The value of alternative data:

The case for media sentiment return, risk, and sustainability return, risk, and sustainability

Contributors:

Svetlana Borovkova

Corne Reniers

Kevin Rojer

Abstract

This white paper addresses the question of the value of alternative data in the investment process. We propose a quantitative framework to assess the added value of an alternative data set on the basis of a backtesting process, in combination with the so-called GH1 measure, which takes into account both return enhancement and risk reduction, with respect to a particular benchmark.

We illustrate this framework on the example of LSEG News Analytics data, which is used as a standalone investment signal as well as combined with the traditional multifactor investment strategy. We demonstrate that this alternative data set provides significant value to investors: using media sentiment as the single factor achieves the same investment results as the full-blown multifactor strategy.

We show that the value of alternative data depends on the investor type, and calculate the monetary value of this data set for various types of investors. Finally, we show that, even for a small size fund and simple strategies, typical recovery factors from purchasing alternative data make a compelling business case for including alternative data (such as sentiment) in your investment strategies.

LSEG DATA & ANALYTICS

1. Alternative data for investing

'Data is the new oil' – this phrase summarises that data is becoming the world's most valuable commodity. The buzz around alternative data has increased considerably in the financial and investment community, as asset managers try to extract signals from alternative data to transform them into added wealth for their clients.

Information retrieved from non-traditional sources is increasingly used as input for investment decisions. It can give investors insights that go beyond traditional data such as earnings, credit reports, and company and industry statistics. For example, alternative data can come from ingenious new data sources such as sensors or GPS locations. Other examples of alternative data sources are social media platforms, news wires, satellite imagery, sustainability-related data, online spending patterns and more. The usefulness of alternative data lies in providing almost real-time information and information which was recently unavailable.

There are several formal characterisation methods for alternative data sets, based, for example, on a data set's origin, how and by whom it is generated, and others. These characterisation schemes do not consider how to determine the value of alternative data for investors. This is exactly the subject of this paper. Here we suggest an alternative data valuation framework, which distinguishes various investment styles and takes into account the existing investment philosophy.

An interesting (and unorthodox) source of alternative data is satellite imagery. Satellites orbit regularly and frequently over the same spots around the earth, providing regular updates on a location. The cost of satellite imagery has diminished significantly in the last few years, while the resolution has improved. Examples of satellite imagery applications in investment decisions include the number of cars in parking lots of retailers to forecast revenue, or ships passing through ports to proxy the volume and rate of commercial activities. Another interesting example is the use of images of shadow lengths from a real estate project's construction site to determine the project's pace.

A popular source of alternative data – and the example we use in this paper – is the sentiment of news and social media content. Research in this area dates back to the seminal papers by Tetlock (2007, 2008) and is further developed by numerous papers by Borovkova (e.g., Borovkova and Xiaobo (2015)) and others. In these papers, various applications of sentiment in finance are investigated, such as systemic risk monitoring, commodity trading, sector rotation and so on. It appears that sentiment can be a powerful tool in reducing risk and enhancing investor returns. These and other related papers justify the use of sentiment as an alternative (and additional) signal for investment strategies.

Challenges with alternative data sets are related to their volume and complexity: how to process the data, how to turn it into actionable signals and how to join it to more traditional data typically used in investing (such as asset prices or company's fundamentals). In the satellite imaging of parking lots example, how does one determine whether the lot belongs to Walmart or Target or Costco? Textual data sets such as media content are challenging too, since these are typically unstructured and at first glance it seems hard to determine what asset(s) a text refers to for example, whether 'apple' refers to a fruit or a rather large tech company. These examples are related to mapping the data into an investable universe, and the data set which has been already mapped (or tagged) into companies, commodities or other assets is clearly more valuable to users, as it saves them a lot of processing time and energy.

This brings us to another major challenge with alternative data: how to assign value to it. This question is interesting from the perspective of both asset managers (who can spend substantial amounts on alternative data), and data vendors. Data buyers and data providers have different perspectives when it comes to pricing. For a data vendor, it is essential to recover the costs associated with the creation and distribution of the data set. However, the value of alternative data for asset managers lies predominantly in its monetising potential. Our paper proposes a practical framework to quantify the value of alternative data in quantitative investing.

Specifically, we address the following questions:

1. What is the impact of an alternative data set on portfolio performance?
2. Is there a systematic way to capture and evaluate the benefits achieved by using alternative data?
3. Is the performance of a data set sensitive to an investor profile and strategy?

