U.S. Filings: No News is Good News

Textual Consistency in Corporate Filings Signals Outperformance

Authors

Frank Zhao
617-530-8107
fzhao@spglobal.com

Company annual filings are a vital but often under-analyzed source of information for investors. Market moving content is buried within an ever-growing body of text that on average is equivalent to a 240-page novel. The filings contain subtle revisions making a computational linguistic approach imperative. **Faced with this voluminous amount of text and the minute number of changes, investors have historically overlooked the newly embedded information and the implications of those additions.**

This paper extends the first Quantamental Research paper[1] on corporate filings, which considered the *Risk Factors* section by exploring the five commonly shared sections between Form 10-K and Form 10-Q.[2] Key insights for the U.S. market include:

Consistency Matters and the Market Rewards It: Firms with the greatest year-over-year textual similarity outperformed those with the least similarity by 4.18% in the *MD&A* section and by 5.26% in the *Risk Factors* section annually after accounting for commonly used risk strategies (Exhibit 1).

Consistency in the Auxiliary Sections Matter Too: Firms with fewest changes in the *Controls & Procedures, Legal Proceedings and Quantitative & Qualitative Disclosures about Market Risk* outperformed those with the least similarity by 3.09% annually after accounting for commonly used risk strategies (Exhibit 2).

Supplementing the textual similarity scores from the Risk Factors section with those from the *Quantitative & Qualitative Disclosures about Market Risks* and *Controls & Procedures* sections improved the historical performance of the active long-side of the strategy to 2.37% from 1.75% annually with a better risk-and-reward profile (Exhibit 2).

Filers with the Most Consistency Exhibit Positive Earnings Surprises and Momentum: Firms with the most (least) similarity in the *MD&A and Risk Factors* sections were historically those that have outperformed (underperformed) over the past 12-months; have experienced positive (negative) earnings surprises; and have their growth prospects revised upwards (downwards) by sell-side analysts. (Exhibit 3).

Even Small Textual Inconsistency Matters for Performance and Volatility: Strategies that purchased firms with extreme similarity and sold those with extreme dissimilarity had higher cumulative dollar growth historically without taking on additional risk (Exhibit 5).

[1] See Yang and Oyeniyi (2021) in the reference section.
[2] The five sections are: i) Management Discussion & Analysis (MD&A), ii) Risk Factors, iii) Controls & Procedures, iv) Legal Proceedings and v) Quantitative & Qualitative Disclosures about Market Risks. See appendix A.1.

1. Introduction

Form 10-K and Form 10-Q[3] are financial reports filed annually and quarterly by publicly traded firms in the U.S. as mandated by the Securities Exchange Commission (SEC). The filings detail a company's latest financial information and operations in great length. In the rest of this section, an overview is provided of the layout of the major sections accompanied with their historical descriptive statistics.

Overview of the Layout and the Major Sections of Filings:

Every Form 10-K is structured similarly with 4 major parts and 15 items. Form 10-Q has 2 major parts and 10 items.[4] In the context of this brief, we examine the five commonly shared, major items between the two forms: i) Management Discussion & Analysis (MD&A) ii) Risk Factors iii) Controls & Procedures, iv) Legal Proceedings and v) Quantitative & Qualitative Disclosures about Market Risks. See appendix A.1 for more details.

Descriptive Statistics of the Items:

The two sections that are the longest are the *MD&A* and the *Risk Factors* section with historical averages of 10,000 and 8,000 words, respectively. The average length[5] of all but the *Risk Factors* section have been broadly unchanged since 2006[6]. The length of the *Risk Factors* section, however, has increased almost 3-fold to 11,000+ words presently. It appears that firms append incremental disclosures to this section over time as new risks surface. See appendix A.7.

The following sections discuss the construction of the textual similarity strategy, the intuition behind it, and the empirical results.

2. Signal Construction & Intuition

There are a number of documented approaches[7] from literature in the domain of linguistics and natural language processing to measure textual similarity. This paper uses cosine similarity.

Cosine Similarity

Cosine similarity quantifies the geometric angle between two texts using their numerical representation in an n-dimensional Euclidean space where the n denotes the number of unique words.[8] The range of the scores in our context is between 0 and 1 where the first value

[3] An abbreviated and unaudited version of Form 10-K,
[4] https://www.sec.gov/fast-answers/answersreada10khtm.html
[5] Using the number of words as a proxy
[6] The US Filings data set starts in 2006; there are tentative plans of extending back to 1994 in Q1 2022
[7] Others are Jaccard similarity, minimum edit distance and simple similarity. See Yang and Oyeniyi (2021) and Cohen, Malloy and Nguyen (2020).
[8] See Zhao (2020)

denotes the two texts as diametrically opposite and the latter denotes the two texts as the same.[9]

Inputs of Cosine Similarity

Each value in the two numerical vectors represents a word score. In this context, the analysis employs a concept from information retrieval (e.g., internet searches), *term frequency-inverse document frequency* (TF-IDF) to generate the word-level scores. See appendix A.3 for details on the construction, its intuition and the advantages of its selection.

3. Empirical Results

Textual similarity in the context of this paper measures the year-over-year similarity of a section between two adjacent filings for a firm to account for seasonality.[10]

The textual similarity investment strategy buys a portfolio of filers with the greatest year-over-year textual similarity and sells a portfolio of filers with the least similarity.[11] **The ex-ante hypothesis is that the path of least resistance is for filers to maintain much of the structure and content in their filings year-over-year. Larger textual changes occur when a filer needs to justify financial softness or fulfill its legal or fiduciary responsibilities, which is viewed negatively in our narrative.** In Section 4, we explore empirically why filers with the most (least) textual similarity outperform (underperform) historically.

3.1 Historical Performance in the MD&A and Risk Factors Sections

Historically, firms with the greatest textual similarity in the *MD&A and the Risk Factors* sections outperformed those with the least similarity by 4.18% and 5.26% per year since 2008, respectively, with significance at the 1% level (Exhibit 1). The results suggest that the level of textual similarity in both sections has historically shown the ability to differentiate between future winners and losers in the cross-section at the one-month forward horizon.

The second noteworthy observation is that the long-sides of these strategies contributed meaningfully and significantly to the overall long-short strategies (compare values in column [6] to column [9] in Exhibit 1). This is highly desirable since there are impediments against shorting stocks.

