GLOBAL

Quantamentals

I just called to say I'm bullish

Global analyst conference calls and stock returns

The main goal of this report is to analyse the *soft information* that is revealed during earnings conference calls in addition to the hard numbers such as revenues and earnings. Academic research has often argued that investors focus excessively on whether the reported numbers beat or miss expectations. If this is true, then we can expect the soft information which emerges from the call to be incorporated slowly in prices, which would lead to return predictability.

Mining conference call transcripts

This report applies textual analysis to earnings conference calls. We build on our *Quantamentals: Positively Persuasive* note expanding the coverage to the global universe and applying new methods of sentiment classification. Our tools allow us to process a large amount of text and measure whether the discussion among analysts and company management had a bullish or a bearish tone.

Tone matters

Our results suggest that tone does predict returns beyond the immediate price reaction that characterises the first few days after the event. In particular, the effect is present over horizons of one to three months and even longer. We emphasise that this is *not* a high frequency signal that requires frequent trading with holding periods of just a few days.

How can investors use this signal?

First, we argue that fundamental analysts can use our signals as a screening variable to identify, among thousands of earnings conference calls, the ones with most bullish or bearish tone – or the ones which display a marked difference between direction of the surprise and tone. Second, our backtest suggests that a simple strategy based on the change in tone from the previous call has more attractive risk-adjusted returns than a traditional earnings surprise strategy.

Global Stock Screen

Worst EPS Surprise - Worst Q&A Tone Change

Name	Fiscal Period	Call Date	Sector	EPS Surprise	Q&A Tone Change	Rec
Lundin Petroleum AB	Q4 2014	04/02/2015	Energy	-1.1	-8.2%	
Assurant, Inc.	Q4 2014	13/02/2015	Financials	-0.5	-5.0%	UP
International Consolidated Airlines Group SA	Q4 2014	27/02/2015	Industrials	-0.6	-4.7%	
Rexel SA	Q4 2014	12/02/2015	Industrials	0.0	-4.7%	
Norwegian Cruise Line Holdings Ltd.	Q4 2014	18/02/2015	Consumer Discretionary	0.0	-4.6%	
Noble Corp. Plc	Q4 2014	05/02/2015	Energy	0.0	-4.6%	
Natixis SA	Q4 2014	19/02/2015	Financials	-0.1	-4.6%	
Fossil Group, Inc.	Q4 2014	17/02/2015	Consumer Discretionary	0.0	-4.2%	N
Royal Bank of Scotland Group Plc	Q4 2014	26/02/2015	Financials	-0.6	-4.2%	UP
Randgold Resources Ltd.	Q4 2014	09/02/2015	Materials	-0.3	-4.2%	N

Best EPS Surprise - Best Q&A Tone Change

Ticker	Fiscal Period	Call Date	Sector	EPS Surprise	Q&A Tone Change	Rec
Airbus Group NV	Q4 2014	27/02/2015	Industrials	0.2	5.5%	
Boliden AB	Q4 2014	12/02/2015	Materials	0.2	5.2%	
Agnico Eagle Mines Ltd.	Q4 2014	12/02/2015	Materials	3.4	5.1%	N
Liberty Media Corp.	Q4 2014	25/02/2015	Consumer Discretionary	0.3	4.6%	OP
Host Hotels & Resorts, Inc.	Q4 2014	19/02/2015	Financials	0.2	4.4%	
Total System Services, Inc.	Q4 2014	27/01/2015	Information Technology	0.1	4.0%	N
Bristol-Myers Squibb Co.	Q4 2014	27/01/2015	Health Care	0.1	3.9%	
Moody's Corp.	Q4 2014	06/02/2015	Financials	0.2	3.8%	
PartnerRe Ltd.	Q4 2014	05/02/2015	Financials	0.2	3.6%	N
ONEOK, Inc.	Q4 2014	24/02/2015	Energy	0.2	3.6%	

Source: FactSet, Macquarie Research, April 2015

Inside

I just called to say I'm bullish 2
Academic insights 3
The data: Global conference call transcripts 5
Measuring tone 9
Measuring complexity 13
Results 14
Conclusion 39
References 40
Appendices 42

Analyst(s)
Macquarie Capital (Europe) Limited
Giuliano De Rossi , PhD
+44 20 3037 1997 giuliano.derossi@macquarie.com
Jakub Kolodziej
+44 20 3037 2016 jakub.kolodziej@macquarie.com
Gurvinder Brar
+44 20 3037 4036 gurvinder.brar@macquarie.com
Ioan Mirciov , PhD
+44 20 3037 1939 ioan.mirciov@macquarie.com
Macquarie Securities (Australia) Limited
John Conomos, CFA
+61 2 8232 5157 john.conomos@macquarie.com
Werner Fortmann
+612 82379642 werner.fortmann@macquarie.com
Jeremy Lamplough
+612 8232 1060 jeremy.lamplough@macquarie.com
Thomas Guillet
+61 2 8232 8931 thomas.guillet@macquarie.com
Macquarie Capital Securities Limited
Woei Chan
+852 3922 1421 woei.chan@macquarie.com
Anthony Ng
+852 3922 1561 anthony.ng@macquarie.com
Jason Zhang
+852 3922 1168 jason.zhang-msg@macquarie.com
Per Gullberg, CFA
+852 3922 1478 per.gullberg@macquarie.com
Macquarie First South Securities (Pty) Ltd
Josiah Rudolph, FRM
+27 11 583 2210 josiah.rudolph@macquarie.com

20 April 2015

I just called to say I'm bullish

Do analysts add value?

This report explores the question whether analysts help investors unveil important information about a company by participating in conference calls with the company's management. In particular, we argue that the Q&A session contains, in addition to the hard numbers announced when a company publishes its results, a wealth of *soft information* which can be used to predict returns.

The main explanation for this effect is related to investor inattention. More and more evidence available from academic studies suggests that investors focus excessively on earnings and on whether the reported numbers beat or miss expectations. If this is true, then we can expect the soft information which emerges from the call to be incorporated slowly in prices, thereby creating the conditions for return predictability.

Mining conference call transcripts

We access a large number of conference call transcripts for global companies which are available to us as individual documents in PDF format. The methodology described in the next sections allows us to measure the *tone* of the discussion between analysts and management. In practice, by using simple algorithms we try to mimic a human reader who forms an impression on whether the general tone of the call was bullish or bearish.

Tone matters

Our results suggest that there is a relation between sentiment and future stock returns, even after correcting for the well know effect of earnings surprises. Suppose that two companies have both beaten expectations but one has shown a marked improvement in analyst call sentiment, the other a deterioration. Our analysis finds that, on average, the company with improved conference call tone tends to outperform over a period of three calendar months after the call has taken place. The effect is robust and significant both statistically and economically.

How can investors use this signal?

Sentiment, as calculated from the language used in earnings conference calls, can be used as a screening variable. In this way it is possible to identify, among thousands of earnings conference calls, the ones with most bullish or bearish tone. Alternatively, it could be used to identify the companies which display a marked difference between direction of the surprise (i.e. whether it missed or beat expectations) and tone (positive or negative). An earnings beat with bearish conference call tone might signal that the good news is over and there may be a turning point in the near future.

Moreover, our backtest suggests that a simple long-short strategy based on the change in tone from the previous conference call generates alpha. The strategy is built by sorting first on earnings surprises and then, among positive surprises, selecting those with positive tone. Our work indicates that such a strategy can be viewed as an enhancement of the traditional earnings surprise effect in that it results in a significant improvement in risk-adjusted returns.

Academic insights

Vast recent literature on text mining in Finance

Textual analysis in Finance has become by now a vast academic topic, as evidenced by as many as three recent surveys: Li (2010a), Kearney and Liu (2014) and Loughran and McDonald (2015).

Many different categories of financial texts have been used in natural language processing: news, message board posts and social media, financial disclosures, press releases, conference call transcripts, IPO prospectuses and even SEC comment letters (Ryans, 2014).

Typically the goal is to predict returns

In most of the available literature the main goal is to predict earnings or abnormal stock returns at least since the early papers like Tetlock et al. (2008). However, several contributions focus on other aspects, e.g. Borochin et al. (2015), Wang et al. (2013) and Kogan et al. (2009) try to measure risk from textual disclosures. Buehlmaier and Zechner (2014) try to predict if a merger offer will be completed successfully using a sample of 130,000 articles.

Several papers focus on conference call sentiment and stock returns…

Turning to the subject of this report, a number of academic papers have analysed systematically the transcripts of conference calls where companies comment on their results and take questions from analysts. The main result is that tone predicts excess returns (McKay Price et al., 2012; Chen et al., 2014; Davis et al., 2014; Druz et al., 2015), particularly the tone of Q&A sessions for non dividend paying companies (McKay Price et al. 2012). In most cases the authors focus on relatively short time horizons, of a few days at most. In contrast, our goal is to find a predictive signal that works over longer horizons.

…complexity of the language used by management…

In addition, Bushee et al. (2014) found that complexity in management discussions decreases the effect of the disclosures on returns. In a related paper, Cicon (2014) argued that when the language of management changes between the Discussion and the Q&A session the firm tends to underperform. Firms focusing more on the short term during the call tend to underperform (Brochet et al., 2014). The importance of tone and complexity in company disclosures is also highlighted in a series of papers on press releases and financial disclosures. An early example is Henry (2008) who tried to examine how tone and complexity metrics affect investors' reaction to the earnings press releases.

… or on the contributions of the individual analysts

Three recent papers focus on the language and level of participation of the individual analysts. Chen et al. (2015) argue that a complete lack of questions from the audience is a very negative signal with strong implications for subsequent price performance. Mayew et al. (2013) identified the questions asked by individual sell-side analysts and investigated the relation between number of questions and forecast accuracy. Their main conclusion is that analysts who participate actively in the call are more likely to have superior information. Druz et al. (2015) argue that experienced analysts adjust their forecasts in line with the predictive content of managerial tone. In contrast, inexperienced analysts tend to overreact when the tone of the management discussion is unexpectedly negative and underreact when the tone used by the management in answering analysts' questions is surprisingly negative.

Chen et al. (2013) analyse how the tone of the Q&A session varies with the time of day. They find that tone becomes increasingly negative with time and attribute this effect to accumulating mental and physical fatigue. The deterioration in tone is shown to have negative effects on abnormal returns.

Larcker and Zalkolyukina (2012) devise a model to detect deceptive statements

Larcker and Zalkolyukina (2012) propose a classification model to detect deceptive statements made my management during a conference call. The model is calibrated on a set of companies that had significant financial restatements after the call due to serious accounting issues. This allows the authors to identify the typical features of deceptive managerial language, including references to general knowledge and lack of references to shareholder value and value creation.

A simple bag of words method is typically used…

All the papers about conference calls mentioned above, without exception, use a simple *bag of words* approach to process text. As we shall explain in more detail later, the approach consists of counting how many positive and how many negative words appear in the text.

… even though more sophisticated methods are available

However, more sophisticated techniques than the 'bag of words' approach are available in the literature. Examples of such methods that have been used in financial applications are: Jegadeesh and Wu (2013), the naïve Bayes approach (Antweiler and Frank, 2004; Das and Chen, 2007; Li, 2010b; Huang et al, 2014; Ryans, 2014; Buehlmaier and Zechner, 2014; Buehlmaier and Whited, 2014), the support vector regression approach (Antweiler and Frank, 2004; Kogan et al. 2009). Neural networks have also been used (Heston and Sinha, 2014). A survey of the methodological aspects can be found in Kearney and Liu (2014). One of the goals of our analysis is to understand whether a statistical approach to text mining can lead to better results compared to the typical methods employed in the existing work. Loughran and McDonald (2015) argue that there is currently no compelling evidence that more refined measures of sentiment deliver more accurate results.

Previous Macquarie quant papers using textual analysis

The Macquarie quant team used text mining techniques in the past to analyse company disclosures in the US stock market. Our report *Quantamentals: Camouflaged in Complexity, February 2013*, used changes in complexity, measured from 10-K reports, to predict stock returns. In *Quantamentals: Positively Persuasive, May 2013*, we analysed 6,000 quarterly conference call transcripts for US companies. In *Quantamentals: A surprising tone, July 2014*, we considered earnings press releases of Russell 3000 companies since 2004. Finally, in *Quantamentals: How are you really feeling? November 2014*, we scanned the front page of single stock Macquarie reports.

The data: Global conference call transcripts

Data Source

We obtain transcripts of global analyst calls from Factset

We obtained conference call transcripts in PDF format from FactSet using FactSet's API, which allows batch file download.

All transcripts available in English for the period from Q4 2001 to Q2 2014 – over 58,000 files – were downloaded. We eliminated duplicates and retained only the corrected version if more than one version of the same transcript was available. Our sample was restricted to the historical constituents of our broad global developed countries universe in order to avoid look-ahead biases.

Only calls on earnings announcements are retained

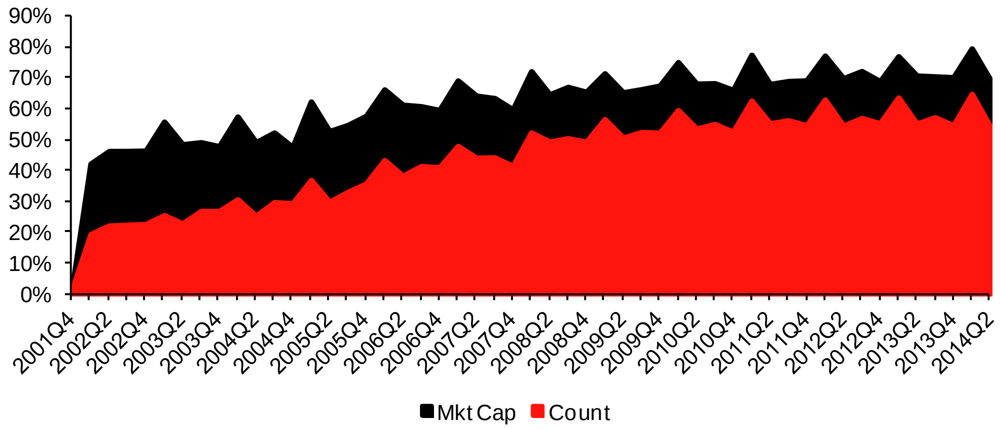
We selected only earnings calls (dropping any update on trading results, merger announcements etc.) and matched them with EPS surprises from the I/B/E/S database. At the end of the whole procedure we are left with roughly 38,000 transcripts for which I/B/E/S estimates are available.

Coverage

About 70% of our global universe is covered

In Fig 1 we show the global universe coverage by number of stocks and market capitalization. The coverage improves over time and stabilizes in the most recent periods around 55% and 70% for count and market capitalization, respectively. The "zigzag" pattern on the coverage graphs is caused by differences in analyst calls' frequency. Most American large cap companies hold analyst calls quarterly while in other regions many companies follow a semi-annual or even an annual earnings call schedule.

Fig 1 Percentage of the Global Universe Covered

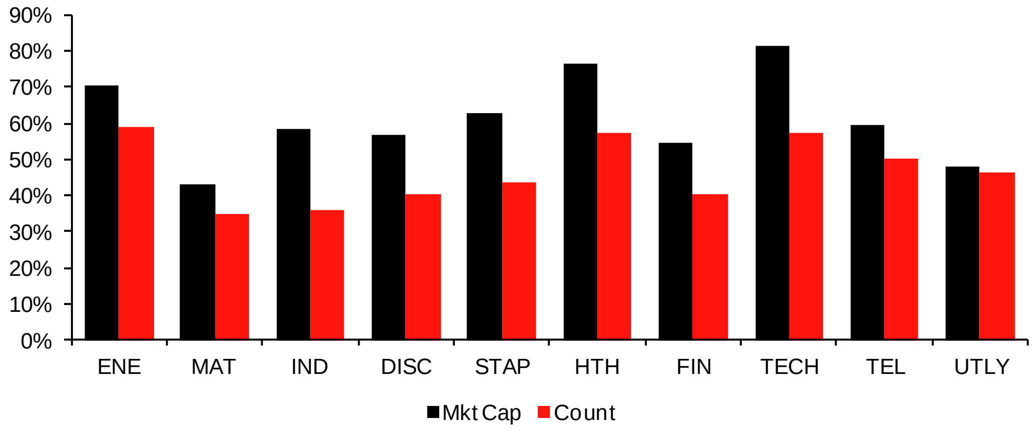


Source: FactSet, Macquarie Research, April 2015

No obvious sector biases in the coverage

In Fig 2 we report the average distribution of the sample across industries. We present it relative to sector share in the universe over time. In general our sample appears to be relatively balanced with the average coverage based on market capitalization exceeding 40% for each sector. In our sample Materials and Industrials are underrepresented and Technology, Energy and Healthcare sectors overrepresented. These deviations are more pronounced for market capitalization weighted coverage.

Fig 2 Average Percentage of Sector Covered in the Global Universe



Source: FactSet, Macquarie Research, April 2015.
Key: ENE = Energy, MAT = Materials, IND = Industrials, DISC = Consumer Discretionary, STAP = Consumer Staples, HLTH = Healthcare, TECH = Information Technology, TEL = Telecommunication Services .

Coverage is more extensive in North America and lower in the Asia Pacific region

In Fig 3 and Fig 4 we break down our sample by three regions and show the proportion of the universe covered based on the number of stocks (Fig 3) and on market capitalization (Fig 4).

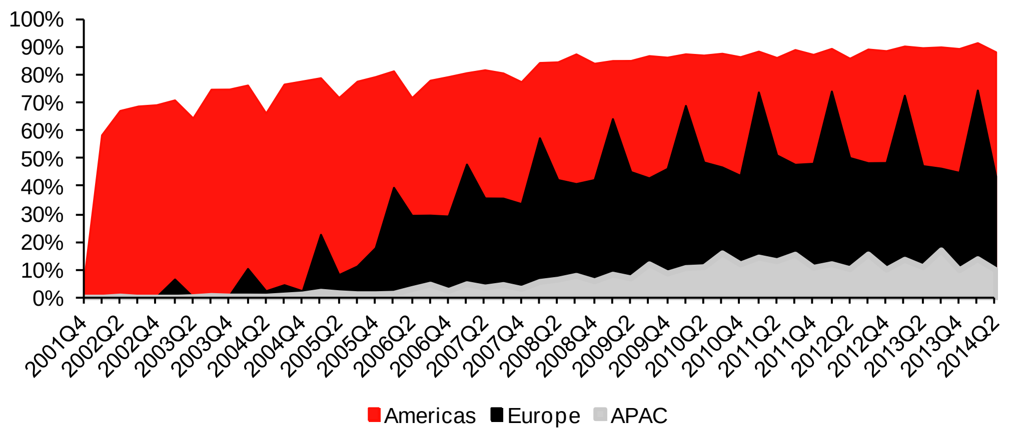
Our sample is dominated by the North America region in terms of the number of reporting companies per quarter. The average per quarter for North America is 535 companies compared to 168 and 32 for Europe and APAC, respectively.

The average coverage for North America is over 80% based on the stock number and around 90% based on the market capitalization. It is also very stable across quarters as most of companies hold quarterly calls.

Europe with over 8,500 calls has an average coverage of only 42% based on the market capitalization. Nevertheless the coverage increases significantly after 2005. In Europe, the “zigzag” pattern is the most pronounced with a difference of over 20 percentage points between the coverage available for the end of the year reporting compared to end of other quarters. The coverage for the end of the year reporting exceeds 70% after 2007 based on the market capitalization.

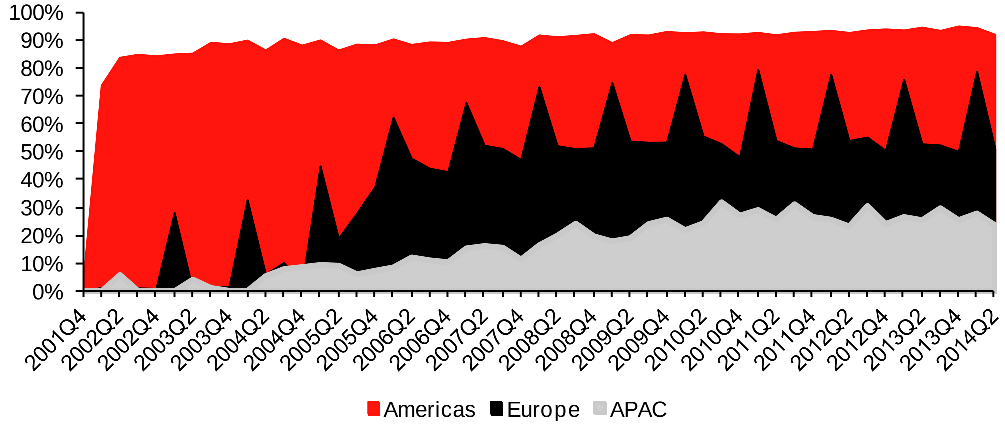
APAC has the worst coverage among regions and the market capitalization based coverage exceeding 20% only after year 2007.

Fig 3 Percentage of Global Universe Covered by Region (By Number of Stocks)



Source: FactSet, Macquarie Research, April 2015

Fig 4 Percentage of Global Universe Covered by Region (By Market Capitalization)



Source: FactSet, Macquarie Research, April 2015

Any sample selection bias is likely, if anything, to affect negatively our results

Before we proceed further, we shall comment on potential biases in the sample. FactSet provided transcripts are likely subject to the survivorship bias, as some documents might be dropped from the database for delisted companies. Moreover FactSet limits its event coverage using market capitalization cut-off points. We believe that both of these limitations would rather work against us and lead to underestimation and not to overestimation of the abnormal returns, as companies in our sample are arguably the most efficient – largest and most successful companies in the world, covered by several analysts.

Processing the data

Convert PDF to text

After obtaining the final list of transcript files, we convert PDF files to text. We use a proprietary PDF converter that yielded best results in our sample, but it is worth noting that free PDF converter software is available online.

Removing headers, footers, disclaimers

We clean each text file by eliminating any surviving headers and footers, removing numbers, words representing quantities (millions, billions etc.), hyperlinks and punctuation except for end of sentence marks.

Separating management discussion from O&A

Next, we break the text into three sections – management discussion, Q&A and disclaimer, using regular expressions on headlines between paragraphs. We drop all disclaimers and save section passages separately. Unfortunately, the converted text did not allow us to separate efficiently questions from answers in the Q&A section.