We take the popular factor model investment strategy (*Fama and French, 2016*) – the main workhorse of modern investing – as the main competitor and use news sentiment data as the example of an alternative data set. We show that sentiment data has a significant value for asset managers: portfolios constructed using sentiment signals significantly outperform the benchmark, but, most importantly, using sentiment as the sole portfolio construction factor performs as well as, and in some cases better than, traditional multifactor portfolios. We also show that the value of alternative data depends on the investor's profile and the fund's size.

2. Factor investing and investment styles

Factor investing is a widely used active investment approach that targets quantifiable firm characteristics, called factors, that explain differences in stock returns. Usually, factor-based portfolios are built on the basis of security characteristics such as size, growth and momentum.

Academic research on factor risk premium goes back to the famous Fama and French model, which incorporates the size and value factors (as well as the market factor). The Fama and French model can be further extended in several ways, such as by adding the momentum and low volatility factors. This is the extended version of the factor model we will use here.

An investor profile or style describes an individual's or an organisation's preference in investment decisions. These preferences are related to the risk attitude, the investment horizon and rebalancing frequency, the asset mix, investing belief, whether to use derivatives, whether to allow short positions, the choice of domestic versus international stocks, active or passive investing, asset allocation and the particular investment strategy.

Here we will not consider the classic investment question of the proportion of stocks versus bonds, but focus instead on equity investors (or the equity part of their portfolios). We will distinguish between different rebalancing frequencies and between long-only and long-short portfolios. We run our backtests for three investment profiles:

- **Investor 1:** quarterly rebalancing frequency, long positions only
- **Investor 2:** monthly rebalancing frequency, long positions only
- **Investor 3:** monthly rebalancing frequency, long and short positions

3. Valuation framework

Does an alternative data set help in generating profitable trading signals? Our framework quantifies the value of alternative data from a data consumer's perspective. We focus on the data's monetising potential rather than on the ease and convenience of use (which, of course, also contributes to the data set's value).

Note that the financial return – or the so-called alpha, the excess return over the benchmark – is not the only performance measure used in investments. Risk is equally significant in determining the performance of an investment strategy, and many alternative data sets can help reduce the risk of investments rather than directly enhance returns. Hence, a sound quantification framework should combine both these measures.

The primary method for quantifying the value of any investment strategy is through backtesting. Essentially, backtesting determines how a manager would have performed had it incorporated a specific investment strategy in the past. Managers base their analysis on historical data, and they assume that the results are transferable to the future. Thus, the methodology employs sample data from a relevant period reflecting various market conditions, to judge whether the backtested result is a robust strategy or a poor prospect. A backtest should typically consider trading costs, as these can tally up throughout the backtesting period and influence a strategy's actual performance.

Following such a backtesting framework, we determine the value of alternative data by learning how a strategy would have performed had it incorporated alternative data in the past. The value of alternative data depends on the choice of the associated investment strategy. Therefore, it is expected that the value will be different for different investment parameters. Accordingly, we construct historical portfolios for each investor profile and compare them to portfolios without alternative data.

We need to agree on an appropriate benchmark which is compared to portfolios constructed using alternative data. This benchmark can be anything, such as a passive index such the S&P500 or another active strategy currently employed by the fund. Once the benchmark (or a set of them) has been chosen, we need to make the portfolio and benchmark comparable. For that, we can leverage/deleverage the benchmark to match the portfolio volatility. Similarly, we could lever up/down the portfolio to match the volatility of the benchmark. Matching the portfolio volatility with the benchmark volatility allows us to compare the returns at that specific level of risk.

This approach also allows us to determine the added value of a data set, even if it does not (apparently) contribute to enhancing the returns, but it does decrease the risk of the portfolio. Consider a hypothetical case: Portfolio 1, which does not use alternative data (i.e., the benchmark) has the (annualised) return of 10% and volatility of 20%. Portfolio 2, which is built using alternative data, exhibits the return of 7% with volatility of just 10%. Although it seems that using alternative data lowers the return, it lowers the volatility so much that it is still profitable to use it. This can be seen by simple calculations: for each dollar to be invested into Portfolio 2, borrow an additional dollar (let's assume that the borrowing rate is 2%) and invest the double amount into Portfolio 2. The resulting volatility will be 2*10% = 20%, so the same as the benchmark, but the total return will be 2*7%−2% = 12% which is 2% per annum higher than the benchmark. This additional return of 2% can be then translated into the added monetary value of data, by multiplying it with the invested assets under management (AUM).

