings as of fiscal end-period for 10-K and 10-Q common sections

## **MRF Positive-to-Negative Score Coverage**

<b>Year</b>	<b>MD&amp;A</b>	<b>RF</b>	<b>Q&amp;QDMR</b>
2007	32%	15%	9%
2008	32%	18%	10%
2009	33%	19%	11%
2010	44%	27%	15%
2011	58%	33%	20%
2012	61%	34%	22%
2013	62%	34%	23%
2014	65%	36%	24%
2015	68%	38%	26%
2016	73%	43%	29%
2017	80%	45%	31%
2018	84%	49%	34%
2019	88%	51%	35%
2020	93%	82%	38%

- Data
  - [S&P Global Market Intelligence Machine Readable Filings](#)
  - [S&P Global Compustat®](#)
- Reference Documentation
  - S&P Global Market Intelligence TDA User Guide
  - S&P Global Market Intelligence Machine Readable Filings
- Python Sample Code
  - Refer to Workbench Notebook
- Tools
  - Alteryx Designer
  - Python
  - [S&P Global Market Intelligence ClariFI®](#)
  - [S&P Global Marketplace Workbench](#)

## About S&P Global Market Intelligence

At S&P Global Market Intelligence, we understand the importance of accurate, deep and insightful information. We integrate financial and industry data, research and news into tools that help track performance, generate alpha, identify investment ideas, perform valuations and assess credit risk. Investment professionals, government agencies, corporations and universities around the world use this essential intelligence to make business and financial decisions with conviction.

S&P Global Market Intelligence is a division of S&P Global (NYSE: SPGI), the world's foremost provider of credit ratings, benchmarks and analytics in the global capital and commodity markets, offering ESG solutions, deep data and insights on critical business factors. S&P Global has been providing essential intelligence that unlocks opportunity, fosters growth and accelerates progress for more than 160 years. For more information, visit [www.spglobal.com/marketintelligence](http://www.spglobal.com/marketintelligence).

### CONTACT US

#### The Americas

+1-877-863-1306

#### Europe, Middle East & Africa

+44-20-7176-1234

#### Asia-Pacific

+852-2533-3565

Copyright © 2021 S&P Global Market Intelligence Inc. All rights reserved. No content, including by framing or similar means, may be reproduced or distributed without the prior written permission of S&P Global Market Intelligence or its affiliates. The content is provided on an "as is" basis.