Computing sentiment and complexity scores

Finally, we move to computing sentiment and complexity scores. We divide text into sentences and tokenize them into words. In calculation of scores we use word count after dropping stop words. We apply an amended stop words list. We take a default list of English stop words available in the tm library and exclude all words that occur in all of the sentiment dictionaries.

For sentiment measures based on dictionaries we generate a term frequency matrix for the entire sample, in which columns represent words in the dictionary and rows represent text files. Naïve Bayes classifiers require generating a term matrix for each text file, as the classifier is trained to assess the sentiment of the sentence and not the entire text. In case of this term matrix, columns represent words used for classification and rows represent sentences. The other difference is that the term matrix used in training the classifier represents binary occurrences of words in sentences when for “bag of words” method is based on number of occurrences.

Software

We used several packages available in R for textual analysis and parallel computing

Despite the popularity of Python and NTLK library in natural language processing, we decided to code the entire project in R. We find standard R packages sufficient for all text processing tasks.

Natural Language processing can be very computationally expensive, but the problem lends itself naturally to parallel computing, as each text file might be processed separately. We use a standard R package for parallel computing on multiple CPU threads. Please contact the authors if you would like to discuss our approach.

Measuring tone

This section introduces the methods we employed in order to measure sentiment from each conference call transcript. The 'bag of words' approach is the most commonly used one – indeed all the academic papers cited on page 3 rely on it.

"Bag of words" approach

Counting positive and negative words

This method requires two word lists: one for positive and one for negative terms. We calculate the net tone of a document using the formula:

$$Net\ tone = \frac{Count\ of\ \ Positive\ Words - Count\ of\ \ Negative\ Words}{Count\ of\ \ All\ Words\ Excluding\ \ Stop\ Words}$$

Separating the effects of earnings surprises from the effects of tone

We also compute measures of abnormal tone to capture the residual component of sentiment metric isolated from the earning surprise effect. We avoid introducing a forward-looking bias by regressing, for each reporting date, all preceding calls' net tone on EPS surprise that we winsorise at -50% and 50%. The resulting slope and intercept estimates are then used to compute residuals for that date calls' tone metrics following the formula:

$$Net\ tone_{i,t} = \alpha_{t-1} + \beta_{t-1}\ EPS\ \ Surprise_{i,t} + Abnormal\ tone_{i,t}$$

Details on the variables used to measure EPS surprise can be found on page 24. As some managers might be inherently more optimistic about their companies than others, selection of calls with highest net tone may have undesirable bias towards companies with more optimistic managers.

Word classifications are subjective

From a purely methodological point of view, the main criticism of the bag of words approach concerns its subjectivity, as discussed by Li (2010b). Words are typically included in the two lists based on a subjective assessment which may be affected by previous empirical results. As a consequence, it may be difficult to obtain genuine out of sample evidence on the performance of this approach. Nevertheless, it is very easy to implement with little or no computational requirement.

Available dictionaries

Following our earlier work on Quantamentals: Positively Persuasive, we employ the following dictionaries:

- Loughran and McDonald (2011) – Positive and Negative Tone lists
- Diction – Positive and Negative Emotions lists
- Linguistic Inquiry and Word Count – Satisfaction, Inspiration, Praise, Hardship, Denial and Blame lists
- Combination of all dictionaries – Positive and Negative Tone lists excluding words in the intersection of both lists

Diction and LIWC word lists come from the off-the-shelf software packages and both have been extensively used in academic research to assess text sentiment based on word count. The disadvantage of these dictionaries is that they were designed to capture emotional states in broad ranges of text types, what might make them less suitable for specific purposes such as measuring tone of financial disclosures. For instance, Diction includes word 'weapon' and 'warfare' as negative, what would inadvertently lead to low scores for the entire defence industry. Loughran and McDonald's dictionary (2011) should avoid this problem of misclassification, as it was developed specifically for financial applications.

In the results section we compare the performance for all dictionaries individually and for the combination.

A naïve Bayes approach

Modelling the probability that a sentence is bullish given the words that appear in it

The naïve Bayes classifier is an alternative to "bag of words" approach. This probabilistic classifier is based on Bayes' theorem with an assumption of independence between the discriminant features, hence its "naïve" name. As a classifier, it is a supervised learning method and it requires a training set of documents or sentences.

Our approach consists of manually scoring a subsample of the text (a set of sentences) randomly derived from the entire sample. We then tokenize each sentence into single words and create a term matrix with rows representing sentences and columns representing all unique words in sentences. In the matrix we record the occurrence of words sentence by sentence using a binary TRUE/FALSE variable. In the last step we then build a simple model of the probability that, given that certain words appear in the sentence, the sentence is positive, negative or neutral.

Model setup

More formally, consider the random variables $D_1, \dots D_n$ which represent the classification of n documents or sentences, so that $D_i \in \{pos, neg, neu\}$ - i.e. each document can be classified as positive, negative or neutral.

Applying Bayes' theorem

In addition, consider a dictionary made up of *m* words. The random matrix *W* has elements $W_{1,1}, \dots W_{n,m}$ which take on the value TRUE if word *j* appears in document *i*, FALSE otherwise.

Applying Bayes' theorem, the probability that document D_1 is of class 'pos' given the random variables $W_{1,1}, \dots W_{1,m}$ (which identify which words are present and which ones are absent in the document) is described by the expression:

$$\Pr\left(D_1 = pos \mid W_{1,1}, \dots W_{1,m}\right) = \frac{\Pr\left(W_{1,1}, \dots W_{1,m} \mid D_1 = pos\right)\Pr\left(D_1 = pos\right)}{\sum_{e_C \in \{pos,neg,neu\}} \Pr\left(W_{1,1}, \dots W_{1,m} \mid D_1 = e_C\right)\Pr\left(D_1 = e_C\right)}$$

"Naïve" assumption that words occur independently

Assuming that the conditional probabilities that, if the document is of a certain class, each of the *m* words will appear in it are independent from each other, we can simplify the above expression to:

$$\Pr\left(D_1 = pos \mid W_{11}, \dots W_{1m}\right) = \frac{\Pr\left(D_1 = pos\right)\prod_j \Pr\left(W_{1j} \mid D_1 = pos\right)}{\sum_{e_C \in \{pos,neg,neut\}} \Pr\left(D_1 = e_C\right)\prod_j \Pr\left(W_{1j} \mid D_1 = e_C\right)}$$

The required probabilities can be estimated from the frequencies in the training set as follows.

Probabilities are estimated from sample frequencies

The marginal probability that a document is of class 'pos' is estimated as:

$$\hat{P}\left(D = pos\right) = \frac{\#\left(D_i = pos\right)}{n}, \quad i = 1, \dots, n$$

An estimate of the conditional probability that word j appears in a document, if the document is of class 'pos', is obtained as:

$$\hat{P}\left(W_j \mid D = pos\right) = \frac{\#\left(D_i = pos, W_{i,j}\right)}{\#\left(D_i = pos\right)}, \quad i = 1, \dots, n$$

All required frequencies can be calculated by summing occurrences in columns and rows of the term matrix. This makes the naïve Bayes classifier computationally inexpensive compared to other classification methods such as Support Vector Machine (SVM) or Neural Networks that require training of parameters.

If a sentence contains words that were not available in the set of m terms used to train the model then the missing words are ignored.

The sentence is assigned to the category that has the highest probability

Once we have estimated the required probabilities, we classify each sentence by choosing the category with highest conditional probability, formally

$$\underset{e_C \in \{pos,neg,neut\}}{argmax} \quad \hat{P}(D = e_C \mid W_1, \dots, W_m)$$

For a whole document we count the number of sentences classified as positive and the number of those classified as negative and compute the difference of the corresponding ratios. We refer to this difference as the net tone.

Naïve Bayes with Informative Priors

Typically the prior conditional probabilities are assumed to be equal across words

The Naïve Bayes classifier for discrete features can be interpreted as a Maximum A Posteriori (MAP) estimator of the marginal probabilities $\hat{P}(D = e_C)$ and the conditional probabilities $\hat{P}(W_j|D = e_C)$ for each $e_C \in \{pos, neg, neu\}$ and $j = 1, \ldots, m$. The Appendix contains a more detailed description of the argument. Implicitly, this amounts to imposing an uninformative prior on each conditional probability, i.e. for each word j the prior distributions of $\hat{P}(W_j|D = pos)$, $\hat{P}(W_j|D = neg)$ and $\hat{P}(W_j|D = neu)$ are the same.

We impose informative priors by incorporating information from the dictionaries

However, as we argue in the Appendix, it is well known that one can easily incorporate informative priors about marginal and conditional probabilities into the model. An example would be to set the priors based on one of the existing dictionaries. It is a simple way to incorporate the subjective belief that 'positive' words from the dictionary are more likely to appear in positive sentences than a negative ones.

This approach is computationally very straightforward as it corresponds to adding artificial sentences to the sample e.g. including a sentence with "positive" words from the dictionary and classifying it as positive. Once the new sample, augmented with artificial observations that reflect the prior, is obtained, the Naïve Bayes estimation proceeds exactly as in the standard case.

Using informative priors provides several advantages. It allows us to attribute some discriminant ability to words that have not appeared in the training set and would normally be completely omitted without using priors. This should help generalising models to very large and diverse text sets.

Moreover, using informative priors with predefined word lists might be seen as a form of regularisation for dictionaries, as words with low classification ability will be shrunk according to their probabilities in the training set. In the "bag of words" approach we implicitly assume that all words in the list have equal discriminant ability.

How many words? The mutual information criterion

Danger of overfitting if all available words are included in the model

As the training set with scored sentences has for obvious reasons a limited size, using all unique words in the set to train the Naïve Bayes classifier is likely to result in overfitting due to words with spurious classification ability. This statement might seem at odds with our discussion in the previous section, which advocated the use of a larger set of words. It is important to stress that the danger of overfitting is caused by the fact that many of the words available in the training set will only appear a few times and therefore we do not have enough information to draw conclusions on their predictive power.

Mutual information gives an indication of the usefulness of a word in classifying sentences

A simple way to avoid overfitting is to select only a subset of words for the analysis using a ranking for words based on mutual information. The mutual information between two random variables, one of which represents a classification and the other a binary feature, is defined as:

$$MI = \sum_{e_W \in \{True, False\}} \sum_{e_C \in \{Pos, Neg, Neu\}} \Pr(W = e_W, D = e_C) \log_2 \frac{\Pr(W = e_W, D = e_C)}{\Pr(W = e_W)\Pr(D = e_C)}$$

It is a summation over all possible combinations of the values of the two random variables, which involves joint probabilities (e.g. $Pr(W = True, D = Pos)$, the probability that a sentence is positive and the word is present) and marginal probabilities (e.g. $Pr(D = Neu)$,the probability of observing a neutral sentence). In practice we compute the sample counterpart of the quantity defined above by substituting sample frequencies for probabilities.

To gain some intuition on the definition, imagine a situation where the presence of a word does not provide any useful information on the likely classification of the sentence (*D*). In that case the argument of the logarithm is equal to one (because the joint probability at the numerator is equal to the product of the marginal probabilities at the denominator) and therefore the whole expression equals zero. In practice we compute the sample counterpart of the quantity defined above by substituting sample frequencies for probabilities.

We can keep the model parsimonious by selecting the words with highest MI

Mutual information provides an indication of the discriminant ability of a given feature. Its main limitation is that it is univariate, i.e. by using MI for the features selection we ignore interactions among them. This method also requires setting a minimal acceptable cut-off level for MI. A simple approach, described in the next section, consists of testing the predictive ability of a set of candidate models out of sample. Starting with a single feature model, we add one feature at a time by choosing the one with the highest MI among those that are not yet included in the model. The model which maximises predictive accuracy out of sample is then selected.

Measuring complexity

Language complexity can be measured quantitatively

In addition to tone, complexity is one of the measurable features of conference call language that have been used to extract predictive signals in the literature, for example by Bushee et al. (2014). Intuitively, the use of complex language by management could indicate that the speaker is reluctant to present the results clearly or answer the question directly.

In practice, we measure complexity in our data by using a range of simple measures:

- Total number of words
- Total number of sentences
- Average length of a sentence (number of words)
- Average number of syllables per sentence
- Average number of syllables per words
- Fraction of complex words (defined as words with more than 2 syllables)

A related concept is readability, which refers to the suitability of a text for different categories of readers. We adopted two alternative measures that are widely used in computational linguistics to compare the complexity of different textbooks.

Combining complexity scores into a readability measure

The Flesch reading ease statistic is defined as

$$206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right).$$

Complex language tends to be characterised by long sentences and long words, which result in relatively high values of the two ratios that appear in the formula. Since both ratios have negative weight, lower scores correspond to more complex (i.e. less readable) texts. A score between 0 and 30 is deemed appropriate for university graduates, while books meant for 13 to 15-years-old students should score between 60 and 70 and so on.

An alternative statistic maps directly complexity onto US school grade levels:

$$0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right)$$

In this case a higher scores indicates higher complexity. While we acknowledge that readability measures are designed to work on written text, rather than oral discussion, we calculate the statistic for both sessions of each conference call.

Results

Can the machine tell bulls from bears?

Assessing the performance of our classification methods

In this section we assess whether the classification methods described in the previous section can successfully identify positive and negative sentiment. This point is crucial as the potential for return predictability (and therefore alpha generation) depends on being able to score automatically a large number of documents.

Training the Naïve Bayes model

The naïve Bayes model is trained on 10,000 sentences

As detailed above, the Bayesian approach to classification requires a training set of sentences. To this end, we manually scored 10,000 sentences randomly derived from the entire sample - we randomize over text files and sentence locations in files. Each sentence is assigned a score of *positive*, *negative* or *neutral*.

The next step is to choose the set of features, i.e. in our case the set of words that can be used to predict the score of each sentence. We adopt two alternative solutions:

1) Using all words in the Loughran-McDonald (2011) dictionary
2) Estimating the optimal number of features by cross validation.

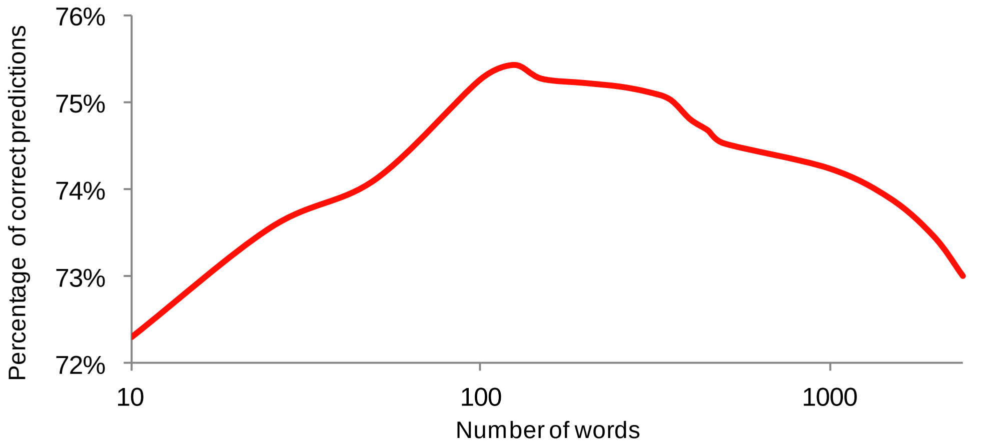
Choosing the words for our model

In order to implement solution 2) we divide the sample into five equal subsets and in each trial we select four of them for the training set and leave out one as pseudo-out-of-sample. In each trial we train a Bayes classifier on the training set and compute the percentage of correct predictions for the cross-validation set. By resampling five times the cross-validation sets cover all observations in the sample. We then sum the correct predictions for five cross-validation sets.

The whole procedure is repeated for different alternative lengths of the word list, from a minimum of one to a maximum of 2,424 (the total number of unique words that appear more than once). In each trial we simply select the n words with highest MI.

In Fig 5 we present the accuracy of predictions for the different number of words (log scale) included in the model corresponding to different levels of MI cut-off points. Our sample consists of 2,424 unique words, but the Naïve Bayes classifier reaches the best forecasting rate with only 125 words. The results show that training using all available words may lead to significant overfitting.

Fig 5 Optimal number of words for Naïve Bayes classifier



Source: FactSet, Macquarie Research, April 2015

Comparison of Classifier Performance

It pays to go Bayes!

In Fig 6 we compare the correct prediction rates for different classifiers calculated on our 10,000 sentence training set. It is worth clarifying that the 'count of words' approach is designed to obtain a sentiment score from a whole document, not to classify individual sentences into a discrete set of possible outcomes. However, in this section we derive a simple classification method based on word lists by just classifying as positive any sentence with positive net tone, as neutral any sentence with net tone of zero and as negative any sentence with negative net tone. In order to check the robustness of our results, we also consider a variant where the definition of neutral sentence is widened to include those with a net count between -1 and 1.

Cross validation analysis

Fig 6 shows that the Naïve Bayes classifier outperforms both "bag of words" methods and the naïve benchmarks in our training set. It proves also that the count of words methods that assume that all words have the same sentiment loadings perform poorly on the sample with short messages, as single word appearance in the sentence can drive the classification.

Fig 6 Comparison of Classifier Performance on the Training Set

Method	Calibration	Correct predictions	Proportion Correct
Naïve Bayes	Word list chosen in CV	7544	75.4%
Naïve Bayes	LM words	7347	73.5%
Count of words	LM words, neutral = (0)	6444	64.4%
Count of words	LM words, neutral = (-1,1)	6940	69.4%
Naïve	Naïve (all sentences are neutral)	6702	67.0%
Naïve	Pick random category	5163	51.6%

Source: FactSet, Macquarie Research, April 2015

In particular, the naïve Bayes model with optimised word list (first row in Fig 6) predicted correctly 90.9% of the neutral sentences, 51% of the bullish ones and just 23.9% of the bearish ones. Because neutral sentences make up 67% of our sample this translates in an overall proportion of 75.4% correct predictions. Bullish sentences are more frequent than bearish ones (24.5% and 8.5% respectively) and this affects the relative predictive accuracy.

In the next section we will investigate whether the Naïve Bayes classifier generalizes well to the entire sample and whether it is superior to the count of words approach in predicting abnormal returns.

Informative priors

A priori a positive word should be more likely to appear in positive than in negative sentences

We have argued in the previous section that it is possible to impose informative priors which reflect the information we have about the meaning of each word. A simple approach, detailed in the Appendix, consists of assuming that 'positive' words (according to one of the dictionaries) are more likely to appear in the sentence if we know that the sentence is bullish. In particular, we have imposed on the conditional probabilities a Dirichlet prior such that the median of the distribution is equal to:

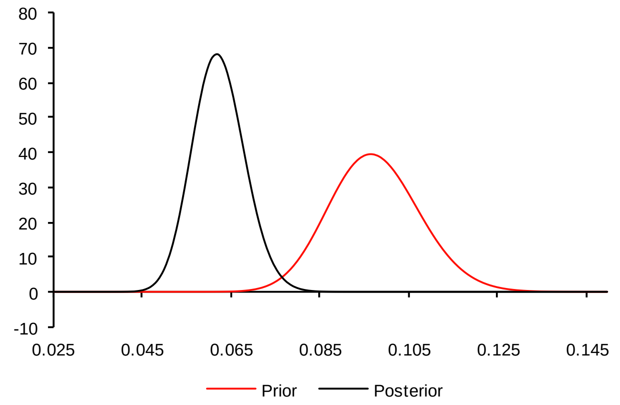
- the marginal probability times a constant $\kappa > 1$, when the word is positive and we condition on the sentence being positive or the word is negative and we condition on the sentence being negative
- the marginal probability divided by κ, when the word is positive and we condition on the sentence being negative or the word is negative and we condition on the sentence being positive.

Finally, the median of the prior distribution of the probability conditional on a neutral sentence is set equal to the marginal distribution.

Example: Is the word "question" negative?

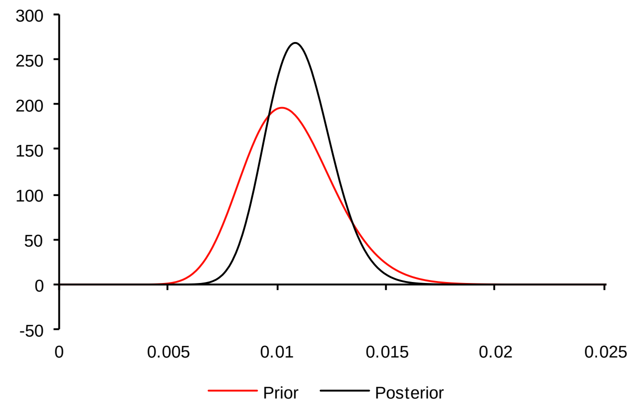
A good example of the shrinkage effect is a word "question" that appears in the Loughran McDonald (2011) dictionary as negative. We use the word list to impose informative priors in the Naive Bayes model as described above, for various values of κ. Fig 7 through Fig 9 depict the prior and posterior distributions of the conditional probabilities that the word *question* appears in a sentence (for $\kappa = 3$). The prior implies a relatively high probability to find the word in a negative sentence (Fig 7), which reflect the fact that the word *question* is included in the negative word list. However, by inspecting the posterior we conclude that the negativity of the word "question" is significantly "neutralized" compared to the prior, i.e. the distribution concentrates on lower values. Similarly, the posterior distribution implies a higher probability of finding the word in a bullish and, particularly, in a neutral sentence.

Fig 7 Probability that the word *question* appears conditional on the sentence being bearish



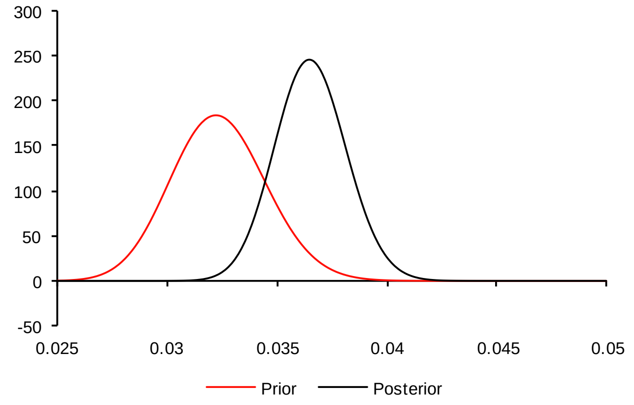
Source: Macquarie Research, April 2015

Fig 8 Probability that the word *question* appears conditional on the sentence being bullish



Source: Macquarie Research, April 2015

Fig 9 Probability that the word *question* appears conditional on the sentence being neutral



Source: Macquarie Research, April 2015

The tone in global conference calls

A first characterisation of the properties of the proposed tone measures can be obtained by analysing:

- Trends in tone measures over time,
- Correlations among the different tone measures,
- Correlations of the tone measures with earnings surprise
- Correlations of the tone measures with stock characteristics,
- In our analysis we compare results for levels as well as for changes in the tone measures.