3.2 Auxiliary Sections Have Supplementary Information

Standalone textual similarity strategies in the *Controls & Procedures, Legal Proceedings* or *Quantitative & Qualitative Disclosures about Market Risk* showed promise, but their signal

[9] The lower bound for cosine similarity scores for this paper is zero because we are working with non-negative numbers (i.e., the numerator of the score has frequency of words, which is non-negative). The theoretical range of cosine similarity scores is between -1 and 1.
[10] For example, a cosine similarity score is generated for the *MD&A* section from two adjacent years of a filer.
[11] See appendix A.4 for details on portfolio construction and our analysis framework.

scores did not exhibit enough dispersion in the cross-section. That is because many filers have the textual similarity score of 1 as filers leverage identical text from the previous filing.

Exhibit 1: Textual Similarity Strategy in the MD&A and Risk Factors Sections
Russell 3000; January 2008 – December 2020[12]

		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
		Signal Sort Order	Strategy Start Date	Average Firm Count in Each Quintile Bin	Average Monthly Spearman Correlation	Hit Rate Monthly Spearman Correlation	Annualized Average Monthly Long - Market	Annualized Monthly IR (Long - Market)	Hit Rate Average Monthly Long - Market	Annualized Average Monthly Long - Short	Annualized Monthly IR (Long - Short)	Hit Rate Average Monthly Long - Short
[1]	Cosine Similarity - MD&A	Desc	200801	476	0.010	58.3%	1.77%	0.68	62.2%	4.18%	1.09	66.7%
[2]	p-value	NaN	NaN	NaN	0.004	0.030	0.018	NaN	0.003	0.000	NaN	0.000
[3]	Cosine Similarity - Risk Factors	Desc	200801	448	0.012	65.4%	1.75%	0.67	55.1%	5.26%	1.48	67.3%
[4]	p-value	NaN	NaN	NaN	0.000	0.000	0.004	NaN	0.230	0.000	NaN	0.000

Note: Returns are Carhart Four-Factor Adjusted. Values that are shaded in green (red) are statistically significant at least at the 10% level and are consistent (inconsistent) with our ex-ante hypothesis. Desc = descending sort and Asc = ascending sort. Source: S&P Global Market Intelligence Quantamental Research. All returns and indices are unmanaged, statistical composites and their returns do not include payment of any sales charges or fees an investor would pay to purchase the securities they represent. Such costs would lower performance. It is not possible to invest directly in an index. Past performance is not a guarantee of future results. Data as of 03/01/2021.

One potential use for those scores is to supplement the ones from the *MD&A and the Risk Factors* sections to differentiate further between the filers that have maintained consistency.[13] For example, supplementing the textual similarity scores from the *Risk Factors* section with those from the *Quantitative & Qualitative Disclosures about Market Risks and Controls & Procedures* sections improved the economic performance of the active long side to 2.37% from 1.75% per year. The annualized information ratio[14], a measure of economic performance after taking into account risk, improved to 0.90 from 0.67 (Exhibit 2 row [1] column [7]).

The second potential use is to combine the textual similarity scores from the three auxiliary sections jointly in a composite score with equal weighting. We see historically that firms with the most similarity in the three sections outperformed the market benchmark and the firms with the least similarity by 1.68% and 3.09% per year, respectively, with significance at the 5% level (row [3] in Exhibit 2). While the economic performance was meaningful in the portfolios containing filers with the most extreme textual similarity scores, the composite signal's ability to differentiate between the winners and the losers in the cross-section was not significant, however (row [3] column [4] in Exhibit 2).[15]

[12] See coverage results in the appendix A.7.
[13] Creating a composite score by combining scores from different sections in an equal-weighting scheme.
[14] Information ratio measures the economic performance of a strategy after taking into its risk. It is calculated as the average return from a long-short equity strategy divided by the volatility of those returns.
[15] This reinforces the notion that the texts in these sections rarely change. When they do, the revisions are market moving.

Exhibit 2: Textual Similarity Strategy in the Controls & Procedures, Q and Q[16]
Disclosures about Market Risk, Legal Proceedings
Russell 3000; January 2008 – December 2020[17]

		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
		Signal Sort Order	Strategy Start Date	Average Firm Count in Each Quintile Bin	Average Monthly Spearman Correlation	Hit Rate Monthly Spearman Correlation	Annualized Average Monthly Long - Market	Annualized Monthly IR (Long - Market)	Hit Rate Average Monthly Long - Market	Annualized Average Monthly Long - Short	Annualized Monthly IR (Long - Short)	Hit Rate Average Monthly Long - Short
[1]	Cosine Similarity - Composite of Risk Factors, QoQ and Ctrls	Desc	200801	478	0.008	61.5%	2.37%	0.90	62.2%	4.86%	1.33	67.9%
[2]	p-value	NaN	NaN	NaN	0.002	0.003	0.002	NaN	0.003	0.000	NaN	0.000
[3]	Cosine Similarity - Composite of QoQ, Ctrls and Legal	Desc	200801	480	0.004	55.8%	1.68%	0.61	60.3%	3.09%	0.78	62.2%
[4]	p-value	NaN	NaN	NaN	0.193	0.128	0.036	NaN	0.013	0.007	NaN	0.003

Note: Returns are Carhart Four-Factor Adjusted. Values that are shaded in green (red) are statistically significant at least at the 10% level and are consistent (inconsistent) with our ex-ante hypothesis. D = descending sort and A = ascending sort. Source: S&P Global Market Intelligence Quantamental Research. All returns and indices are unmanaged, statistical composites and their returns do not include payment of any sales charges or fees an investor would pay to purchase the securities they represent. Such costs would lower performance. It is not possible to invest directly in an index. Past performance is not a guarantee of future results. Data as of 03/01/2021.

4. Textual Similarity & Filers' Financial Characteristics

Why did firms with the most (least) textual similarity in the *MD&A* or the *Risk Factors* section historically outperform (underperform)? It is as likely that firms with the most pronounced textual revisions outperform. One avenue of exploration is to examine the correlation dynamics between the textual similarity strategy and those that are documented in empirical literature and have a direct anchor to information from the filers' financials, the market or the active-covering analysts. Examples include cheaper valuation, earnings surprise, more productive assets and analyst bullishness/bearishness among others.

