Finding Value in Earnings Transcripts Data with AlphaSense

Vinesh Jha
ExtractAlpha
info@extractalpha.com

John Blaine
Will Montague
AlphaSense

In this research note, written as a collaboration between ExtractAlpha and AlphaSense, we examine sentiment changes in earnings call transcripts and find that portfolios which are long stocks with improving transcript tone and short stocks with decreasing transcript tone tend to outperform. The outperformance is not explained by exposures to common risk and return factors such as Momentum and EPS Revisions, and is additive to a baseline quantitative strategy.

Introduction

Historically, quantitative investors based their strategies on a handful of types of data sources: financial statements; market data; sell side earnings estimates; and structured data from financial filings, such as insider trades. As conventional structured numerical data has become increasingly commoditized among institutional investors, the possibility of gaining an investment edge using data sets with wide adoption among systematic investors has decreased. As a result, quants have turned their attention to differentiating themselves along a variety of dimensions, including an increased focus on sophisticated modeling and on execution efficiency.

There has also been an industry wide push into alternative sources of data. Unstructured text data has attracted a particular amount of attention, in part due to its sheer volume, which dwarfs that of structured numerical data. Data vendors have used sentiment analysis and natural language processing to convert text, from a variety of sources – news, blogs and microblogs, financial filings, earnings transcripts, and so on – into structured data which can be ingested by a systematic investment process.

Filings and transcripts are particularly appealing because there is no ambiguity concerning their *relevance*. For example, a Tweet may obliquely mention a product or company but may not be really about it; and a news story may mention several companies but may not reflect the sentiment in that news story as pertains to a particular name. Filings and transcripts, by contrast, are decidedly about the company in question.

For a given text document, one can infer its *sentiment*, which is typically described as the degree to which the document reflects positively or negatively on the company. For example, if an earnings transcript contains a disproportionate number of terms like “shortfall” and “decline” then it is reasonable to think that its sentiment is negative. To quantify sentiment, we need to do several things:

- Determine which words or phrases are positive or negative; this is called building a *dictionary*
- Combine the positive and negative word frequencies into the sentiment, or *tone*, of a document
- Determine whether this tone is *significant* relative to other documents or stocks

Finally, we need to determine the *value relevance* of a document with a significant positive or negative tone. Does the publication of a document with high or increasing tone generally predict outperformance or underperformance of the stock? Does it do so in a way that is not captured by other common risk or return factors?

In this paper we review some of our findings on the value relevance of sentiment from text documents, with a particular emphasis on earnings call transcripts. The research was a collaboration between AlphaSense and ExtractAlpha, a research firm specializing in innovative data and quantitative analytics for institutional investors

Financial document types

Academic work on the text analysis of financial documents has covered filings such as 10K's and 10Q's (Loughran and McDonald, 2011) as well as earnings transcripts (Henry, 2008); Kearney and Liu (2014) provides an excellent survey. AlphaSense enables users to search across these documents and all other financial filings.

For many document types, however, inferring sentiment is problematic. The vast majority of annual and quarterly financial statements (10K's and 10Q's) consist of numerical data and boilerplate risk statements which may throw off false positives in a sentiment analysis; the management discussion and analysis (MD&A) section, which is most relevant to our needs, is a small portion. Furthermore, quarterly discussions tend to be fairly terse.

Forms 8K, pertaining to significant corporate developments, are interesting and abundant, but many of them are purely factual and may not include anything relating to sentiment, so we are likely to get more noise than signal.

Transcripts, on the other hand, are more spontaneous and less scripted than financial filings, particularly in the question and answer section. In a live setting, management has little choice but to respond to a question that has been asked, and their extemporaneous replies often reveal their sentiment. Therefore in this paper we have chosen to use the transcripts of earnings calls to better capture the sentiment of management in response to specific items which shareholders care about – and, in fact, the sentiment embedded in the questions themselves. As noted in several papers (Loughran and McDonald, 2011; Henry and Leone, 2009), a dictionary needs to be relevant to the style of language used in a particular document, and AlphaSense has built positive and negative word dictionaries calibrated to terms which capture sentiment in earnings call transcripts.

Transcript data in this paper is sourced from Thomson Reuters and is available starting in 2005. In separate analysis, we have confirmed that our results are qualitatively similar using FactSet as our transcript data source. AlphaSense is agnostic to the underlying dataset; data derived from either one can be provided for analysis and testing.

In this paper we restrict our analysis to U.S. stocks which are sufficiently large and liquid for institutional investors. Our test period runs from 2005 through 2014. We require stocks to meet the following minima:

- market capitalization of $500mm USD
- average daily trading volume of $1mm USD
- nominal price of $4

These universe requirements are updated monthly. We impose a lag sufficient to account for the latency between an earnings call and its transcription, and then the dissemination of that transcript to end users. Lags of up to 9 days have been observed in both data sources but the large majority of transcripts are available from the data providers (Thomson Reuters and FactSet) within three business days of the earnings call; AlphaSense does not introduce any additional latency beyond what is imposed by the underlying data suppliers. We have observed that our results are fairly robust to the length of the lag assumed.