Tone over time

The distribution of tone displays a consistent pattern over time, shifts negatively in 2008

We begin our analysis of the tone measures by looking at the sample distribution over time. In Fig 10 to Fig 15 we plot the percentiles of the cross sectional distribution of tone over time, using the entire global universe since 2001. The different charts refer to the six alternative measures discussed above. The pattern is very consistent across alternative approaches and it shows the dependence of overall tone on business cycle fluctuations: Sentiment improves after the end of the tech bubble, deteriorates significantly in 2008 as the credit crisis unfolds and recovers from 2009 onwards. The overall distribution appears remarkably stable over time.

The distribution of changes in tone

In Fig 16 to Fig 21 we present similar results for changes in tone. In this case the distribution is positively skewed during the positive phase of the business cycle and negatively skewed in periods of crisis, notably during the credit crisis. Dispersion in sentiment changes is high at the beginning of the sample period due to the limited availability of data. Overall, mean and median appear very stable over time and oscillating around zero, with only one sizable downtick during the financial crisis.

Fig 10 LM Net (MD) Level

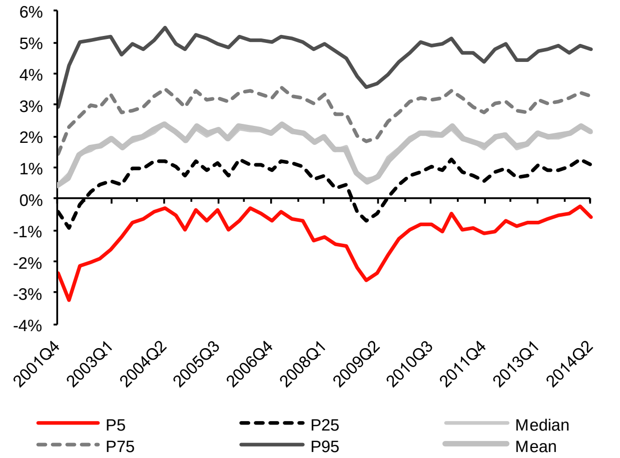


Fig 11 Diction Net (MD) Level

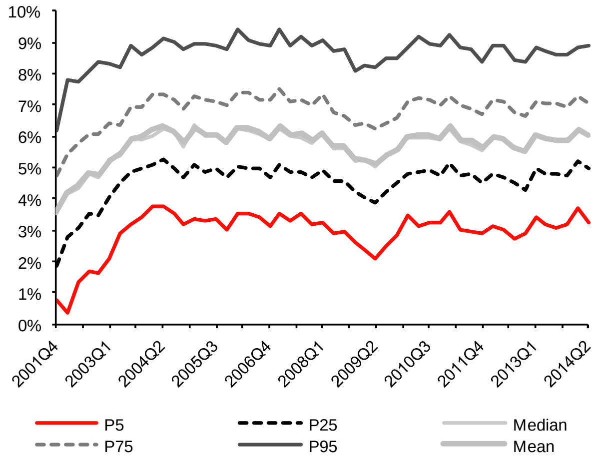


Fig 12 LIWC Net (MD) Level

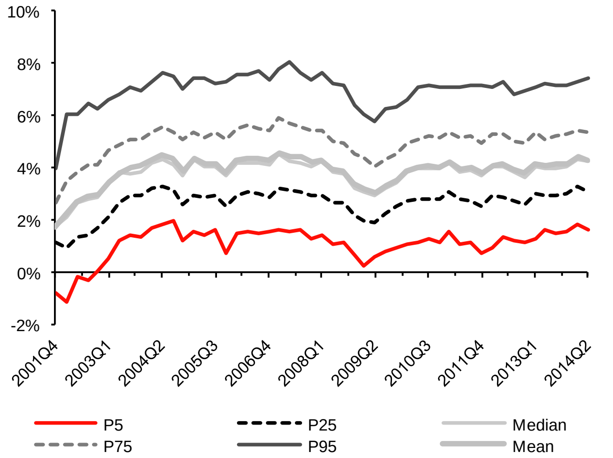


Fig 13 All Dictionaries Net (MD) Level

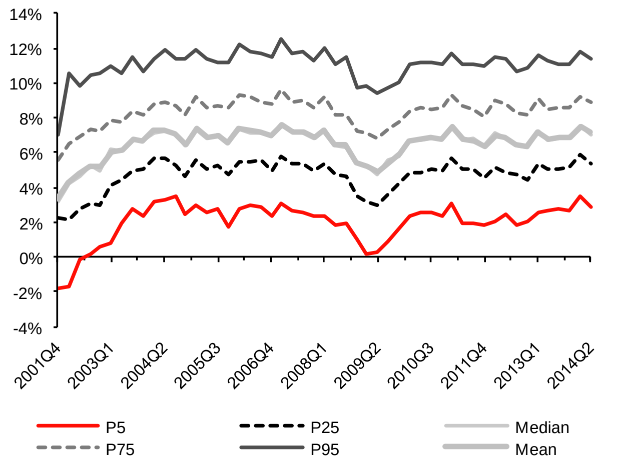


Fig 14 Bayes with LM Net (MD) Level

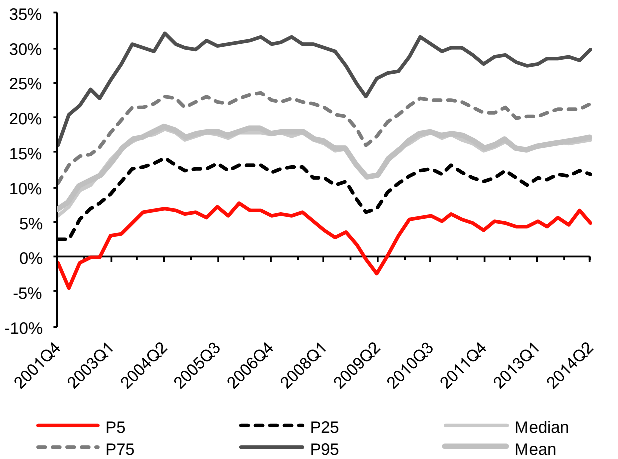
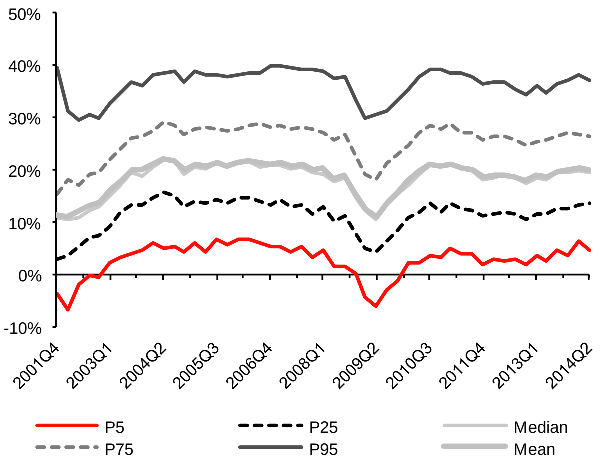


Fig 15 Bayes w/o List Net (MD) Level



Source: FactSet, Macquarie Research, April 2015

Fig 16 LM Net (MD) Change

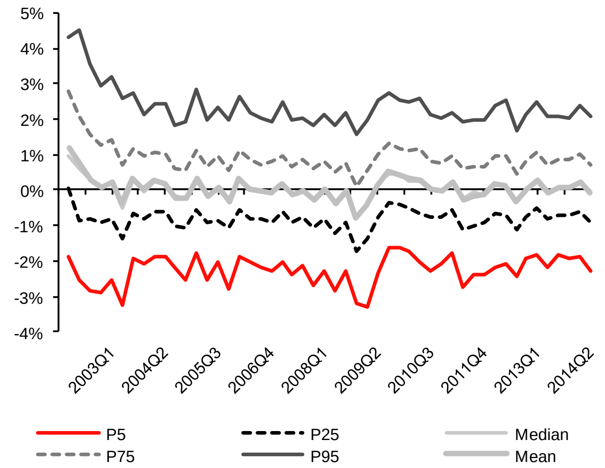


Fig 17 Diction Net (MD) Change

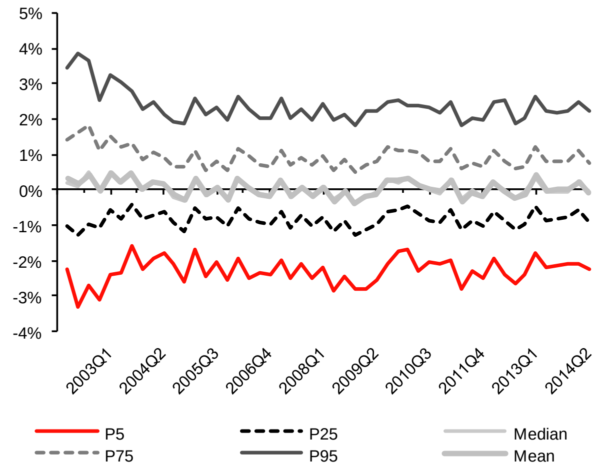


Fig 18 LIWC Net (MD) Change

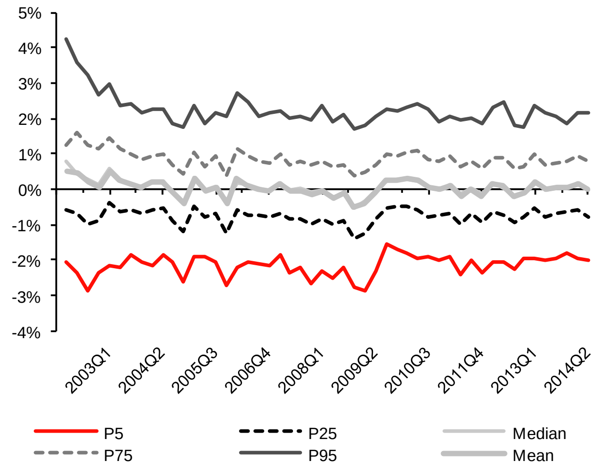


Fig 19 All Dictionaries Net (MD) Change

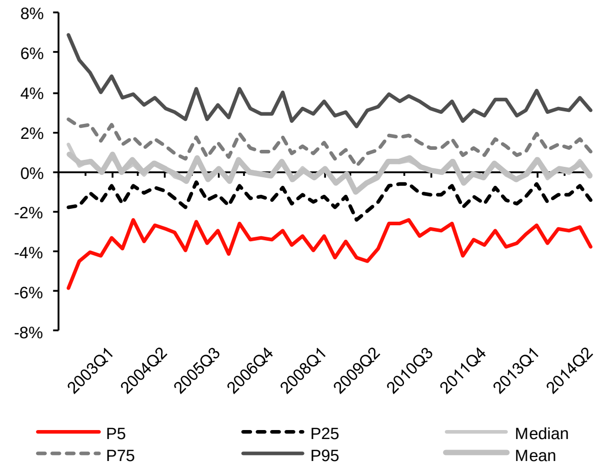


Fig 20 Bayes with LM Net (MD) Change

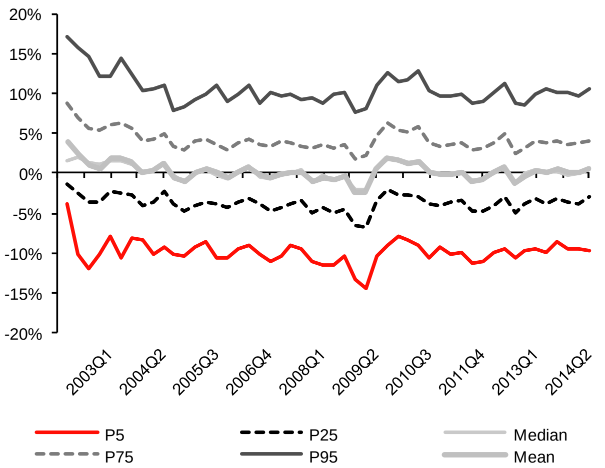
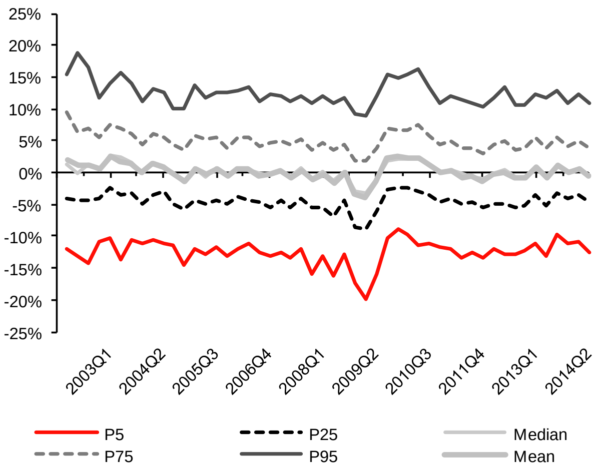


Fig 21 Bayes w/o List Net (MD) Change



Source: FactSet, Macquarie Research, April 2015

Correlations among the Tone Measures

Sentiment extracted from the MD section and tone of the Q&A session convey different messages

In Fig 22 and Fig 23 we examine the correlation among different measures of tone and between tone and complexity measures. The statistics are calculated on the pooled sample. The first two blocks of factors represent sentiment measures obtained from the management discussion (MD) and Q&A sessions respectively. The third block refers to complexity measures. The first two blocks along the diagonal, i.e. the correlations among sentiment measures calculated on the same section of the call, suggest that alternative measures agree on the tone of the call. Lower values are found in the off diagonal blocks that capture the correlation between MD and Q&A tone. This result indicates that the discussion among management and analysts can have a different tone compared to the initial presentation. This may be either because analysts do not share the optimism of the management or because analysts push executives to comment on topics they are reluctant to bring up and therefore the language becomes more negative. Intuitively, our argument would suggest that the signal obtained from the Q&A session should be more effective.

Sentiment and complexity

Correlations between complexity measures and sentiment are low, as we would expect given that they capture very different features of the call.

Fig 22 Correlation among the Tone Measures in Levels

	LM Net (MD)	Diction Net (MD)	LIWC Net (MD)	All Dictionaries Net (MD)	Bayes with LM Net (MD)	Bayes w/o List Net (MD)	LM Net (Q&A)	Diction Net (Q&A)	LIWC Net (Q&A)	All Dictionaries Net (Q&A)	Bayes with LM Net (Q&A)	Bayes w/o List Net (Q&A)	Word Count (MD)	Flesch Kincaid GL (MD)	Word Count (Q&A)	Flesch Kincaid GL (Q&A)
LM Net (MD)	1	0.66	0.67	0.83	0.78	0.67	0.40	0.25	0.26	0.28	0.35	0.41	0.08	0.00	0.03	0.03
Diction Net (MD)	0.66	1	0.70	0.85	0.60	0.47	0.26	0.30	0.24	0.27	0.26	0.26	0.05	-0.05	0.04	-0.03
LIWC Net (MD)	0.67	0.70	1	0.85	0.61	0.57	0.26	0.23	0.31	0.27	0.27	0.34	0.05	-0.06	0.06	0.02
All Dictionaries Net (MD)	0.83	0.85	0.85	1	0.66	0.63	0.33	0.29	0.30	0.33	0.28	0.37	0.12	-0.02	0.03	0.02
Bayes with LM Net (MD)	0.78	0.60	0.61	0.66	1	0.73	0.30	0.24	0.25	0.25	0.37	0.35	0.05	0.14	0.08	0.06
Bayes w/o List Net (MD)	0.67	0.47	0.57	0.63	0.73	1	0.36	0.24	0.32	0.31	0.38	0.49	0.03	0.16	0.09	0.07
LM Net (Q&A)	0.40	0.26	0.26	0.33	0.30	0.36	1	0.50	0.52	0.72	0.69	0.55	-0.01	0.03	0.13	0.00
Diction Net (Q&A)	0.25	0.30	0.23	0.29	0.24	0.24	0.50	1	0.69	0.80	0.49	0.30	-0.06	0.05	-0.02	-0.17
LIWC Net (Q&A)	0.26	0.24	0.31	0.30	0.25	0.32	0.52	0.69	1	0.87	0.52	0.39	-0.03	0.06	0.04	-0.10
All Dictionaries Net (Q&A)	0.28	0.27	0.27	0.33	0.25	0.31	0.72	0.80	0.87	1	0.53	0.38	-0.01	0.07	0.11	-0.09
Bayes with LM Net (Q&A)	0.35	0.26	0.27	0.28	0.37	0.38	0.69	0.49	0.52	0.53	1	0.67	-0.02	0.05	0.05	0.14
Bayes w/o List Net (Q&A)	0.41	0.26	0.34	0.37	0.35	0.49	0.55	0.30	0.39	0.38	0.67	1	0.03	0.02	0.02	0.24
Word Count (MD)	0.08	0.05	0.05	0.12	0.05	0.03	-0.01	-0.06	-0.03	-0.01	-0.02	0.03	1	0.04	0.02	0.12
Flesch Kincaid GL (MD)	0.00	-0.05	-0.06	-0.02	0.14	0.16	0.03	0.05	0.06	0.07	0.05	0.02	0.04	1	0.03	0.15
Word Count (Q&A)	0.03	0.04	0.06	0.03	0.08	0.09	0.13	-0.02	0.04	0.11	0.05	0.02	0.02	0.03	1	0.08
Flesch Kincaid GL (Q&A)	0.03	-0.03	0.02	0.02	0.06	0.07	0.00	-0.17	-0.10	-0.09	0.14	0.24	0.12	0.15	0.08	1

Source: FactSet, Macquarie Research, April 2015

Fig 23 Correlation among the Tone Measures in Changes

	LM Net (MD)	Diction Net (MD)	LIWC Net (MD)	All Dictionaries Net (MD)	Bayes with LM Net (MD)	Bayes w/o List Net (MD)	LM Net (Q&A)	Diction Net (Q&A)	LIWC Net (Q&A)	All Dictionaries Net (Q&A)	Bayes with LM Net (Q&A)	Bayes w/o List Net (Q&A)	Word Count (MD)	Flesch Kincaid GL (MD)	Word Count (Q&A)	Flesch Kincaid GL (Q&A)
LM Net (MD)	1	0.59	0.54	0.78	0.66	0.50	0.19	0.11	0.10	0.15	0.13	0.14	0.03	0.01	0.00	0.00
Diction Net (MD)	0.59	1	0.61	0.82	0.46	0.35	0.14	0.15	0.10	0.15	0.10	0.10	0.03	-0.04	0.01	-0.01
LIWC Net (MD)	0.54	0.61	1	0.77	0.46	0.38	0.12	0.13	0.15	0.15	0.11	0.11	-0.06	-0.04	0.00	-0.01
All Dictionaries Net (MD)	0.78	0.82	0.77	1	0.52	0.46	0.17	0.14	0.13	0.18	0.11	0.13	0.07	-0.02	0.01	0.00
Bayes with LM Net (MD)	0.66	0.46	0.46	0.52	1	0.62	0.13	0.12	0.11	0.13	0.16	0.14	-0.08	0.15	0.00	0.04
Bayes w/o List Net (MD)	0.50	0.35	0.38	0.46	0.62	1	0.12	0.11	0.13	0.13	0.14	0.16	-0.10	0.13	0.01	0.04
LM Net (Q&A)	0.19	0.14	0.12	0.17	0.13	0.12	1	0.41	0.35	0.64	0.58	0.44	0.02	-0.01	0.03	0.07
Diction Net (Q&A)	0.11	0.15	0.13	0.14	0.12	0.11	0.41	1	0.57	0.75	0.40	0.26	-0.12	0.00	-0.14	-0.14
LIWC Net (Q&A)	0.10	0.10	0.15	0.13	0.11	0.13	0.35	0.57	1	0.78	0.38	0.28	-0.13	0.01	-0.13	-0.10
All Dictionaries Net (Q&A)	0.15	0.15	0.15	0.18	0.13	0.13	0.64	0.75	0.78	1	0.40	0.33	-0.06	0.00	-0.05	-0.06
Bayes with LM Net (Q&A)	0.13	0.10	0.11	0.11	0.16	0.14	0.58	0.40	0.38	0.40	1	0.59	-0.07	0.03	-0.05	0.23
Bayes w/o List Net (Q&A)	0.14	0.10	0.11	0.13	0.14	0.16	0.44	0.26	0.28	0.33	0.59	1	-0.01	0.02	-0.01	0.24
Word Count (MD)	0.03	0.03	-0.06	0.07	-0.08	-0.10	0.02	-0.12	-0.13	-0.06	-0.07	-0.01	1	-0.01	0.01	0.06
Flesch Kincaid GL (MD)	0.01	-0.04	-0.04	-0.02	0.15	0.13	-0.01	0.00	0.01	0.00	0.03	0.02	-0.01	1	0.01	0.10
Word Count (Q&A)	0.00	0.01	0.00	0.01	0.00	0.01	0.03	-0.14	-0.13	-0.05	-0.05	-0.01	0.01	0.01	1	0.11
Flesch Kincaid GL (Q&A)	0.00	-0.01	-0.01	0.00	0.04	0.04	0.07	-0.14	-0.10	-0.06	0.23	0.24	0.06	0.10	0.11	1

Source: FactSet, Macquarie Research, April 2015

Correlation with EPS Surprises

Surprises and tone are positively correlated…

Our goal is to understand whether soft information, in the form of management and analyst sentiment, provides any incremental information over hard numbers in financial disclosures. We present correlations between tone and complexity measures (and their changes) with EPS surprises in Fig 24 and Fig 25.

Firstly, correlations for all net tone metrics have as expected positive sign. Correlations for most of the tone metrics are slightly higher in levels than in changes.

…particularly for levels and for the MD section

Sentiment measures derived from Management Discussion sessions appear to be more strongly correlated with EPS surprise than measures derived from Q&A sessions. This might be related to the fact that analysts' questions may encourage the management to disclose additional information, not provided in the financial statement. In particular, it is natural to assume that analysts are likely to ask the management to comment on future earnings and long term issues faced by the firm, whereas the financial statement is by its nature mostly backward looking.