Exhibit 3 shows ten strategies. The first two rows ([1] and [2]) are textual similarity strategies in the *MD&A* and *Risk Factors* sections. The remaining eight strategies (rows [3] – [10]) are the: [3] low-beta anomaly (BETA), [4] size anomaly (SMB), [5] relative valuation anomaly (HML), [6] price momentum anomaly (PMOM), [7] asset growth anomaly (AG), [8] gross profitability anomaly (GPA), [9] analyst revision anomaly (ANA) and [10] earnings surprise anomaly (SUE).

Values in Exhibit 3 are pairwise average monthly correlations between the two strategies in question.[18] Historically, the two textual similarity strategies have strong positive correlation to the low beta strategy; the price momentum strategy; the analyst revision strategy; the earnings surprise strategy and have strong negative correlation to the small-cap and relative valuation strategies.

[16] Quantitative and Qualitative Disclosures about Market Risk
[17] See coverage results in the appendix A.7.
[18] When the sign is positive (negative), it signifies the two strategies historically have moved in the same (opposite) direction. The magnitude signifies the strength of the moves where 1 suggests they move in perfect unison.

The correlations suggest that firms with the most (least) year-over-year textual similarity are firms that: are larger-(smaller-) caps; have been winners (losers) in the past 12-month; have experienced positive (negative) earnings surprises; and are positively (negatively) viewed by the sell-side analysts. The results reinforce the hypothesis that the path of least resistance is for filers to retain much of the structure and the content in their newest filings when the underlying business is strong and growing. When the underlying business exhibits weakness, the filers need to explain away the softness. One way this manifests is via textual amendments and additions.

Exhibit 3: Correlations of Long-Short Quintile Return Spreads
Russell 3000; January 2008 – December 2020[19,20]

	Signal Name	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
[1]	Cosine Similarity - MD&A	0.67	-0.40	-0.12	0.43	0.07	-0.24	0.39	0.29
[2]	Cosine Similarity - Risk Factors	0.62	-0.39	-0.29	0.46	0.02	0.02	0.36	0.32
[3]	Low Beta Anomaly [Low - High]		-0.68	-0.59	0.72	-0.29	-0.02	0.62	0.45
[4]	Small Cap Effect [Low - High]			0.74	-0.74	0.43	-0.30	-0.72	-0.61
[5]	Relative Valuation [High - Low]				-0.85	0.60	-0.46	-0.79	-0.65
[6]	Price Momentum [High - Low]					-0.42	0.23	0.88	0.66
[7]	Asset Growth YoY [Low - High]						-0.37	-0.40	-0.46
[8]	Gross Profitability to Assets [High - Low]							0.21	0.46
[9]	Analyst Revision 3-Mth FY1 EPS [High - Low]								0.67
[10]	Earnings Surprise [High - Low]								

Note: Lower (higher) correlations are denoted by green (red) colors. Source: S&P Global Market Intelligence Quantamental Research. Data as of 03/01/2021.

5. Textual Similarity Strategy Has Additive Predictive Value

The previous section shows that the historical performance of a textual similarity strategy is correlated to the performance of a number of other strategies that explain stock returns in the cross-section. This begs the question whether the performance of textual similarity strategies is largely or solely driven by their underlying correlations to the other strategies.

Exhibit 4 details the performance of the textual similarity strategies after controlling for the previous eight discussed strategies jointly via a regression framework.[21] After controls, the results indicate that the textual similarity strategy in the *MD&A* section produced approximately 4.00% per year of performance historically with statistical significance at the 1% level (see row [1], column [1] in Exhibit 4). In a joint framework, firms with the most (least) textual similarity in the *MD&A* section are lower (higher) beta firms, have been winners (losers) in the past 12-month and are cheaper (expensive) on a relative valuation basis.

The textual similarity strategy in the *Risk Factors* section added approximately 4.88% per year to performance historically with statistical significance at the 1% level (see row [3], column [1]

[19] See coverage results in the appendix A.7.
[20] Sector-neutral correlations are in appendix A.8.
[21] With the exception of the low-beta anomaly where it is substituted for market risk premium (MRP), which is defined as the difference between the market performance and the risk-free rate. Following literature with the substitution.

in Exhibit 4). In a joint framework, firms with the most (least) textual similarity in the *Risk Factors* section are lower (higher) beta firms which have been winners (losers) in the past 12-month.

Exhibit 4: Estimated Additive Performance and Strategy Exposures
Russell 3000; January 2008 – December 2020[22,23]

		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
	Signal	Alpha	MRP	SMB	HML	PMOM	AG	GPA	ANA	SUE
[1]	Cosine Similarity - MD&A	4.00%	**-0.15**	-0.06	**0.36**	**0.23**	0.05	-0.01	0.05	-0.04
[2]	T-Stat	3.39	-6.53	-1.58	6.70	6.68	0.67	-0.20	0.68	-0.44
[3]	Cosine Similarity - Risk Factors	4.88%	**-0.11**	**0.04**	**0.08**	**0.13**	0.07	0.04	-0.03	-0.02
[4]	T-Stat	4.55	-7.06	1.89	1.79	3.30	1.41	1.06	-0.55	-0.25

Note: Values that are shaded in green (red) are statistically significant at least at the 10% level. The performance values in column [1] are annualized from monthly averages. S&P Global Market Intelligence Quantamental Research. All returns and indices are unmanaged, statistical composites and their returns do not include payment of any sales charges or fees an investor would pay to purchase the securities they represent. Such costs would lower performance. It is not possible to invest directly in an index. Past performance is not a guarantee of future results. Data as of 03/01/2021.

Both strategies tend to do well when the stock market underperforms the short-term Treasury market. This usually occurs during periods of economic uncertainty, which implies that the two strategies are more defensive in nature. That is consistent with a flight-to-quality approach during uncertain times. These lower-beta filers are most likely the ones with the healthiest balance sheets and have the best risk controls, resulting in the least amount of textual revisions year-over-year.

