January 23, 2018

TONE AT THE TOP? QUANTIFYING MANAGEMENT PRESENTATION

QES Handbook of Text Mining, Part II

- **Text Mining Management Presentations and Conference Calls.** We continue our quest for unique and orthogonal sources of information. In the New Year of 2018, we shift our attention to an interesting textual database – S&P Capital IQ's Call Transcripts. We showcase how our suite of proprietary and sophisticated NLP (Natural Language Processing), machine learning (e.g., deep learning via Convolutional Neutral Networks), linguistic, and psychological research are used to extract salient information from conference calls.
- **Extracting Signals via NLP and Machine Learning.** Given that most analytical techniques are designed to analyze numerical data rather than unstructured textual information, the first challenge is how to make such data machine readable via data parsing, cleaning, and transformation. By harnessing the best of NLP algorithms and our cloud based platform, we can leverage a variety of text based features, such as language complexity, tone analysis, VADER, executive personalities, and topic modeling. We observe that the company executives are positively biased in their choice of words and muddle conversations with subjectivity, while analysts are more critical. More importantly, CEO conversations need to be closely monitored, as they have remarkable implications for future company performance and share price.
- **Introducing the SMEC (Systematic Mining of Earnings Calls) Model.** We explore a wide range of machine learning algorithms such as deep learning/CNN and LDA for topic modelling to better quantify information from earnings calls. Finally, we create a composite stock selection model – SMEC, by combining various text mining signals. Our SMEC model covers more than 2,000 US companies and nearly 1,200 stocks outside of the US (mostly in Europe, Canada, ANZ, Japan, and EM). It generates a Sharpe ratio of 1.5x, with low or even negative correlation to other fundamental and quantitative strategies.
- **An Effective Tool for Fundamental Managers.** During earnings season, there are dozens or hundreds of company presentations and conference calls on the same day. Investors' attention is limited, but more importantly, we are naturally not very good at deciphering the hidden mood and messages from management conversations. The SMEC and the associated signals can help fundamental portfolio managers to make timely and accurate investment decisions.

Source: Wolfe Research Luo's QES

Gaurav Rohal, CFA
GRohal@wolferesearch.com

Yin Luo, CFA, CPA
YLuo@wolferesearch.com

Miguel Alvarez
MAlvarez@wolferesearch.com

Javed Jussa
JJussa@wolferesearch.com

Sheng Wang
SWang@wolferesearch.com

QES Desk Phone: 1.646.582.9230
Luo.QES@wolferesearch.com

Table of Contents

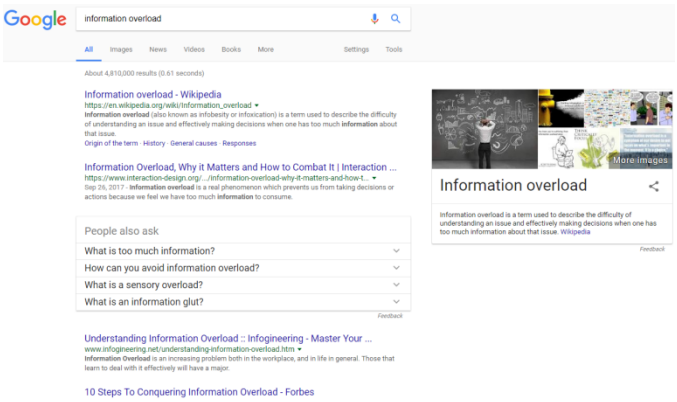
A LETTER TO OUR READERS

As argued in our previous research (see Luo, et al [2017a, 2017b, and 2017c]), the future of active investing rests on how to best incorporate unique data sources with sophisticated modeling techniques. In this research, we continue our quest for orthogonal data sources. In the New Year of 2018, we shift our attention on an interesting textual database – S&P Capital IQ's Call Transcripts.

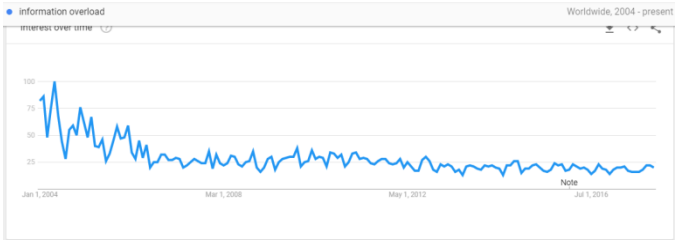
If you google "information overload", you will be immediately overloaded with information. There are almost five million hits within less than a second (see Figure 1 A). The amount of data available certainly exceeds human ability to read and comprehend. Interestingly, people seem to have adapted and found faster and more efficient ways to cope with the problem of too much information. As shown in Figure 1 (B), the relative search volume of "information overload" declined considerably from 2004 to 2008 and has remained relatively stable since then.

Figure 1 Information Overload

A) Google Search



B) Google Trends



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

Most of the information related to investing is in the form of unstructured text (e.g., documents, books, health records), images, audios, and videos. Given that most analytical techniques are designed to analyze numerical data rather than unstructured information, the first challenge is how to make such intelligence machine readable. Natural Language Processing (NLP) is a great example – it quantifies and transforms textual data into tangible information. Analyzing textual information, however, goes far beyond NLP. The process requires a suite of integrated technology and analytical systems, across various facets spanning web scraping, pruning, distributed parallel computing, NLP algorithms, topic modelling and machine learning.

Our analysis begins with basics, by gauging the complexity of language used in the earnings calls via readability tests. These tests generally produce the grade level of formal education needed to comprehend the underlying text. We find more complex language (i.e., poor readability) is linked to inferior future performance. Technical jargon, unnecessarily long words and long sentences used by management in conference calls are perceived as a fog to hide subpar underlying performance.

Then we move on to traditional sentiment or tone analysis, using predefined lexicons/dictionaries (e.g., Harvard IV-4). We have utilized the Loughran and McDonald [2011] dictionary extensively in our

previous research (see Rohal, et al [2017a]). The Loughran and McDonald dictionary is based on US corporate regulatory filings; therefore, leveraging it to call transcripts provides a true out-of-sample test. We find that company executives are more selective in their choice of words in order to paint a positive picture, while analysts are far more critical. Therefore, spontaneous nature of the Q&A section leads to a better capture of the overall sentiment.

We also implement a more sophisticated VADER (Valence Aware Dictionary and Sentiment Reasoner) algorithm of sentiment analysis. Unlike the traditional bag-of-words approach using lexicons, VADER incorporates word-order sensitive relationships between terms such as degree modifiers in both formal and informal communications. VADER incorporates punctuation, capitalization of words, degree modifiers (e.g., "extremely"), and contrastive conjunction (e.g., "but") to increase or decrease the intensity of the sentiment.

Next, we infuse some of the latest research in management psychology in our analysis of earnings calls. To quantify the five personality traits (extraversion, emotional stability, agreeableness, conscientiousness, and opening to experience), we apply a syntactic parser and POS (Part-of-Speech) tagging techniques to extract word tokens to their respective part of speech classes (e.g., nouns, verbs, and adjectives).

Lastly, we shift our focus to another vast field of NLP – topic modeling. The conventional approach to topic modeling is tokenization, by breaking down a document into a bag-of-words and counting for the frequency of each unique word (unigram) or phrase (bigrams, trigrams, and n-grams). To reduce dimensionality, we apply a cutting-edge algorithm – Word2Vec or word embedding. The vectorization of words and phrases facilitates the application of machine learning techniques, such as deep learning via CNN (Convolutional Neutral Networks). In particular, we develop a powerful $LDA\ (Latent\ Dirichlet\ Allocation) + Herfindahl\ Index + Logit$ model that can effectively extract positive and negative topics from textual documents.

In the end, we introduce our combined SMEC (Systematic Mining of Earnings Calls) signal. The SMEC model covers over 2,000 stocks in the US and nearly 1,200 stocks outside of the US (Europe, Canada, ANZ, Japan, and EM). The SMEC model delivers a Sharpe ratio of 1.5x, with low correlation to common quantitative and fundamental strategies.

The SMEC model and its associate tools and signals are particularly useful for fundamental analysts and portfolio managers. During earnings season, there could be hundreds of companies reporting and presenting operational performance via investors' presentations and conference calls on the same day. Investors' attention is limited, but more importantly, we are naturally not very adept at deciphering the hidden mood and message from management. Training the computers to shift through mountains of call transcripts and make investment recommendations sounds very enticing. However, it poses considerable challenges to our data storage configuration and computational framework. More importantly, making insightful decisions from textual data requires a thorough understanding of linguistics, psychology, NLP (Natural Language Processing), machine learning, and cloud computing. In this regard, our NLP based models come to the rescue and make a highly complex process accessible to non-specialists.

We offer regular data feeds and a web portal based on the SMEC model and its underlying components. Interested readers should contact your sales representative or Luo.QES@wolferesearch.com for more details.

Regards,

Yin, Gaurav, and Luo's QES Team

INTRODUCTION

As we discussed in Luo, et al [2017a, 2017b, and 2017c], the underwhelming performance of conventional factors and technological advancements have unleashed an arms race to acquire and process alternative data sources. In this research, we demonstrate how NLP and machine learning techniques can be harnessed to quantify information from corporate presentations and Q&A's.

In this paper, we showcase how qualitative textual data, natural language processing (NLP), and machine learning techniques can be all brought together to forecast future company fundamentals and stock returns. We utilize the call transcripts provided by S&P Capital IQ as our data source. The focus of the research rests particularly on an introduction to various breakthrough NLP techniques and their application to financial information processing using the call transcript textual dataset.

A BRIEF HISTORY OF NLP

Traditionally, investors focused on numeric data, using either fundamental analysis or quantitative models. However, the vast majority of the available information is in textual format. For example, there are almost 5,000 documents filed by public companies in the US every day. On average, there are over 150,000 words in a typical 10-K – the standard annual filing of a company's performance required by the SEC (Securities and Exchange Commission). Companies communicate their performance and strategies to the public via investors' presentations and conference calls. Investors do not have enough time to shift through this mountain of written material. Furthermore, human interpretation is subjective and certainly exposed to behavioral biases.

Das and Chen [2001] and Antweiler and Frank [2002] are among the first researchers to study textual information through NLP. The authors used linguistic methods to examine the effect of messages posted on Yahoo! Finance and Raging Bull, for the companies in the Dow Jones Index. Other well-known early works include Li [2006] and Tetlock [2007]. Li [2006] analyzes sentiment from corporate filings and finds that certain words in firms' annual reports predict low annual earnings and stock returns. Tetlock [2007] quantifies sentiment based on news stories rather than annual reports. He suggests that the fraction of negative words in firm-specific news stories leads to low subsequent earnings. For the tone analysis part, Tetlock [2007] study is based on a general-purpose Harvard dictionary, which is then further refined by Loughran and McDonald [2011]. They suggest that the Harvard psychological dictionary may not be suitable for finance and accounting applications, because the meaning of positive and negative words may have very different connotations in a financial context. Loughran and McDonald [2011] develop a finance-specific sentiment dictionary. They are also among the first to extensively apply NLP on the SEC Form 10-K filings and link the signals to stock returns, trading volume, return volatility, fraud, material weakness, and unexpected earnings. We have taken their work further and have developed a comprehensive suite of algorithms and signals based on SEC's EDGAR filing system (see Rohal, et al [2017] and Wang, et al [2017b]).

On company call transcripts, Ahmad and Zinzalian [2010] is one of the original studies. Ahmad and Zinzalian find that, with large enough training data, historical volatility together with n-gram features and POS tagging, can improve stock volatility forecasting over pure statistical models. Price, et al [2011] propose that conference call linguistic tone is a significant predictor of abnormal return and trading volume. The Q&A section of the call has incremental explanatory power for the Post-

Earnings-Announcement-Drift (PEAD)[1]. Huang, et al [2014] fit a topic model to analyst conference call transcripts and reports; to examine sell-side equity analysts' information interpretation and discovery roles. Zhao [2017] finds interesting patterns resulting from language complexity and sell-side research participation in earning call. Call, et al [2017] show that buy-side appearances on earnings conference calls are associated with subsequent decreases in sell-side coverage, lower stock returns, and increases of bid-ask spreads, implied volatility, and short interest.

In this paper, we analyze the earnings call transcripts data provided by S&P Capital IQ. We introduce a suite of our cutting-edge NLP techniques and their application in deciphering call transcripts.

[1] PEAD is one of the most thoroughly studied market anomalies in empirical finance. However, we find the magnitude of PEAD has declined substantially in recent years (see Wang, et al [2017a]). PEAD also disappears in two or three days after earnings announcement.