Low correlations suggest potential to derive a new signal

Overall relatively low correlation of sentiment measures with EPS surprise may indicate that information coming from earnings call tone is quite distinct from information about beat or miss of consensus estimates.

Complexity and surprises

Correlation coefficients for complexity metrics are relatively low and negative, suggesting that, if anything, negative surprises may prompt the management to adopt a more complex language, in line with the literature on company disclosures.

Fig 24 Correlation of Tone and Complexity Measures in Levels with EPS Surprise

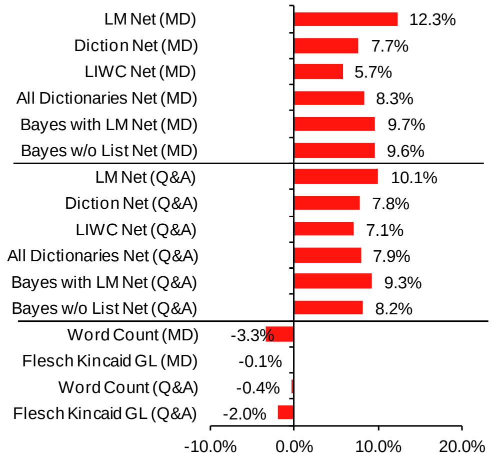
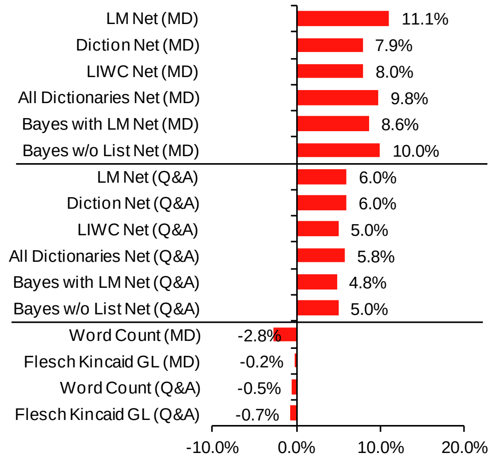


Fig 25 Correlation of Tone and Complexity Measures in Changes with EPS Surprise



Source: FactSet, Macquarie Research, April 2015

Correlation of Tone Measures with Stock Characteristics

Tone is related to price momentum and low risk

Next, we examine the correlation between measures derived from call transcripts and stock characteristics. In Fig 26 and Fig 27 we report correlation of sentiment metrics and changes in metrics with selected factors. Sentiment appears to be moderately positively correlated with momentum and negatively with return volatility. This result suggests that, unsurprisingly, companies which have experienced a period of high volatility or disappointing price performance tend to score poorly in terms of sentiment.

Changes in tone are uncorrelated with any quant factor

A striking result is the almost complete lack of correlation between changes in sentiment and any of the typical quant factors (Fig 27). This result is encouraging because it implies that the signal can contribute significantly to the breadth of a multifactor model. The next section will assess whether the signal can also be employed to generate alpha.

Fig 26 Correlation of Tone Measures in Levels with Stock Characteristics

	Earnings Yield	Book Yield	3M Momentum	6M Momentum	12-1M Momentum	ROE	ROA	250D Volatility	EPS Revisions 1M	EPS Revisions 3M	Mkt Cap
LM Net (MD)	-0.03	-0.17	0.13	0.18	0.21	0.01	0.05	-0.21	0.00	0.02	0.08
Diction Net (MD)	0.00	-0.10	0.09	0.13	0.16	0.02	0.01	-0.20	0.00	0.02	0.03
LIWC Net (MD)	0.01	-0.15	0.08	0.12	0.16	0.02	0.08	-0.25	0.00	0.02	0.12
All Dictionaries Net (MD)	0.00	-0.16	0.10	0.15	0.19	0.03	0.06	-0.27	-0.01	0.02	0.10
Bayes with LM Net (MD)	-0.01	-0.11	0.10	0.16	0.19	0.00	0.00	-0.20	0.01	0.03	0.06
Bayes w/o List Net (MD)	0.00	-0.19	0.09	0.16	0.22	0.03	0.09	-0.26	0.01	0.04	0.11
LM Net (Q&A)	-0.02	-0.17	0.13	0.16	0.16	0.03	0.08	-0.14	0.01	0.02	0.02
Diction Net (Q&A)	0.01	-0.09	0.10	0.14	0.13	0.03	0.04	-0.12	0.00	0.01	0.01
LIWC Net (Q&A)	-0.01	-0.13	0.09	0.13	0.13	0.03	0.07	-0.11	0.01	0.02	0.03
All Dictionaries Net (Q&A)	0.00	-0.14	0.11	0.14	0.14	0.03	0.07	-0.13	0.00	0.01	0.00
Bayes with LM Net (Q&A)	-0.01	-0.11	0.10	0.14	0.13	0.01	0.04	-0.09	0.01	0.03	0.04
Bayes w/o List Net (Q&A)	-0.03	-0.16	0.08	0.12	0.13	0.03	0.07	-0.15	0.01	0.03	0.10
Word Count (MD)	0.02	0.04	-0.03	-0.03	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	0.11
Flesch Kincaid GL (MD)	0.00	0.00	-0.01	0.00	0.00	-0.01	0.00	0.02	0.01	0.01	-0.03
Word Count (Q&A)	0.03	-0.01	0.00	0.01	0.03	0.01	0.04	-0.05	0.00	0.00	0.06
Flesch Kincaid GL (Q&A)	-0.01	0.02	-0.02	-0.02	-0.02	-0.01	-0.01	-0.02	0.01	0.01	0.03

Source: FactSet, Macquarie Research, April 2015

Fig 27 Correlation of Tone Measures in Changes with Stock Characteristics

	Earnings Yield	Book Yield	3M Momentum	6M Momentum	12-1M Momentum	ROE	ROA	250D Volatility	EPS Revisions 1M	EPS Revisions 3M	Mkt Cap
LM Net (MD)	-0.02	0.00	0.10	0.07	0.01	0.00	-0.02	0.02	0.00	-0.01	0.00
Diction Net (MD)	-0.02	0.00	0.07	0.05	0.01	-0.01	-0.02	0.01	0.00	0.00	0.00
LIWC Net (MD)	-0.02	0.00	0.07	0.06	0.02	0.00	-0.02	0.01	0.01	0.00	0.00
All Dictionaries Net (MD)	-0.02	0.00	0.10	0.08	0.03	-0.01	-0.03	0.01	0.00	0.00	0.00
Bayes with LM Net (MD)	-0.02	0.00	0.08	0.07	0.01	0.00	-0.02	0.02	0.01	0.00	0.00
Bayes w/o List Net (MD)	-0.02	-0.01	0.10	0.10	0.05	0.00	-0.03	0.01	0.00	0.01	0.00
LM Net (Q&A)	-0.01	0.00	0.07	0.03	-0.02	0.00	-0.01	0.02	0.00	-0.01	0.00
Diction Net (Q&A)	-0.02	0.00	0.03	0.01	-0.04	0.00	-0.01	0.04	-0.01	-0.02	-0.01
LIWC Net (Q&A)	-0.02	0.00	0.04	0.02	-0.03	0.00	-0.02	0.04	0.00	-0.01	-0.01
All Dictionaries Net (Q&A)	-0.02	0.00	0.05	0.02	-0.04	0.00	-0.02	0.04	-0.01	-0.02	-0.01
Bayes with LM Net (Q&A)	-0.01	0.00	0.04	0.03	-0.03	0.00	-0.01	0.03	0.00	0.00	0.00
Bayes w/o List Net (Q&A)	-0.01	-0.01	0.06	0.05	0.01	0.00	-0.01	0.01	0.01	0.01	0.00
Word Count (MD)	0.00	0.00	-0.03	-0.02	-0.01	0.00	0.00	0.01	-0.01	0.00	0.00
Flesch Kincaid GL (MD)	0.00	0.00	0.01	0.01	0.00	0.00	0.01	-0.01	0.00	0.00	0.00
Word Count (Q&A)	0.00	0.00	-0.02	-0.01	0.00	-0.01	0.00	0.00	-0.01	0.00	0.00
Flesch Kincaid GL (Q&A)	0.00	0.00	0.00	0.00	0.01	0.00	0.00	-0.01	0.01	0.01	0.00

Source: FactSet, Macquarie Research, April 2015

Event study: Does tone affect future returns?

Measuring the effect of tone on returns after the conference call

This section presents the results of an event study that analyses the relation between sentiment (as measured from the language used in conference call) and returns. We compute excess returns to the stocks with best and worst sentiment, defined as those in the top and bottom tercile, respectively, after ranking all available observations by net tone. It is important to note that our thresholds change over time in such a way that we compare each sentiment observation against the distribution obtained from the sample available up to that point.

Stocks are sorted in terciles without look-ahead biases

Formally, we stack all sentiment observations in a vector u:

$$u = \left(T_{11}, T_{21}, \dots, T_{n_1 1}, T_{12}, T_{22}, \dots, T_{n_2 2}, \dots, T_{n_p p}\right)'$$

where T_{ij} is the sentiment score (net tone) obtained from the i-th conference call of firm j, p is the total number of firms in the sample and n_j is the total number of documents available for firm j. Similarly, we build a vector that holds the corresponding excess returns between time $\underline{\tau}$ and time $\overline{\tau}$ after the call has taken place

$$v = \left(r_{11}, r_{21}, \dots, r_{n_1 1}, r_{12}, r_{22}, \dots, r_{n_2 2}, \dots, r_{n_p p}\right)'$$

where each element is defined as

$$r_{ij} = \frac{P_{j, t_{ij}+\overline{\tau}}}{P_{j, t_{ij}+\underline{\tau}}} - \frac{I_{j, t_{ij}+\overline{\tau}}}{I_{j, t_{ij}+\underline{\tau}}},$$

t_{ij} is the date of firm j's i-th call and P_{it}, I_{it} are the total return index for the stock and its benchmark respectively, expressed in US dollars. Let us call u_i^* the second tercile (i.e. the $66.\overline{6}\%$ percentile) of the distribution of the subset of u consisting of calls which took place prior to the i-th observation.The average return to the top tercile of stocks sorted by sentiment is then defined as $\frac{1}{n^*}\sum_{i:u_i \geq u_i^*} v_i$ where $n^*$is the total number of observations that fall in the top tercile (i.e. for which $u_i \geq u_i^*$). The returns to the middle and bottom terciles are defined along the same lines.

Returns are measured in excess of sector and region

The benchmark I_{jt} is an equally weighted index selected according to stock j's region (North America, Europe and APAC) and sector (10 GICS sectors) classification. Our definition of abnormal return is equivalent to the return on a strategy in which the position in a stock is hedged as of time *t+1* (one day after the call) and never readjusted over the investment period. Tercile buckets are constructed based on call metrics known as of time 0 and the allocation to buckets is based on tercile cut-off points calculated for all previous calls, thus avoiding any forward looking bias.

We trade at close the day after the call

We consider two alternative horizons in our return calculations: $\overline{\tau} = 21$ and $\overline{\tau} = 63$ trading days, roughly corresponding to one and three calendar months. The starting point for the first two horizons is one day after the call ($\underline{\tau} = 1$). It is worth emphasising that the call typically occurs before market open on day 0, therefore our choice to use the closing price of day 1 implies that we are consider returns from two trading days after the event.

The columns headed T1, T2 and T3 display the excess returns of each tercile of the observations sorted by tone, equally weighted (as detailed above). Finally, the "spread" columns report the spread between terciles 3 and 1.

Comparison with the effect of earnings surprises

As a benchmark for event studies, we start off by computing market reaction to earnings surprises. We use two alternative measures, both based on I/B/E/S data. "Surprise" is defined as the difference between the actual EPS number and the average forecast obtained the day before the announcement, scaled by the absolute value of the average forecast. The SUE (Standardized Unanticipated Earnings) measure takes forecast dispersion into account:

$$SUE = \frac{actual\ EPS - average\ estimate}{standard\ deviation\ of\ estimates}.$$

Fig 28 documents the well known post earnings announcement drift (PEAD), i.e. the outperformance of stocks that have beaten analyst expectations. Spread are, on average, between 25 and 30 basis points over the one month horizon.

Fig 28 Summary of Market Reaction to EPS Surprise

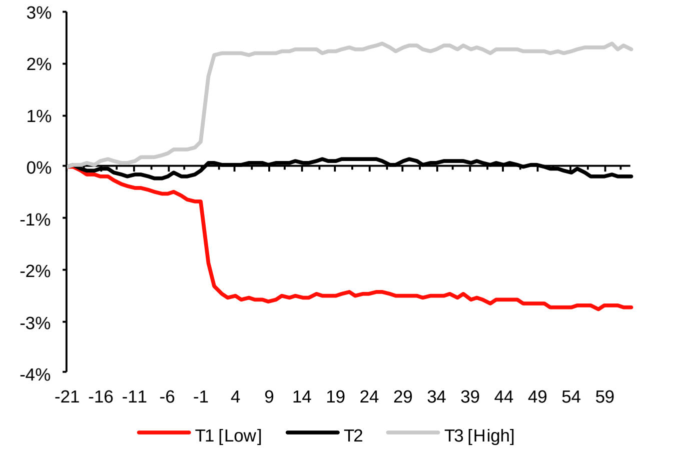
EPS Surprise								
	T1 to T21 Drift Returns in bps				T1 to T63 Drift Returns in bps			
Variable	T1	T2	T3	Spread	T1	T2	T3	Spread
Surprise	-15.6	9.2	11.1	26.7	-34.5	-27.7	0.2	34.8
SUE	-15.4	6.9	12.8	28.2	-28.7	-18.5	-17.5	11.1

Source: FactSet, Macquarie Research, April 2015

The post earnings announcement drift

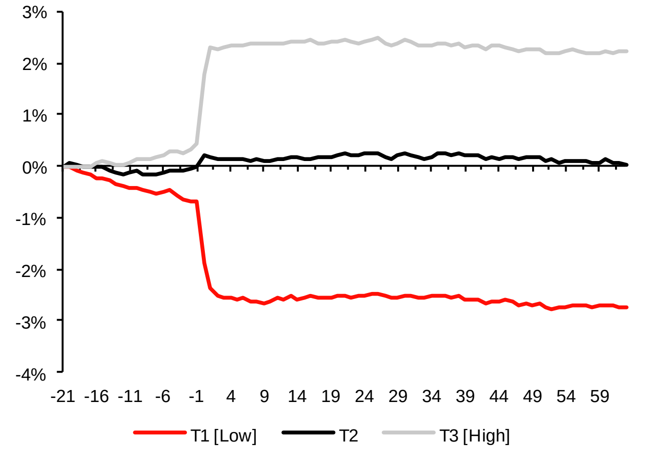
Fig 29 to Fig 32 illustrate the phenomenon graphically, using the two alternative measures of surprise and sorting stocks in terciles (Fig 29 and Fig 30) or quintiles (Fig 31 and Fig 32). The scale on the horizontal axis refers to days in event time, time zero being the date of the announcement. The drift begins to materialise *before* the announcement, typically a reflection of guidance from the management or information spillover from other firms in the same sector. Most of the price effect is concentrated immediately after the conference call, when companies that have surprised most positively outperform by 2 to 3%. The post announcement spread is represented in the chart by the further divergence between stocks with high and low surprise scores, which is mostly due to the underperformance of the lowest quantile.

Fig 29 Marker Reaction to Earnings Surprises (Terciles)



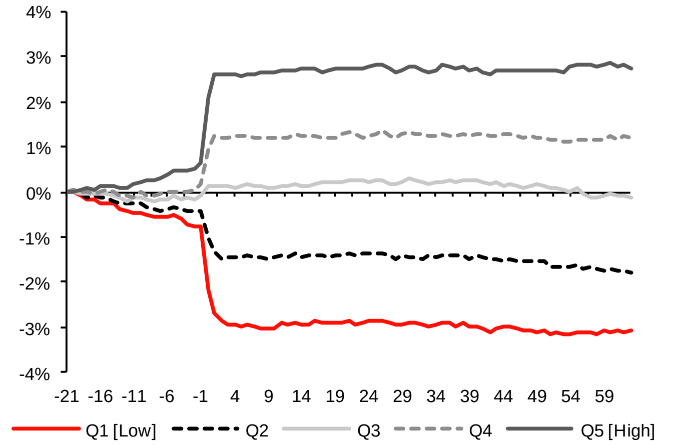
Source: FactSet, Macquarie Research, April 2015

Fig 30 Marker Reaction to Earnings Surprises (SUE) (Terciles)



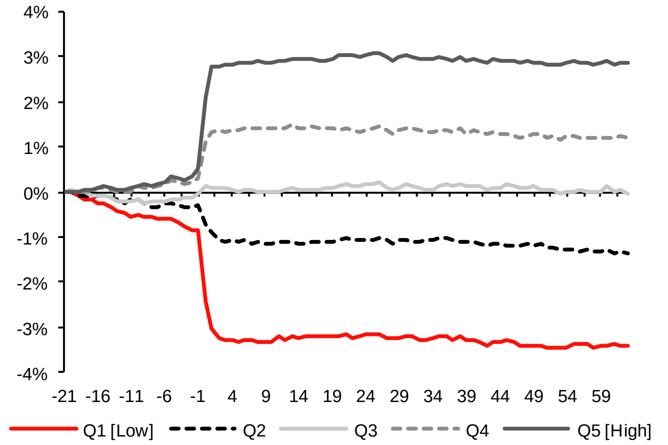
Source: FactSet, Macquarie Research, April 2015

Fig 31 Marker Reaction to Earnings Surprises (Quintiles)



Source: FactSet, Macquarie Research, April 2015

Fig 32 Marker Reaction to Earnings Surprises (SUE) (Quintiles)



Source: FactSet, Macquarie Research, April 2015

Conference call tone and returns

Fig 33 displays the results for conference call sentiment. The table is split into two panels, which refer to two alternative signals: the net tone detected in the call transcript and the change in tone from the previous available call. The latter was found to have better predictive power in *Positively Persuasive*, although it has never, to the best of our knowledge, been used in academic work.

Sentiment predicts returns, particularly changes in tone from the Q&A section

The rows in each table refer to alternative measures of net tone. “MD” and “Q&A” identify the Management Discussion and Q&A session respectively. “All dictionaries” is the union of all word lists from the three sources mentioned on page 9. Two implementations of the Naïve Bayes approach are also considered: One where the list of words is optimised by cross validation (as detailed on page 14) and one where the list of available terms is the union of the positive and negative word lists from Loughran and McDonald (2011).

Several results can be highlighted from the table:

1. The spreads between best and worst sentiment terciles tend to be positive on both horizons.
2. The overall results for changes in sentiment are stronger than those for the corresponding levels.

3. In the bottom panel, where we use changes as a signal, the spread tends to widen between 21 and 63 days after the call (column 8 vs. column 4). This is true in particular if we use signals derived from the Q&A session.
4. The same does not hold for sentiment used as a signal in levels.
5. We find no evidence that our Bayesian approach predicts returns better than the simple dictionary-based method.

Fig 33 Summary of Market Reaction to Tone

Tone Measures - Levels								
	T1 to T21 Drift Returns in bps				T1 to T63 Drift Returns in bps			
Variable	T1	T2	T3	Spread	T1	T2	T3	Spread
LM Net (MD)	-4.4	-9.0	16.6	21.0	-21.6	-40.0	-0.1	21.5
Diction Net (MD)	-20.0	9.2	11.5	31.5	-40.5	-14.0	-9.8	30.7
LIWC Net (MD)	-11.6	-1.2	14.2	25.8	-24.6	-38.1	-0.2	24.4
All Dictionaries Net (MD)	-5.1	-9.5	16.7	21.8	-25.3	-39.1	1.3	26.6
Bayes with LM Net (MD)	-4.7	-2.1	9.3	14.0	-25.6	-29.3	-8.0	17.6
Bayes w/o List Net (MD)	-3.4	-1.8	8.0	11.5	-17.3	-26.7	-17.7	-0.5
LM Net (Q&A)	-2.4	0.4	4.5	6.9	-11.9	-22.1	-22.9	-11.1
Diction Net (Q&A)	-3.5	3.9	1.9	5.4	-30.7	-15.0	-14.2	16.4
LIWC Net (Q&A)	-5.8	-0.2	8.8	14.6	-21.2	-18.7	-18.2	3.0
All Dictionaries Net (Q&A)	0.2	-5.2	7.6	7.4	-27.4	-14.0	-18.1	9.3
Bayes with LM Net (Q&A)	-5.2	10.4	-2.8	2.4	-16.1	-3.9	-40.0	-23.9
Bayes w/o List Net (Q&A)	-5.1	7.3	0.9	6.0	-13.1	-16.4	-30.2	-17.1

Tone Measures - Changes								
	T1 to T21 Drift Returns				T1 to T63 Drift Returns			
Variable	T1	T2	T3	Spread	T1	T2	T3	Spread
LM Net (MD)	-23.0	6.1	22.7	45.8	-40.7	-12.8	5.8	46.6
Diction Net (MD)	-22.7	3.4	25.8	48.5	-46.6	-10.8	10.4	57.0
LIWC Net (MD)	-6.1	-10.4	23.8	29.9	-34.3	-16.1	3.6	37.9
All Dictionaries Net (MD)	-16.2	1.6	20.8	37.0	-36.9	-13.7	3.0	39.9
Bayes with LM Net (MD)	-19.3	10.6	15.0	34.3	-39.6	-12.1	5.0	44.6
Bayes w/o List Net (MD)	-10.5	-0.7	17.7	28.3	-19.7	-26.4	-0.9	18.8
LM Net (Q&A)	-15.6	6.8	14.9	30.5	-55.3	-2.9	13.0	68.3
Diction Net (Q&A)	-1.4	1.8	6.1	7.5	-43.8	-8.2	8.8	52.7
LIWC Net (Q&A)	-10.4	8.7	8.2	18.6	-34.2	-15.4	7.5	41.8
All Dictionaries Net (Q&A)	-8.8	1.5	15.1	23.9	-51.4	-17.1	29.4	80.8
Bayes with LM Net (Q&A)	-9.2	7.1	7.9	17.1	-38.3	-8.4	2.2	40.5
Bayes w/o List Net (Q&A)	-9.3	4.2	10.9	20.2	-41.5	-3.2	-1.9	39.6

Source: FactSet, Macquarie Research, April 2015

Disentangling the effects of tone and surprises

In the results presented so far the effects of sentiment cannot be distinguished from the effects of earnings surprises. To what extent does the drift detected in Fig 33 overlap with the well known PEAD documented in Fig 28? It is natural to assume that the sentiment detected during a conference call tends to be positive when earnings beat expectations. In Fig 34 we show average drift returns for tercile buckets constructed on *abnormal* tone metrics and changes in abnormal tone metrics for both sections of the call (page 9). The idea is to strip out the effect of the surprise and concentrate on a pure sentiment measure. The results indicate, once again, that changes in sentiment are positively associated to subsequent returns, both at a one-month and at a three-month horizon.