Let us formalise this example. There are two possibilities. First, it could be that the benchmark has lower volatility than the portfolio: ($\sigma_B < \sigma_{port}$). In this case, the investor needs to create a leveraged benchmark, to match the portfolio's volatility. This is done by investing σ^{port}/σ^{B} into the benchmark (and borrowing $\sigma^{port}/\sigma^{B}-1$ against the risk-free rate). Figure 1 illustrates this situation.

In this case, we have:

$$R_{adj.benchmark} = \frac{\sigma_{port}}{\sigma_B}(R_B - r_f) + r_f \quad (1)$$

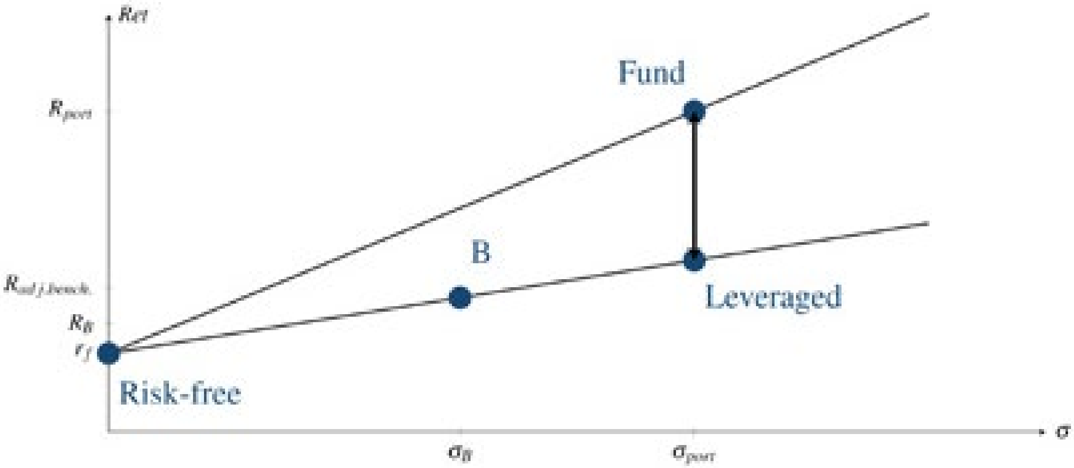

Figure 1: The benchmark has a lower volatility than the portfolio: ($\sigma_B < \sigma_{port}$).

Alternatively, if the benchmark has higher volatility than the portfolio ($\sigma_B > \sigma_{port}$) – this was the situation in the above example – then the adjusted benchmark is a mix of the proportion σ_{port} / σ_B invested in the benchmark and the rest in the risk-free asset. The case of deleveraging the σB benchmark is illustrated in Figure 2, and the return on the deleveraged benchmark is again given by (1).

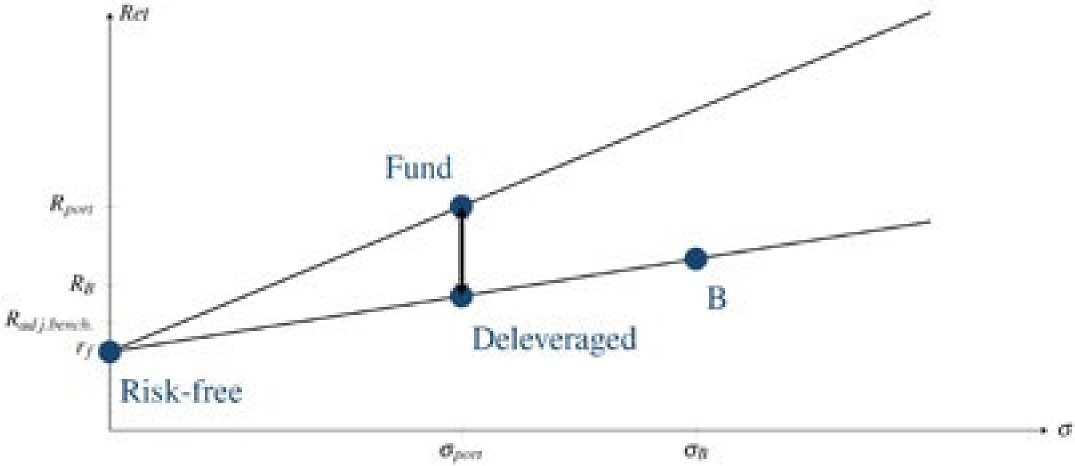

Figure 2: The benchmark has a higher volatility than the portfolio: ($\sigma_B > \sigma_{Fund}$).