S&P CAPITAL IQ CALL TRANSCRIPT DATA

Regular readers of our research should know that we spend significant amount of our time searching for new and unique datasets (see Jussa, et al [2017b]). Many of our recent research is sourced from boutique data vendors, e.g., Estimize (see Wang, et al [2017a]), Ravenpack (see Luo, et al [2017a, 2017b]), Wind (see Wang, et al [2017c]), RS Metrics (see Jussa, et al [2017a]). However, many of the large well-known vendors may also have interesting hidden jewels. We have been partnering with S&P Capital IQ for many years and studied a few of unique datasets, e.g., Key Developments and Future Events (KDFE), Global Premium Financials. In this research, we focus on an exciting database from S&P Capital IQ – the Transcript Database, which provides current and historical call transcript data covering approximately 7,000 public companies globally.

The Transcript Database provides textual translation of earnings calls, guidance/update, shareholder/analyst calls, analyst/investor day, M&A calls, and operating results calls, and fixed income calls, etc. In this research, we focus primarily on earnings calls, as other types of calls are limited in numbers.

Each earnings call is typically split in two parts: "Presentation Section" and "Q&A Section". The presentation section usually includes a speech by company executives, e.g., CEO, CFO. The Q&A section contains conversations between company management and buy/sell side research analysts, investors, or potentially, the media. Each sentence of the call is tagged to an executive or analyst. It also provides detailed meta data such as speaker name, speaker type and associated company for speaker. Combined with two other closely related databases offered by S&P Capital IQ – the Professionals (background information on company executives, board members and investment professionals) and KDFE, we can generate even greater insights.

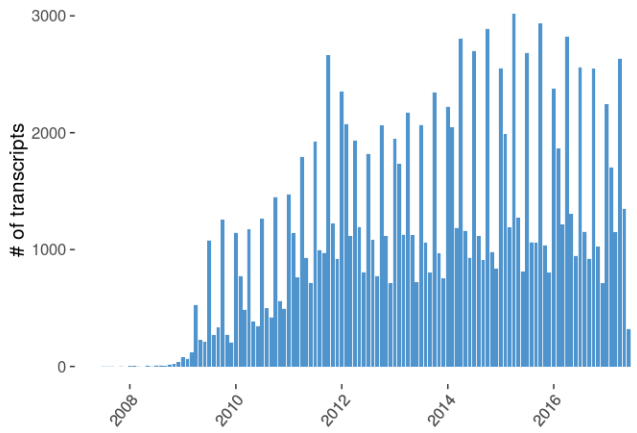
Presentations in general are well rehearsed and convey management's message to the investment community. While the Q&A section tends to be more spontaneous and at times unexpected, company management still maintains the control of information flow. We expect both sections to provide substantial and complementary information for the future business outlook of a company (financial and operational performance, stock return/volatility).

COVERAGE

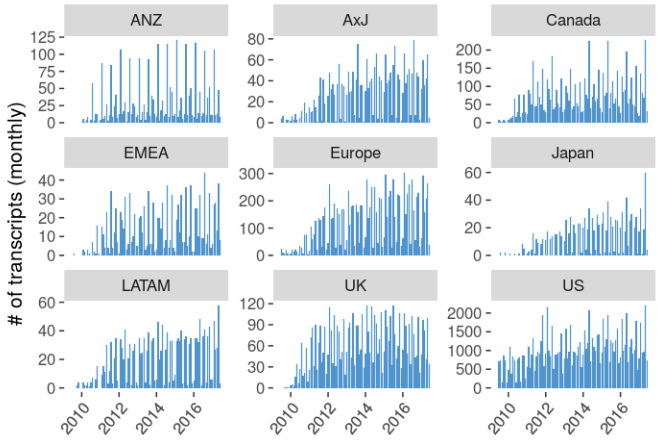
S&P Capital IQ provides multiple copies of transcripts with different vintage of edits at different points of time, including "proofed copy", "edited copy", "spell-checked copy" and "audited copy". We create our sample – a union of proofed and edited copies to balance between timeliness and coverage. The Transcript database begins in 2009. Currently, it covers around 2,500 stocks in the Russell 3000 universe (see Figure 2 A). Outside the US, coverage is somewhat limited (around 1,200 companies), but has improved over time (see Figure 2 B).

Figure 2 Monthly coverage of call transcripts data across the world

A) Monthly number of earnings calls



B) Number of earnings calls, by region



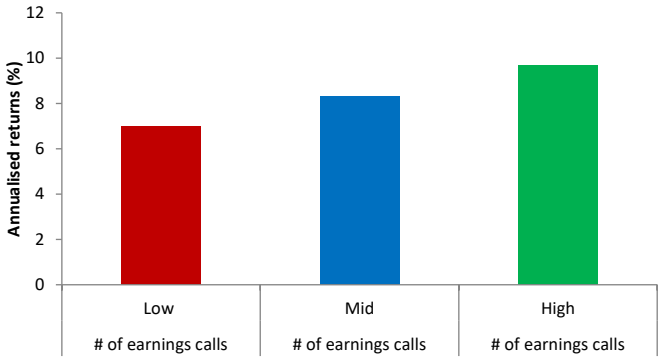
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

The frequency of earnings call transcripts obviously coincides with company financial reporting (quarterly, semi-annually, or annually) announcements (see Figure 4 A). During earnings season, there could be up to 400 conference calls on the same day. Attention is a limited, scarce, and valuable resource. When individuals perform multiple tasks or process multiple sources of information, performance tends to suffer. Limited attention, therefore, is likely to affect how investors process information and how markets react to news (see Hirshleifer, et al [2009]). If investors' attention is indeed distracted on the earnings announcement date, the immediate market reaction is likely to be muted; therefore, the alpha contained in the transcripts is likely to last longer.

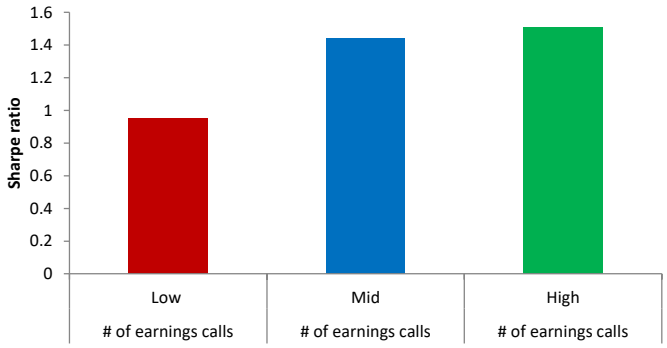
Based on the "limited attention" hypothesis, investors should have a more difficult time to interpret news on busy days. In this case, busy days are the days with multiple earnings calls. Performance of our Call transcript based our composite SMEC model is much stronger for earnings call released on busy/high news days relative to low news days (see Figure 3). Details on the SMEC model will be explained in the following sections.

Figure 3 SMEC composite L/S performance on high/low news days, Russell 3000 universe

A) Annualized returns



B) Sharpe ratio

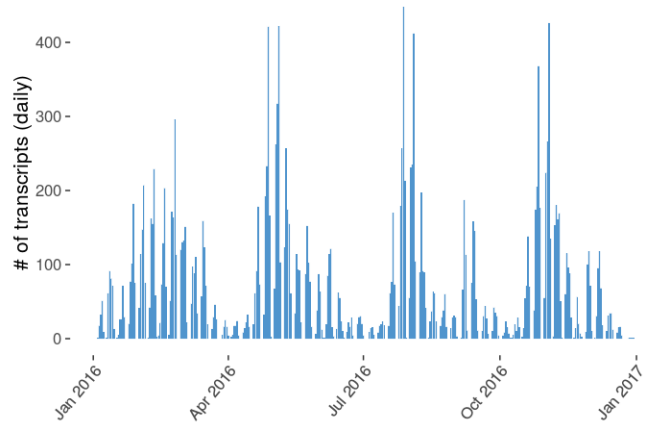


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

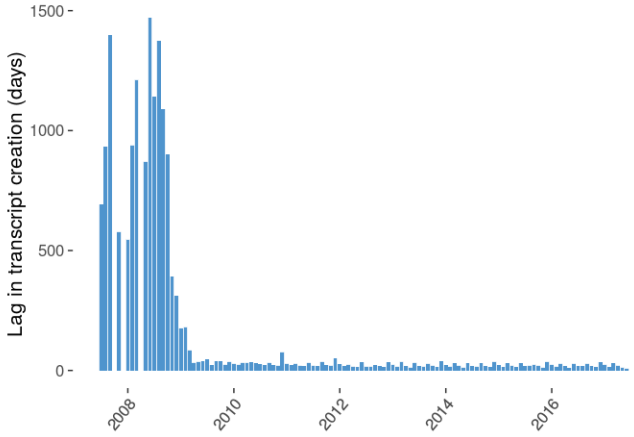
S&P Capital transforms audio/video calls into written transcripts. Part of the process can be done by computer algorithms, but human intervention such as checking, editing, and verification are almost always needed. As shown in Figure 4 (B), live production of call transcripts started in mid-2009. S&P Capital IQ backfilled a few extra years of historical data. Historically, there are a few days of delay from the time when the call was conducted to the date when the transcript data became available in the database. Currently, S&P states that it aims to provide an edited copy of most transcripts within three hours after the end of a call[2].

Figure 4 Seasonality and processing delay in call transcript data

A) Number of earnings call in the year of 2016



B) Delay in data processing



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

[2] To be conservative and avoid potential look-ahead bias, we add a three-day lag to the call transcript data for this research.

DATA PARSING, CLEANING, AND TRANSFORMATION

As discussed in Luo, et al [2017a, 2017b], data science (such as data cleaning, outlier and seasonality adjustment, and missing value handling) is a critical part of investment research enterprise. For text processing, and in particular call transcripts, we parse all documents by section type and then speaker type. Presentation text is separated into "presentation by CEO" and "presentation by other executives". Q&A section is separated into "questions", "answers by CEO" and "answers by other executives". For the Q&A section, we assume that any communication from executives is an answer; and analyst communication is always a question. We remove any section with less than 50 characters (approximately 10 words). Text is cleaned for extra white spaces and converted to lower case. We do not tokenize or remove special characters at this stage as these treatments vary with each of the NLP algorithms we employ in the research. For example, sentiment analysis does not need numbers, but we would use a special token to identify all the numbers for topic modelling. Although most NLP techniques are based on tokenized words, some rely on sentence structure, which requires text to have special characters for the end of sentence/line.

Stemming and Lemmatization

Before we can perform sentiment analysis, we need to convert text into individual words called tokens. Furthermore, depending upon the lexicon/dictionary used, we may also need to stem or lemmatize words or tokens. The aim of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

For example, words such as "produce", "produces", "produced", and "producing" can be reduced to "produce". In addition, words with similar meanings, such as "profit", "profitable", and "profitability" can be summarized in one word "profit".

Stemming is a cruder approach, which simply chops off the end of words to their roots. The best-known and most popular stemming approach for English is the Porter stemming algorithm – a collection of rules designed to reflect how English handles inflections. For example, both "apple" and "apples" are chopped down to "appl", while both "meanness" and "meaning" are reduced to "mean", which creates a false equivalence.

Lemmatization usually needs a vocabulary and morphological analysis of words, removing inflectional endings only and returning the base form of a word. In simple terms, stemming returns the non-changing forms of words, while lemmatization returns the dictionary form. In the case of "meanness" and "meaning", a good lemmatization algorithm should map the two words to themselves, while still reducing both "apple" and "apples" to "apple". Lemmatization is generally more accurate than stemming, but also has its own downside. Lemmatization often results in more base words and it can't handle unknown words (e.g., iWatch).

After stemming and/or lemmatization, we can perform more useful tasks such as word frequency count and sentiment analysis.

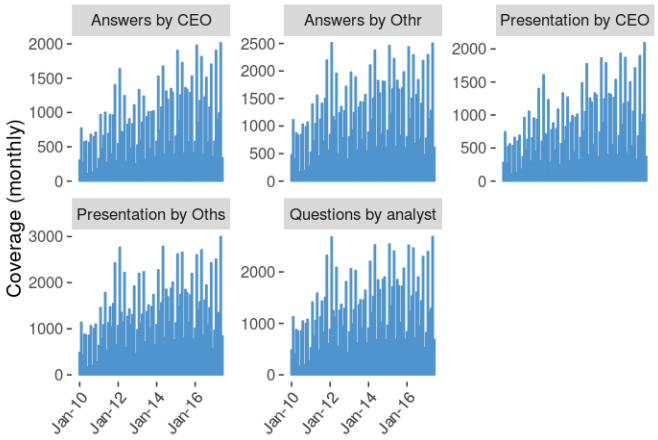
CEOs Taking the Center Stage

Figure 5 (A) and (B) show the monthly average number of calls and the median number of words during each call, by each section and speaker types. Obviously, CEO's participate in the vast majority of analyst calls and answer questions. Most conference calls allow analysts to ask questions and analysts do ask questions whenever possible.