Fig 34 Summary of Market Reaction to Abnormal Tone

Tone Measures Adjusted for Surprise - Levels								
	T1 to T21 Drift Returns				T1 to T63 Drift Returns			
Variable	T1	T2	T3	Spread	T1	T2	T3	Spread
LM Net (MD)	-8.5	-1.9	15.4	23.9	-31.3	-30.9	2.7	34.0
Diction Net (MD)	-13.3	5.9	13.7	27.0	-35.6	-18.0	-4.8	30.8
LIWC Net (MD)	-11.0	1.1	15.6	26.6	-36.1	-24.0	1.8	37.9
All Dictionaries Net (MD)	-6.8	-5.0	17.2	23.9	-28.7	-32.3	1.9	30.5
Bayes with LM Net (MD)	-0.8	-1.9	7.2	8.0	-27.0	-30.4	-1.0	26.0
Bayes w/o List Net (MD)	-1.3	-3.9	9.4	10.7	-8.7	-40.3	-14.4	-5.7
LM Net (Q&A)	0.4	-0.8	3.5	3.1	-13.9	-19.1	-24.9	-11.0
Diction Net (Q&A)	-2.4	4.9	1.4	3.8	-31.8	-5.0	-17.9	13.9
LIWC Net (Q&A)	-5.8	2.6	12.5	18.3	-27.6	-13.3	-11.7	15.9
All Dictionaries Net (Q&A)	-8.1	10.3	3.6	11.7	-29.6	-2.6	-25.9	3.7
Bayes with LM Net (Q&A)	0.7	5.8	-4.6	-5.4	-9.7	-9.4	-44.4	-34.7
Bayes w/o List Net (Q&A)	-2.0	3.2	1.7	3.7	-13.2	-18.5	-27.0	-13.9

Tone Measures Adjusted for Surprise - Changes								
	T1 to T21 Drift Returns				T1 to T63 Drift Returns			
Variable	T1	T2	T3	Spread	T1	T2	T3	Spread
LM Net (MD)	-15.5	2.1	15.5	31.0	-39.2	-13.8	-2.0	37.1
Diction Net (MD)	-22.9	0.2	25.4	48.3	-47.1	-15.2	7.9	55.0
LIWC Net (MD)	-7.5	-9.3	20.2	27.8	-37.5	-14.1	-2.8	34.7
All Dictionaries Net (MD)	-17.9	-0.3	20.7	38.6	-46.1	-8.3	-0.4	45.7
Bayes with LM Net (MD)	-15.4	3.3	14.6	30.0	-34.4	-21.8	2.1	36.5
Bayes w/o List Net (MD)	-13.7	6.9	9.4	23.2	-30.6	-12.0	-11.6	19.0
LM Net (Q&A)	-12.6	3.9	12.5	25.1	-55.9	-6.8	11.5	67.4
Diction Net (Q&A)	-5.1	4.9	3.8	8.9	-43.0	-14.0	6.6	49.6
LIWC Net (Q&A)	-11.1	10.0	3.7	14.7	-33.9	-15.0	-1.7	32.2
All Dictionaries Net (Q&A)	-7.7	5.1	5.9	13.7	-53.3	-13.1	15.6	68.9
Bayes with LM Net (Q&A)	-10.4	6.7	7.1	17.5	-39.4	-18.5	8.2	47.6
Bayes w/o List Net (Q&A)	-8.6	0.6	11.9	20.5	-43.1	-14.9	7.3	50.3

Source: FactSet, Macquarie Research, April 2015

Tone predicts returns even after correcting for surprises

In Fig 35 to Fig 40 we present the average drift in the event window from *t-21* to *t+63* for terciles constructed based on abnormal levels of tone measures extracted from the Q&A section of the call. On the next page, in Fig 41 to Fig 46 we show the average drift in the event window for terciles constructed based on changes in abnormal levels of tone measures. It is worth noting that the drifts are calculated with the same methodology used for the tables except for a slight difference in the calculation of the excess return. To be more precise, in the charts in Fig 41 to Fig 46 we subtract the return to the benchmark assuming a long short investment in the stock and in the relevant index of identical amount at time *t-21*. In all other excess return calculations the reference point is time *t+1*.

Abnormal returns for all tone metrics in the event window are clearly differentiated across tercile buckets. Both the long and the short side appear to contribute to the drift.

Fig 35 LM Net (Q&A) Abnormal Level

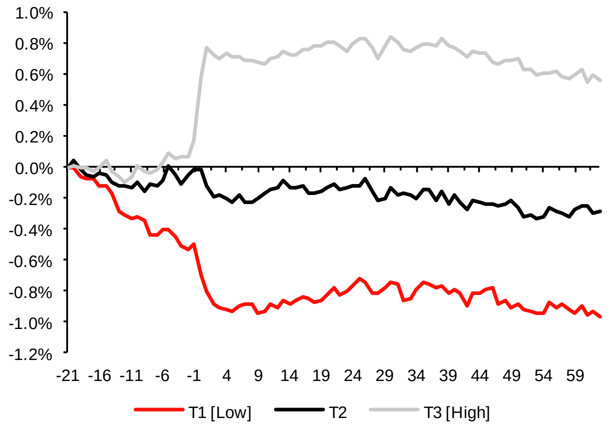


Fig 36 Diction Net (Q&A) Abnormal Level

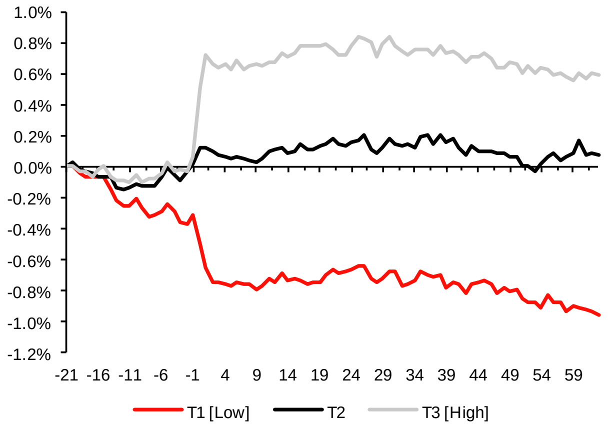


Fig 37 LIWC Net (Q&A) Abnormal Level

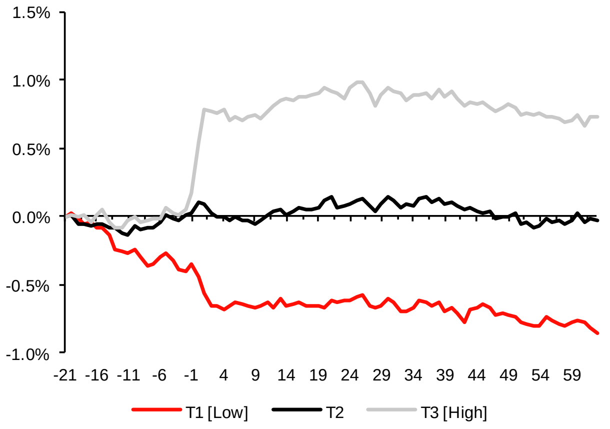


Fig 38 All Dictionaries Net (Q&A) Abnormal Level

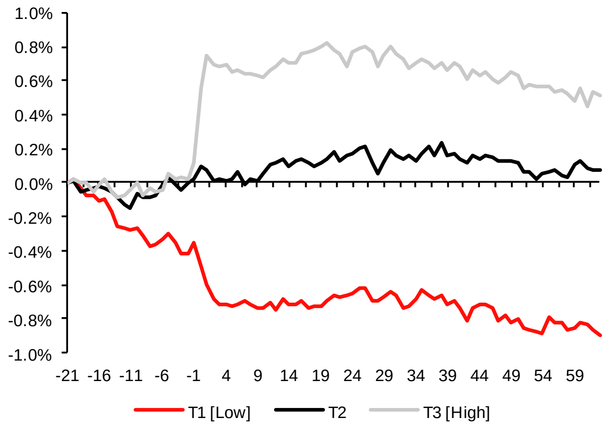


Fig 39 Bayes with LM Net (Q&A) Abnormal Level

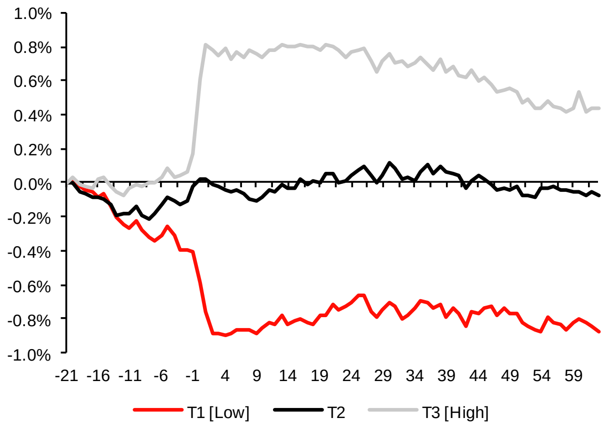
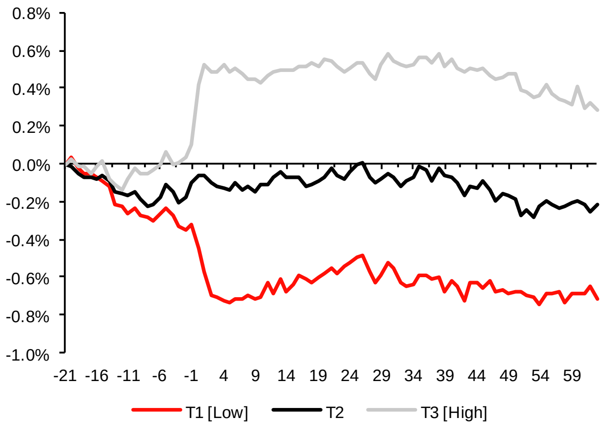


Fig 40 Bayes w/o List Net (Q&A) Abnormal Level



Source: FactSet, Macquarie Research, April 2015

Fig 41 LM Net (Q&A) Abnormal Change

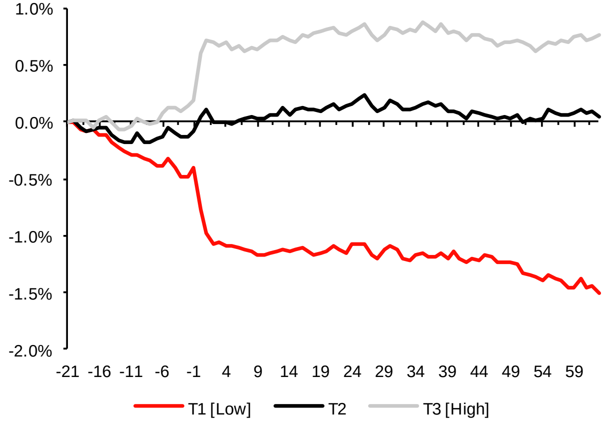


Fig 42 Diction Net (Q&A) Abnormal Change

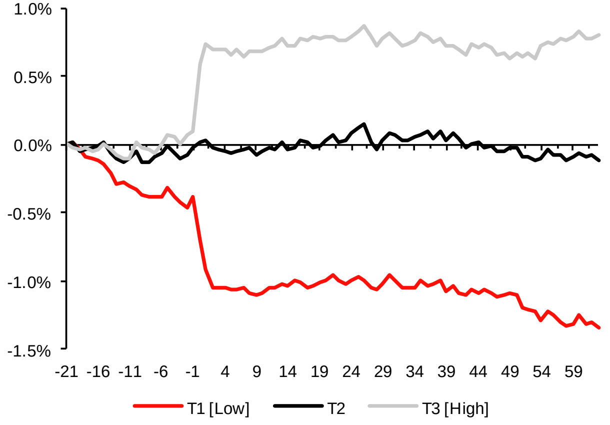


Fig 43 LIWC Net (Q&A) Abnormal Change

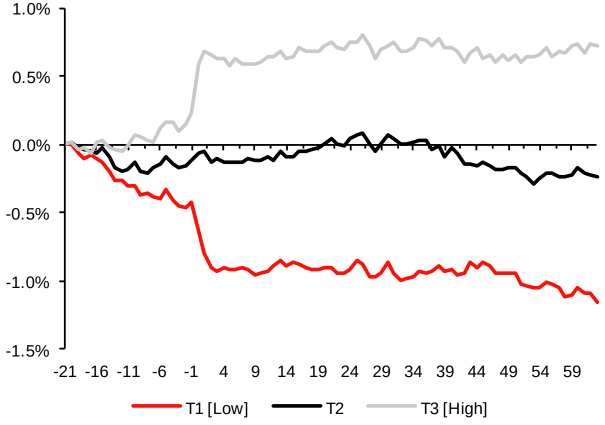


Fig 44 All Dictionaries Net (Q&A) Abnormal Change

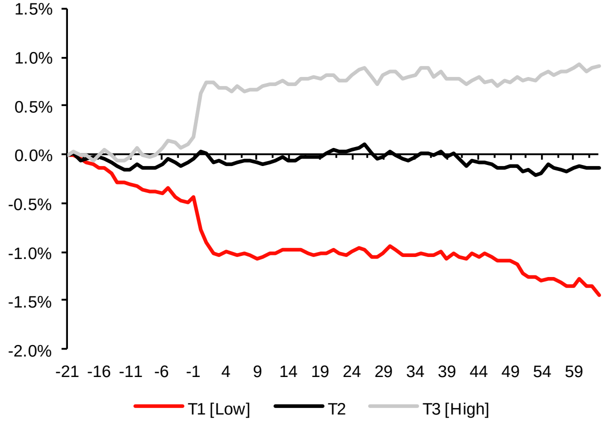


Fig 45 Bayes with LM Net (Q&A) Abnormal Change

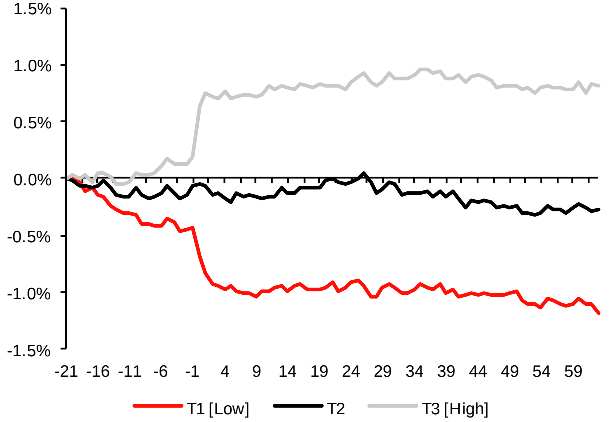
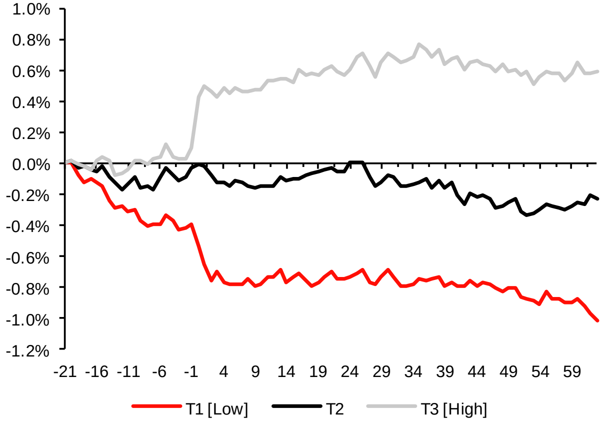


Fig 46 Bayes w/o List Net (Q&A) Abnormal Change



Source: FactSet, Macquarie Research, April 2015

Double Sorting

An alternative method to disentangle the two effects

An alternative method to control for the effect of earnings surprise on tone levels is to use dependent double sorts. Our approach is to sort on earnings surprises first, forming three terciles, and then on sentiment within each tercile. By comparing the average returns of each portfolio we can assess whether, for example, an earnings beat followed by a bullish conference call has significantly different implications on returns compared to an earnings beat with negative sentiment. In Fig 47 we report abnormal returns for terciles formed by ranking within the top and bottom tercile of EPS surprises.

Sort on surprises first and then on tone

In addition, in the last column we compute the average return between the “best surprise / best tone” and the “worst surprise / best tone” portfolios minus the average return between the “best surprise / worst tone” and the “worst surprise / worst tone” portfolios. This statistic can be interpreted as the effect of sentiment on return net of the PEAD.

The average spread is mostly positive until a calendar month after the call has happened, however it appears to be weak or even negative beyond the one-month horizon.

Positive changes in tone predict outperformance within each tercile sorted by earnings surprises

In Fig 48 we repeat the same analysis for change in sentiment variables. In this case, at least if we restrict our attention to the Q&A session, the spread is almost invariably positive. It is also interesting to note that the drift accrues gradually over the three-month period and therefore we detect a positive drift even between 21 and 63 calendar days after the call. Signals based on the Management Discussion session do give rise to a drift but the effect is exhausted within the first 21 trading days. The results in the two columns headed “spread” can be used to gauge whether the average return is driven mostly by companies with positive (column 4) or negative (column 8) surprises. While both categories do tend to make a positive contribution to the spread, it is clearly among stocks that miss expectations that most of the effect manifests itself. In other words, sentiment drives returns mostly because within the bottom tercile of earnings surprises companies that had a strong improvement in sentiment outperform.

Fig 47 Return drift after sorting stocks by tone

Tone Measures - Levels - T1 to T21 Drift Returns in bps									
	Top Tercile of Surprises Terciles of Tone				Bottom Tercile of Surprises Terciles of Tone				Average
Variable	T1	T2	T3	Spread	T1	T2	T3	Spread	Spread
LM Net (MD)	7.9	-5.1	30.6	22.7	-34.1	-9.0	-4.6	29.5	26.1
Diction Net (MD)	-15.0	23.7	20.7	35.7	-32.4	-10.8	-5.5	27.0	31.3
LIWC Net (MD)	-2.0	-1.2	34.2	36.2	-31.5	-12.6	-5.1	26.4	31.3
All Dictionaries Net (MD)	3.4	0.5	28.3	24.9	-25.9	-20.3	-1.9	24.0	24.5
Bayes with LM Net (MD)	1.7	10.5	19.8	18.0	-19.2	-18.8	-9.3	9.9	13.9
Bayes w/o List Net (MD)	8.3	6.0	18.3	10.0	-20.3	-21.9	-4.8	15.6	12.8
LM Net (Q&A)	8.7	7.0	19.0	10.3	-24.2	-11.3	-10.2	14.0	12.1
Diction Net (Q&A)	0.1	17.8	15.0	14.9	-20.0	-9.0	-16.6	3.4	9.1
LIWC Net (Q&A)	-11.3	15.6	28.8	40.1	-8.9	-21.6	-13.5	-4.6	17.8
All Dictionaries Net (Q&A)	6.1	7.4	20.1	14.0	-21.1	-14.2	-10.6	10.5	12.3
Bayes with LM Net (Q&A)	-3.5	26.2	11.8	15.3	-29.3	5.9	-22.6	6.7	11.0
Bayes w/o List Net (Q&A)	4.4	12.6	19.5	15.1	-33.3	1.2	-11.7	21.6	18.3

Tone Measures - Levels - T1 to T63 Drift Returns in bps									
	Top Tercile of Surprises Terciles of Tone				Bottom Tercile of Surprises Terciles of Tone				Average
Variable	T1	T2	T3	Spread	T1	T2	T3	Spread	Spread
LM Net (MD)	12.7	-46.7	35.0	22.3	-43.8	-34.0	-26.4	17.4	19.8
Diction Net (MD)	-14.3	11.7	1.5	15.8	-52.3	-30.3	-23.3	29.0	22.4
LIWC Net (MD)	-9.4	-27.7	36.0	45.4	-29.7	-51.0	-22.4	7.4	26.4
All Dictionaries Net (MD)	2.1	-44.3	42.5	40.4	-48.7	-38.5	-18.2	30.5	35.4
Bayes with LM Net (MD)	5.1	-17.8	13.6	8.4	-37.8	-37.9	-28.5	9.3	8.9
Bayes w/o List Net (MD)	5.4	-14.6	9.8	4.4	-13.0	-48.8	-41.4	-28.4	-12.0
LM Net (Q&A)	12.0	-3.7	-3.5	-15.5	-50.2	-12.2	-40.2	9.9	-2.8
Diction Net (Q&A)	-9.2	1.0	9.6	18.9	-60.6	-22.5	-22.7	37.9	28.4
LIWC Net (Q&A)	-26.4	6.4	21.2	47.7	-40.4	-25.0	-37.1	3.3	25.5
All Dictionaries Net (Q&A)	-16.7	7.4	8.9	25.5	-66.4	-14.8	-25.5	40.9	33.2
Bayes with LM Net (Q&A)	0.0	21.5	-20.0	-20.0	-43.4	5.9	-68.1	-24.8	-22.4
Bayes w/o List Net (Q&A)	10.9	-12.7	5.7	-5.2	-43.9	-9.0	-51.0	-7.0	-6.1