6. Small Textual Revisions Are Meaningful for Returns and Volatility

Historically, the stock-level cosine similarity scores in the *MD&A* section are tightly clustered around the average of 0.93[24] with a small dispersion[25] of 0.07. This raises the question whether minute changes in textual similarity are meaningful.

Exhibit 5 highlights the average returns and volatilities of five portfolios sorted along the spectrum of textual similarity levels from the *MD&A* section. There is a general monotonic relationship between the portfolios' textual similarity levels and their returns and volatilities. Firms with the most (least) textual similarity have generally the highest (lowest) average returns and the lowest (highest) volatility historically despite the small differences in the textual similarity levels (see appendix A.6).

This leads to the question of whether strategies with portfolios that have a smaller number of filers with more extreme textual similarity would improve performance. Exhibit 5 has the

[22] See coverage results in the appendix A.7.
[23] Annualized alphas from sector-neutral results are 2.86% and 3.92% annually for the MD&A and Risk Factors section, respectively, with significance at the 1% level.
[24] The range of cosine similarity score is between 0 and 1.
[25] One standard deviation

cumulative dollar performance for three strategies where the universe in the cross-section is divided into 5, 10 and 25 portfolios with equal number of stocks.

Historically the strategy with 10 portfolios grew to $2.08 at the end of 2020, 22% higher than the strategy with 5 portfolios. The strategy with 25 portfolios grew to $2.55, 26% and 50% higher than the respective cumulative performance of the strategies with the 10 and 5 portfolios. The additional performance did not come at the expense of taking on more risk.

Exhibit 5: Performance of the Textual Similarity Strategy in the MD&A Section Russell 3000; 2008 - 2020[26]

	25 Portfolios	10 Portfolios	5 Portfolios
Cumulative $ Growth from $1 Base	2.55	2.08	1.70
Annualized Information Ratio	1.149	1.140	1.091
Monthly Hit Ratio > 0	65.4%	66.7%	66.7%

Note: The long-short return spreads are Carhart Four Factor Adjusted. Source: S&P Global Market Intelligence Quantamental Research. All returns and indices are unmanaged, statistical composites and their returns do not include payment of any sales charges or fees an investor would pay to purchase the securities they represent. Such costs would lower performance. It is not possible to invest directly in an index. Past performance is not a guarantee of future results. Data as of 03/01/2021.

7. Data

The Machine Readable U.S. Filings data begins in 2006. The feed provides all of the textual portions of public filings, is 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 and machine readable format through Xpressfeed and Snowflake. The filings have been parsed and are stored under individual sections (e.g., *MD&A*), effectively adding a structure to the unstructured textual data.

8. Conclusion

Investors have historically overlooked the implications of year-over-year textual revisions in the newest filings due to the voluminous amount of text and the small amount of changes.

Historically in the U.S. equity market, a textual similarity strategy using cosine similarity in the *MD&A* and in the *Risk Factors* section yielded an addition 4.00% and 4.88% per year, respectively, after taking into account commonly used stock selection and risk strategies. The textual similarity scores from the *Controls & Procedures, Legal Proceedings and Quantitative & Qualitative Disclosures about Market Risk* could be used as a composite strategy or used to supplement the scores from the *MD&A* and *Risk Factors* to drive additional performance.

Historically, firms with the most (least) textual similarity in the *MD&A* and *Risk Factors* section are winners (losers) in the past 12-month. They are defensive strategies in nature with lower (higher) betas that tend to outperform during uncertain economic times.

[26] See coverage results in the appendix A.7.

Appendix

A.1: Section Descriptions

Section Header	Description	10K: Part – Item	10Q: Part – Item
MD&A	A filer discusses the operations of the company in detail in the context of prior periods and provide an overview of the operations that led to increases, decreases or stagnation of the business.	Part II – Item 7	Part I – Item 2
Risk Factors	A filer lays out various risks its business faces. It is the second largest of the five sections in terms of number of words. Out of the five sections, it is the only one that has grown (e.g., almost 3-fold since 2006)	Part I – Item 1A	Part II – Item 1A
Q & Q Disclosures about Market Risks[27]	A filer discloses potential changes in the market value of financial assets and liabilities (e.g., interest rate risk, foreign exchange risk)	Part II – Item 7A	Part I – Item 3
Controls & Procedures	A filer discloses whether there have been changes in its internal control over its financial reporting.	Part II – Item 9A	Part I – Item 4
Legal Proceedings	A filer discloses any significant pending lawsuit or other legal proceedings.	Part I – Item 3	Part II – Item 1

Source: S&P Global Market Intelligence Quantamental Research. Data as of 03/01/2021.

A.2 Descriptive Statistics of Sections

January 2006 – December 2020

	Average Historical U.S. Coverage Russell 3000 Universe		Average Length Number of Words	
Section Header	**Form-10K**	**Form-10Q**	**Form-10K**	**Form-10Q**
MD&A	88%	87%	10.3K	6.5K
Risk Factors	90%	81%	7.9K	2.2K
Q & Q Disclosures about Market Risks	91%	91%	553	362
Controls & Procedures	91%	92%	498	286
Legal Proceedings	92%	80%	359	238

Source: S&P Global Market Intelligence Quantamental Research. Data as of 03/01/2021.

[27] Quantitative & Qualitative Disclosures about Market Risks

A.3 TF-IDF Equation – Details & Intuition

TF-IDF has two inputs: I) term frequency (TF) II) inverse document frequency (IDF). TF measures how often a word appears in a text. IDF measures whether a word appears scarcely or ubiquitously in a corpus, a data science parlance for dictionary. In our case, the corpus is a section of two year-over-year adjacent filings for a firm. The IDF term is the weighting scheme. A word that appears scarcely (ubiquitously) across a corpus gets a higher (lower) weight. **The intuition is that the scarcity (ubiquity) of a term in a corpus is deemed to have more (less) information content.**

Term	Calculation	Intuition
TF	TF = A / B • A = Frequency of a word in a section of a filing • B = Number of unique non-stop words in a corpus, which is a section in two adjacent filings of a firm	• The numerator is how frequently a word appears in a section of a filing and the denominator is the number of unique non-stop words in a corpus such that the TF term and the IDF term are both consistently using a particular corpus as a reference point. • Since we normalize the numerical representation of a section in a filing to a unit vector, the TF-IDF score is independent of whether the denominator uses the number of unique words from a text or in a corpus as long as the selection is consistently being applied.
IDF	IDF = natural log((C+1) / (D+1)) + 1 • C = Number of filings in a corpus • D = Number of filings that a word appears in a corpus	• The weighting scheme • Places a bigger weight on a word that appears scarcely in a corpus
Normalizing	Euclidean Norm	• Normalize for a section's varying length

Source: S&P Global Market Intelligence Quantamental Research. Data as of 03/01/2021.