Within this investible universe of approximately 1400 to 1800 names, the majority have earnings transcripts once per quarter, and per the chart below, most of that universe has had an earnings transcript in any given 3 month period. There are two common reasons for a firm within the universe to be missing transcript data: first, because the company does not do an earnings call (for example, Berkshire Hathaway); and second, because of some gaps which exist in the treatment of the identifiers of defunct companies in the underlying database, supplied by Thomson Reuters. These gaps do not seem to follow any particular patterns, but users should be aware of them. We observed similar gaps with an alternative data set supplied by FactSet.

Figure 1 shows the number of firms in our total universe, and the number with an available transcript in the trailing calendar quarter, over time since the beginning of our sample.

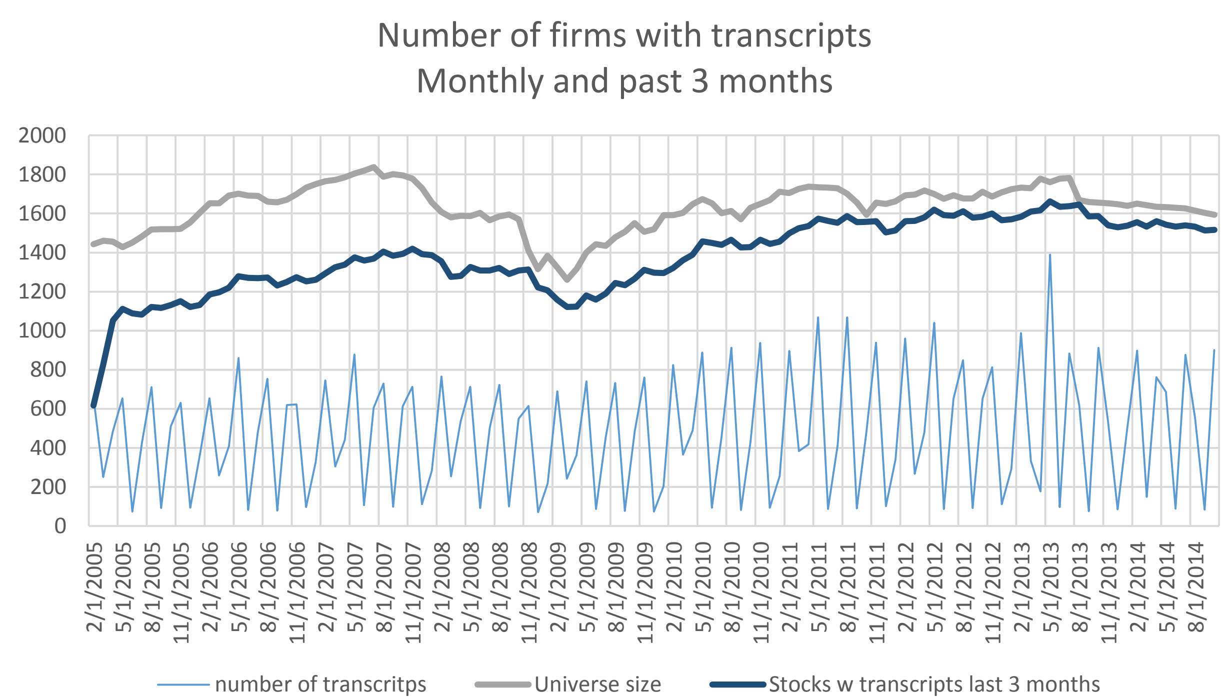

Figure 1. Number of firms in the universe and number with transcripts in the last 1 and 3 months.

When measuring daily stock returns, we use returns that have been residualized to common risk factors including industry, size, and value. This allows us to disentangle the effects of risk, including market risk – an important factor given the dramatic swings in the market during the time period in question – from idiosyncratic stock-level alpha. See the Appendix for details on the residualization of returns, which follows the tone change residualization.

Measuring transcript tone

Academic literature on text analysis of financial documents has focused on a variety of attributes of the documents in question: their sentiment, length, and readability (Feldman et al, 2009; Lee, 2012; Loughran and McDonald, 2013). Greater length and complexity are assumed to be proxies for management obfuscation. While this is an interesting topic, here we focus more on a direct measure of text sentiment, namely *tone.* Here we address how we use AlphaSense to measure the tone of earnings transcripts.

Text data has become interesting to investors, but accessing this data in an efficient manner has been a challenge to date. AlphaSense has addressed the challenge by:

- **Collecting and categorizing** historical and current text documents relating to a company in a centralized database, including financial filings of all types as well as earnings call transcripts
- Allowing for **efficient searching** of words and phrases across all documents, or across documents of particular types and in particular date ranges
- Designing **custom dictionaries and topic searches** of interest to investors
- Allowing for **flexible combinations of words and phrases**
- Providing **summary statistics** at the document level, in a clean flat file structure, including the number of hits for the words and phrases in question and the complete length of a document

An AlphaSense user can design a set of database *queries* to assess the characteristics of text documents. These queries combine a set of keywords and phrases (for example, the words from our negative dictionary) with document filters, including document type, timeframe, and industry. One can then save this query configuration and use it as the basis to run an historical batch job. This job outputs a flat file containing security and document identifiers, date stamps, and summary statistics for each document. These statistics include the document length and the count of the number of times the words or phrases from the query appear in the document.