Now the adjusted benchmark shows the return that the benchmark would have achieved if it had the same risk as the portfolio. The difference between the actual return of the portfolio and the return of the volatility-matched benchmark is called the GH1 measure (introduced by Graham and Harvey (1997)), and it tells us whether or not there was overperformance of the portfolio with respect to the benchmark, depending on whether this measure is positive or negative.

Using this GH1 measure, one can determine whether the combination of the index and the risk free asset generates higher return at the same level of risk than the fund that uses alternative data. If the fund generates higher returns than the adjusted portfolio, investors should act on the alternative data set. However, there is another, last step necessary, since the GH1 measure needs to be translated into monetary terms. To quantify the value of a particular data set for a particular fund, we multiply the fund's AUM with this GH1 number. This value is the expected additional yearly income for the fund that uses the alternative data. If this expected annual increase is higher than the cost of purchasing, processing and onboarding the alternative data set, then the asset manager can decide that it is profitable to buy and use this in their investment process. We will illustrate this approach on a specific example at the end of the paper.

4. Methodology: sentiment signals and factor models

News sentiment investment signals

The data set we use as the example of an alternative data source is the news sentiment data from News Analytics. This data is powered by an engine, based on complex natural language processing (NLP) algorithms, which 'reads' and interprets (in real time) all news that reaches the Reuters news wire. Each news item is tagged in relation to a company or a commodity – currently 45,000 listed companies are incorporated, as well as all commodities. Moreover, the engine returns, for each news item, a wealth of quantitative characteristics, the most important being the sentiment score – the probability that a particular news item conveys a positive, negative or neutral outlook on the price of the underlying asset (the one which the news item was tagged for). The relevance score indicates how relevant a particular news item is for that asset. Furthermore, the novelty indicator gives the number of similar news items that have been already observed before.

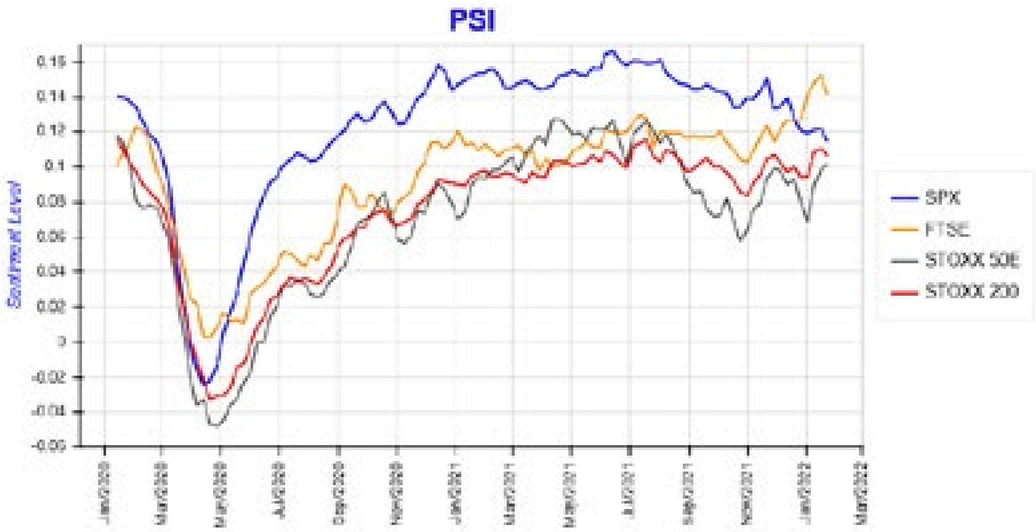

Figure 3: The sentiment indicators for S&P 500, FTSE and STOXX indices, January 2020 to March 2022.