A.4 Portfolio Construction and Back-Test Framework

Filings are events[28] that have various submission dates. In order to achieve breadth in the cross-section, we apply a look-back window of six calendar months to construct each of the monthly rebalancing strategies. For example, at the end of April when we are rebalancing and constructing our portfolio, we look at the stock-level cosine similarity scores that are available since January of the same year to form our portfolio. At the end of May, we construct our portfolio by looking back as far as February. If there are two or more scores for a firm in the window, we take the latest score.

All long-only and long-short returns in our back-test are equal-weighted, are rebalanced monthly at month end and are binned into quintiles where the top (bottom) quintile or the long (short) portfolio contains the 20% of stocks with the highest (lowest) signal scores in the cross-

[28] Generally, event-driven signals lack a sufficient number of stocks for portfolio construction and the timing of the next event is unknown.

section. Definitions of the columns from the following exhibits containing back-test results are in appendix A.2.

A.5 Meaning of Columns in Tables Containing Empirical Results

Signal Name, Sort Order, Start Date, Firm Count

- Column 1: the sort order of a signal where 'D' or -1 is descending and 'A' or 1 is ascending
- Column 2: the date back-tests commenced for a signal
- Column 3: the average number of firms in a quintile bucket in our sample period;

Signal Strength Metrics

- Column 4: the average monthly information coefficient (i.e., Spearman correlation) that is used to assess a signal's historical predictive strength
- Column 5: the monthly hit rate for column 4 - the percent of the months where the IC > 0

Active Long Metrics

- Column 6: the annualized average monthly market-adjusted return of the long portfolio
- Column 7: annualized information ratio of column 6
- Column 8: the monthly hit rate for column 6 where the market-adjusted return of the long-side > 0

Long-Short Metrics

- Column 9: the annualized long-short returns
- Column 10: annualized information ratio of column 9
- Column 11: the monthly hit rate for column 9 where the monthly long-short return > 0

A.6 Raw Returns and Volatilities of Portfolios Across the Spectrum of Textual Similarity Level in the *MD&A* Section

Russell 3000; 2008 – 2020[29]

		[1] Most Similarity	[2]	[3] Neutral	[4]	[5] Least Similarity
[1]	Avg Language Similarity Scores	0.980	0.964	0.948	0.925	0.847
[2]	Annualized Mthly Return / Volatility	0.60	0.53	0.55	0.58	0.43
[3]	Avg Monthly Returns	1.01%	0.95%	1.04%	0.99%	0.86%
[4]	Avg Monthly Volatility	5.89%	6.20%	6.49%	6.81%	6.94%

S&P Global Market Intelligence Quantamental Research. All returns and indices are unmanaged, statistical composites and their returns do not include payment of any sales charges or fees an investor would pay to purchase the securities they represent. Such costs would lower performance. It is not possible to invest directly in an index. Past performance is not a guarantee of future results. Data as of 03/01/2021.

[29] See coverage results in the appendix A.7.

A.7 Annual Descriptive Statistics on the Five Major Commonly Sections

Russell 3000; January 2006 – December 2020

MD&A		count 10k	count 10q	pct 10k	pct 10q	pct 10kq	avg 10k words	med 10k words	avg 10q words	med 10q words
	avg	2633	7860	88%	87%	87%	10289	9184	6528	5740
	median	2722	7910	91%	88%	89%	10470	9308	6515	5763
	2006	2368	7006	79%	78%	78%	9595	8454	6185	5471
	2007	2458	7463	82%	83%	83%	9797	8773	6152	5350
	2008	2387	7341	80%	82%	81%	10256	9134	6436	5600
	2009	2365	7241	79%	80%	80%	10976	9620	6779	5887
	2010	2573	7757	86%	86%	86%	10680	9445	6482	5635
	2011	2716	8194	91%	91%	91%	10470	9209	6455	5631
	2012	2740	8103	91%	90%	90%	10495	9308	6479	5630
	2013	2756	8297	92%	92%	92%	10481	9291	6617	5764
	2014	2806	8473	94%	94%	94%	10576	9381	6616	5830
	2015	2824	8401	94%	93%	94%	10485	9409	6609	5827
	2016	2783	8260	93%	92%	92%	10553	9511	6589	5825
	2017	2722	7801	91%	87%	88%	10406	9399	6515	5763
	2018	2735	8319	91%	92%	92%	10412	9532	6546	5862
	2019	2758	7910	92%	88%	89%	9939	9012	6246	5611
	2020	2510	7336	84%	82%	82%	9219	8288	7220	6418

Risk Factors		count 10k	count 10q	pct 10k	pct 10q	pct 10kq	avg 10k words	med 10k words	avg 10q words	med 10q words
	avg	2714	7285	90%	81%	83%	7892	6797	2230	131
	median	2781	7419	93%	82%	85%	7612	6639	2051	76
	2006	2535	5961	85%	66%	71%	4258	3602	1526	92
	2007	2624	6693	87%	74%	78%	4736	3976	1601	92
	2008	2517	6703	84%	74%	77%	5259	4463	1690	120
	2009	2458	6600	82%	73%	75%	5863	5141	1651	103
	2010	2657	7077	89%	79%	81%	6175	5414	1753	103
	2011	2776	7476	93%	83%	85%	6633	5810	1854	90
	2012	2802	7419	93%	82%	85%	7187	6241	1931	76
	2013	2818	7619	94%	85%	87%	7612	6639	2051	74
	2014	2866	7840	96%	87%	89%	8364	7263	2330	75
	2015	2878	7842	96%	87%	89%	9048	7782	2545	74
	2016	2837	7804	95%	87%	89%	9674	8366	2586	76
	2017	2781	7379	93%	82%	85%	10062	8643	2473	66
	2018	2793	7972	93%	89%	90%	10574	9018	2903	75
	2019	2806	7543	94%	84%	86%	11358	9692	2963	69
	2020	2568	7340	86%	82%	83%	11584	9899	3591	785