To measure tone, a user can set up AlphaSense queries with keyword lists consisting of positive and negative dictionaries. Previous academic work has often used generic word lists, such as the Harvard Psychological Dictionary, which do not necessarily pertain to negative and positive tone in a financial context; for example, “beat” may sound like a negative in some contexts but is typically positive in a financial context. More recent academic studies such as Loughran and McDonald (2011) have addressed this issue by creating dictionaries which are designed with the language of financial documents in mind. AlphaSense analysts have extended this work by reading thousands of earnings call transcripts and curating custom dictionaries of positive and negative words and phrases. We then use these dictionaries, in combination with document type filters, to build queries to assess tone.

We can use the resulting summary statistics to define the tone of a document as simply:

$$tone = \frac{\#\, sentences\ with\ positive\ words - \#\, sentences\ with\ negative\ words}{\#\, sentences}$$

Schematically, our process is as shown in Figure 2.

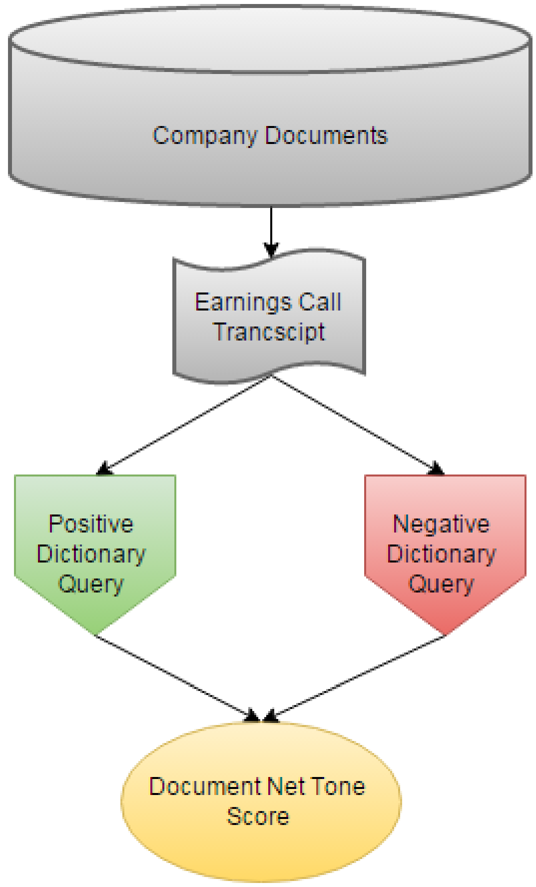

Figure 2. Schematic diagram of transcript tone calculation.

Measuring tone over time

The widest range of values tone takes over our sample is between -0.04 and +0.62, with the majority falling between +0.10 and +0.35 and the median around 0.25. That is, for every 100 sentences in an earnings call, there will be 25 more sentences containing positive words than sentences containing negative words. It is rare for the number of sentences with negative words in a transcript to exceed the number of sentences with positive words. The values have been fairly constant over time, with a noticeable dip during the financial crisis. In Figure 3 we plot the 10th, 25th, 50th (median), 75th, and 90th percentiles of tone by month.

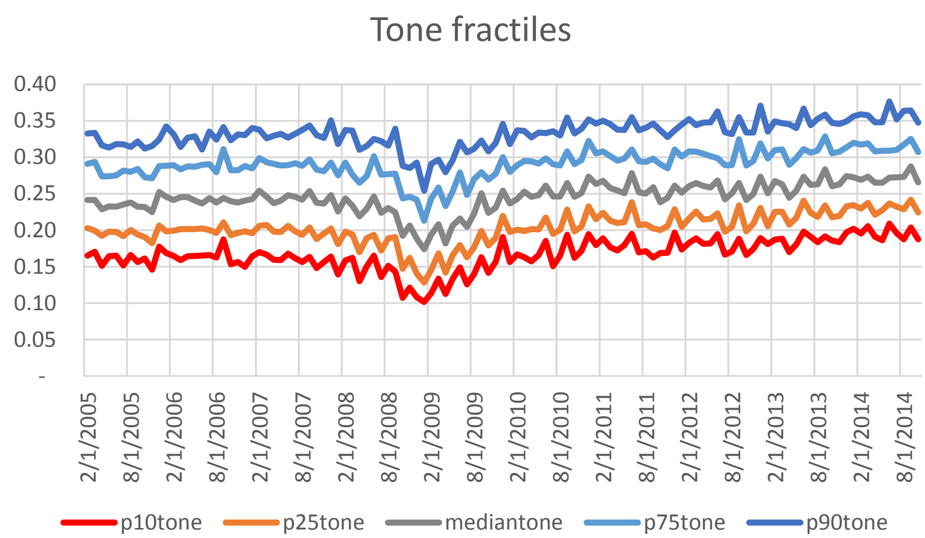

Figure 3. Average tone value over time.