Tone Measures - Levels - T21 to T63 Drift Returns in bps									
	Top Tercile of Surprises Terciles of Tone				Bottom Tercile of Surprises Terciles of Tone				Average
Variable	T1	T2	T3	Spread	T1	T2	T3	Spread	Spread
LM Net (MD)	17.9	-37.3	1.8	-16.1	18.3	-17.1	-19.8	-38.1	-27.1
Diction Net (MD)	15.8	-11.0	-19.0	-34.8	3.8	-4.7	-17.5	-21.3	-28.1
LIWC Net (MD)	5.8	-25.7	3.8	-2.0	35.4	-33.5	-15.6	-51.1	-26.5
All Dictionaries Net (MD)	15.1	-45.2	14.3	-0.8	4.7	-7.1	-16.2	-20.9	-10.9
Bayes with LM Net (MD)	15.7	-23.6	-7.3	-23.0	10.6	-12.7	-16.1	-26.7	-24.9
Bayes w/o List Net (MD)	10.9	-18.0	-9.2	-20.2	37.8	-23.4	-33.2	-71.0	-45.6
LM Net (Q&A)	17.5	-9.8	-22.9	-40.4	-7.7	17.7	-27.0	-19.3	-29.9
Diction Net (Q&A)	5.0	-14.1	-6.7	-11.7	-19.7	-0.6	0.4	20.1	4.2
LIWC Net (Q&A)	-1.1	-5.3	-11.1	-10.0	-4.6	5.9	-19.8	-15.1	-12.6
All Dictionaries Net (Q&A)	-7.7	4.9	-14.9	-7.2	-27.9	17.9	-10.9	17.0	4.9
Bayes with LM Net (Q&A)	13.5	-1.7	-31.5	-44.9	9.9	6.5	-37.3	-47.2	-46.1
Bayes w/o List Net (Q&A)	17.6	-22.7	-15.2	-32.8	17.8	-8.5	-31.7	-49.5	-41.2

Source: FactSet, Macquarie Research, April 2015

Fig 48 Return drift after sorting stocks by change in tone

Tone Measures - Changes - T1 to T21 Drift Returns in bps									
	Top Tercile of Surprises Terciles of Tone				Bottom Tercile of Surprises Terciles of Tone				Average
Variable	T1	T2	T3	Spread	T1	T2	T3	Spread	Spread
LM Net (MD)	-19.1	18.3	36.8	55.9	-38.8	-13.1	7.1	46.0	50.9
Diction Net (MD)	-12.7	17.9	30.6	43.2	-38.9	-31.8	28.5	67.4	55.3
LIWC Net (MD)	-0.3	3.4	33.7	34.0	-27.8	-31.4	16.7	44.4	39.2
All Dictionaries Net (MD)	-4.5	-1.8	42.5	47.0	-44.7	-9.8	10.6	55.3	51.2
Bayes with LM Net (MD)	-14.2	17.4	33.4	47.5	-37.4	-8.7	2.2	39.6	43.5
Bayes w/o List Net (MD)	3.8	10.2	21.6	17.8	-39.8	-12.4	8.3	48.1	33.0
LM Net (Q&A)	-13.9	23.5	28.6	42.6	-33.5	-15.7	7.7	41.2	41.9
Diction Net (Q&A)	7.2	14.4	17.1	9.9	-23.4	-10.1	-8.9	14.5	12.2
LIWC Net (Q&A)	-5.1	18.2	27.2	32.3	-30.8	-3.6	-8.7	22.1	27.2
All Dictionaries Net (Q&A)	-2.9	8.9	36.4	39.3	-31.2	-12.1	2.2	33.4	36.3
Bayes with LM Net (Q&A)	8.1	10.3	20.2	12.1	-40.0	3.0	-7.2	32.8	22.5
Bayes w/o List Net (Q&A)	0.2	27.9	7.4	7.1	-38.3	-14.7	9.8	48.1	27.6

Tone Measures - Changes - T1 to T63 Drift Returns in bps									
	Top Tercile of Surprises Terciles of Tone				Bottom Tercile of Surprises Terciles of Tone				Average
Variable	T1	T2	T3	Spread	T1	T2	T3	Spread	Spread
LM Net (MD)	-16.3	17.2	17.0	33.3	-75.1	-26.1	1.2	76.4	54.8
Diction Net (MD)	-1.5	-0.4	20.7	22.2	-70.8	-41.7	15.5	86.3	54.3
LIWC Net (MD)	8.6	10.2	-1.8	-10.4	-77.4	-27.0	7.3	84.7	37.2
All Dictionaries Net (MD)	-15.5	-0.6	35.3	50.8	-84.2	-8.6	-6.0	78.2	64.5
Bayes with LM Net (MD)	-18.0	14.7	22.4	40.4	-70.4	-32.8	6.0	76.4	58.4
Bayes w/o List Net (MD)	-10.4	20.9	6.9	17.4	-48.9	-38.0	-11.7	37.2	27.3
LM Net (Q&A)	-45.1	33.3	32.9	78.0	-88.7	-30.4	22.8	111.4	94.7
Diction Net (Q&A)	-5.6	4.2	25.6	31.2	-76.1	-26.5	6.9	83.0	57.1
LIWC Net (Q&A)	-26.8	19.3	33.3	60.1	-67.7	-24.4	-4.1	63.7	61.9
All Dictionaries Net (Q&A)	-37.3	15.8	51.5	88.8	-98.0	-24.3	29.1	127.1	107.9
Bayes with LM Net (Q&A)	-8.4	12.1	18.9	27.3	-96.7	-11.0	8.3	105.0	66.2
Bayes w/o List Net (Q&A)	-8.6	9.1	21.9	30.4	-89.0	-15.6	3.6	92.6	61.5

Tone Measures - Changes - T21 to T63 Drift Returns in bps									
	Top Tercile of Surprises Terciles of Tone				Bottom Tercile of Surprises Terciles of Tone				Average
Variable	T1	T2	T3	Spread	T1	T2	T3	Spread	Spread
LM Net (MD)	16.3	3.3	-22.3	-38.6	-19.1	-1.9	3.2	22.3	-8.1
Diction Net (MD)	21.7	-15.9	-7.2	-29.0	-19.1	13.4	-12.5	6.6	-11.2
LIWC Net (MD)	23.5	9.7	-38.1	-61.6	-29.0	12.2	-0.6	28.3	-16.6
All Dictionaries Net (MD)	1.4	8.3	-11.4	-12.8	-22.6	22.0	-17.6	5.1	-3.9
Bayes with LM Net (MD)	4.5	2.8	-9.3	-13.8	-18.5	-8.0	10.0	28.5	7.3
Bayes w/o List Net (MD)	-3.5	11.5	-10.3	-6.8	17.9	-25.1	-10.3	-28.2	-17.5
LM Net (Q&A)	-20.3	15.5	2.4	22.8	-48.8	-4.0	38.3	87.1	54.9
Diction Net (Q&A)	-2.9	-3.9	7.0	9.8	-36.2	-12.7	37.1	73.3	41.6
LIWC Net (Q&A)	-10.4	7.1	3.0	13.4	-26.0	-4.9	17.3	43.2	28.3
All Dictionaries Net (Q&A)	-21.2	12.7	9.6	30.9	-46.7	-3.4	38.4	85.1	58.0
Bayes with LM Net (Q&A)	-7.5	9.1	-2.9	4.6	-50.4	-13.0	50.1	100.5	52.6
Bayes w/o List Net (Q&A)	1.5	-15.0	15.8	14.3	-45.3	12.6	14.3	59.6	36.9

Source: FactSet, Macquarie Research, April 2015

The drift accrues gradually over time after the call…

To investigate these findings in more detail we plot the return, again in event time, for six portfolios obtained by double sorting. Out of the total of nine portfolios, we focus on the ones with high and low earnings surprises (Fig 49 and Fig 50). Also, we consider the signal obtained from the Q&A section of the call, which appears to deliver a stronger signal.

The results show that, after the initial market reaction at the time of the call, companies that delivered the most positive surprises and the most bullish tone (gray line in Fig 49) display a mild positive drift in returns. Similarly, in Fig 50 the companies with negative surprises and positive tone tend to overshoot more than those which disappointed and also displayed negative conference call tone.

…Particularly when the signal is based on changes in tone

The same pattern emerges, but with a more significant impact on returns, when we analyse the signal based on *changes* in sentiment. Fig 51 shows that, among companies that beat expectations, the ones with improved conference call sentiment continue to outperform. In contrast, positive surprise and negative tone tend to result in a lack of post earnings announcement drift (red line in Fig 51) or even a mild reversal of the gains obtained immediately after the call. The effect is even stronger for negative surprises (Fig 52). A company that disappointed and also displayed negative conference call tone tends to underperform steadily during our event window. In contrast, a company that disappointed but had a positive analyst call tends to rebound after the initial negative price effect (gray line in Fig 52).

Fig 49 High EPS Surprise with All Dictionaries Net (Q&A) Level

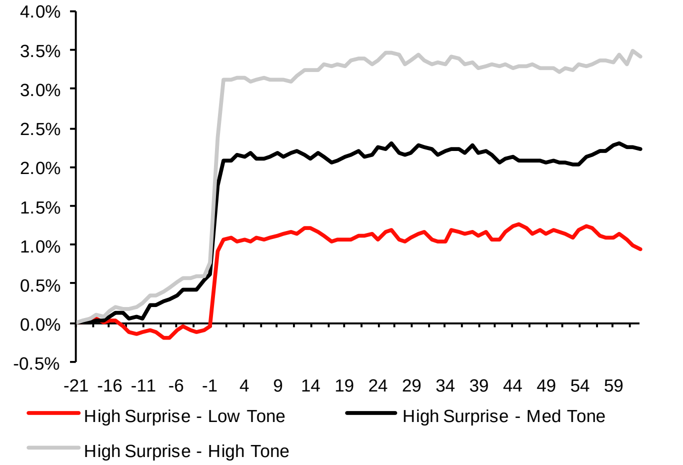
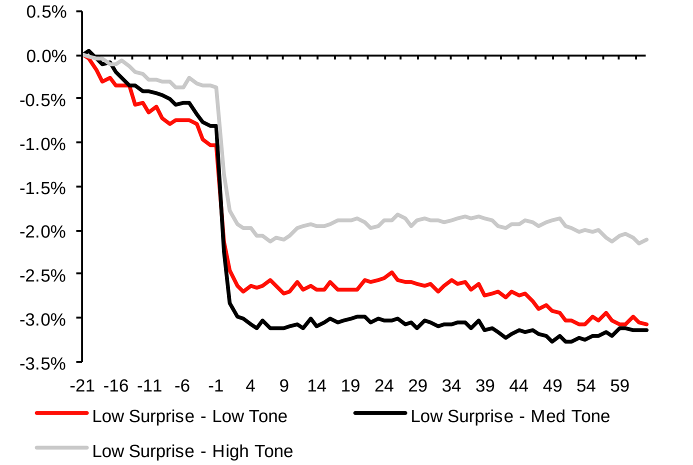


Fig 50 Low EPS Surprise with All Dictionaries Net (Q&A) Level



Source: FactSet, Macquarie Research, April 2015

Fig 51 High EPS Surprise with All Dictionaries Net (Q&A) Change

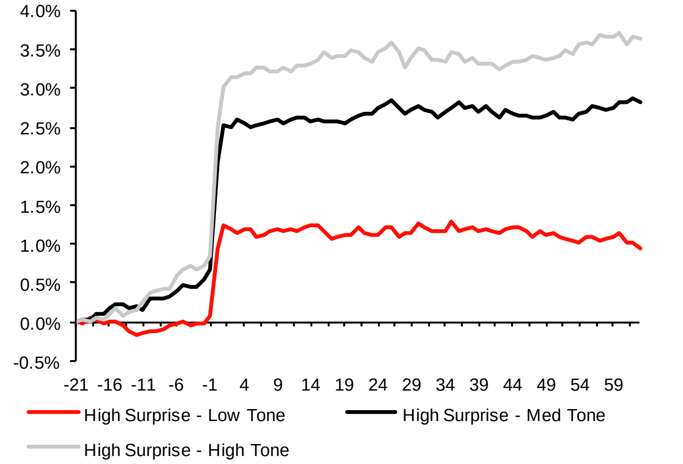
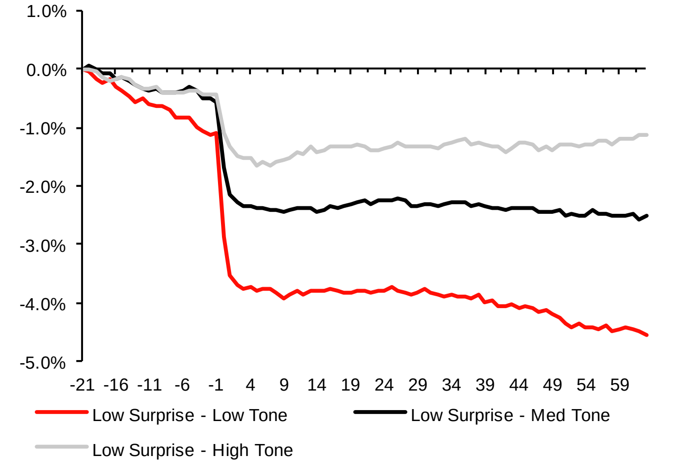


Fig 52 Low EPS Surprise with All Dictionaries Net (Q&A) Change



Source: FactSet, Macquarie Research, April 2015

A simple trading strategy

We run a simple backtest using buy and hold rules with a fixed horizon

In the last section we move from event studies to a simple backtest. Designing a simple trading strategy in our framework is particularly challenging given the seasonal nature of the signal, which tends to be available only during relatively short earnings seasons.

Each position is opened at the end of the second trading day after the call

We adapt the methodology of Gatev et al. (2006), who analysed a simple pairs trading strategy, to our problem. The main idea is to allocate the available capital evenly to n potential trades at the beginning of the backtest, keeping n constant over time. If no conference call has taken place at time (i.e. day) $t-1$ then we do not execute any trade. If a call did take place at time $t-1$ then we use the thresholds calculated as of $t-1$ to classify it as belonging to tercile 1, 2 or 3 and open a position if the call is in the top tercile. The position is assumed to be opened at the closing price of day t. The position remains open for τ days, where τ is also kept constant over time.

The maximum number of portfolio holdings is fixed in advance

Once all the n potential trades, each with the same share of the initial capital, have been activated we leave the portfolio unchanged until one of the trades reaches day τ since it was opened. At that point we sell the stock and check again if a new conference call with positive tone is available. If that is the case, then the proceeds of the sale are used to purchase shares in the new stock that has just been selected.

When there are not enough trading opportunities unused funds are invested in the index

In some periods there will be very few conference calls and therefore the allocated capital will not be fully used. Trades that are inactive for lack of available conference calls are assumed to hold an equally weighted index rather than cash. In other periods, particularly in the middle of the earnings season, there may be more than n companies reporting at the same time. If we need to fill m potential trades that are inactive at time t but there are more than m top tercile calls at time $t-1$, we randomise the inclusion in a way that each stock has the same probability of being chosen.

The short portfolio follows similar rules

A similar procedure is used to include bottom tercile trades in the short portfolio. At the beginning of the sample period it is assumed that the same amount C is held in cash as collateral and allocated to both the short and the long portfolios.

Finally, we rebalance on a monthly basis by calculating the net position of the strategy up to that point (i.e. the value of the long portfolio plus cash minus the value of the short one) and reallocating $1/n$ of that value to each potential trade. Returns are computed by marking to market the portfolio.

In practice this setup is equivalent to having n separate trading desks and a *passive* desk. Each of the active desks can only buy one stock at a time and hold it for τ days. The passive desk plays two roles:

1) It receives the proceeds of the sale when an active desk unwinds a position and switches to inactive

2) It allocates a proportion of the passively managed fund to any inactive desk which is about to open a new position. In this case the size of the position is equal to the amount held in the index divided by the number of inactive desks at that point in time.

Two signals: Single sort on surprises vs. double sort on surprises and tone

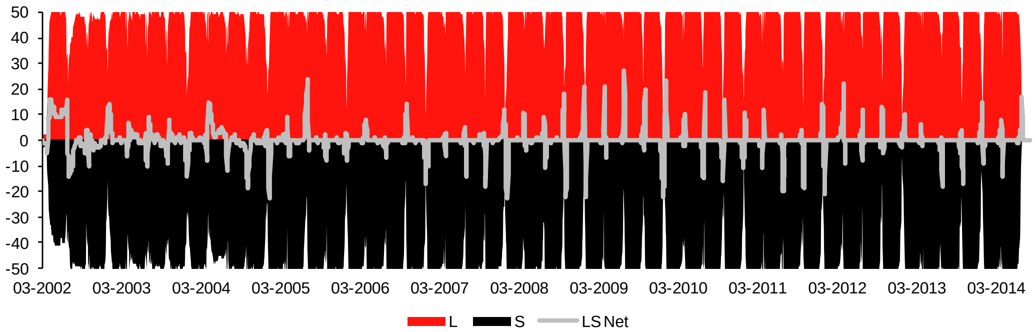
Throughout the backtest analysis we use two alternative signals. The first is based on a double sort on earnings surprise and tone – candidates for inclusion in the long portfolio must be in the top tercile sorted by tone of the top tercile sorted by surprise. The second signal, used as a benchmark, relies solely on surprises. In this case we sort stocks into six portfolios instead of three and, for inclusion in the long basket, we require a stock to be in the top bucket.

A maximum number of 50 stocks in each basket would work well for US data

How many positions should we aim for? To obtain some intuition on the effect of setting alternative values of n we run two backtests with $n = 50$ and $n = 100$ using US data only. Fig 53 shows the results for $n = 50$ when using a simple earnings surprises signal. The pattern over time shows that every quarter, just before the new earnings season starts, the number of active bets drops significantly. A similar pattern emerges in Fig 54 when we backtest our main signal based on tone, i.e. we identify the stocks with top tercile surprises and top tercile improvement in tone. However, the limited availability of data has an effect on the results at the beginning of the sample period.

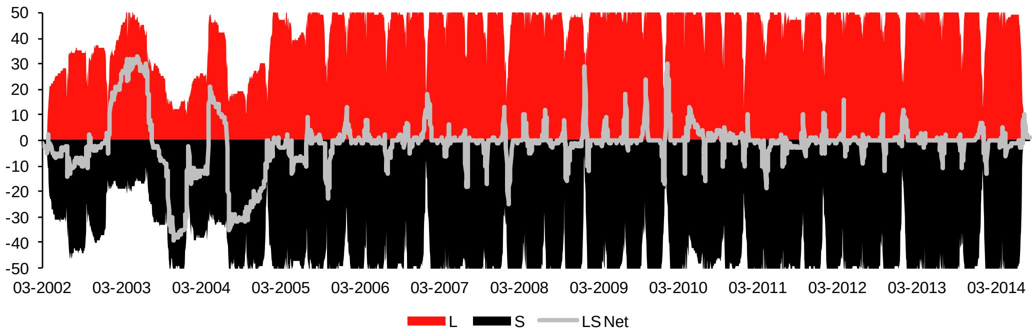
If we increase n to 100 (Fig 55 and Fig 56) the number of active positions decreases slightly at the beginning of the sample for surprises (Fig 55) while for our conference call sentiment signal (Fig 56) the full allocation of 100 positions is rarely reached even at the end of the sample period.

Fig 53 Number of active positions, earnings surprises, n=50



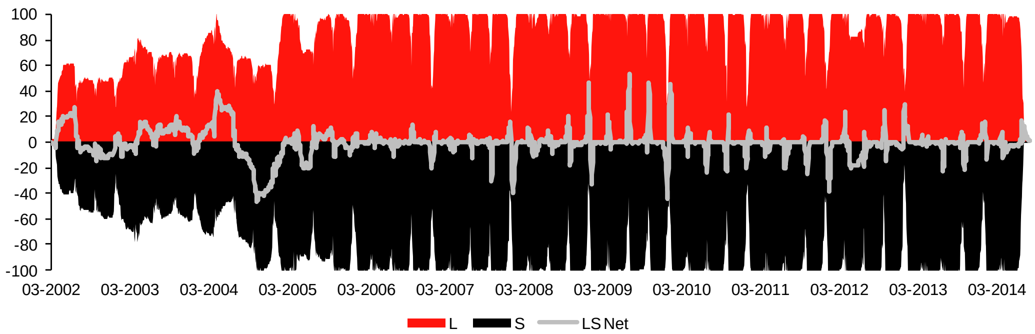
Source: FactSet, Macquarie Research, April 2015. The chart refers to the case with 50 slots and a holding period of 62 trading days, US stocks

Fig 54 Number of active positions, changes in Q&A tone, n=50



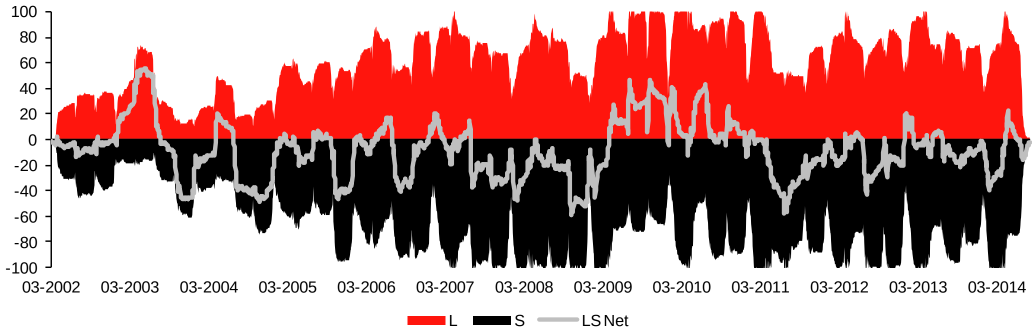
Source: FactSet, Macquarie Research, April 2015. The chart refers to the case with 50 slots and a holding period of 62 trading days, US stocks

Fig 55 Number of active positions, earnings surprises, n=100



Source: FactSet, Macquarie Research, April 2015. The chart refers to the case with 100 slots and a holding period of 62 trading days, US stocks

Fig 56 Number of active positions, changes in Q&A tone, n=100



Source: FactSet, Macquarie Research, April 2015. The chart refers to the case with 100 slots and a holding period of 62 trading days, US stocks

The long short strategy generates strong risk adjusted returns

Fig 57 and Fig 58 display the information ratios of different strategies based on the conference call tone signal extracted from the management discussion (Fig 57) and Q&A section (Fig 58), for a range of values of n (number of slots) and τ (holding period in trading days). The long side outperforms consistently the short side. While drawing any definite conclusion on the optimality of the strategies would be difficult, given the risk of data mining, it is reassuring to see that the risk adjusted returns are strongest for holding periods up to roughly three calendar months and a number of potential trades of at least 50.

As expected, the information obtained from the Q&A section of the calls seems to generate a more profitable signal compared to the management discussion.