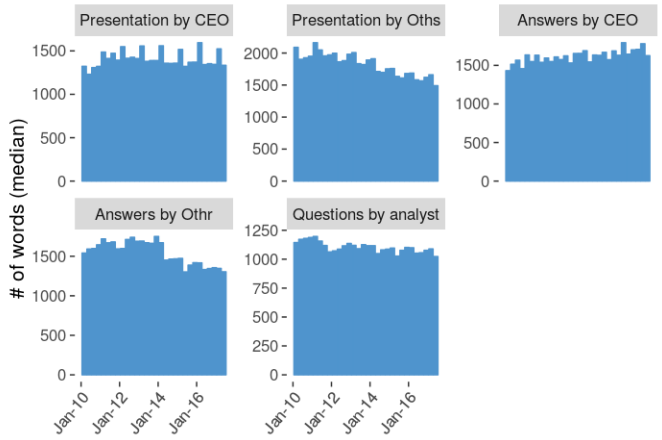
During the main presentation, CEO's speak as many words as all the other executives combined. Actually, the number of words by CEO's has been steady over time, while other members of the management team have become quieter over time (see Figure 5 B). Interestingly, in the past eight years, CEO's have become more and more likely to answer questions, at the expense of other executives.

Figure 5 Coverage and number of words by section and participants

A) Number of earnings calls by participants



B) Median number of words by participants



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

WORDCLOUD ANALYSIS

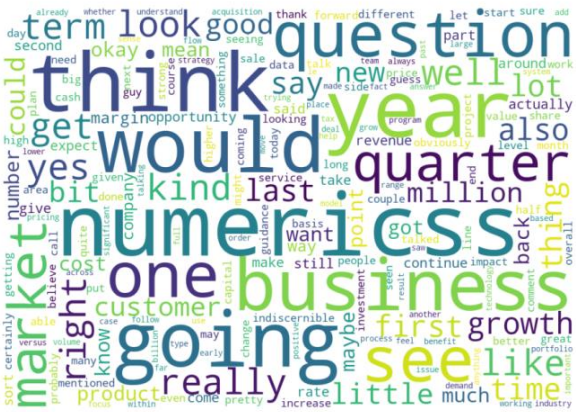
Figure 6 shows the WordCloud based on all call transcripts in 2017. WordCloud analysis is simply a frequency count of each word appearing in the text. More frequently appeared words are more prominently presented with larger fonts. We remove stop words[3] and lemmatize words to their roots before running the word frequency count. More details on linguistic analysis will be described in the next few sections. As expected, in the presentation section, the most frequent words are mostly trivial such as *business*, *quarter*, *year*, *one*, *market*. Similarly, in the Q&A section, management tends to make not-so-insightful comments like "that's a good question" and use words such as *would*, *going*, *think* and *growth*, etc[4]. In its current form, WordCloud does not appear to be overly informative.

[3] In computer science, a stop word is a commonly used word (such as "a", "the") that most NLP algorithms can safely ignore. Stop words are so frequently used that they are generally not very useful for our purpose of analysis. Most search engines (e.g., Google) also automatically remove stop words when indexing entries for searching and when retrieving search results.
[4] Here "numericss" is a tag we use to represent numbers.

Figure 6 WordCloud based on recent call transcripts

A) Presentation section

B) Q&A section



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

We can improve our WordCloud analysis by further removing common words that don't alter actual meanings. We also discard words which are very sector specific such as "energy", "oil" and "retail". As shown in Figure 7, the results appear to be slightly more insightful about the representative topics being discussed. At a first glance, the Q&A section looks more informative with words such as *acquisition*, *deal*, and *lower*, etc. Still, most of these words carry little information on a standalone basis. Repeating the same procedure only exposes us to the next layer of trivial words without taking us much further. Obviously, more sophisticated algorithms are required.

Figure 7 WordCloud based on recent call transcripts (post filtering common words)

A) Presentation section

B) Question and Answer section

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

QUANTIFYING TRANSCRIPT DATA, FEATURE ENGINEERING

Most statistical tools and machine learning algorithms are designed to process numerical data. Call transcripts are, however, textual in nature. Therefore, optimally transforming words into numbers is the next challenge we must tackle. In machine learning language, feature engineering is the process of using domain knowledge of the data to create features (or factors, in investment jargon) that computer algorithms can analyze. Feature engineering is fundamental to the success of any model, and is both difficult and expensive.

In the case of call transcript data, we apply our NLP techniques to perform feature extraction and generate structured, timely, and informative signals from this unstructured data.

READABILITY INDEX AND LANGUAGE COMPLEXITY

During earnings season, hundreds of companies around the world can all report on the same day. Most public companies conduct analyst conference calls in conjunction with their press releases. Fundamental analyst read press releases, supplementary documents, and the subsequent regulatory filings (with detailed financial statements, disclosures, and management discussion and analysis). Sell-side analysts, buy-side analysts/portfolio managers, and the media also participate in management calls\ in the attempt to gauge additional insights from the company presentation and Q&A. As we have showed in previous research (see Luo, et al [2014]) and also covered in academia (e.g., Hirshleifer, et al [2009]), investors' attention is limited. The human ability to process volumes of data and information is limited. We are also often biased by our prior views of a company and overconfidence can further shadow our judgment. Therefore, investors' ability to listen to hundreds of earnings calls, to read multiple call transcripts, and to derive their investment conclusions, all over a short period of time, will be highly imprecise.

One of the most basic NLP techniques is the readability test. We can gauge the language complexity using readability indices. These indices generally output a number, which approximates the grade level of education needed to comprehend the underlying text. In other words, the higher the score, the higher the complexity.

Most of these indicators are based on two factors. One factor relates to the sentence structure, e.g., the average number of words per sentence. The other factor relates to word structure or complexity and is usually based on either the proportion of easy words (defined by a lexicon/dictionary) or the average number of syllables per word. Some of the popular readability Indices include:

- **Gunning Fog Index [1952].** The index is attributed to a textbook writer, Robert Gunning. He claimed that newspapers and business documents were full of "fog" and unnecessary complexity. The Gunning fog score estimates the years of formal education a person needs to understand the text on the first reading.

$$Grade\ Level = 0.4\ (ASL + PHW)$$

Where,

ASL = Average Sentence Length (i.e., number of words divided by the number of sentences)

PHW = Percentage of Hard Words

The ideal range of the Gunning fog index should be around seven or eight. The fog Index defines "Hard Word" based on the number of syllables (i.e., three or more). Proper nouns, familiar jargon, or compound words, as well as common suffixes (e.g., -es, -ed, -ing) are not counted toward the number of syllables. The fog index is simple to compute, but not necessarily accurate. Not all complex words are difficult to understand. Similarly, short but uncommon words can be difficult for most people.

- **Automated Readability Index [1967].** Like other popular readability formulas, the ARI formula outputs a number, which approximates the grade level needed to comprehend the text. Unlike the other indices, the ARI relies on the number of characters per word, instead of the usual syllables per word. The number of characters is more readily and accurately counted than syllables.

$$Grade\ Level = 4.71\left(\frac{characters}{words}\right) + 0.5\left(\frac{words}{sentences}\right) - 21.43$$

Where,

$characters = number\ of\ letters\ and\ numbers,$

$words = number\ of\ words$

$sentences = number\ of\ sentences$

- **Dale-Chall Readability Score [1948, revised 1995].** The Dale-Chall Formula calculates the US grade level of a text sample based on sentence length and the number of "hard" words. "Hard' words are anything outside of a list of 3000 words that are not familiar to most 4th-grade students. Therefore, the Dale-Chall index is based on a pre-defined Word Familiarity list, rather than the number of syllables or characters.

$$Raw\ Score = 0.1579 \times PDW + 0.0496 \times ASL$$

Where,

$PDW = Percentage\ of\ Difficult\ Words$

$ASL = Average\ Sentence\ Length\ in\ Words$

If (PDW) is greater than 5%, then

$$Adjusted\ Score = Raw\ Score + 3.6365$$

Otherwise

$$Adjusted\ Score = Raw\ Score$$

There are many other readability indices. Most of them are simple variations of the above three examples. In our analysis, we use the Automated Readability and Dale-Chall Readability indices to represent the two broad categories.

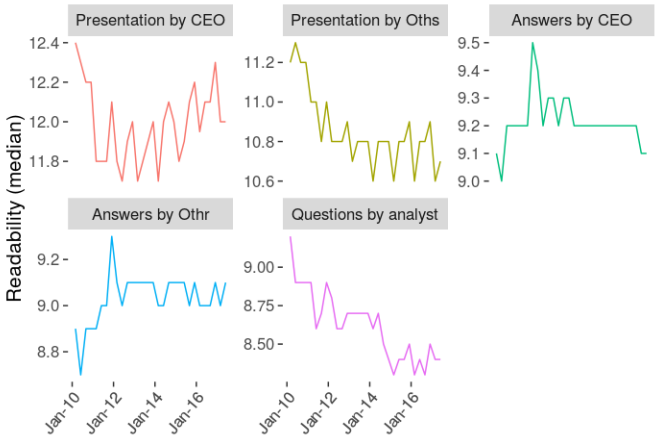
Presentations are Getting Fogged with Complex Language

Figure 8 shows the median readability scores by each section/participant type. Larger scores correspond to higher education grades required to comprehend and hence lower readability. CEO presentation has become more complex over time, while the readability of analyst questions gets better in recent years. The average readability is better for the Q&A than the main presentation – spontaneously spoken language tends to be simpler than heavily scripted presentations. We also

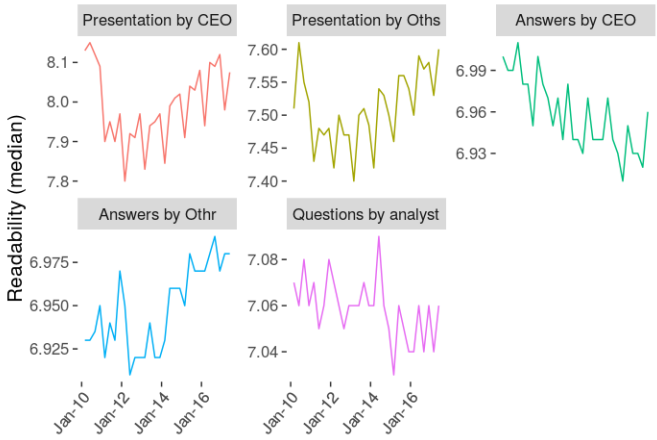
observe a strong seasonal pattern in the readability index, which coincides with annual reporting season. Company management generally spends more time to deliver their annual results than quarterly/interim updates.

Figure 8 Median readability score, by each section and speaker type

A) Automated Readability Score



B) Dale-Chall Readability Score



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

How does the Market React to Complex Language?

Now, we turn our attention to investigate whether readability indices predict future stock returns, using the Spearman Rank IC[5]. As documented in previous research (see Rohal, et al [2017a]), most factors derived from text mining tend to be weak but persistent. The forecasting power of readability scores is mostly negative, but weak (see Figure 9). The negative relationship between readability and future stock returns is in line with expectations. Complex language is poorly understood by investors. Furthermore, many investors associate unnecessarily complicated words with management obscurity and uncertainty.

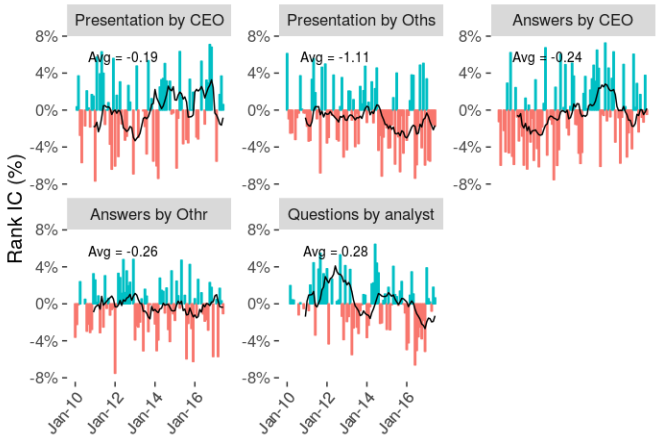
Interestingly, more complex questions by sell-side analysts lead to slightly positive future stock returns. Elaborated questions might be indicative of a more constructive engagement between analysts and company executives.