The raw sentiment scores for a particular stock are extremely volatile. So to put the sentiment in an investment context, it is necessary to 'de-noise' sentiment data, to obtain a meaningful investment signal. The methodology for this is based on the work of Borovkova and Mahakena (2015). The first step is to aggregate stock-specific sentiment scores of individual news items into, for example, daily or weekly sentiment scores, accounting for relevance and novelty. The next, noise filtering step is based on the so-called Local News Sentiment Level (LNSL) model, which extracts the unobserved sentiment from the observed noisy sentiment series using the Kalman filter methodology. The final step in generating sentiment investment signals for a particular stock universe (such as S&P 500 index) is aggregating filtered stock-specific sentiment signals into the market-wide sentiment using an appropriate weighting scheme. The resulting historical sentiment indicators (which we call PSI: Probability Sentiment Indicators) for the last two years for major stock indices are shown in Figure 3.

Factor models

Investment strategies that incorporate alternative data, such as sentiment, can be compared (in terms of their return, volatility or Sharpe ratio) to a benchmark index such as S&P 500. However, this would be an unfair comparison, as index investing is a passive strategy, while a sentiment-based strategy requires periodic rebalancing and is, hence, active. So we compare the investment performance not only to the benchmark, but also to a multifactor strategy, based on the classic factor model – the main workhorse of modern active investing. To build a multifactor portfolio, we estimate and apply the five-factor Fama-French model to all stocks in the S&P 500 during the period from January 2010 through December 2019. We set the target variable as the future one-month excess stock return. The five-factor model is then:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta 1_{,i} (R_{M,t} - R_{f,t}) + \beta_{2,i} SMB_{i,t} + \beta_{3,i} HML_{i,t} + \beta_{4,i} WML_{i,t} + \beta_{5,i} SMV_{i,t} + \varepsilon_{i,t} \quad (2)$$

The factors are the market return, size (SMB), value (HML), momentum (WML) and volatility (SMV). Our proxy for the risk-free rate is the three-month treasury bill rate. We construct the factor portfolios according to the well-known Fama and French methodology (see e.g., Fama and French (2016)), forming the size and value factors yearly at the end of June. The momentum and volatility factors are formed at the beginning of each month on the basis of 12-month price momentum and 60-day volatility. The factor model is estimated on a rolling basis, using the previous two years of data.

Our alternative data investment strategy is based on the new factor: the sentiment. We construct the monthly long-short sentiment factor (on the basis of the monthly changes in sentiment) in the same way as the Fama-French volatility factor. The sentiment factor can be used on its own, or can be combined with the other factors.

Table 1 shows the correlations between the traditional and the sentiment factors for the US market.

	MKT	Size	Value	Momentum	Volatility	Sentiment
MKT	1.00					
Size	0.24	1.00				
Value	0.29	0.17	1.00			
Momentum	-0.25	-0.12	-0.15	1.00		
Volatility	-0.78	-0.27	-0.36	0.42	1.00	
Sentiment	-0.08	-0.01	-0.08	0.02	0.09	1.00

Table 1: Correlations between factor portfolio returns created by Fama and French methodology.

The sentiment factor shows low correlations to the other factors. This indicates that sentiment is almost orthogonal to other factors and hence carries additional information not found in the well-known investment factors.

We base our portfolio construction on the so-called alpha-momentum strategy by Hühn and Scholz (2018). This strategy ranks stocks based on their alphas from the factor model (2). The target variable is the lagged monthly or quarterly stock return (depending on the investor profile). The long-only investment portfolio consists of the top 10 stocks per sector, in terms of their past alphas. The sectors are then given weights identical to the sector weights in S&P 500. If short sales are allowed, the short part of the portfolio is similar to the long portfolio, only it contains the 10 lowest ranking stocks per sector. For long-short portfolios, we consider several leverage options, explained below.

The portfolios for each investor profile are evaluated: those built with the traditional factors and with the sentiment factor. We also include portfolios that combine the traditional and sentiment factors. Recall that portfolios for different investor preferences differ by their rebalancing frequency and presence of short positions.

In our backtests, we do not directly report transaction costs. This is because our main comparison is to the multifactor strategy, which is an active strategy, also incurring transaction costs. So instead, we compare the turnover of (active) strategies with and without sentiment data. It turns out (see Results section below) that the turnovers of strategies with and without sentiment are practically the same (in fact, the sentiment-based strategy leads to lower turnover in some cases). So any slippage or deterioration of performance due to the transaction costs would be the same, whether one uses alternative data or not.

5. Results

Tables 2 and 3 show the performance of the strategies in the period 2010 to 2019 for long-only investors (Investors 1 and 2).