Fig 57 Information ratio, double sort on earnings surprises and changes in MD tone, US market

Holding Period	Number of Slots: 10	20	50	80	100	150	200
20	0.21	0.54	0.67	0.67	0.69	0.67	0.65
41	0.57	0.33	0.78	0.71	0.72	0.68	0.56
62	0.46	0.52	0.77	0.84	0.77	0.65	0.51
124	0.17	0.19	0.39	0.26	0.31	0.31	0.22
249	-0.30	0.15	0.36	0.37	0.41	0.56	0.42

Source: FactSet, Macquarie Research, April 2015

Fig 58 Information ratio, double sort on earnings surprises and changes in Q&A tone, US market

Holding Period	Number of Slots: 10	20	50	80	100	150	200
20	0.46	0.87	0.95	0.85	0.83	0.84	0.84
41	-0.11	0.30	0.57	0.62	0.65	0.57	0.58
62	0.49	0.57	0.88	0.99	0.92	0.74	0.76
124	0.40	0.44	0.42	0.49	0.49	0.32	0.37
249	0.03	0.15	0.16	0.23	0.27	0.40	0.32

Source: FactSet, Macquarie Research, April 2015

By comparison, a similar strategy based on earnings surprises alone would not have been as profitable (Fig 59).

Returns are significant even adjusting for transaction costs

The numbers we have presented so far are gross of transaction costs but we argue that the amount of turnover is not so high as to render the strategy unprofitable. For example, in the case of $n = 50$ and a holding period of 62 trading days if we assume 10bps of transaction costs one way then the information ratio reduces from 0.88 to approximately 0.58. The same assumption implies, in the case $n = 100$, a reduction in IR from 0.92 to approximately 0.56.

Fig 59 Information ratio, earnings surprises, US market

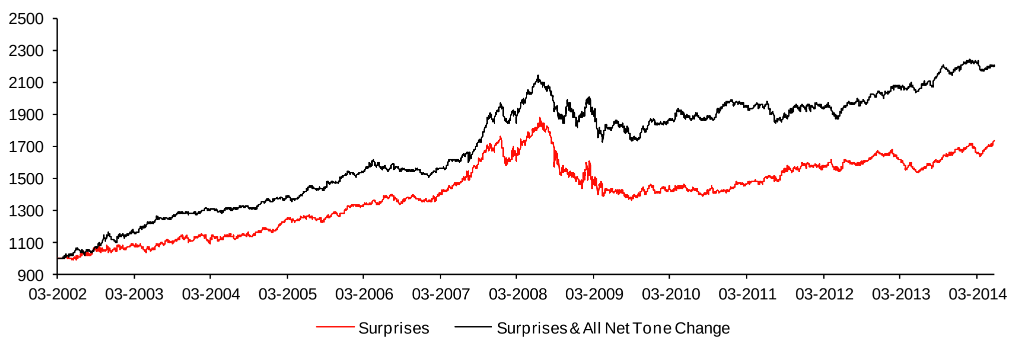
Holding Period	Number of Slots						
	10	20	50	80	100	150	200
20	-0.04	0.25	0.33	0.30	0.16	0.38	0.45
41	-0.33	-0.10	0.18	0.23	0.20	0.15	0.33
62	0.32	0.20	0.49	0.34	0.44	0.42	0.33
124	-0.21	-0.05	0.07	0.11	0.25	0.27	0.25
249	0.17	-0.12	0.22	0.11	0.19	0.22	0.36

Source: FactSet, Macquarie Research, April 2015

Sorting on conference call sentiment would have avoided the worst drawdowns

To compare the two strategies in more detail we plot the corresponding total return indices in Fig 60. The conference call sentiment signal in this case is the change in Q&A tone. Our strategy would have improved on the earnings surprises portfolio by mitigating some of the worst drawdowns and, in general, accruing profits in a smoother way over time.

Fig 60 Total return, changes in Q&A tone and earnings surprises, US market



Source: FactSet, Macquarie Research, April 2015. The chart refers to the case with 50 slots

Returns are consistently positive in Europe…

The results of the backtests for European data are reported in Fig 61, Fig 62 and Fig 63. Risk adjusted returns to the strategies based on conference call tone (Fig 61 and Fig 62) are almost invariably positive, suggesting that the long side outperforms the short one. The values of n which generate the best results are lower than in the US backtest, probably due to the more limited availability of data (particularly in the early part of the sample period).

… although the strategy is less profitable

Information ratios are generally lower than in the US backtest. We attribute this result to the fact that our long and short portfolios in Europe are not as well diversified as in the US, due to data limitations (the coverage is less extensive at the beginning of the sample) and the lower frequency of conference calls. Both factors reduce the availability of potential trades.

By comparing the results in Fig 61 and Fig 62 with those in Fig 63 we conclude that, in Europe, the evidence that we can improve on the risk adjusted return of a simple earnings surprise strategy is weaker than in the US.

Fig 61 Information ratio, double sort on earnings surprises and changes in MD tone, European market

Holding Period	Number of Slots: 10	20	50	80	100	150	200
20	0.34	0.69	0.68	0.62	0.56	0.48	0.48
41	0.37	0.53	0.49	0.43	0.38	0.29	0.28
62	0.08	0.40	0.47	0.37	0.30	0.23	0.22
124	0.24	0.43	0.36	0.43	0.27	0.30	0.23
249	0.00	0.25	0.50	0.51	0.56	0.54	0.49

Source: FactSet, Macquarie Research, April 2015

Fig 62 Information ratio, double sort on earnings surprises and changes in Q&A tone, European market

Holding Period	Number of Slots: 10	20	50	80	100	150	200
20	0.40	0.69	0.66	0.57	0.55	0.52	0.51
41	0.04	0.26	0.59	0.42	0.36	0.35	0.35
62	0.37	0.51	0.73	0.59	0.41	0.38	0.38
124	0.24	0.30	0.36	0.48	0.50	0.42	0.27
249	0.24	-0.12	0.09	0.44	0.42	0.52	0.50

Source: FactSet, Macquarie Research, April 2015

Fig 63 Information ratio, earnings surprises, European market

Holding Period	Number of Slots: 10	20	50	80	100	150	200
20	-0.21	0.38	0.33	0.60	0.68	0.53	0.49
41	0.03	0.17	0.26	0.54	0.68	0.43	0.26
62	0.02	0.14	0.48	0.64	0.55	0.41	0.26
124	0.36	0.29	0.34	0.56	0.51	0.62	0.41
249	-0.01	0.07	0.21	0.42	0.32	0.50	0.70

Source: FactSet, Macquarie Research, April 2015

Conclusion

This report has analysed a large sample of global analyst conference call transcripts. In particular, we have focused on the language used by management and analysts in order to obtain a quantitative measure of sentiment, or tone. We reach two main conclusions.

First, the *soft information* available in conference calls does predict future stock returns. This is arguably due to the fact that investors are overly focused on whether a company beats or misses analyst expectations, which implies underreaction to the broader information content of the call.

Second, we showed that it is possible to exploit systematically this predictability by analysing sentiment through simple text mining techniques applied to conference call transcripts. While simple, these methods are sufficiently accurate to generate profitable trading signals. The resulting strategies have typical investment horizons of one month or longer and therefore do not require frequent trading with holding periods of just a few days.

Further work

As we obtained access to a new database of transcripts that allow us to parse questions and answers separately as well as recognise speakers' names and positions, we identify potential areas of further research:

- The impact of participants on calls' tone and informativeness – *Cicon (2014)* finds that CEO participation inhibits information discovery, in contrast number of analysts active on the call contributes positively.
- Measures of tone controlled for manager-specific tone – *Davis et al. (2014)* find that part of the abnormal tone of conference call has a significant manager-specific component.
- The impact of analysts' tone and participation on earnings forecast accuracy – *Mayew et al. (2013)* find that analysts who participated actively in conference call issue more accurate forecasts.

References

Antweiler, W. and Frank, M., "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards" (2004). Journal of Finance 59, 1259-1294.

Borochin, P., Cicon, J., DeLisle, J., and McKay Price, S., "The Effects of Conference Call Content on Market Perceptions of Firm Risk" (2015). Available at SSRN: http://ssrn.com/abstract=2579907.

Brochet, F., Loumioti, M., and Serafeim, G., "Short-termism, Investor Clientele, and Corporate Performance" (2014). HBS Working Paper, 12-073

Buehlmaier, M., and Whited, T., "Looking for Risk in Words: A Narrative Approach to Measuring the Pricing Implications of Financial Constraints" (2014). Simon School Working Paper No. FR 14-11. Available at SSRN: http://ssrn.com/abstract=2435116

Buehlmaier, M., and Zechner, J., "Slow-Moving Real Information in Merger Arbitrage" (2014). Available at SSRN: http://ssrn.com/abstract=2187144

Bushee, B., Gow, I., and Taylor, D., "Linguistic Complexity in Firm Disclosures: Obfuscation or Information?" (2014). Available at SSRN: http://ssrn.com/abstract=2375424.

Chen, J., Demers, E.A., and Lev, B., "Oh What a Beautiful Morning! The Time of Day Effect on the Tone and Market Impact of Conference Calls" (2013). Darden Business School Working Paper No. 2186862. Available at SSRN: http://ssrn.com/abstract=2186862.

Chen, J., Nagar, V. and Schoenfeld, J., "Sources of Analyst Expertise" (2014). Ross School of Business Paper No. 1197. Available at SSRN: http://ssrn.com/abstract=2084488

Chen, S., Hollander, S., and Law, K., "The Price of Silence: When No One Asks Questions During Conference Calls" (2015). Available at SSRN: http://ssrn.com/abstract=2449341

Cicon, J., "Say it again Sam: The Idiosyncratic Information Content of Corporate Conference Calls" (2014). Available at SSRN: http://ssrn.com/abstract=2389391

Das, S. and Chen, M., "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web" (2007). Management Science 53, 1375-1388.

Davis, A., Ge, W., Matsumoto, D. and Zhang, J., "The Effect of Manager-Specific Optimism on the Tone of Earnings Conference Calls" (2014). Review of Accounting Studies, Forthcoming.

Druz, M., Wagner, A. and Zeckhauser, R., "Tips and Tells from Managers: How Analysts and the Market Read Between the Lines of Conference Calls" (2015). HKS Faculty Research Working Paper, RWP15-006.

Gatev, E., Goetzmann, W.N., and Rouwenhorst, K.G., "Pairs trading: Performance of a relative-value arbitrage rule" (2006). Review of Financial Studies 19, 797-827.

Henry, E., "Are Investors Influenced by How Earnings Press Releases are Written?" (2008). International Journal of Business Communication 45, 363-407.

Heston, S. and Sinha, N., "News versus Sentiment: Comparing Textual Processing Approaches for Predicting Stock Returns" (2014). Robert H. Smith School Research Paper. Available at SSRN: http://ssrn.com/abstract=2311310

Huang, A.H., Zang, A.Y., and Zheng, R., "Evidence on the Information Content of Text in Analyst Reports" (2014). Accounting Review 89, 2151-2180.

Jegadeesh, N. and Wu, A., "Word Power: A New Approach for Content Analysis" (2013). Journal of Financial Economics 110, 712-729.

Kearney, C. and Liu, S., "Textual Sentiment in Finance: A Survey of Methods and Models" (2014). International Review of Financial Analysis 33, 171-185.

Kogan, S., Levin, D., Routledge, B., Sagi, J. and Smith, N., "Predicting Risk from Financial Reports with Regression" (2009). Available online: http://www.cs.cmu.edu/~nasmith/papers/kogan+levin+routledge+sagi+smith.naacl09.pdf

Larcker, D.F., and Zalkolyukina, A.A., "Detecting deceptive discussions in conference calls" (2012). Journal of Accounting Research 50, 495-540.

Li, F., “Textual Analysis of Corporate Disclosures: A Survey of the Literature” (2010a). Journal of Accounting Literature 29, 143-165.

Li, F., “The Information Content of Forward-looking Statements in Corporate Filings - A Naive Bayesian Machine Learning Approach” (2010b). Journal of Accounting Research 48, 1049-1102.

Loughran, T. and McDonald, B., “Textual Analysis in Accounting and Finance: A Survey” (2015). Available at SSRN: http://ssrn.com/abstract=2504147

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

Mayew, W., Sharp, N., and Venkatachalam, M., “Using Earnings Conference Calls to Identify Analysts with Superior Private Information” (2013). Review of Accounting Studies 18, 386-413.

McKay Price, S., Doran, J., Peterson, D. and Bliss, B., “Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone” (2012). Journal of Banking and Finance 36, 992-1011.

Ryans, J., “Textual Classification of SEC Comment Letters” (2014). Berkeley Haas Working Paper Series. Available at SSRN: http://ssrn.com/abstract=2474666

Tetlock, P., Saar-Tsechansky, M., and Macskassy, S., “More than Words: Quantifying Language to Measure Firms' Fundamentals” (2008). Journal of Finance 63, 1437-1467.

Wang, Ch., Tsai, M., Liu, T., and Chang, Ch., “Financial Sentiment Analysis for Risk Prediction” (2013). Proceedings of the International Joint Conference on Natural Language Processing. Available online: http://www.aclweb.org/anthology/I13-1097.

Appendices

The Naïve Bayes approach as a MAP estimator

This section describes in more detail the Bayesian framework typically used to derive the Naïve Bayes estimator as a Maximum A Posteriori (MAP) estimator. Assume that we have a dictionary of m words and a document which can be classified as positive, neutral or negative.

A random variable distributed as a Dirichlet with parameter α has pdf

$$f_{Dir}(x) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

where $\Gamma(x)$ is the gamma function, $0 \leq x_i \leq 1$ for each $i = 1, \ldots, K$ and $\sum_i x_i = 1$.

The full setup consists of a hierarchy of random variables:

$$W_j | D = z, \phi_j \sim Categorical(\phi_{jz}), j = 1, \ldots, m, z \in \{pos, neg, neu\}$$

where $D|\theta \sim Categorical(\theta)$ and in turn

$$\theta \sim Dirichlet(\alpha), \phi_j \sim Dirichlet(\beta_j)$$

The $2 \times m$ hyperparameter matrix B and the 3×1 hyperparameter vector α govern the shape of the priors.

The posteriors are

$$\theta \big| D \sim Dirichlet(c + \alpha), \phi_j \big| D = z, W \sim Dirichlet(d_{jz} + \beta_j)$$

where c is a vector that contains the counts of observations falling in the three categories while d_{jz} is a vector which contains the counts of occurrences of word j in sentences of category z.

The mode of a Dirichlet random variable is

$$\frac{\alpha_i - 1}{\sum_i \alpha_i - K}$$

which implies that

$$\hat{\phi}_{jz,MAP} = \frac{d_{jz} + \beta_j - 1}{\sum_z d_{jz} + 2\beta_j - 2}$$

Typically the prior is chosen with $\beta_j = 1$ for any j, which avoids the issue of dealing with estimated conditional probabilities that are equal to zero. Each β_j can be interpreted as a pseudo-count of observations.

In our experiments we use data-driven priors that depend on a hyperparameter $\kappa \geq 1$. In particular, we set the mean of our priors according to the following rules. Suppose that word j appears in the positive word list. Then we set the prior mean of $\phi_{j,neu}$ equal to the overall frequency of the word, the prior mean of $\phi_{j,pos}$ equal to κ times the frequency and the prior mean of $\phi_{j,neg}$ equal to the frequency divided by κ. Priors for words on the negative list are set along the same lines.

Results for sentiment measures based on Management Discussion

Fig 64-Fig 75 show the average drifts in event time for terciles of stocks sorted by sentiment measures calculated from the management discussion section of the call, as opposed to the Q&A section which is used in the rest of the report.

Fig 64 LM Net (MD) Abnormal Level

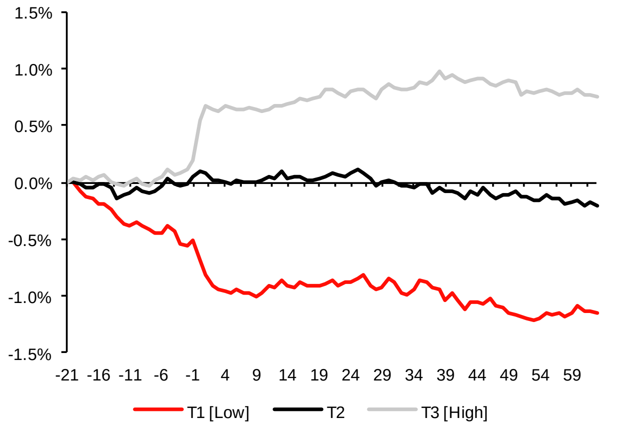


Fig 65 Diction Net (MD) Abnormal Level

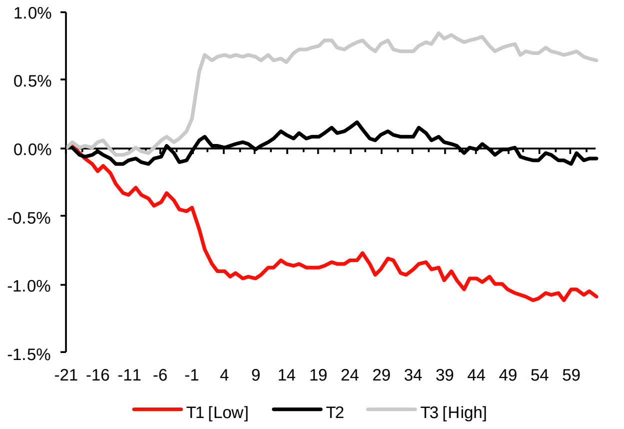


Fig 66 LIWC Net (MD) Abnormal Level

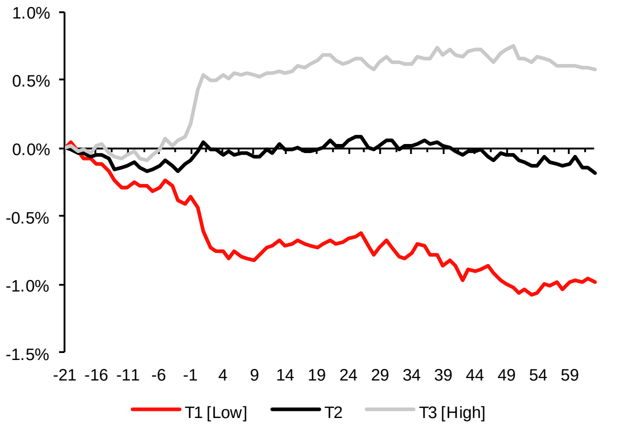


Fig 67 All Dictionaries Net (MD) Abnormal Level

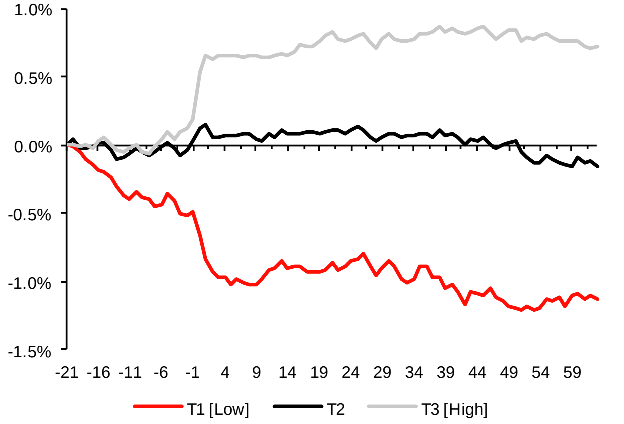


Fig 68 Bayes with LM Net (MD) Abnormal Level

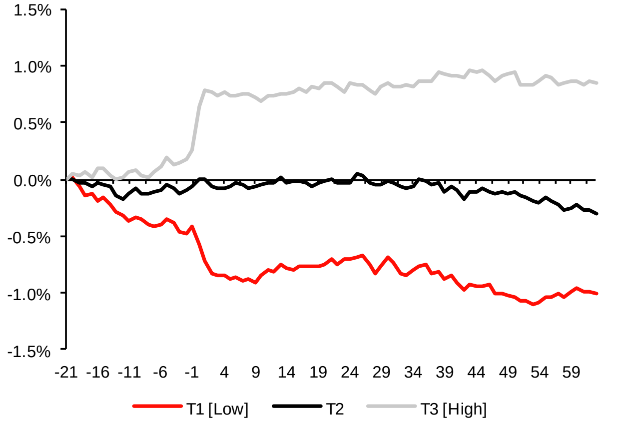
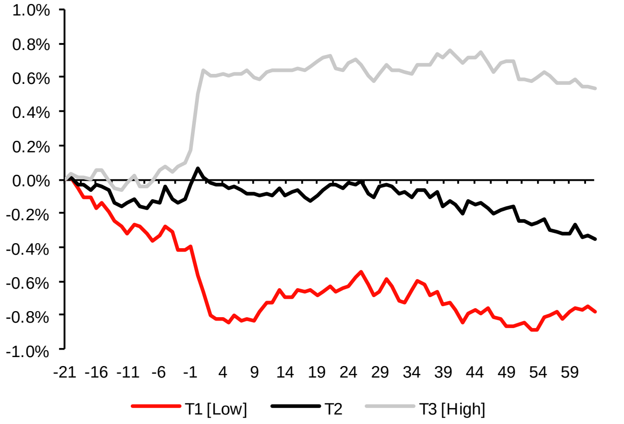