[5] As a reminder, we use Rank IC to measure the predictive power of a factor in selecting/ranking stocks. It is computed as the rank correlation between current month's signal and the following month's return, among all stocks in our investment universe.

Figure 9 The predictive power of readability score on future stock returns

A) Automated Readability score

B) Dale-Chall Readability Score

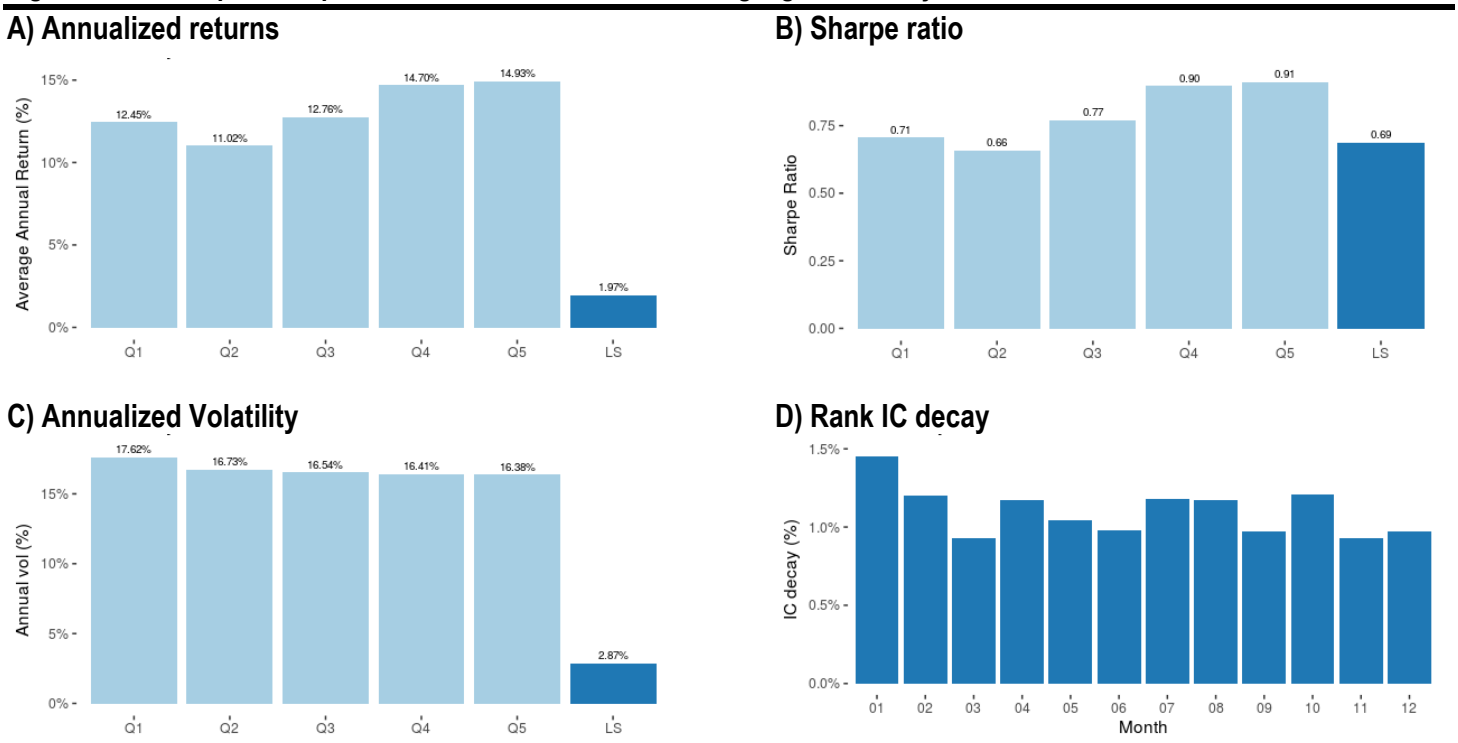


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

We further construct a simple composite readability factor, by combining the CEO and other executive presentation sections. As shown in Figure 10, a simple long/short quintile portfolio[6] delivers a modest return of 2%, but a reasonable Sharpe ratio of 0.7x. Performance comes mostly from the short side. Companies with poor readability in earnings calls tend to produce lower returns with higher volatility than firms using simple language. Mild returns with low volatility and slow signal decay are common traits of NLP based signals. Textual information is sparsely distributed, but signals are not crowded; therefore, returns are persistent.

[6] In this example, we use Russell 3000 as our investment universe. Stocks in each quintile are equally weighted. Portfolios are rebalanced monthly. Results for other countries are reported at the end of this paper.

Figure 10 Quintile portfolio performance based on executive language readability, in the Russell 3000 universe



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

SENTIMENT OR TONE ANALYSIS BASED ON LEXICONS

Next, we move to the most popular approach for NLP – sentiment or tone analysis. Sentiment analysis attempts to objectively characterize the message conveyed by the underlying textual document. A lexicon or dictionary is commonly used to calculate the tone of each individual word. The tone is then aggregated for the complete text to arrive at an overall sentiment of the message. We have done significant work on sentiment analysis in previous publications (see Rohal, et al [2017a]).

Why can sentiment around earnings calls predict future stock returns? We have two hypotheses. First, it is related to the "limited attention" and under-reaction argument; limited ability of human analysts to listen, read, and analyze management conference calls, investors may not allow them to immediately react to the news delivered by management, which causes post-announcement drift (mostly likely in the same direction, i.e., positive tone in the call leads to higher subsequent returns and vice versa). Furthermore, we expect the tone of management to contain more useful information about the underlying fundamentals of the company (e.g., profitability, solvency, business prospects) than what is presented in the written materials (e.g., press releases, supplemental packages, regulatory filings, financial statements). Our empirical results show evidences on both fronts.

The most common approach for sentiment analysis is to count the number of positive and negative words, using a pre-defined dictionary. The relative proportion of positive/negative words is used as the positive/negative tone measure for a document. The polarity score is defined as:

$$Polarity = \frac{(Numer\ of\ positive\ words - Number\ of\ negative\ words)}{Total\ numer\ of\ words}$$

In addition to the positive/negative sentiment, we can classify a given text into objective versus subjective. The subjectivity score can be computed as:

$$Subjectivity = \frac{(Numer\ of\ subjective\ words + Number\ of\ objective\ words)}{Total\ numer\ of\ words}$$

In this research, we use a generic (Harvard IV-4) and a finance-oriented (Loughran and McDonald) lexicon for our sentiment analysis.

Harvard IV-4 Dictionary

The Harvard IV-4 is one of the most extensively used dictionaries employing well-known semantic text analysis program called the General Inquirer. Figure 11 shows the WordCloud (frequency count) the words appeared in the call transcript database, using the Harvard IV-4 dictionary. Harvard dictionary categorizes each word into active (passive), strong (weak), overstated (understated), and positive (negative). We tag each individual word using two large valence categories labeled positive (negative) from Harvard IV-4. This is then aggregated for the complete text to compute the overall polarity score.

Figure 11 Harvard IV-4 dictionary words appeared in the call transcripts

A) Presentation section

B) Q&A section

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

CEOs are Overwhelmingly Bullish in their Tone

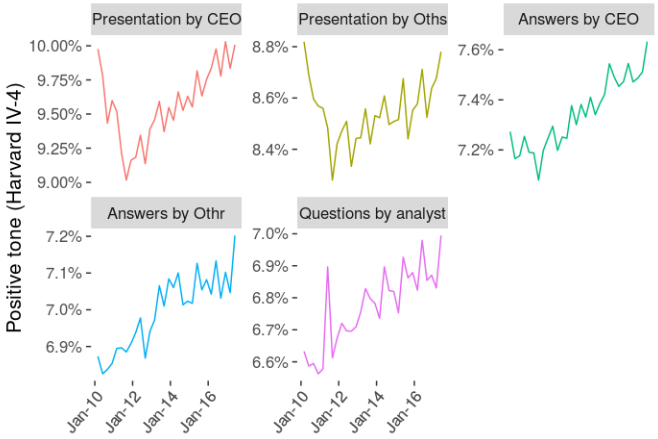
Figure 12 shows the percentage of positive/negative words in our call transcript data, based on the Harvard IV-4 dictionary. Overall, the use of positive words is much more prevalent than negative ones in management communication. Furthermore, the percentage of positive words has been rising, while the ratio of negative words has been declining in the past eight years.

We find that company executives (especially CEO's) tend to be selective in their choice of words, while analysts are more likely to be more critical. Analysts' questions are more cautious than

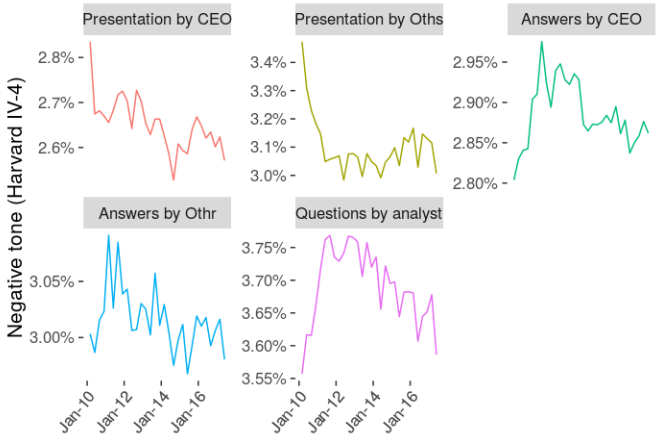
management's answers. The management presentation section is more bullish than the Q&A section. Moreover, CEO's are more positive than other executives of the same company.

Figure 12 Positive/Negative tone of call transcripts based on the Harvard IV-4 dictionary

A) % Positive Words



B) % Negative Words

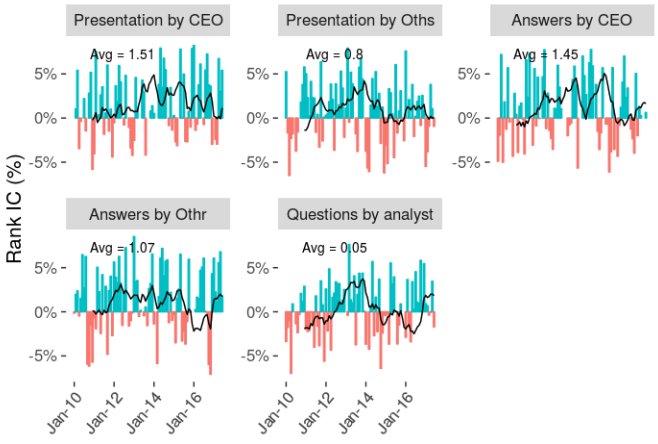


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

The sentiment factors based on the Harvard dictionary does show noteworthy predictive ability of future stock returns, particularly for positive tone (see Figure 13). More interestingly, CEO speech, both from the presentation and Q&A sections, has much stronger performance than other executives/analysts.

Figure 13 The performance of factors based on the based on the Harvard IV-4 dictionary, Rank IC

A) % Positive Words



B) % Negative Words

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

Loughran and McDonald [2001] Dictionary

The Harvard IV dictionary is designed for general purpose; it is not specialized for financial applications. The meaning of positive or negative in a financial context may have different connotations to the context in general language. Loughran and McDonald [2011] show that the standard English dictionaries are poorly specified for accounting and finance topics. In a large sample of 10 Ks, almost 75% of the words identified as negative by the widely used Harvard dictionary are words typically not considered negative in a financial context.

For example, words such as *mine*, *liability*, *cancer* and *vice* are classified as negative words by the Harvard dictionary. They appear frequently in finance, but have very different implications. These words are not predictive of the tone of documents related to accounting and finance and add noise to the sentiment measure, which will negatively affect a model's predictive ability.

In this section, we re-compute all factors with Loughran and McDonald [2011] dictionary. We have used this dictionary extensively to conduct sentiment analysis on 10-K/10-Q filings in a recent research report (see Rohal, et al [2017a]). In addition to positive and negative tone, we also compute more features such as "uncertainty", "litigious", "constraining", "superfluous", and "interesting". These are additional sentiment types based on the Loughran and MCDonald [2011] lexicon. For example, "Constraining" includes words such as *commit*, *compulsion*, and *limit*; while "Uncertainty" includes words such as *ambiguous*, *anticipate* and *approximate*.

Figure 14 shows the WordCloud using the Loughran and McDonald lexicon for the earnings calls.