Controls & Procedures		count 10k	count 10q	pct 10k	pct 10q	pct 10kq	avg 10k words	med 10k words	avg 10q words	med 10q words
	avg	2741	8236	91%	92%	91%	498	391	286	222
	median	2806	8156	94%	91%	92%	497	394	290	220
	2006	2612	8104	87%	90%	89%	543	352	301	207
	2007	2641	8124	88%	90%	90%	535	358	302	210
	2008	2547	7696	85%	86%	85%	497	358	274	211
	2009	2484	7584	83%	84%	84%	485	371	259	209
	2010	2638	8037	88%	89%	89%	492	372	256	212
	2011	2806	8551	94%	95%	95%	467	374	268	214
	2012	2834	8447	94%	94%	94%	475	381	273	217
	2013	2850	8632	95%	96%	96%	483	394	280	220
	2014	2892	8786	96%	98%	97%	483	399	286	225
	2015	2902	8707	97%	97%	97%	503	412	294	228
	2016	2862	8547	95%	95%	95%	504	416	295	230
	2017	2803	8061	93%	90%	91%	510	421	304	229
	2018	2816	8610	94%	96%	95%	500	418	301	236
	2019	2838	8156	95%	91%	92%	505	426	303	239
	2020	2587	7494	86%	83%	84%	493	417	290	238

QoQ Disclosures on Mkt Risk		count 10k	count 10q	pct 10k	pct 10q	pct 10kq	avg 10k words	med 10k words	avg 10q words	med 10q words
	avg	2719	8179	91%	91%	91%	553	398	362	221
	median	2786	8081	93%	90%	91%	559	407	360	219
	2006	2574	7958	86%	88%	88%	505	319	339	219
	2007	2613	8040	87%	89%	89%	476	335	335	221
	2008	2538	7776	85%	86%	86%	539	389	385	259
	2009	2467	7624	82%	85%	84%	580	416	397	257
	2010	2625	8069	88%	90%	89%	547	407	378	232
	2011	2783	8448	93%	94%	94%	551	401	368	228
	2012	2818	8355	94%	93%	93%	556	400	358	222
	2013	2827	8536	94%	95%	95%	559	401	361	215
	2014	2871	8689	96%	97%	96%	575	406	360	215
	2015	2882	8639	96%	96%	96%	563	410	366	223
	2016	2843	8488	95%	94%	94%	578	419	360	214
	2017	2786	8018	93%	89%	90%	572	415	355	212
	2018	2800	8558	93%	95%	95%	565	419	352	208
	2019	2802	8081	93%	90%	91%	545	407	343	189
	2020	2557	7408	85%	82%	83%	588	427	372	198

Legal Proceedings		count 10k	count 10q	pct 10k	pct 10q	pct 10kq	avg 10k words	med 10k words	avg 10q words	med 10q words
	avg	2758	7212	92%	80%	83%	359	104	238	65
	median	2817	7300	94%	81%	84%	354	101	239	64
	2006	2659	6335	89%	70%	75%	463	143	304	84
	2007	2666	6595	89%	73%	77%	477	134	317	81
	2008	2584	6513	86%	72%	76%	465	139	299	75
	2009	2507	6422	84%	71%	74%	434	118	266	66
	2010	2701	6866	90%	76%	80%	395	110	262	66
	2011	2817	7300	94%	81%	84%	413	127	255	68
	2012	2846	7291	95%	81%	84%	385	110	244	65
	2013	2858	7569	95%	84%	87%	354	101	239	64
	2014	2897	7806	97%	87%	89%	349	92	224	60
	2015	2911	7837	97%	87%	90%	313	87	204	59
	2016	2872	7769	96%	86%	89%	295	86	207	59
	2017	2812	7371	94%	82%	85%	288	83	192	58
	2018	2824	7979	94%	89%	90%	270	79	203	58
	2019	2835	7572	95%	84%	87%	251	77	189	58
	2020	2588	6961	86%	77%	80%	236	75	172	57

A.8: Sector-Neutral Correlations of Long-Short Quintile Return Spreads

Russell 3000; January 2008 – December 2020[30]

	Signal Name	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
[1]	Cosine Similarity - MD&A	0.61	-0.54	-0.53	0.56	-0.11	0.62	0.46	0.53
[2]	Cosine Similarity - Risk Factors	0.27	0.02	-0.03	0.09	0.09	0.29	0.02	0.00
[3]	Low Beta Anomaly [Low - High]		0.71	0.82	0.73	0.31	0.69	0.66	0.54
[4]	Small Cap Effect [Low - High]			0.80	-0.73	0.38	-0.49	-0.70	-0.59
[5]	Relative Valuation [High - Low]				-0.92	0.47	-0.58	-0.85	-0.65
[6]	Price Momentum [High - Low]					-0.38	0.56	0.87	0.68
[7]	Asset Growth YoY [Low - High]						-0.17	-0.33	-0.40
[8]	Gross Profitability to Assets [High - Low]							0.47	0.51
[9]	Analyst Revision 3-Mth FY1 EPS [High - Low]								0.68
[10]	Earnings Surprise [High - Low]								

Note: Lower (higher) correlations are denoted by green (red) colors. Source: S&P Global Market Intelligence Quantamental Research. Data as of 03/01/2021.

[30] See coverage results in the appendix A.7.

References

Bernard, V. L., and Jacob K. Thomas, 1989, Post-earnings-announcement drift: Delayed price response or risk premium? Journal of Accounting Research Supplement 27, 1-48.

Black, Fischer, 1972. “Capital Market Equilibrium with Restricted Borrowing,” Journal of Business 45 (3), 444-455

Carhart, M. M., 1997, “On Persistence in Mutual Fund Performance,” Journal of Finance, 52, 57-82.

Chan, Louis K. C., Jegadeesh, N., and Lakonishok, J. 1996, Momentum strategies, Journal of Finance 51, 1681-1713.

Cohen, L., Malloy, C., and Nguyen, Q. “Lazy Prices." *Journal of Finance* 75 (2020): 1371-1415.