At the individual company level, tone is also very stable. In Figure 4 we divide tone into 10 groups each month, and then plot, on average, what group those companies' transcripts fall into in future quarters.

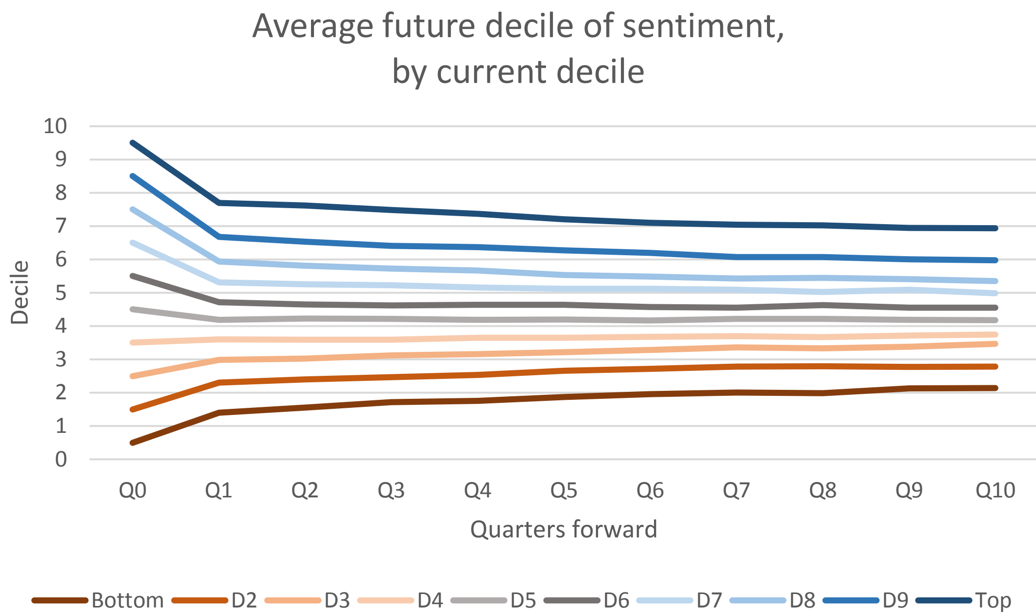

Figure 4. Mean reversion of tone deciles.

Although there is some mean reversion in tone apparent in the more extreme deciles for the first quarter, the tone levels do not converge even after ten quarters. As an alternate measure, the quarter over quarter autocorrelation in tone is 0.69. So it would appear that transcript tone does not vary significantly over time, perhaps owing to the general upbeat or downbeat outlook of management, as the company's transcript participants generally do not change drastically from quarter to quarter.

As a result of this stability, we choose to focus on the *change* in tone: how upbeat or downbeat is the current earnings call, relative to prior earnings calls *for that same company*? The level of prior tone is well known to the market, but innovations or changes in tone represent surprises relative to the market's expectations, similarly to how the level of quarterly earnings is only interesting when measured relative to past earnings or expectations. We measure tone change as simply the current tone, minus the average tone in transcripts during the prior two years.

Details on the tone change methodology, and a "recipe" for following our analysis precisely, can be found in the Appendix.

Sentiment results

We start with an event study that examines the cumulative residual returns from an earnings event, conditioned on whether the earnings call's tone was higher or lower than the tone of recent quarters. The event study methodology allows us to partition earnings events by positive and negative tone as well as other factors such as earnings surprise, and to understand the alpha realization that follows these events. Figure 5 shows the cumulative residual returns by tone change for various tone change magnitudes.