	S&P 500	Factor model	Sentiment only	Factor model	+ Sentiment
Return (%)	11.21	13.79	13.46		14.28
Volatility (%)	15.55	17.32	17.09		17.41
Tracking error (%)		4.27	4.54		4.23
Sharpe ratio	0.70	0.78	0.77		0.80
Information ratio		0.61	0.49		0.73
Turnover	0	48.50	50.16		50.98

Table 2: Portfolio performance measures for Investor 1.

	S&P 500	Factor model	Sentiment only	Factor model	+ Sentiment
Return (%)	11.21	13.49	13.35		13.24
Volatility (%)	15.55	17.27	17.06		17.28
Tracking error (%)		4.29	4.54		4.23
Sharpe ratio	0.70	0.76	0.76		0.75
Information ratio		0.53	0.47		0.48
Turnover	0	59.83	59.31		61.10

Table 3: Portfolio performance measures for Investor 2.

We see that for both investor types, the sentiment-only strategy is on par with the full-blown multifactor strategy, both providing between 2% and 2.5% excess return over the S&P benchmark. This is remarkable, since it shows that the same investment results as for multifactor investing can be achieved using only this alternative data source. For quarterly rebalancing, adding sentiment to the multifactor strategy further increases the return by 0.5% and improves the Sharpe ratio.

	S&P 500	Factor model	Sentiment
Return (%)	11.21	13.86	13.77
Volatility (%)	15.55	17.29	17.08
Tracking error (%)		4.86	5.23
Sharpe ratio	0.70	0.78	0.79
Information ratio		0.55	0.49
Turnover	0	63.80	63.18

Table 4: Portfolio performance measures for Investor 3, 10 % short.

For long-short portfolios (Investor 3), we compare the multifactor to the sentiment-only strategy. We look at different leverage possibilities, parameterised by the percentage of the total portfolio value that can be in short sales (for example, 100% leverage means that the value of long positions is twice the value of short ones, i.e., the weight on long sub-portfolio is 200% and on short sub-portfolio -100%). Tables 4 to 7 show the strategies' performance for four different levels of leverage.

	S&P 500	Factor model	Sentiment
Return (%)	11.21	14.41	14.39
Volatility (%)	15.55	17.41	17.20
Tracking error (%)		5.85	6.39
Sharpe ratio	0.70	0.81	0.82
Information ratio		0.55	0.50
Turnover	0	72.90	72.25

Table 5: Portfolio performance measures for Investor 3, 25 % short.

	S&P 500	Factor model	Sentiment
Return (%)	11.21	15.28	15.38
Volatility (%)	15.55	17.81	17.65
Tracking error (%)		7.86	8.5
Sharpe ratio	0.70	0.84	0.85
Information ratio		0.53	0.49
Turnover	0	88.27	87.38

Table 6: Portfolio performance measures for Investor 3, 50 % short.

The returns and Sharpe ratios in the four tables above show that the true difference between performance of different investment types lies in the ability to do short sales (rather than in the rebalancing frequency), as the excess return over the benchmark is now between 3% and 5.5%. The short positions contribute significantly to the excess return over the benchmark (for both factor and sentiment-based strategies), and the sentiment-based strategy always outperforms the factor strategy in these cases.

	S&P 500	Factor model	Sentiment
Return (%)	11.21	16.88	17.18
Volatility (%)	15.55	19.28	19.41
Tracking error (%)		11.66	12.90
Sharpe ratio	0.70	0.86	0.87
Information ratio		0.49	0.46
Turnover	0	118.82	117.63

Table 7: Portfolio performance measures for Investor 3, 100 % short.

In all, we see that alternative data can provide significant value for investors – in the case of media sentiment, we are able to generate as good investment results as when using the full-blown multifactor model. However, the question remains: what is the absolute monetary value of such an alternative data set? In our final section, we calculate this according to the methodology presented in Section 3.

6. The value of alternative data: the case for sentiment

Tables 2 to 7 above summarise the performance of investment strategies with and without sentiment as the alternative data source. To determine the value of this data set, we proceed as in Section 3, by calculating the GH1 measure. Recall that this measure gives the overperformance of the strategy using alternative data, for the same level of risk as the (de)leveraged benchmark.