Figure 14 Loughran and McDonald dictionary positive/negative sentiment words in the call transcripts

A) Presentation section

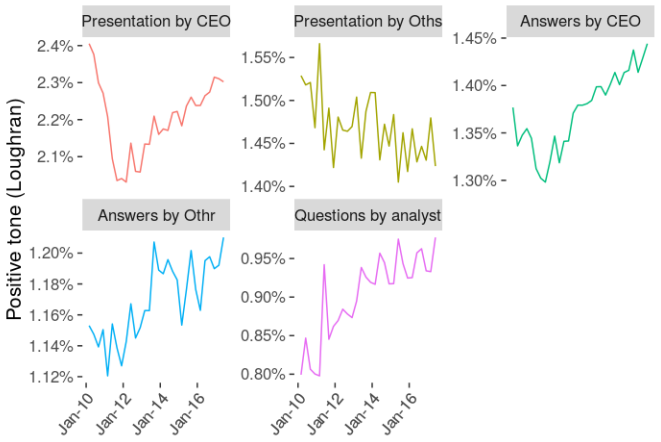
B) Q&A section

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

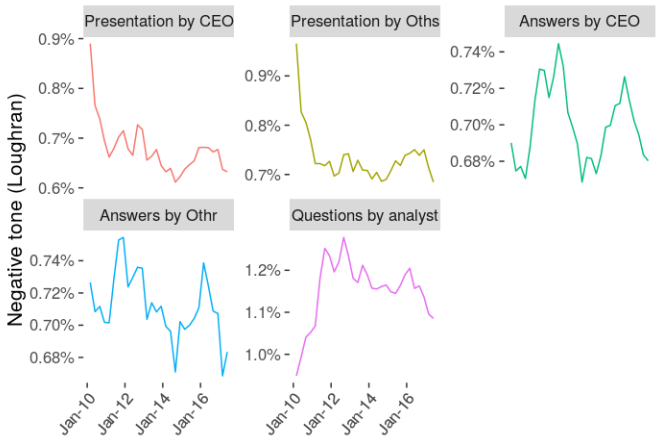
Figure 15 shows the aggregate positive/negative tone (as well "uncertainty" and "constraining") of call transcripts based on the Loughran and McDonald dictionary. The overall pattern is similar to using the Harvard dictionary (see Figure 12). Company executives are consistently more positive than analysts. Within company executives, CEO's are more confident than other executives. This indicates CEO's are particularly assertive in the message they want to convey to the public.

Figure 15 Median tone of call transcripts, based on the Loughran and McDonald dictionary

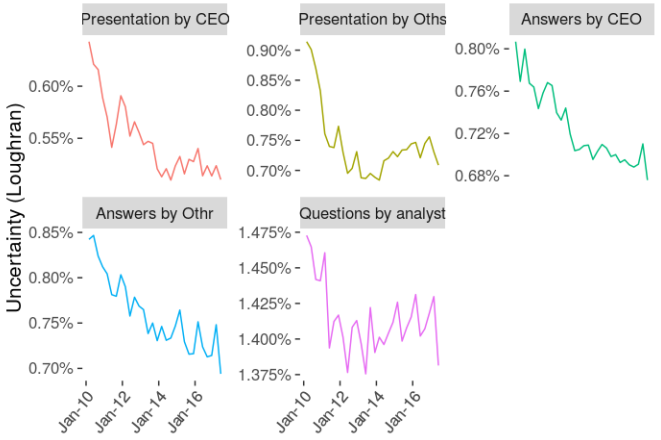
A) % of Positive Words



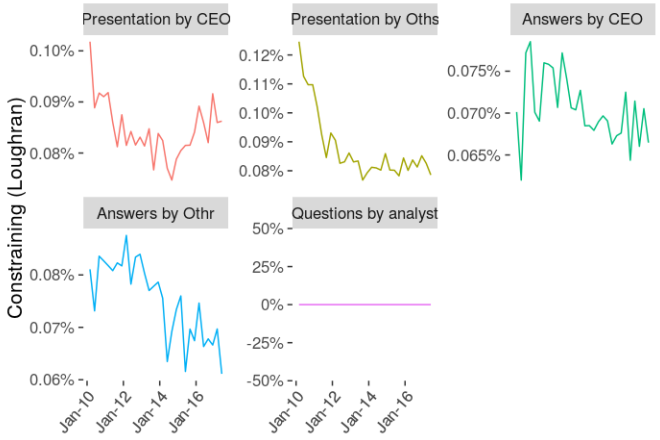
B) % of Negative Words



C) Uncertainty



D) Constraining



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

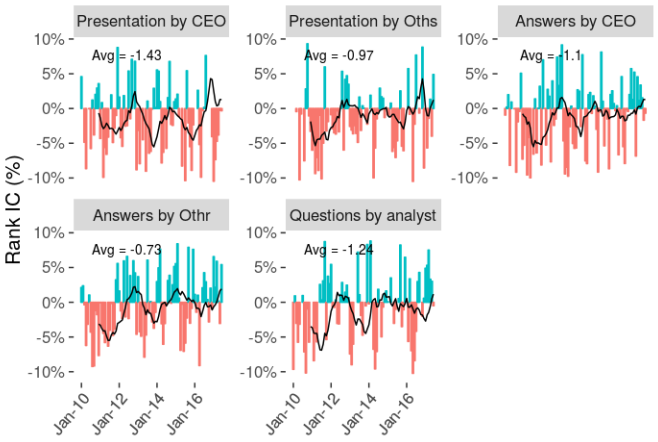
The Loughran and McDonald dictionary is based on the words appearing in EDGAR US corporate filings. Therefore, extending the dictionary to earnings calls is a true out-of-sample test. As expected, the performance of our signals using the Loughran and McDonald dictionary nearly doubles vis-a-vis the Harvard IV-4, as measured by Rank IC (see Figure 16). Thus, the financial context of the lexicon is indeed an important consideration for measuring the sentiment of call transcripts. Price, et al [2011] also observe that when quantifying earnings calls, a context specific dictionary is more powerful than a more widely used generic dictionary (i.e., Harvard IV-4).

In summary, positive tone measures are more predictive of future stock returns than negative tones. CEO speeches, both the main presentation and Q&A, have far more significant influence on future stock performance than the rest of the management team. The other auxiliary measures such as uncertainty or constraining, have weaker predictive ability (see Figure 16 C and Figure 16 D).

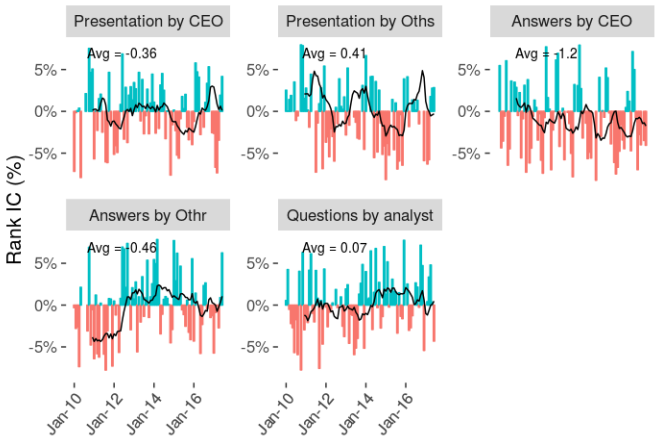
Figure 16 The performance of factors based on the Loughran and McDonald dictionary, Rank IC

A) Positive Tone

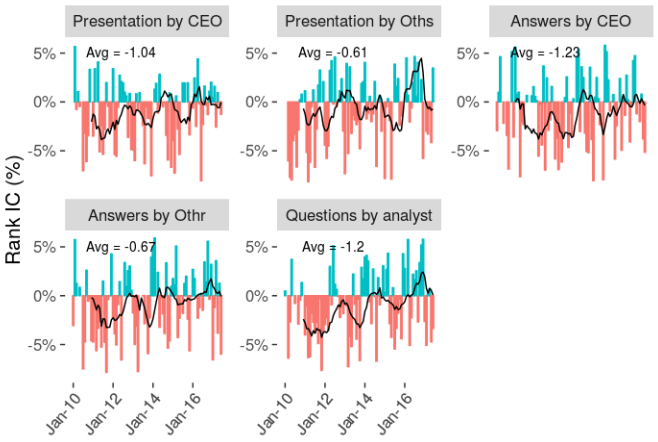
B) Negative Tone



C) Uncertainty



D) Constraining



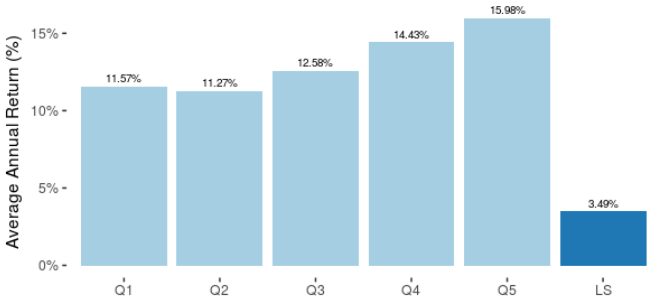
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

Next, we create a net sentiment measure or the "Polarity" score using the ratio of positive and negative words based on the Loughran and McDonald dictionary. We also merge the "uncertainty" and "constraining" words with the negative words for this "Polarity" signal. We further construct a composite net sentiment signal for the Presentation and the Q&A sections by combining signals for each of the speaker types for each section.

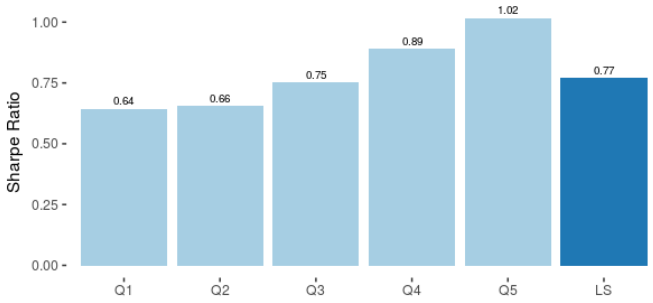
The long/short quintile portfolio (see Figure 17) based on this net sentiment composite performs much better than the readability index (see Figure 10). The performance of the five quintile portfolios forms a desirable monotonic pattern. More importantly, performance comes mostly from the long side, i.e., companies expressing bullish sentiment, especially in the Q&A section (see Figure 17 C and D).

Figure 17 Quintile portfolio performance, Loughran and McDonald Net Sentiment (Polarity)

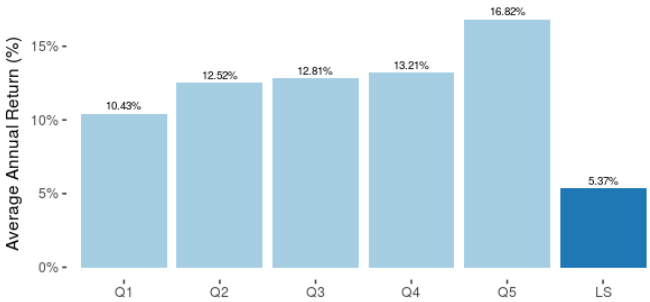
A) Presentation section, annualized returns



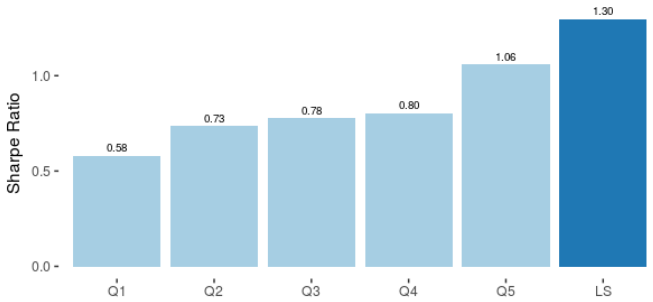
B) Presentation section, Sharpe ratio



C) Q&A section, annualized returns



D) Q&A section, Sharpe ratio

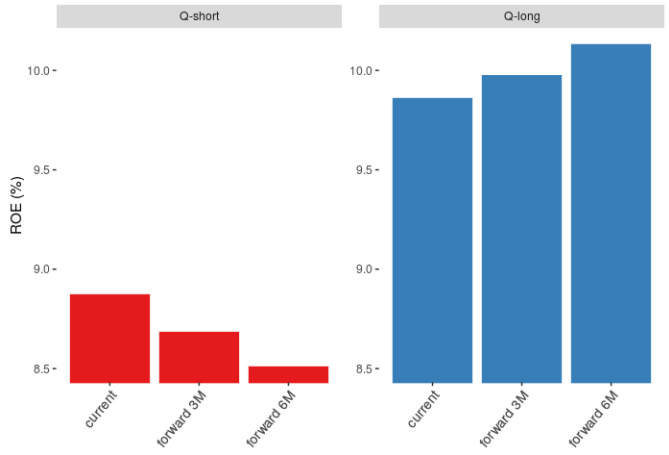


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

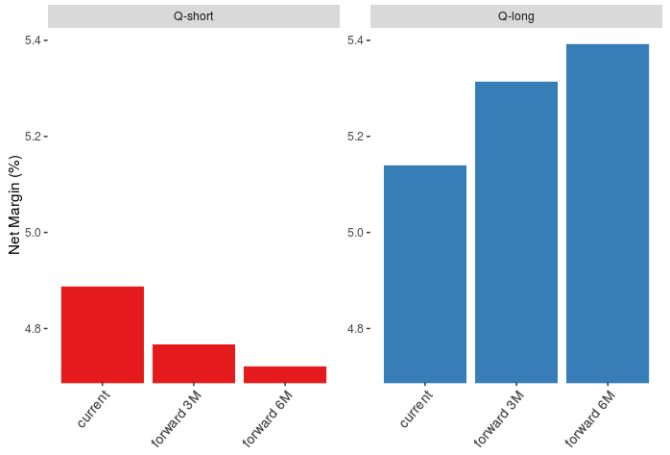
We also observe an interesting pattern on the progression of company fundamentals. As shown in Figure 18, companies with positive sentiment on the earnings calls are far more profitable (in terms of both ROE and net profit margin). Furthermore, management confidence actually translates into even higher profitability from operations in the subsequent quarters. On the opposite end, gloomy comments from less profitable firms lead to deteriorating operating performance in the future.