Fig 69 Bayes w/o List Net (MD) Abnormal Level



Source: FactSet, Macquarie Research, April 2015

Fig 70 LM Net (MD) Abnormal Change

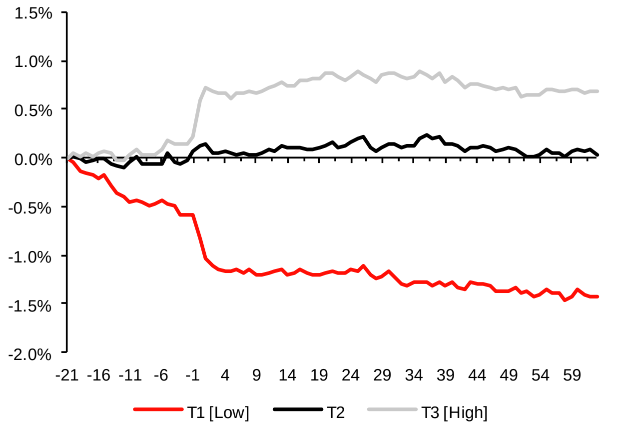


Fig 71 Diction Net (MD) Abnormal Change

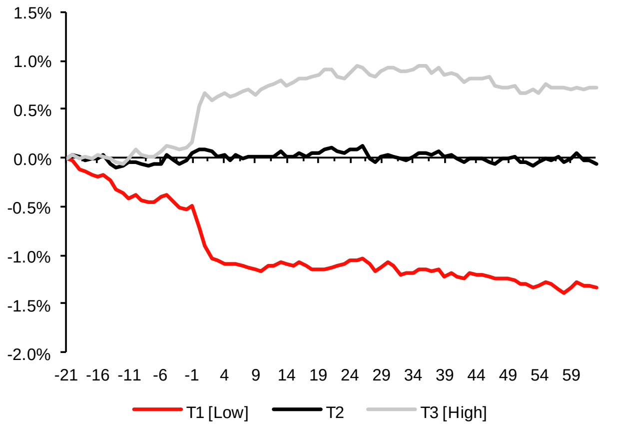


Fig 72 LIWC Net (MD) Abnormal Change

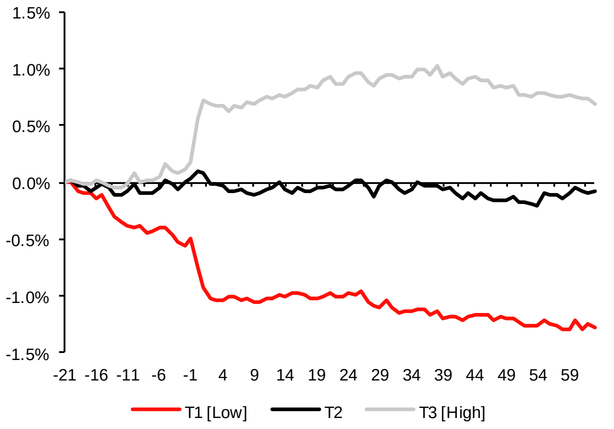


Fig 73 All Dictionaries Net (MD) Abnormal Change

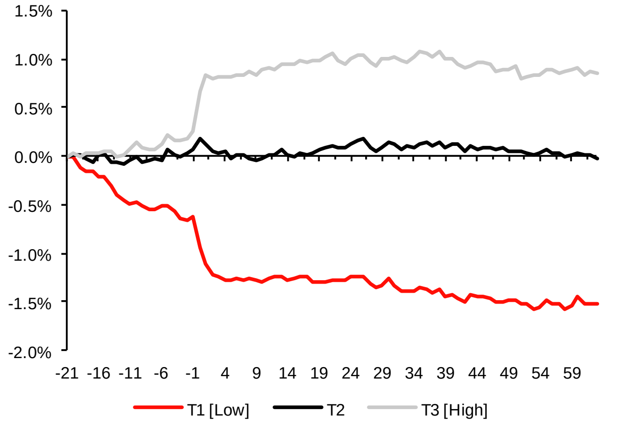


Fig 74 Bayes with LM Net (MD) Abnormal Change

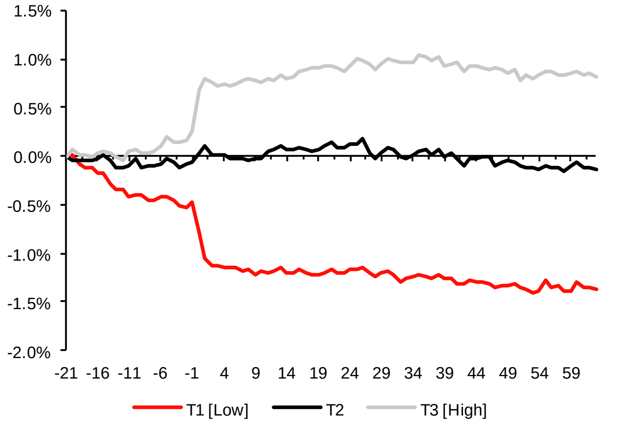
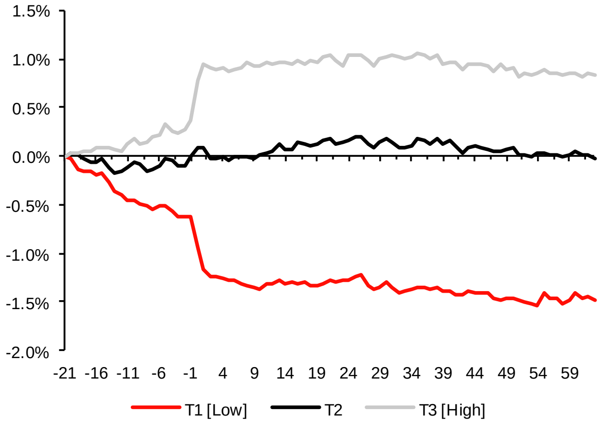


Fig 75 Bayes w/o List Net (MD) Abnormal Change



Source: FactSet, Macquarie Research, April 2015

Important disclosures:

Recommendation definitions

Macquarie - Australia/New Zealand
Outperform – return >3% in excess of benchmark return
Neutral – return within 3% of benchmark return
Underperform – return >3% below benchmark return

Benchmark return is determined by long term nominal GDP growth plus 12 month forward market dividend yield

Macquarie – Asia/Europe
Outperform – expected return >+10%
Neutral – expected return from -10% to +10%
Underperform – expected return <-10%

Macquarie First South - South Africa
Outperform – expected return >+10%
Neutral – expected return from -10% to +10%
Underperform – expected return <-10%

Macquarie - Canada
Outperform – return >5% in excess of benchmark return
Neutral – return within 5% of benchmark return
Underperform – return >5% below benchmark return

Macquarie - USA
Outperform (Buy) – return >5% in excess of Russell 3000 index return
Neutral (Hold) – return within 5% of Russell 3000 index return
Underperform (Sell)– return >5% below Russell 3000 index return

Volatility index definition*

This is calculated from the volatility of historical price movements.

Very high–highest risk – Stock should be expected to move up or down 60–100% in a year – investors should be aware this stock is highly speculative.

High – stock should be expected to move up or down at least 40–60% in a year – investors should be aware this stock could be speculative.

Medium – stock should be expected to move up or down at least 30–40% in a year.

Low–medium – stock should be expected to move up or down at least 25–30% in a year.

Low – stock should be expected to move up or down at least 15–25% in a year.
* Applicable to Asia/Australian/NZ/Canada stocks only

Recommendations – 12 months
Note: Quant recommendations may differ from Fundamental Analyst recommendations

Financial definitions

All "Adjusted" data items have had the following adjustments made:
Added back: goodwill amortisation, provision for catastrophe reserves, IFRS derivatives & hedging, IFRS impairments & IFRS interest expense
Excluded: non recurring items, asset revals, property revals, appraisal value uplift, preference dividends & minority interests

EPS = adjusted net profit / efpowa*
ROA = adjusted ebit / average total assets
ROA Banks/Insurance = adjusted net profit /average total assets
ROE = adjusted net profit / average shareholders funds
Gross cashflow = adjusted net profit + depreciation
*equivalent fully paid ordinary weighted average number of shares

All Reported numbers for Australian/NZ listed stocks are modelled under IFRS (International Financial Reporting Standards).

Recommendation proportions – For quarter ending 31 March 2015

	AU/NZ	Asia	RSA	USA	CA	EUR	
Outperform	48.99%	59.51%	49.30%	43.79%	59.59%	52.20%	(for US coverage by MCUSA, 7.42% of stocks followed are investment banking clients)
Neutral	34.12%	26.62%	35.21%	50.29%	34.93%	31.32%	(for US coverage by MCUSA, 5.68% of stocks followed are investment banking clients)
Underperform	16.89%	13.87%	15.49%	5.93%	5.48%	16.48%	(for US coverage by MCUSA, 0.87% of stocks followed are investment banking clients)

Company-specific disclosures:

Important disclosure information regarding the subject companies covered in this report is available at www.macquarie.com/disclosures.

Analyst certification:
The views expressed in this research accurately reflect the personal views of the analyst(s) about the subject securities or issuers and no part of the compensation of the analyst(s) was, is, or will be directly or indirectly related to the inclusion of specific recommendations or views in this research. The analyst principally responsible for the preparation of this research receives compensation based on overall revenues of Macquarie Group Ltd ABN 94 122 169 279 (AFSL No. 318062) (MGL) and its related entities (the Macquarie Group) and has taken reasonable care to achieve and maintain independence and objectivity in making any recommendations.

General disclaimers:
Macquarie Securities (Australia) Ltd; Macquarie Capital (Europe) Ltd; Macquarie Capital Markets Canada Ltd; Macquarie Capital Markets North America Ltd; Macquarie Capital (USA) Inc; Macquarie Capital Securities Ltd and its Taiwan branch; Macquarie Capital Securities (Singapore) Pte Ltd; Macquarie Securities (NZ) Ltd; Macquarie First South Securities (Pty) Limited; Macquarie Capital Securities (India) Pvt Ltd; Macquarie Capital Securities (Malaysia) Sdn Bhd; Macquarie Securities Korea Limited and Macquarie Securities (Thailand) Ltd are not authorized deposit-taking institutions for the purposes of the Banking Act 1959 (Commonwealth of Australia), and their obligations do not represent deposits or other liabilities of Macquarie Bank Limited ABN 46 008 583 542 (MBL) or MGL. MBL does not guarantee or otherwise provide assurance in respect of the obligations of any of the above mentioned entities. MGL provides a guarantee to the Monetary Authority of Singapore in respect of the obligations and liabilities of Macquarie Capital Securities (Singapore) Pte Ltd for up to SGD 35 million. This research has been prepared for the general use of the wholesale clients of the Macquarie Group and must not be copied, either in whole or in part, or distributed to any other person. If you are not the intended recipient you must not use or disclose the information in this research in any way. If you received it in error, please tell us immediately by return e-mail and delete the document. We do not guarantee the integrity of any e-mails or attached files and are not responsible for any changes made to them by any other person. MGL has established and implemented a conflicts policy at group level (which may be revised and updated from time to time) (the "Conflicts Policy") pursuant to regulatory requirements (including the FCA Rules) which sets out how we must seek to identify and manage all material conflicts of interest. Nothing in this research shall be construed as a solicitation to buy or sell any security or product, or to engage in or refrain from engaging in any transaction. In preparing this research, we did not take into account your investment objectives, financial situation or particular needs. Macquarie salespeople, traders and other professionals may provide oral or written market commentary or trading strategies to our clients that reflect opinions which are contrary to the opinions expressed in this research. Macquarie Research produces a variety of research products including, but not limited to, fundamental analysis, macro-economic analysis, quantitative analysis, and trade ideas. Recommendations contained in one type of research product may differ from recommendations contained in other types of research, whether as a result of differing time horizons, methodologies, or otherwise. Before making an investment decision on the basis of this research, you need to consider, with or without the assistance of an adviser, whether the advice is appropriate in light of your particular investment needs, objectives and financial circumstances. There are risks involved in securities trading. The price of securities can and does fluctuate, and an individual security may even become valueless. International investors are reminded of the additional risks inherent in international investments, such as currency fluctuations and international stock market or economic conditions, which may adversely affect the value of the investment. This research is based on information obtained from sources believed to be reliable but we do not make any representation or warranty that it is accurate, complete or up to date. We accept no obligation to correct or update the information or opinions in it. Opinions expressed are subject to change without notice. No member of the Macquarie Group accepts any liability whatsoever for any direct, indirect, consequential or other loss arising from any use of this research and/or further communication in relation to this research. Clients should contact analysts at, and execute transactions through, a Macquarie Group entity in their home jurisdiction unless governing law permits otherwise. The date and timestamp for above share price and market cap is the closed price of the price date. #CLOSE is the final price at which the security is traded in the relevant exchange on the date indicated.

Country-specific disclaimers:
Australia: In Australia, research is issued and distributed by Macquarie Securities (Australia) Ltd (AFSL No. 238947), a participating organisation of the Australian Securities Exchange. **New Zealand**: In New Zealand, research is issued and distributed by Macquarie Securities (NZ) Ltd, a NZX Firm. **Canada**: In Canada, research is prepared, approved and distributed by Macquarie Capital Markets Canada Ltd, a participating organisation of the Toronto Stock Exchange, TSX Venture Exchange & Montréal Exchange. Macquarie Capital Markets North America Ltd., which is a registered broker-dealer and member of FINRA, accepts responsibility for the contents of reports issued by Macquarie Capital Markets Canada Ltd in the United States and sent to US persons. Any US person wishing to effect transactions in the securities described in the reports issued by Macquarie Capital Markets Canada Ltd should do so with Macquarie Capital Markets North America Ltd. The Research Distribution Policy of Macquarie Capital Markets Canada Ltd is to allow all clients that are entitled to have equal access to our research. **United Kingdom**: In the United Kingdom, research is issued and

Auckland Tel: (649) 377 6433	**Bangkok** Tel: (662) 694 7999	**Calgary** Tel: (1 403) 294 9541	**Denver** Tel: (303) 952 2800	**Frankfurt** Tel: (069) 509 578 000	**Geneva** Tel: (41) 22 818 7777	**Hong Kong** Tel: (852) 2823 3588
Jakarta Tel: (62 21) 515 1818	**Johannesburg** Tel: (2711) 583 2000	**Kuala Lumpur** Tel: (60 3) 2059 8833	**London** Tel: (44 20) 3037 2000	**Manila** Tel: (63 2) 857 0888	**Melbourne** Tel: (613) 9635 8139	**Mumbai** Tel: (91 22) 6653 3000
Munich Tel: (089) 2444 31800	**New York** Tel: (1 212) 231 2500	**Paris** Tel: (33 1) 7842 3823	**Perth** Tel: (618) 9224 0888	**Seoul** Tel: (82 2) 3705 8500	**Shanghai** Tel: (86 21) 6841 3355	**Singapore** Tel: (65) 6601 1111
Sydney Tel: (612) 8232 9555	**Taipei** Tel: (886 2) 2734 7500	**Tokyo** Tel: (81 3) 3512 7900	**Toronto** **Tel: (1 416) 848 3500**	**Vancouver** Tel: (1 604) 605 3944		

MACQUARIE

EMEA Research

Heads of Equity Research

John O'Connell (Global)	(612) 8232 7544
Shai Hill (Europe)	(44 20) 3037 4232
George Brits (South Africa)	(2711) 583 2223

Consumer Staples

Food & Beverages

Sreedhar Mahamkali (London)	(44 20) 3037 4016

Consumer Discretionary

Retailing

Sreedhar Mahamkali (London)	(44 20) 3037 4016
Charlie Storey (London)	(44 20) 3037 2327

Energy

David Farrell (London)	(44 20) 3037 4465
Giacomo Romeo (London)	(44 20) 3037 4445
Joe Stokeld (London)	(44 20) 3037 4457
Kate Sloan (London)	(44 20) 3037 4453
Aditya Suresh (Hong Kong)	(852) 3922 1265
Gerhard Engelbrecht (Johannesburg)	(2711) 583 2407

Alternative Energy & Utilities

Shai Hill (Europe)	(44 20) 3037 4232
Dominic Nash (Europe)	(44 20) 3037 4239
Jose Ruiz (Europe)	(44 20) 3037 1912
Peter Crampton (Europe)	(44 20) 3037 4559

Financials

Diversified Financials

Neil Welch (London)	(44 20) 3037 4272

Banks

Edward Firth (Europe)	(44 20) 3037 4077
Benjie Creelan-Sandford (London)	(44 20) 3037 4081
Thomas Stoegner (London)	(44 20) 3037 4172
Piers Brown (London)	(44 20) 3037 4044
Patrick Lee (London)	(44 20) 3037 4482
Chantal Sirisena (London)	(44 20) 3037 2190

Insurance

Neil Welch (London)	(44 20) 3037 4272

Industrials

Capital Goods

Rowan Goeller (Johannesburg)	(2711) 583 2131

Autos

Christian Breitsprecher (Frankfurt)	(49 69) 50957 8014
Jens Schattner (Frankfurt)	(49 69) 50957 8026

Transportation – Infrastructure

Rowan Goeller (Johannesburg)	(2711) 583 2131

Materials

Chemicals/Containers, Packaging/Paper & Forest Products, Construction Materials

Rowan Goeller (Johannesburg)	(2711) 583 2131
Al Kabili (New York)	(1 212) 231 0683

Global Metals & Mining

Jeff Largey (London)	(44 20) 3037 4359
Alon Olsha (London)	(44 20) 3037 2637
Daniel Lurch (London)	(44 20) 3037 5329
Kieran Daly (Johannesburg)	(2711) 583 2208
James Oberholzer (Johannesburg)	(2711) 583 2367
Gerhard Engelbrecht (Johannesburg)	(2711) 583 2407
Adrian Wood (Sydney)	(612) 8232 8531

Pharmaceuticals

Ada Li (Johannesburg)	(2711) 583 2181

Real Estate

Property Trusts & Developers

Leon Allison (Johannesburg)	(2711) 583 2209
Nazeem Samsodien (Cape Town)	(2721) 813 2771

TMET

Telecommunications

Guy Peddy (London)	(44 20) 3037 4509

Media

Tim Nollen (New York)	(1 212) 231 0635

Commodities & Precious Metals

Colin Hamilton (Global)	(44 20) 3037 4061
Jim Lennon (London)	(44 20) 3037 4271
Matthew Turner (London)	(44 20) 3037 4340
Stefan Ljubisavljevic (London)	(44 20) 3037 4247
Chris Gadd (London)	(44 20) 3037 1957
Vivienne Lloyd (London)	(44 20) 3037 4530
Yubin Fu (London)	(44 20) 3037 2622
Daryna Kovalska (London)	(44 20) 3037 2732
Graeme Train (Shanghai)	(86 21) 2412 9035
Angela Bi (Shanghai)	(86 21) 2412 9086

European Macro Group

Economics & Strategy

Elna Moolman (Johannesburg)	(2711) 583 2570

Strategy

George Brits (South Africa)	(2711) 583 2223

Quantitative

Gurvinder Brar (Global)	(44 20) 3037 4036
Giuliano De Rossi (London)	(44 20) 3037 1997
Ioan Mirciov (London)	(44 20) 3037 1939
Josiah Rudolph (Johannesburg)	(2711) 583 2210

Find our research at

Macquarie:	www.macquarie.com.au/research
Thomson:	www.thomson.com/financial
Reuters:	www.knowledge.reuters.com
Bloomberg:	MAC GO
Factset:	http://www.factset.com/home.aspx
CapitalIQ	www.capitaliq.com

Contact Gareth Warfield for access (612) 8232 3207

Email addresses

FirstName.Surname@macquarie.com

eg. **john.oconnell@macquarie.com**

Equities

Stevan Vrcelj (Global Head)	**(612) 8232 5999**
Dipesh Patel (London)	(44 20) 3037 4978
Sarah-Jane Wagg (Johannesburg)	(2711) 583 2000

European Execution Services

Anthony Bray (London)	**(44 20) 3037 4603**
Robert Tappin (London)	(44 20) 3037 4827
Wayne Drayton (London)	(44 20) 3037 4980
KC O'Rourke (London)	(44 20) 3037 4910
Danny Want (London)	(44 20) 3037 4847
Peter Homan (London)	(44 20) 3037 4740
Gregorio Esmelian (London)	(44 20) 3037 4973
Will Rogers (London)	(44 20) 3037 1721
Richard McGlashan (London)	(44 20) 3037 4824
Andreas Bergmann (London)	(44 20) 3037 4680
Robert Lenihan (London)	(44 20) 3037 4908
Richard Bateson (London)	**(44 20) 3037 4821**
David Hemming (London)	(44 20) 3037 4909
Nick Bryan (London)	(44 20) 3037 4768
Chris Charbonnier (London)	(44 20) 3037 4760
Marc Crome (London)	(44 20) 3037 4778
George Sampson (London)	(44 20) 3037 1732
Katie Ramsey (London)	(44 20) 3037 4919

ETF Sales Trading

Robin Quinnell (London)	**(44 20) 3023 8885**
Bachir Binebine (Hong Kong)	(852) 3922 2089
Chris Johnson (New York)	(1 212) 231 0881

US Sales Trading

Chris Reale (New York)	**(1 212) 231 2555**

EU Cash Sales

Sam Bygott-Webb (London)	**(44 20) 3037 4767**
Richard Alderman (London)	(44 20) 3037 4875
Luke Ahern (London)	(44 20) 3037 4960
Charles Lesser (London)	(44 20) 3037 4771
John Gilbert (London)	(44 20) 3037 4933
James Lumby (London)	(44 20) 3037 4846
Awais Khan (London)	(44 20) 3037 4967
Roger Lee (London)	(44 20) 3037 4954
Joanna Thompson (London)	(44 20) 3037 4925
Claire Elsdon (London)	(44 20) 3037 4787
Markus Geisbuesch (Frankfurt)	(49 69) 50957 8709
Holger Hoepfner (Geneva)	(41 22) 818 7777
Jan Halaska (Boston)	(1 617) 598 2503
Chris Carr (New York)	(1 212) 231-6398
Doug Stone (New York)	(1 212) 231 2606
Andrea Duguet (New York)	(1 212) 231 0977

South Africa Equity Syndication

Franco Lorenzani (Johannesburg)	**(2711) 583 2014**

South Africa Sales

Roland Wood (Cape Town)	**(2721) 813 2611**
Sven Forssman (Johannesburg)	(2711) 583 2369
Ed Southey (Johannesburg)	(2711) 583 2026
Khensani Mokoena (Johannesburg)	(2711) 583 2016
Jesse Ushewokunze (New York)	(1 212) 231 2504

South Africa Sales Trading

Harry Ioannou (Johannesburg)	**(2711) 583 2015**
Welcome Plessie (Johannesburg)	(2711) 583 2058
Martin Hughes (Johannesburg)	(2711) 583 2019
Marcello Damilano (Johannesburg)	(2711) 583 2018

Commodity Hedge Fund Sales

Chris Looney (New York)	**(1 212) 231 0836**
Iain Lindsay (London)	(44 20) 3037 4825
Guy Keller (Singapore)	(65) 6601 0303

Commodity Corporate Sales

Nael Noueiri (London)	(44 20) 3037 4913
Rohan Khurana (Singapore)	(65) 6601 0308

Commodity Investor Products

Arun Assumall (London)	**(44 20) 3037 4953**
Catherine Littlefield (New York)	(1 212) 231 6348