To calculate the GH1 measure, first we need to choose the benchmark. For the purposes of illustration, we have selected the S&P 500 index. For Investor 1, the sentiment-only strategy GH1 metrics with respect to the S&P 500 is equal to 1% per annum, and the combination of sentiment and multifactor strategy provides a GH1 measure of 1.73%. Translated into monetary units, this means that, for a small equity fund of USD100 million under management, running the simple sentiment strategy can provide USD1 million profit per year, and running it in combination with the multifactor strategy, USD1.73 million.

Note that, for Investor 2, the addition of the sentiment factor to the existing multifactor strategy does not seem to provide additional value. However, recall that this is only the case for this particular use of sentiment, and other ways of using sentiment, for example as a risk overlay or in sector rotation strategies, can still provide significant monetary benefits, as we have shown in our recent white paper 'Adding sentiment to multifactor equity strategies'. Even if the power of sentiment is purely in risk reduction, then the example in Section 3 still shows that there are significant monetary benefits to sentiment use. These can be calculated (and realised) by leveraging the sentiment-based strategy to achieve above-benchmark returns.

For long-short investors with different leverage constraints, the GH1 measures, with respect to the S&P 500 index, are shown in Table 8 below.

Leverage value created (in mln USD)	GH1 measure
10%	1.45%
25%	2%
50%	2.7%
100%	3.2%

Table 8: GH1 measures for a long-short investor (Investor 3).

Calculating the GH1 measure with respect to the S&P 500 as the benchmark is, however, not a fair comparison, since we are comparing an active strategy with a passive one, whose turnover (and hence, transaction costs) are zero.

If, on the other hand, we use the multifactor strategy as the benchmark, then the addition of the sentiment, for example for Investor 1, provides us with a GH1 measure of 0.42%, which, for the same size fund, still gives USD420,000 in additional pure profits per year – recall that there are no additional transaction costs. If the data cost USD100,000 per year (this could also include the costs of processing and additional infrastructure), then the recovery factor from this investment (i.e., the multiplier of the cost to profit) is 4.2. Clearly, for a larger fund, this recovery factor can be many times larger, since it grows linearly with AUM.

So, the numbers such as the GH1 measure and the recovery factor with respect to data cost (which are essentially back-of-the-envelope calculations, once backtesting is done), can make a compelling business case for including alternative data in your investment process.

5. Concluding remarks

Data-driven investment processes can significantly benefit from alternative data sources, especially those containing new information not yet incorporated into traditional investment signals. News and media sentiment is one such alternative data source, that is currently at the forefront of quant investors' attention. The potential added value of such data comes from its ability to instantly capture new price-moving information about companies and other assets, providing an information edge to asset managers that use such data.

In this paper we developed a simple yet illuminating framework for assessing the added value of alternative data with respect to any benchmark, be it a passive index or an active, factor-based investment strategy. We illustrated the steps of this framework – from the choice of the investment strategy to backtesting to the valuation of the data – on the example of the alternative data set of News Analytics, and have shown that such a data set can earn back many multiples of its cost.

We have deliberately chosen simple investment strategies for their robustness: their added value is due to the inclusion of the alternative data and not due to additional model complexity. However, quantitative investment specialists can undoubtedly develop their own, more complex investment models and expand the use of alternative data beyond simple examples we presented here.

References

Borovkova, S., and Mahakena, D. (2015). News, volatility and jumps: the case of natural gas futures, quantitative finance. *Quantiative Finance*, 1217-1242.

Borovkova, S., and Xiaobo, D. (2015). News sentiment, factor models and abnormal stock returns. *Factor Models and Abnormal Stock Returns* (25 November, 2015).

Fama, E.F., and French, K.R. (2016). Dissecting anomalies with a five-factor model. *The Review of Financial Studies*, 29(1), 69–103.

Hühn, H.L., and Scholz, H. (2018). Alpha momentum and price momentum. *International Journal of Financial Studies*, 6(2), 49.

About LSEG News Analytics:

Using cutting-edge Natural Language Processing, News Analytics provides measurement of company sentiment, relevance and novelty of text, as well as other valuable metadata. Supporting use cases such as trade signal detection for quant investment strategies, back-testing and market surveillance, News Analytics covers historical data dating back to 2003 and over 32,000 companies. So that when it comes to making sense of the news, and what it means for your business, News Analytics does the hard work for you. Visit www.lseg.com/en/financial-news-services/machine-readable-news#t-textual-news-services

Contact us to discover more: LSEG.com/en/contact-sales

Contact

Dr. Svetlana Borovkova
svetlana.borovkova@probability.nl