Figure 18 The progression of profitability for companies with bullish/bearish earnings calls

A) ROE, current and next two quarters



B) Net margin, current and next two quarters



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

VALENCE AWARE DICTIONARY AND SENTIMENT REASONER (VADER)

While the "bag-of-words" approach with binary (positive/negative) tone assignment to each word/token is easy to implement and quite effective, we can also experiment with more sophisticated algorithms. As the strength of the sentiment expressed in a word or collection of words normally varies with the use of modifiers such as adjectives and adverbs. A valence aware approach resolves this issue as it incorporates the perceived intensity of sentiment, which would otherwise be almost impossible to capture in a "bag of words" model. VADER implements the grammatical and syntactical rules described in Hutto, et al [2014].

VADER lexicon is sensitive both to the polarity and the intensity of sentiments expressed in a formal document and on social media platform. It builds itself from popular word banks such as LIWC and ANEW (see Bradley, et al [1999]). It uses 7,500 lexical features with validated valence scores that indicate both the polarity (positive/negative), and the intensity on a scale from -4 to $+4$. For example, the word "okay" has a positive valence of 0.9, "good" is 1.9, and "great" is 3.1, whereas "horrible" is -2.5, the frowning emoticon ":(" is -2.2, and "sucks" and its social media slang "sux" are both -1.5.

VADER also incorporates word-order sensitive relationships between terms to improve the sentiment intensity analysis. For example:

- Punctuation. For example, the exclamation point (!), increases the magnitude of the intensity without modifying the sematic orientation. "*This has been a great quarter!*" is more intense than "*This has been a great quarter*".
- Capitalization, especially ALL-CAPS to emphasize a sentiment-relevant word in the presence of other non-capitalized words increases the magnitude of the intensity.
- Degree modifiers (also known as intensifiers, booster words, or degree adverbs). impact the degree of sentiment by either increasing or decreasing the intensity. For example, "This

acquisition is extremely accretive" is more intense than "This acquisition is accretive", while "This acquisition is somewhat accretive" reduces the intensity.

- The contrastive conjunction "but" signals a shift in sentiment polarity, with the sentiment of the text following the conjunction being dominant. For example, "This quarter's earnings were strong, but we want to be cautious on our revenue outlook" has mixed sentiment, with the outlook part dictating the overall sentiment.

Along with valence scores for positive, negative and neutral tone, VADER provides a compound score. This is computed by summing the valence score of each word in the lexicon, adjusting according to the rules, and then normalizing. It is this normalized, weighted composite score that makes up their most useful metric for a single unidimensional measure of sentiment for a given sentence.

In statements #2 to #3 below, we can see how degree modifiers like "less" and "extremely" decreases and increases overall compound score. Also replacing "successful" with "not successful" in statement #4, flips the overall sentiment of the text completely. While the negation such as "not good" and "not bad" can still be captured through bigram bag-of-words in a lexicon based approach, most other modifiers are difficult to handle. These subtle changes are trivial for a human reader, but are extremely difficult or impossible to capture using a bag-of-word based lexicon approach with binary classifications. For instance, conventional dictionaries consider "less" a negative word and "successful" as a positive word; therefore, traditional bag-of-word approach generates a net zero tone for statement #2. While "less successful" to a human reader means, still successful but to a lesser extent, the VADER model produces a net positive sentiment score.

1. "Our R&D pipeline was successful this year."
   "Positive": 0.375, "Negative": 0.0, "Compound": 0.2023

2. "Our R&D pipeline was less successful this year."
   "Positive": 0.274, "Negative": 0.0, "Compound": 0.1298

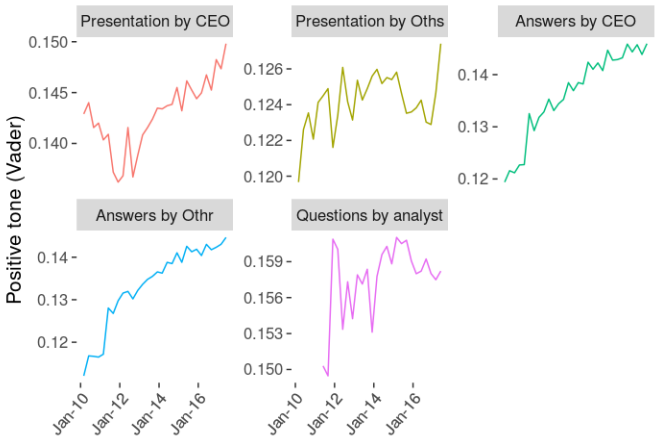
3. "Our R&D pipeline was extremely successful this year."
   "Positive": 0.344, "Negative": 0.0, "Compound": 0.2716

4. "Our R&D pipeline was not successful this year."
   "Positive": 0.0, "Negative": 0.285, "Compound": -0.1511

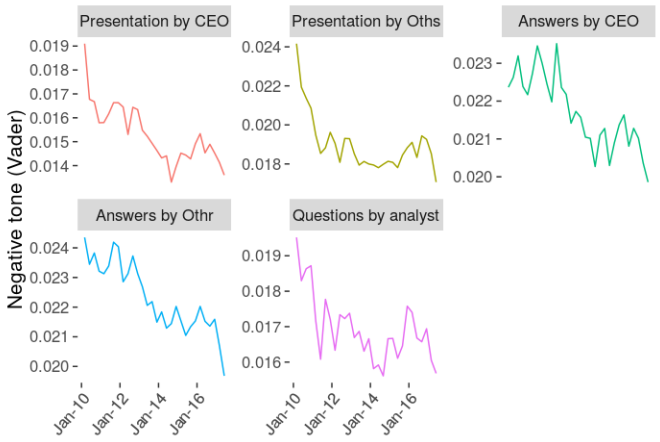
Figure 19 shows the median sentiment of all companies in our transcript database, based on the VADER model. Results are as expected and consistent with the bag-of-words approach.

Figure 19 Median tone of call transcripts, based on the VADER model

A) Positive Tone



B) Negative Tone



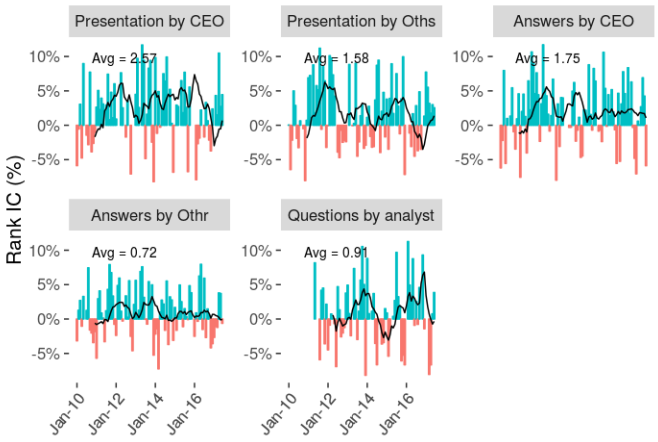
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

Tone at the Top

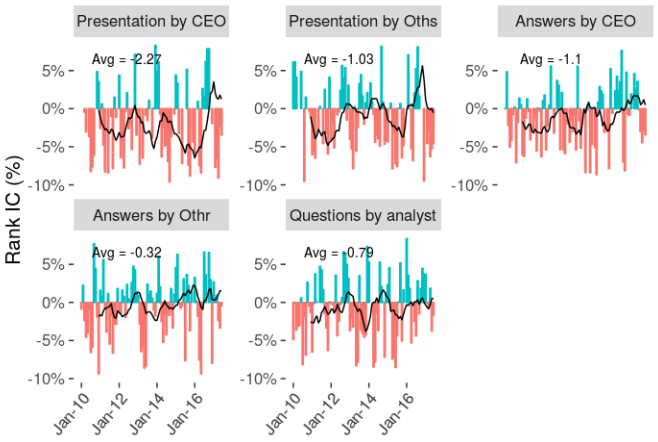
The refined VADER sentiment model also shows stronger predictive power of future stock returns than the conventional bag-of-words approach (see Figure 20). Positive (negative) tone leads to higher (lower) subsequent returns. The compound measure further boosts performance (see Figure 20 C). CEO communication, both from the presentation and the Q&A sections, has a much stronger impact on stock returns than the rest of the management team.

Figure 20 Rank IC of Positive / Negative tone based on VADER sentiment

A) Rank IC Positive tone measure



B) Rank IC Negative tone measure



C) Rank IC compound tone measure

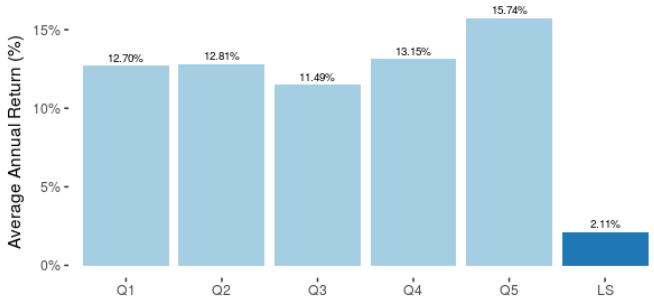
D) Rank IC Neutral tone measure

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

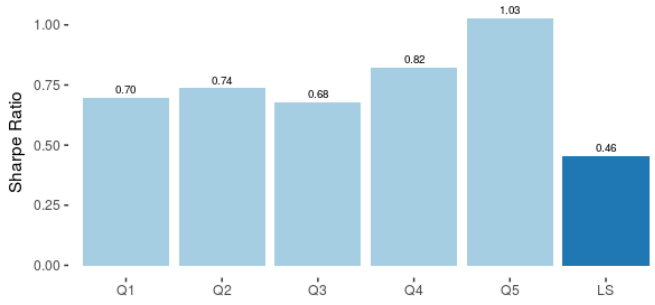
Lastly, we combine the sentiment measures from the CEO and other executives to create an overall VADER sentiment signal for the presentation and Q&A section, respectively. Our VADER composite factor based on the Q&A section delivers superior performance (see Figure 21). This may be due to the fact that VADER is better suited for informal language used in the Q&A discussions rather than the well scripted words used in the main presentation section.

Figure 21 Quintile portfolio performance, VADER Sentiment, Russell 3000 universe

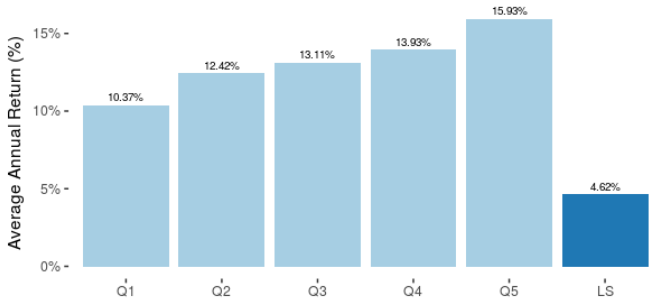
A) VADER presentation composite, annualized returns



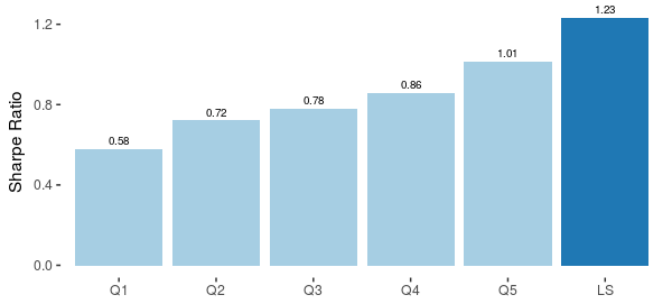
B) VADER presentation composite, Sharpe ratio



C) VADER Q&A composite, annualized returns



D) VADER Q&A composite, Sharpe ratio



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

QUANTIFYING EXECUTIVE PERSONALITIES

One of the cornerstones of today's fundamental research is to meet company management and attempt to get a deep understanding of how management views their own company/industry and the "styles" ("traits" or "personalities") of the executive team. Intuitively speaking, we certainly expect to find interesting relationships between executive traits and firm policies. Available academic research in this space is limited. Malmendier, et al [2005] and Malmendier and Tate [2008] study how management overconfidence impacts a firm's investment decision and merger activity. Graham, et al [2013] documents evidence that CEO behavior is related to measures of overconfidence, optimism, and risk aversion.