Cooper, M. J., Gulen, H., and Schill, M. J., 2008, Asset growth and the cross-section of stock returns, Journal of Finance 63, 1609-1652.

Fama, E. F., and K. R. French, 1993, “Common Risk Factors in the Returns on Bonds and Stocks,” Journal of Financial Economics, 33, 3-53.

Frazzini, Andrea, and Lasse Pedersen, 2014. “Betting Against Beta,” Journal of Financial Economics 111 (1), 1 - 25

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.

Novy-Marx, R., 2013, The other side of value: The gross profitability premium, Journal of Financial Economics 108, 1-28.

Yang, Z., and Oyeniyi, T. “Hiding in Plain Sight – Risks That Are Overlooked.” S&P Global Market Intelligence Quantamental Research March 2021

Zhao, F. “Natural Language Processing – Part III: Feature Engineering Applying NLP Using Domain Knowledge to Capture Alpha from Transcripts.” S&P Global Market Intelligence Quantamental Research January 2020

Our Recent Research

March 2021: Hiding in Plain Sight – Risks That Are Overlooked

This report uses three metrics (Minimum Edit Distance, Jaccard Similarity, and Cosine Similarity) to identify companies that made significant changes to the "Risk Factors" section of their filings. These metrics can serve as alpha signals or be used to quickly identify a pool of companies that require further investigation.

January 2021: Leadership Change That Matters: A Value and Momentum Story

December 2020: Warranted Optimism: Sentiment vs. Supply Chain

December 2020: A Dark Winter for REITS: Trouble Brewing

October 2020: Sweet Spots in the C-Suite: Executive Best Practices for Shareholder Friendly Firms

The Business Roundtable, an association of CEOs of America's leading companies, published a new statement on corporate responsibility in August 2019. The statement identifies five important corporate stakeholders: customers, employees, suppliers, communities and shareholders.1 This report highlights four key types of executive policy that drive value creation for stakeholders: profitability vs. growth decisions, mergers & acquisitions policy, return of cash to shareholders, and insider stock ownership. In it, we demonstrate empirically those practices that increase corporate value over time, thereby rewarding shareholders, employees, and other stakeholders. These practices also form a scorecard by which stakeholders can evaluate whether or not management is undertaking actions likely to increase corporate prosperity.

October 2020: Just the (Build)Fax: Property Intelligence from Building Permit Data

August 2020: The Analyst Matrix: Profiting from Sell-Side Analysts' Coverage Networks

Sell-side analyst coverage data provides a new and rich source of establishing connections between firms, as analysts (given their industry expertise) are likely to cover fundamentally related firms. This report uses sell-side analysts' coverage data to build a connected-firm network (CFN) - a portfolio of companies that are covered by analyst(s) that follow a focal firm. This network has three broad applications: measuring the "strength" of economic relationships between companies; forecasting fundamentals of companies in the network; and as a stock selection signal.

June 2020: The Information Supply Chain Begins Recovering From COVID

May 2020: Never Waste a Crisis: Following the Smart Money Through Beneficial Ownership Filings

May 2020: Risky Business: Foot Traffic, Vacancy Rates and Credit Risks

May 2020: Finding the Healthy Stocks in Health Care During Lockdown

May 2020: No More Walks in the (Office) Park: Tying Foot Traffic Data to REITs

May 2020: Do Markets Yearn for the Dog Days of Summer: COVID, Climate and Consternation

April 2020: Cold Turkey - Navigating Guidance Withdrawal Using Supply Chain Data

April 2020: Data North Star - Navigating Through Information Darkness

March 2020: Long Road to Recovery: Coronavirus Lessons from Supply Chain and Financial Data

COVID-19 continues to disrupt global supply chains in unprecedented ways. Leveraging maritime shipping data from Panjiva, this report includes a review of trade and financial data to analyze the impact of the SARS-CoV-2 / COVID-19 coronavirus outbreak. Findings include:

Second-order supply chain effects are also emerging with the apparel industry now seeing a shortage of materials globally due to earlier outages in China.

Retailers including Costco and Target are gaining from increased sales of health- and personal care products. Yet, supply shortages are rapidly emerging in part due to medical supply export restrictions in several countries.

There is a notable, but not statistically significant, relationship with firms with higher exposure to Asia having seen a weaker sector neutral stock price performance.

February 2020: Ship to Shore: Mapping the Global Supply Chain with Panjiva Shipping Data in Xpressfeed™

January 2020: Natural Language Processing – Part III: Feature Engineering Applying NLP Using Domain Knowledge to Capture Alpha from Transcripts

December 2019: The "Trucost" of Climate Investing: Managing Climate Risks in Equity Portfolios

October 2019: #ChangePays: There Were More Male CEOs Named John than Female CEOs

June 2019: Looking Beyond Dividend Yield: Finding Value in Cash Distribution Strategies

June 2019: The Dating Game: Decrypting the Signals in Earnings Report Dates

May 2019: Bridges for Sale: Finding Value in Sell-Side Estimates, Recommendations, and Target Prices

February 2019: U.S Stock Selection Model Performance Review

February 2019: International Small Cap Investing: Unlocking Alpha Opportunities in an Underutilized Asset Class

January 2019: Value and Momentum: Everywhere, But Not All the Time

November 2018: Forging Stronger Links: Using Supply Chain Data in the Investing Process

September 2018: Their Sentiment Exactly: Sentiment Signal Diversity Creates Alpha Opportunity

September 2018: Natural Language Processing – Part II: Stock Selection: Alpha Unscripted: The Message within the Message in Earnings Calls

July 2018: A Case of 'Wag the Dog'? - ETFs and Stock-Level Liquidity

June 2018: The (Gross Profitability) Trend is Your Friend

May 2018: Buying the Dip: Did Your Portfolio Holding Go on Sale?

March 2018: In the Money: What Really Motivates Executive Performance?