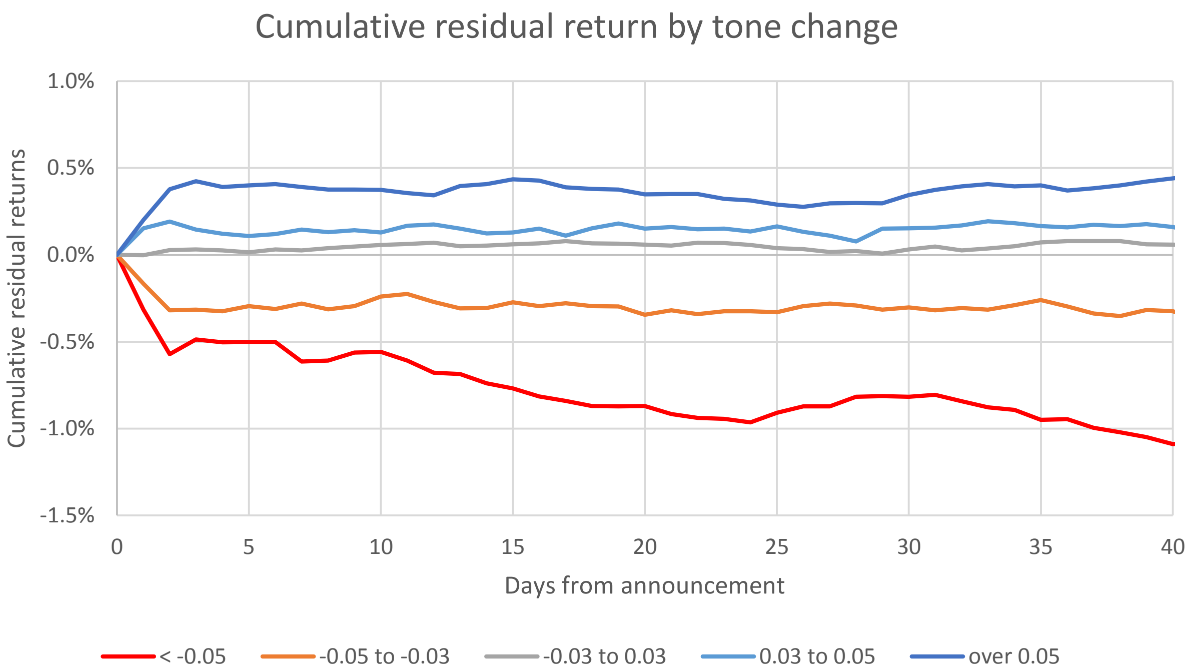

Figure 5. Residual returns from date of transcript, by tone change.

We first observe that on the day of the announcement there is a large event return which lines up with the degree of tone change, with stocks with increasing sentiment showing higher event returns, as we'd expect. Without knowing the timestamp of the earnings call itself, or the latency of the call transcription, however, we cannot assume that we can capture that day's return, or even the subsequent day's return since the call may have happened after the close on the announcement date. After the second day, the drift to tone changes is fairly modest except in the case of very negative tone changes, which continue to drift downwards for two months after the announcement.

We next examine whether tone is incrementally valuable given that market participants know the direction of the earnings surprise. In other words, if a company beats expectations, is there additional value in knowing that the tone of the earnings call also became more or less positive relative to past calls? We measure the earnings surprise magnitude as the stock's residual returns from the day prior to the announcement to the day after the announcement. The advantage of using this measure rather than a beat or miss relative to sell side consensus numbers is that we get a true gauge of market reaction relative to expectations, free of the well-known biases in sell side earnings forecasts. Earnings surprise, by this measure, is not independent of tone; there is a significant 0.18 correlation between earnings surprise and change in sentiment. Figure 6 plots the effects of tone change conditioned on surprise.

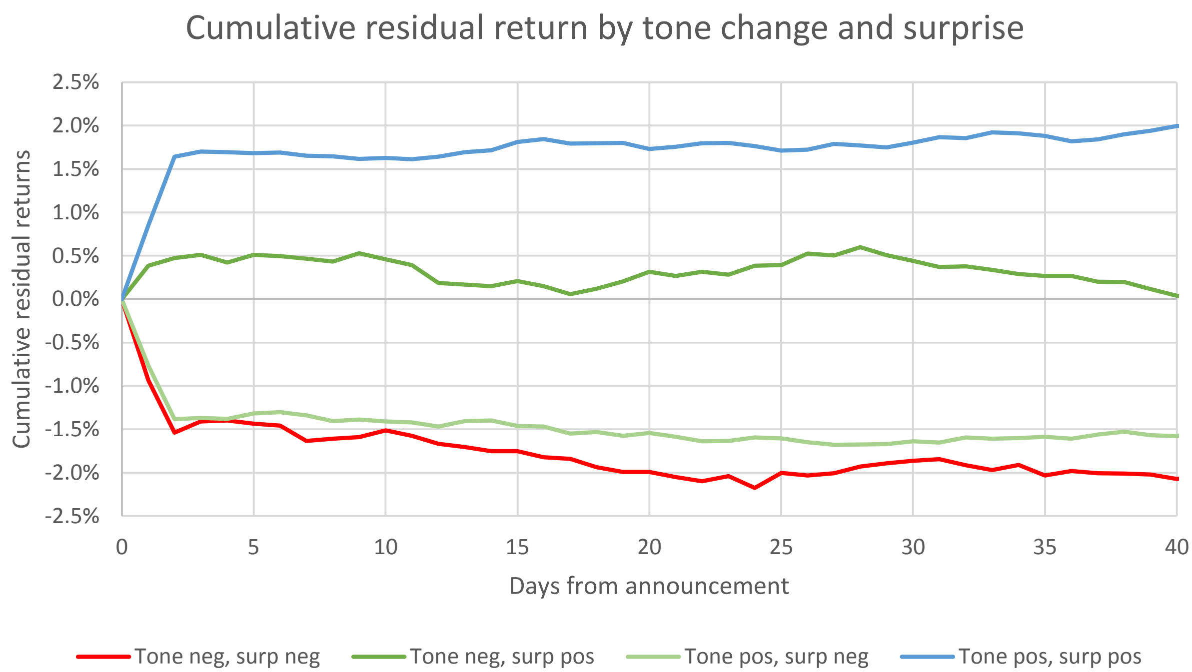

Figure 6. Residual returns from date of transcript, by tone change and EPS surprise.

Here we can see that the surprise effect dominates, but that tone change has incremental power in explaining contemporaneous event returns: negative-surprise stocks with increasing sentiment underperform less, and positive-surprise stocks with decreasing sentiment outperform less.