There are a number of challenges to understanding how executive traits can influence firm policy, operational performance, and stock prices. First, it is overly costly to engage both management and psychologists to conduct such personality tests. Senior management teams neither have the time nor interest to participate in such activities. Second, relying on human interpretation and assessment of CEO personality is neither reproducible nor objective. In a recent paper, Gow, et al [2015] (also see analysis in Jussa, et al [2017]) use linguistic features extracted from conference calls and statistical learning techniques to develop a measure of CEO personality. Gow, et al [2015] find that management personality factors are associated with organizational strategy choices, investment and financial policies, and firm performance.

In research years, psychology research has converged to focus on five personality factors (see Goldberg [1993]):

- Extraversion (versus introversion): network, aggressiveness, fast, enthusiasm, proactive, persuasion
- Emotional stability (versus neuroticism): respect, calm
- Agreeableness: listening skills, open to criticism, teamwork
- Conscientiousness: develops people, removes underperformers, efficiency, organization, commitments, attention to detail, persistence, work ethic, high standards, holds people accountable
- Openness to experience: flexible, brainpower, analytical skills, strategic vision, creativity

SYNTACTIC PARSER AND PART-OF-SPEECH (POS) TAGGING

We can gauge management personality by analyzing the structure of spoken words based on formal grammar rules. In NLP terminology, this is called syntactic parsing. A syntactic parser reads an input sentence and describes its grammatical structure. The parser usually returns a graph of word-word relationships, aiming to extract the reasoning from the underlying text. These parsed trees are useful for grammar checking or extracting meaning from news articles. More importantly, parsed trees are used for semantic analysis and hence play an important role in chat bots.

One aspect of semantic analysis is part-of-speech (POS) tagging. It is the process of tagging words or tokens to their respective part of speech classes such as nouns, verbs and adjectives. These classes are known as lexical categories or parts of speech.

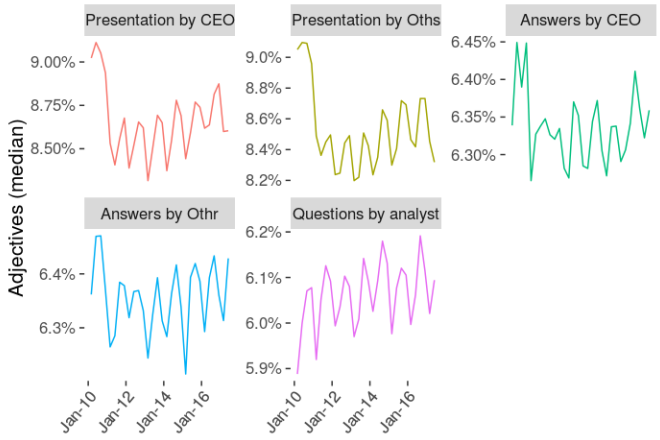
Use of Adjectives/Adverbs

In language, adverbs modify verbs, while adjectives modify other nouns and pronouns. In presuppositions, even if you disagree with the adjective or adverb in the sentence (e.g., "This quarter's earnings were extremely strong."), you still accept what is modified (e.g., "This quarter's earnings were strong"). Because adjectives and adverbs are less objective, the ratio of these words in sentences is a crude way of measuring personality.

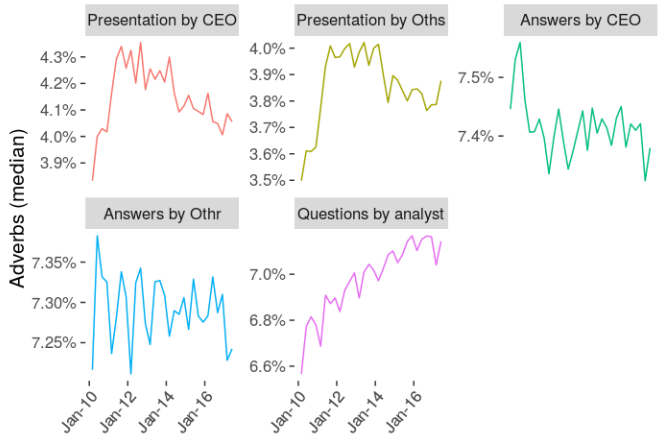
We can simply count the occurrence of adjectives or adverbs in a call transcript as a crude way to assess the subjectivity of the language spoken by the participants. Figure 22 shows the percentage of adjectives and adverbs in the call transcript database. There is no clear trend from the presentation or the answers from company executives. We see research analysts are moving away from simple to-the-point questions, to more subjective language with higher percentage of adjectives and adverbs.

Figure 22 Part-of-Speech distribution

A) Percentage of Adjectives



B) Percentage of Adverbs

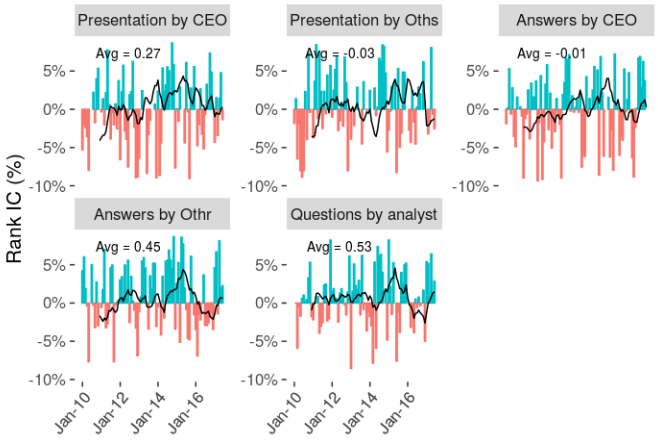


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

These crude measures of personality features are not particularly useful in predicting future stock returns (see Figure 23 A and B). However, we do note that more descriptive/subjective answers by CEO's are negatively correlated to subsequent stock performance (see Figure 23 B).

Figure 23 Percentage of Adjectives/Adverbs, Rank IC

A) Rank IC of Adjectives



B) Rank IC of Adverbs

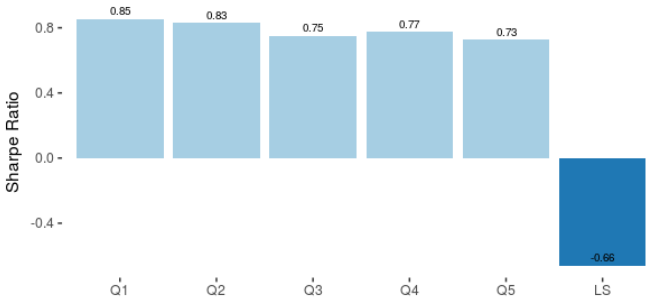
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

Use of Pronouns

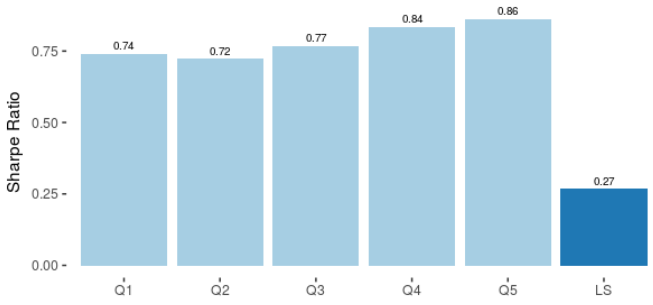
Another basic feature of personalities is the use of first person singular (*I*, *me*, *my, mine*) or second person plural (*we*, *us*, *our*, *ours*) pronouns in the spoken language. This is a subtle hue on the character of the executives as a team player. The performance of the signal is again modest, but the returns are in the intuitive direction. Higher reference to self/first person is negative, while more use of team/second person pronouns is positively associated to future company performance (see Figure 24).

Figure 24 Use of pronouns, quintile portfolio Sharpe ratio

A) Singular first person pronouns



B) Plural first person pronouns



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

RATIO OF DIGITS TO LETTERS

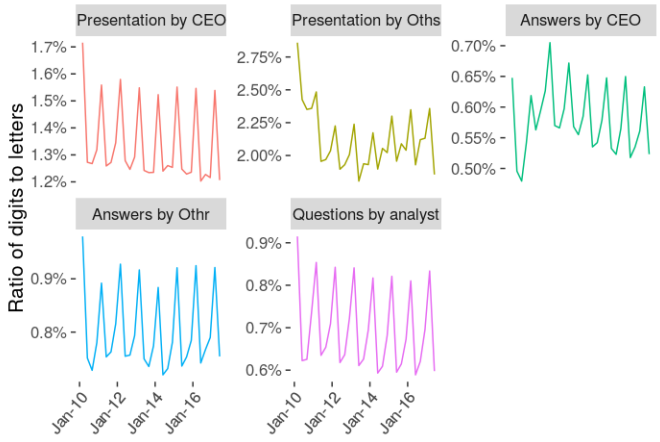
In our previous research (see Rohal, et al [2017]), we found that high rates of numbers over words in a company's 10K/10Q filings strongly predict future stock returns. Numbers are boring, but more concrete and objective. In addition, management tends to spend more time in explanation when the actual numbers are not particularly impressive. In the Q&A section, more usage of numbers might

also be indicative of how well-versed the management is with their financials and operations. Similar to other call transcript data, the ratio of digits to letters shows a strong seasonal pattern, peaking in the first calendar quarter (see Figure 25 A). Most companies use December 31 as their fiscal year end. Therefore, they release their fiscal year end results in Q1 of the following year. The fiscal year end results are far more comprehensive than interim releases. Non-CEO executives are more likely to offer facts than opinions.

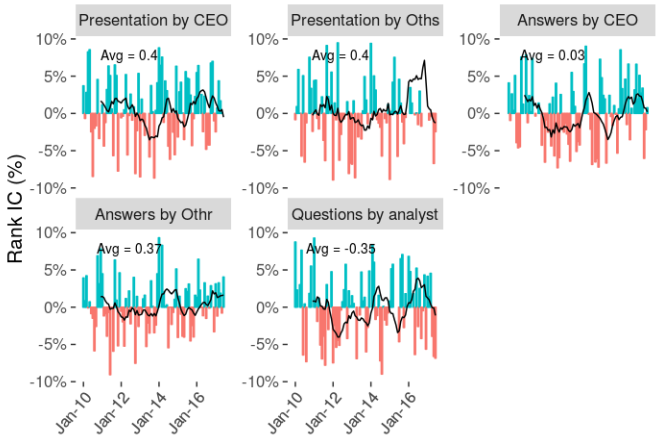
Historical performance of the signal is slightly positive (see Figure 25 B), as use of more numbers reduces the ambiguity of the message and leads to better future performance.

Figure 25 Ratio of digits to letters

A) Median distribution



B) Predictive power of stock returns, Rank IC

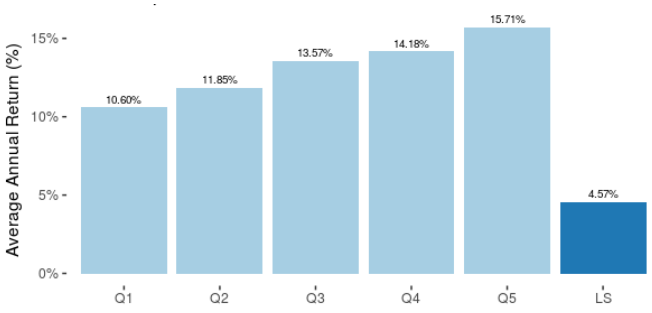


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

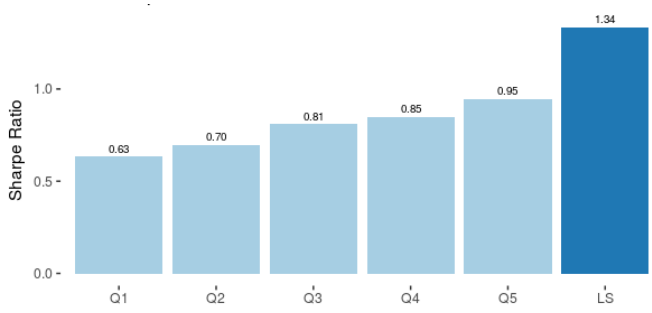
Although IC appears to be fairly weak, long/short portfolios (especially the one based on the presentation section) show decent Sharpe ratios (see Figure 26 B).