February 2018: The Art of the (no) Deal: Identifying the Drivers of Canceled M&A Deals

January 2018: U.S Stock Selection Model Performance Review

September 2017: Natural Language Processing - Part I: Primer

July 2017: Natural Language Processing Literature Survey

June 2017: Research Brief: Four Important Things to Know About Banks in a Rising Rate Environment

April 2017: Banking on Alpha: Uncovering Investing Signals Using SNL Bank Data

March 2017: Capital Market Implications of Spinoffs

January 2017: U.S. Stock Selection Model Performance Review 2016

November 2016: Electrify Stock Returns in U.S. Utilities

October 2016: A League of their Own: Batting for Returns in the REIT Industry - Part 2

September 2016: A League of their Own: Batting for Returns in the REIT Industry - Part 1

August 2016: Mergers & Acquisitions: The Good, the Bad and the Ugly (and how to tell them apart)

July 2016: Preparing for a Slide in Oil Prices -- History May Be Your Guide

June 2016: Social Media and Stock Returns: Is There Value in Cyberspace?

April 2016: An IQ Test for the "Smart Money" – Is the Reputation of Institutional Investors Warranted?

March 2016: Stock-Level Liquidity – Alpha or Risk? - Stocks with Rising Liquidity Outperform Globally

February 2016: U.S. Stock Selection Model Performance Review - The most effective investment strategies in 2015

January 2016: What Does Earnings Guidance Tell Us? – Listen When Management Announces Good News

November 2015: Late to File - The Costs of Delayed 10-Q and 10-K Company Filings

October 2015: Global Country Allocation Strategies

September 2015: Research Brief: Building Smart Beta Portfolios

September 2015: Research Brief – Airline Industry Factors

August 2015: Point-In-Time vs. Lagged Fundamentals – This time i(t')s different?

August 2015: Introducing S&P Capital IQ Stock Selection Model for the Japanese Market

July 2015: Research Brief – Liquidity Fragility

May 2015: Investing in a World with Increasing Investor Activism

April 2015: Drilling for Alpha in the Oil and Gas Industry – Insights from Industry Specific Data & Company Financials

February 2015: U.S. Stock Selection Model Performance Review - The most effective investment strategies in 2014

January 2015: Research Brief: Global Pension Plans - Are Fully Funded Plans a Relic of the Past?

January 2015: Profitability: Growth-Like Strategy, Value-Like Returns - Profiting from Companies with Large Economic Moats

October 2014: Lenders Lead, Owners Follow - The Relationship between Credit Indicators and Equity Returns

July 2014: Factor Insight: Reducing the Downside of a Trend Following Strategy

May 2014: Introducing S&P Capital IQ's Fundamental China A-Share Equity Risk Model

April 2014: Riding the Coattails of Activist Investors Yields Short and Long Term Outperformance

March 2014: Insights from Academic Literature: Corporate Character, Trading Insights, & New Data Sources

February 2014: Obtaining an Edge in Emerging Markets

February 2014: U.S Stock Selection Model Performance Review

January 2014: Buying Outperformance: Do share repurchase announcements lead to higher returns?

October 2013: Informative Insider Trading - The Hidden Profits in Corporate Insider Filings

September 2013: Beggar Thy Neighbor – Research Brief: Exploring Pension Plans

August 2013: Introducing S&P Capital IQ Global Stock Selection Models for Developed Markets: The Foundations of Outperformance

July 2013: Inspirational Papers on Innovative Topics: Asset Allocation, Insider Trading & Event Studies

June 2013: Supply Chain Interactions Part 2: Companies – Connected Company Returns Examined as Event Signals

June 2013: Behind the Asset Growth Anomaly – Over-promising but Under-delivering

April 2013: Complicated Firms Made Easy - Using Industry Pure-Plays to Forecast Conglomerate Returns.

March 2013: Risk Models That Work When You Need Them - Short Term Risk Model Enhancements

March 2013: Follow the Smart Money - Riding the Coattails of Activist Investors

February 2013: Stock Selection Model Performance Review: Assessing the Drivers of Performance in 2012

January 2013: Research Brief: Exploiting the January Effect Examining Variations in Trend Following Strategies

December 2012: Do CEO and CFO Departures Matter? - The Signal Content of CEO and CFO Turnover

November 2012: 11 Industries, 70 Alpha Signals -The Value of Industry-Specific Metrics

October 2012: Introducing S&P Capital IQ's Fundamental Canada Equity Risk Models

September 2012: Factor Insight: Earnings Announcement Return – Is A Return Based Surprise Superior to an Earnings Based Surprise?

August 2012: Supply Chain Interactions Part 1: Industries Profiting from Lead-Lag Industry Relationships

July 2012: Releasing S&P Capital IQ's Regional and Updated Global & US Equity Risk Models

June 2012: Riding Industry Momentum – Enhancing the Residual Reversal Factor

May 2012: The Oil & Gas Industry - Drilling for Alpha Using Global Point-in-Time Industry Data

May 2012: Case Study: S&P Capital IQ – The Platform for Investment Decisions

March 2012: Exploring Alpha from the Securities Lending Market – New Alpha Stemming from Improved Data

January 2012: S&P Capital IQ Stock Selection Model Review – Understanding the Drivers of Performance in 2011

January 2012: Intelligent Estimates – A Superior Model of Earnings Surprise

December 2011: Factor Insight – Residual Reversal

November 2011: Research Brief: Return Correlation and Dispersion – All or Nothing

October 2011: The Banking Industry

September 2011: Methods in Dynamic Weighting

September 2011: Research Brief: Return Correlation and Dispersion

July 2011: Research Brief - A Topical Digest of Investment Strategy Insights

June 2011: A Retail Industry Strategy: Does Industry Specific Data tell a different story?

May 2011: Introducing S&P Capital IQ's Global Fundamental Equity Risk Models

May 2011: Topical Papers That Caught Our Interest

April 2011: Can Dividend Policy Changes Yield Alpha?

April 2011: CQA Spring 2011 Conference Notes

March 2011: How Much Alpha is in Preliminary Data?

February 2011: Industry Insights – Biotechnology: FDA Approval Catalyst Strategy

January 2011: US Stock Selection Models Introduction

January 2011: Variations on Minimum Variance

January 2011: Interesting and Influential Papers We Read in 2010

November 2010: Is your Bank Under Stress? Introducing our Dynamic Bank Model

October 2010: Getting the Most from Point-in-Time Data

October 2010: Another Brick in the Wall: The Historic Failure of Price Momentum

July 2010: Introducing S&P Capital IQ's Fundamental US Equity Risk Model

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.