In order to examine the effects of tone in a portfolio context, we need to move from an events data set to one which compares stocks cross-sectionally. We do so by employing a decay function, which puts more weight on recent events. For two stocks which have had increases in sentiment relative to prior earnings calls, we want to put more weight on the stock with a more recent earnings event. To do this we simply take our change in tone measure and multiply it by a linear decay factor, which gives full weight to events which have just occurred, no weight to events which occurred 45 calendar day ago, and linearly in between. This creates a stock score which dynamically highlights both the recency and the magnitude of the tone change.

We need to assume some latency between the announcement and data availability, so that we can fairly rank stocks according to our decay measure. We employ a lag of 10 weekdays to the earnings call data, which is extremely conservative; the longest observed lag since 2013 has been nine days.

In Table 1 we show the annualized returns to a market neutral portfolio formed by a long portfolio of the top 10% of stocks according to our decayed tone change variable, and a short portfolio of the bottom 10%, rebalanced daily. The portfolio has a daily turnover of 6% each way.

Table 1. Annualized long/short decile return to tone change factor.

	Annualized return	Sharpe ratio
Overall	**3.2%**	**0.48**
Large cap (top 500)	4.0%	0.45
Mid cap (next 500)	9.2%	1.24
Small cap (the rest)	-0.8%	-0.08

These returns are modest, but particularly intriguing for large and mid-cap stocks; small caps do not exhibit significant returns. In separate work we have observed that the autocorrelation of tone from quarter to quarter is much lower for small cap stocks, given the higher variability of the contents of their earnings calls; as such, innovations in tone are more likely to be driven by noise rather than true change in the outlook of call participants.

Controlling for risk

We next examine the degree to which these results are driven by common risk factors. We compute eight classic risk factors, along the lines of frequently used commercial risk models: Yield, Volatility, Momentum, Size, Value, Growth, and Leverage. Each risk factor is normalized to have a cross-sectional mean of zero and a standard deviation of one. We also look at a common return factor with a natural relationship with tone change: sell side EPS estimate revisions. With two exceptions – Momentum and EPS Revisions – these factors appear to be unrelated to change in tone.

In Figure 7 we show the average value of the two factors of interest, by decile of tone change from low (more pessimistic tone) to high (more optimistic tone). Momentum is defined as the 12 month total return of the stock, but excluding the most recent month to account for reversal effects; EPS Revisions is defined as the 30-day change in the consensus quarterly sell side earnings forecast. Note that EPS revisions data is available only from 2007 in our sample, whereas other data starts in 2006.

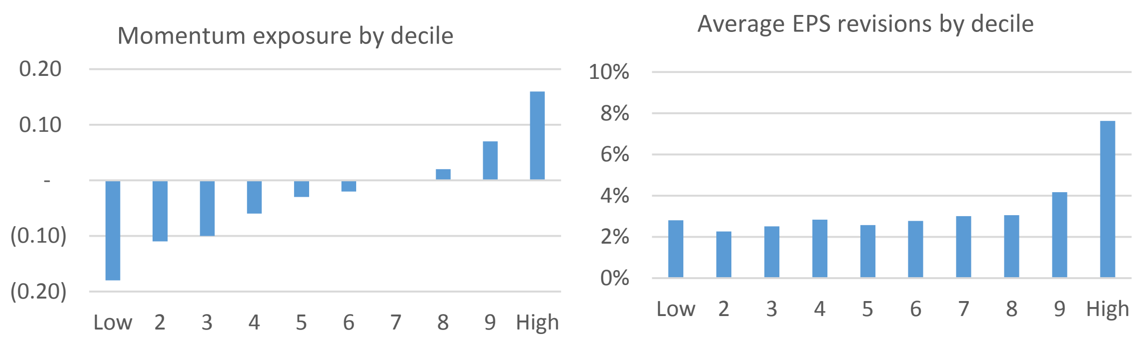

Figure 7. Average risk exposures by tone change decile for two risk factors.

Both factors have a fairly linear relationship to tone change: stocks with increasing tone have also experienced increasing stock prices over the last year, and increasing earnings estimates over the last month. This is not surprising, as the three effects demonstrate that the stock market, sell side analysts, and management are all more upbeat about the firm's future earnings prospects.

Next, we'd like to know whether tone change is telling us anything that we can't glean from factors such as momentum and EPS Revisions alone. To do so, we residualize the tone change factor, stripping out the effects of the risk factors and also the EPS Revisions factor. (Details of the residualization can be found in the Appendix). The result is a purified tone change variable that by construction is uncorrelated to these factors. We can then construct portfolios based on the residualized variable, with the results as shown in Table 2.

Table 2. Annualized long/short decile return to tone change factor, before and after residualization.

	Raw		Residualized	
	Annualized return	Sharpe ratio	Annualized return	Sharpe ratio
Overall	**3.2%**	**0.48**	**3.1%**	**0.62**
Large cap (top 500)	4.0%	0.45	6.9%	1.02
Mid cap (next 500)	9.2%	1.24	4.0%	0.53
Small cap (the rest)	-0.8%	-0.08	-0.6%	-0.06

Our overall results are fairly robust to the residualization, with slightly stronger Sharpe ratios overall and returns of similar magnitude – in fact the results are stronger for large caps after residualization. It seems that our results are independent of Momentum, EPS Revisions, and the other risk factors.

Combining with other factors

While intriguing, these return levels are modest and suggest that tone change variables should be used in conjunction with other factors. Therefore we explore whether layering tone change on top of a simple baseline quantitative strategy can improve the results. We begin with a simple strategy which gives a 2/3^{rds} weight to Value factors (P/E, P/S, P/B) and a 1/3rd weight to an Estimate Revisions factor (30 day change in FQ1 sell side consensus). The baseline strategy is residualized to those factors which it does not explicitly bet on: volatility, momentum, size, growth, and leverage. We do not residualize to the Yield and Growth factors, which have high correlations to our Value factor.

We then adjust this baseline strategy by introducing our residualized tone change variable at a 20% weight, and adjusting the baseline dates pro rata downwards, recognizing that investment managers are unlikely to shift their strategies heavily into a new factor. In Table 3 we show the decile portfolio results before and after the addition of the text-based factor.

Table 3. Annualized long/short decile return to simple quant strategy, with and without tone change.

	Value + Revisions		**With tone change**	
	Annualized return	Sharpe ratio	Annualized return	Sharpe ratio
Overall	**4.2%**	**0.65**	**5.3%**	**0.84**
Large cap (top 500)	2.0%	0.24	4.8%	0.57
Mid cap (next 500)	5.6%	0.56	8.1%	0.84
Small cap (the rest)	1.5%	0.12	1.9%	0.15

As expected, the tone change variables add the most value within large and mid-cap stocks, with the results dramatically improved in the large cap group (albeit off of a lower baseline). The turnover of the strategy increases slightly, from 5% per day to 6% per day, when introducing the tone change variable. Although this is a simplistic implementation, it demonstrates that text-based sentiment change can be additive to an existing quant strategy for liquid U.S. names.

Topic searches

Our initial analysis assessed the value of transcripts as it relates to the tone of management. However, the examination of text data and the tools that AlphaSense provides do not limit themselves to only scrutinizing management sentiment. Other potential avenues of exploration include filtering documents for language which references a particular topic or concept.

Topic filters represent lists of synonymous financial terms and phrases, and can be used to filter for text passages that refer to a particular concept. For example, using the 'Profit margin' topic enables you to filter to language containing reference to terms or phrases such as:

- Margin
- Gross margin
- % of profit
- Income as a % of revenue
- NIM
- Net interest margin

These topic filters can be used in conjunction with each other, or with search strings, to effectively isolate pieces of matching language within documents. We can then use these search results to assess the relative frequency of discussion for a particular document, for all of a company's documents over time, or for an entire groups of companies.

For example, we could extend our prior analysis by combining a 'Profit Margin' topic filter with the negative or positive search string. This would enable us to count the number of times the two concepts occur in conjunction in a particular document, and enable us to answer questions like:

- Is a company more or less positive about profit margins in their most recent call compared to historically?
- Is one company historically more or less positive about their margins than another company?
- Has there been a sudden spike in negativity around profitability for an entire industry?

Schematically testing a Topic Filter with Transcript Tone would look as in Figure 8.

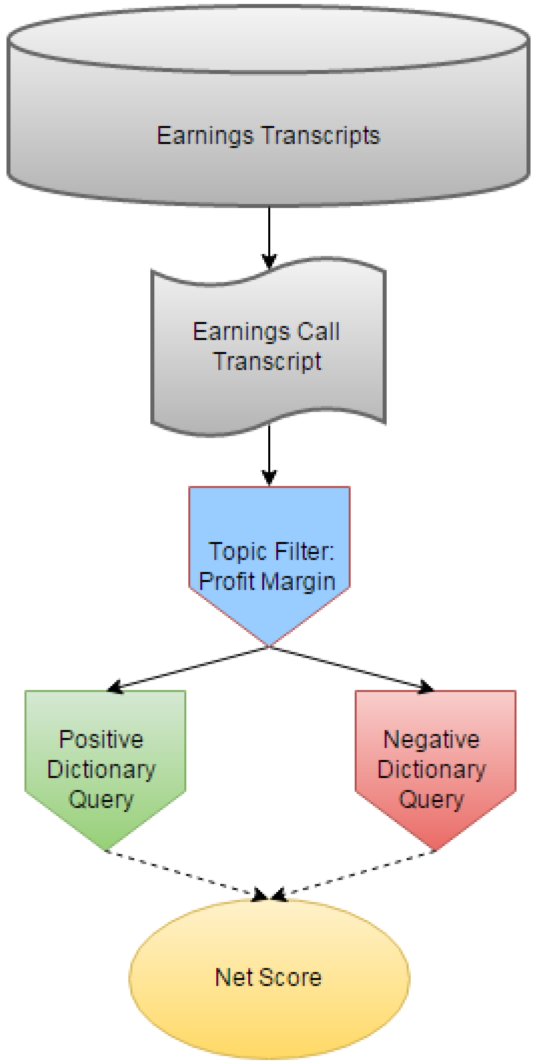

Figure 8. Schematic diagram of tone change applied to a topic filter.

Finally, users are not restricted to the predefined Topic Searches. An AlphaSense user can conduct tests that allow them to measure the impact of any set of text strings of their own choosing, and on the document types which they think would be most relevant.

Summary

AlphaSense provides an open-ended tool which enables investment managers, including quants, to perform topic- and sentiment-based analytics on financial text documents in ways which were previously inefficient. Here we demonstrate that analyzing the tone of earnings call transcripts in particular can lead to outperformance which is not captured by other common risk and return factors. Many further refinements and extensions of this work are possible by using AlphaSense text based queries.

Appendix: building sentiment scores

Here we provide a detailed recipe for building sentiment score and measuring change in sentiment at the stock level.

We first set up two queries using the AlphaSense interactive tool: a negative words query and a positive words query. Each is run on all available dates (2005 to the present) and using earnings transcripts from Thomson Reuters. We run a third query which covers all documents, in order to include the rare transcript which may have no negative or positive words. The dictionaries used for these queries are available to all AlphaSense users. The user is alerted when the historical files are available for download.

The results are presented in a pipe-delimited flat file, with identifiers for the security (ticker, ISIN) and the document (a unique document identifier). Since a particular company may have multiple associated securities, we denormalize the data set to have a document ID/ISIN primary key. We then merge the positive and negative word count data, and the all-document query, by document ID and ISIN. We set the number of negative or positive word counts to zero if they do not appear in our query results. We take the tone as mentioned in the main body of the paper as

$$tone = \frac{\#\ sentences\ with\ positive\ words - \#\ sentences\ with\ negative\ words}{\#\ sentences}$$

We then match those document ID's to our investible universe according to whether the stock meets our liquidity, market cap, and nominal price requirements on the date of the report.

Prior to AlphaSense's live date on 19 April 2013, we did not have accurate datestamps which would tell us exactly when the transcript data was available to users. Therefore prior to that date we apply a lag of 10 weekdays from the earnings call date; this is more conservative than any observed lag since mid-2013. After that date, we use the available date as provided by AlphaSense.

To calculate the benchmark level of *tone* for a particular security, we take the average tone value for any transcript produced in the last 750 calendar days, corresponding to two years plus a short buffer to account for variations in earnings dates. Our tone change variable for a document is simply the difference between *tone* for that document and the moving average of *tone*.

Next we need to transform the document-level tone change score into a stock-level score. We do this by taking into account the recency of the available date for the most recent transcript. We assign a full weight of 1 to a transcript which was available today (using our availability assumptions detailed above) and a zero weight to a 45 calendar day old transcript, and linearly in between. To account for the rare cases in which we have multiple recent transcripts, or have limited history for transcripts, we divide by the total number of transcripts available in the last year. (Normally this number will be identically four for most companies, and therefore the scalar won't make a difference). This gives us our raw stock-level tone change variable.

We are concerned with a tone change measure, and with returns, that are residualized, or adjusted, for the effects of common risk factors. We start with a typical model inspired by the Arbitrage Pricing Theory (Roll and Ross, 1980) to explain the cross section of tone change within our investible universe:

$$TC_S = a + b_1 F_{1,S} + b_2 F_{2,S} + \ldots + b_N F_{N,S} + \varepsilon_S$$

Here, TC_S is the stock's tone change variable on a particular day, and F_1 through F_N are the factor loadings for that stock from a risk model. We specify the risk factors as binary industry variables (a stock is either a member of an industry or it isn't), as well as nine "fundamental" factors which are estimated from company level descriptors:

Beta: Beta of the stock to the S&P 500 index
Value: P/E, P/B, P/S
Growth: Year over year earnings growth, sales growth
Yield: Dividend yield
Leverage: Debt/Equity ratio
Momentum: 11 month stock return, lagged 1 month
Volatility: Daily standard deviation of returns
Size: Total assets, log of Market Capitalization
EPS Revisions: 30-day change in the current quarter (FQ1) Wall Street consensus EPS forecast

A regression is run each day across the stocks in the investible universe. This gives us a residual tone change variable, i.e., the portion of that stock's tone change that is not explained by the risk factors. This is the final variable over which we run our analysis.

References

Feldman, R., Govindaraj, S., Livnat, J., and Segal, B. (2009). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies* 14 (4), 915-953

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *The Journal of Business Communication* 45 (4), 363–407.

Henry, E., and Leone, A. J. (2009). Measuring qualitative information in capital markets research. Available at SSRN. http://ssrn.com/abstract=1470807

Kearney, C., and Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis* 33 (2014) 171–185

Lee, Y.-J. (2012). The Effect of Quarterly Report Readability on Information Efficiency of Stock Prices. *Contemporary Accounting Research* 29, 1137–1170

Loughran, T., and McDonald, B. (2011). When is a liability not a liability? *Journal of Finance* 66, 35–65.

Loughran, T., and McDonald, B. (2014). Measuring Readability in Financial Disclosures. *Journal of Finance* 69 (4), 1643-1671

Roll, R., and Ross, S. (1980). An empirical investigation of the arbitrage pricing theory. *Journal of Finance* 35 (5), 1073–1103