Figure 26 Ratio of digits to letters, quintile portfolio performance

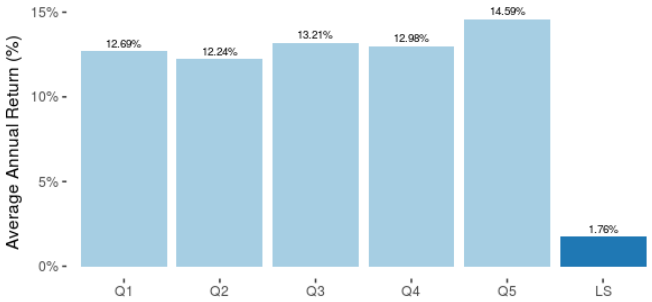
A) Presentation section, annualized returns



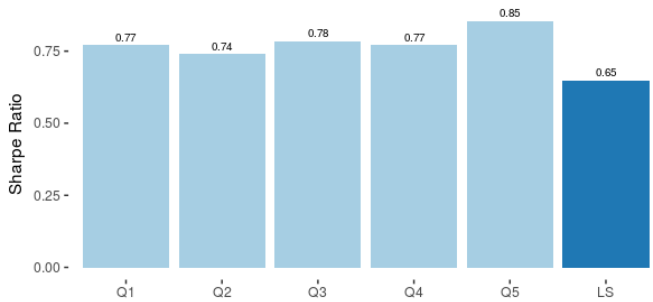
B) Presentation section, Sharpe ratio



C) Q&A section, annualized returns



D) Q&A section, Sharpe ratio



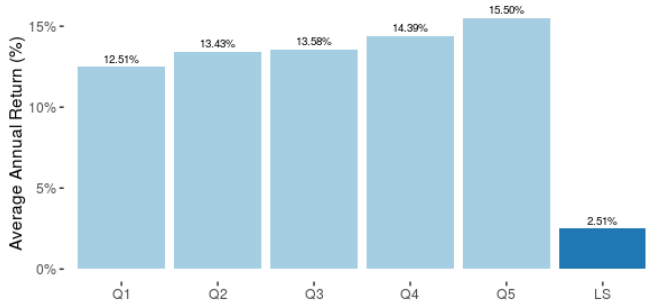
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

DURATION OF THE CALL, SHOULD THE CEO DELIVER THE MESSAGE?

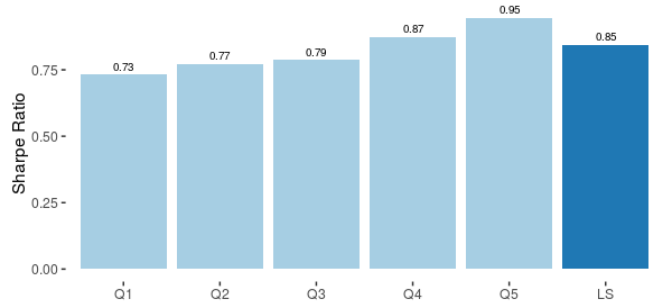
Next, we want to measure the level of active involvement by CEO's. One simple metric is the duration of CEO speech relative to other executives. Interestingly, for the main presentation section, active CEO engagement does lead to a reasonable Sharpe ratio (see Figure 27).

Figure 27 Duration of CEO/Other executives in the presentation section, quintile portfolio performance

A) Annualized returns



B) Sharpe ratio



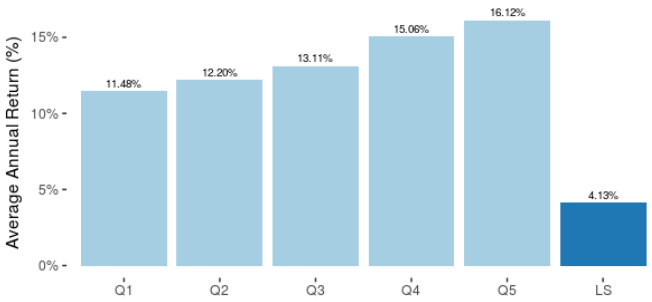
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

CEO SENTIMENT – IT'S ALL ABOUT THE CHANGE

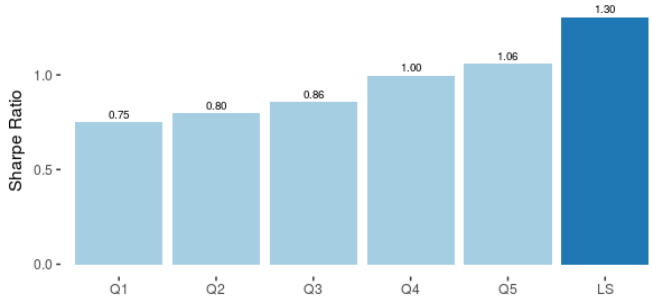
We have studied analyst sentiment and news sentiment extensively in past research (see Wang, et al [2017a] and Luo, et al [2017b]). Sentiment expressed by company management, especially CEO's should be equally important. Throughout the research so far, we have seen that CEO's play a dominant role in investor communications. CEO's tend to use carefully chosen words, with deft control of their tone. What and how CEO's speak matters greatly to company performance and share price. To control for the difference in skills/styles among CEO's and seasonality, in this section, we compute the year-over-year change in CEO sentiment, from both the presentation and Q&A sections. Figure 28 shows the performance of a quintile portfolio based on the CEO sentiment factor. As expected, the signal delivers a rather striking Sharpe ratio of 1.3x, and returns follow a monotonic increasing pattern to signal values.

Figure 28 YoY change in CEO sentiment, quintile portfolio

A) Annualized returns



B) Sharpe ratio

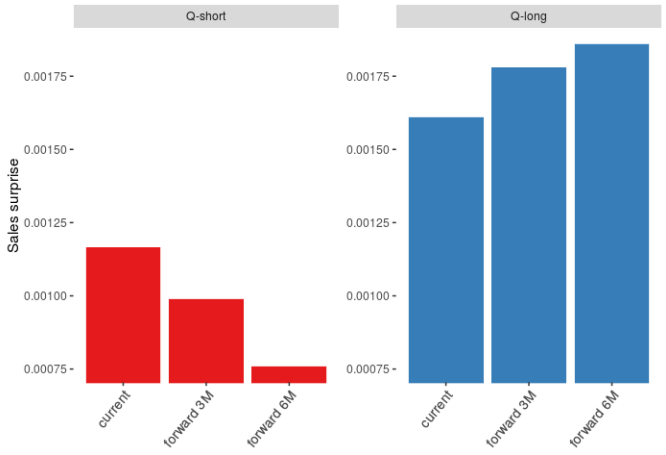


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

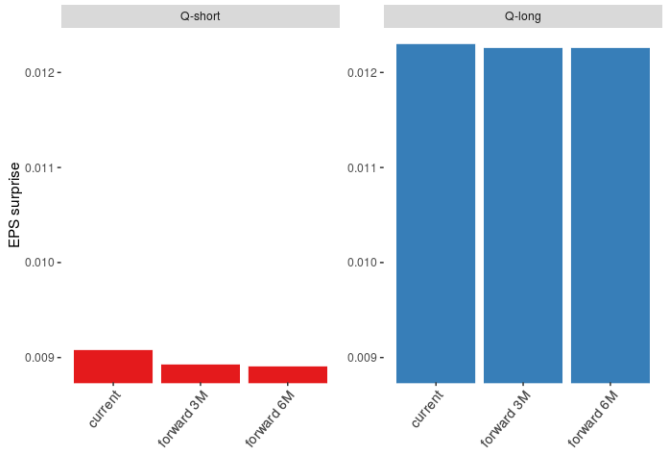
Not only is CEO sentiment highly predictive of returns, it is equally informative of future company fundamentals. As shown in Figure 29, companies with the most positive (negative) CEO sentiment consistently beat (miss) their revenue and earnings expectations, in the current and the next two quarters. Apparently, sell-side analysts can effectively take into account the numerical information in financial reporting and conference calls, but do not incorporate the tone of top in forming their models.

Figure 29 Positive (negative) CEO sentiment strongly predicts revenue/EPS surprises

A) Revenue surprise, current and next two quarters



B) EPS surprise, current and next two quarters



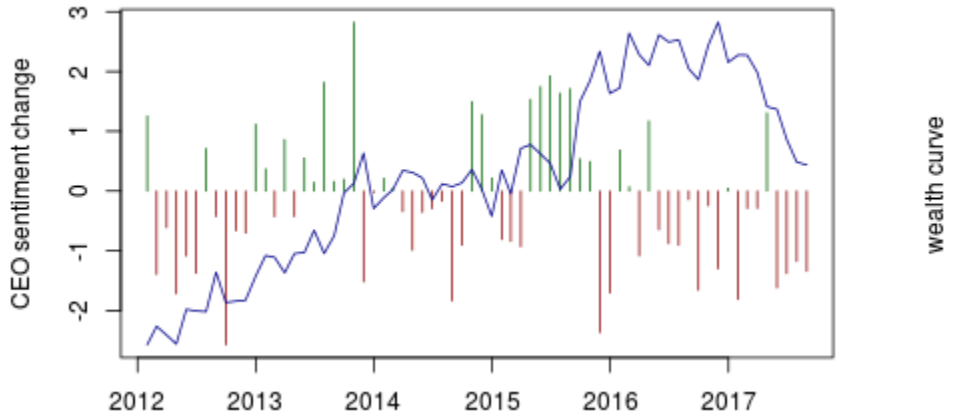
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

A Concrete Example – General Electric

Almost all senior management members have been heavily coached in business communication. However, we also know that effective communication is part art and part science. Some people are just naturally better speakers. One of the most well-known CEOs in effective communication is Jack Welch – former CEO of General Electric (GE). During the 20 years of Jack Welch era, GE's stock price rose close to 2800%.

Figure 30 plots a rather different trajectory of GE in recent years. We overlay GE's share price with CEO sentiment (year-over-year changes). GE has been in huge turmoil in the past few months. In 2017, the overall US equity market was up 20%, while GE stock plunged by 45%, eroding almost $100 billion of shareholder value. It is particularly interesting to note that GE's CEO sentiment was predominantly negative in the past two years.

Figure 30 General Electric CEO sentiment change and stock performance in the last few years



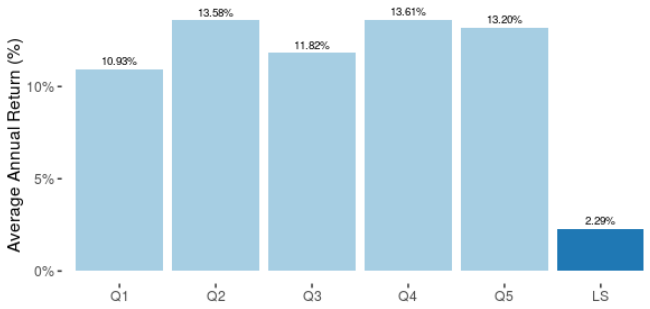
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES.

Sell-side Analyst Participation in Earnings Call

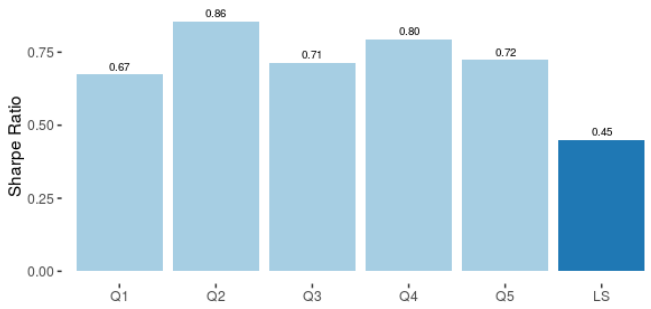
One last thought in this section is related to analyst participation. Sell-side analysts want to both understand what is beyond the written press releases, but also show interest in their coverage companies. Most analysts regularly participate in earnings calls. A lack of analyst participation is a proxy for "popularity". A simple backtesting does suggest that companies with more active analyst participation earn marginally higher returns (see Figure 31).

Figure 31 % of sell-side analyst participation in earnings call (over # of analyst coverage), quintile portfolio

A) Annualized returns



B) Sharpe ratio

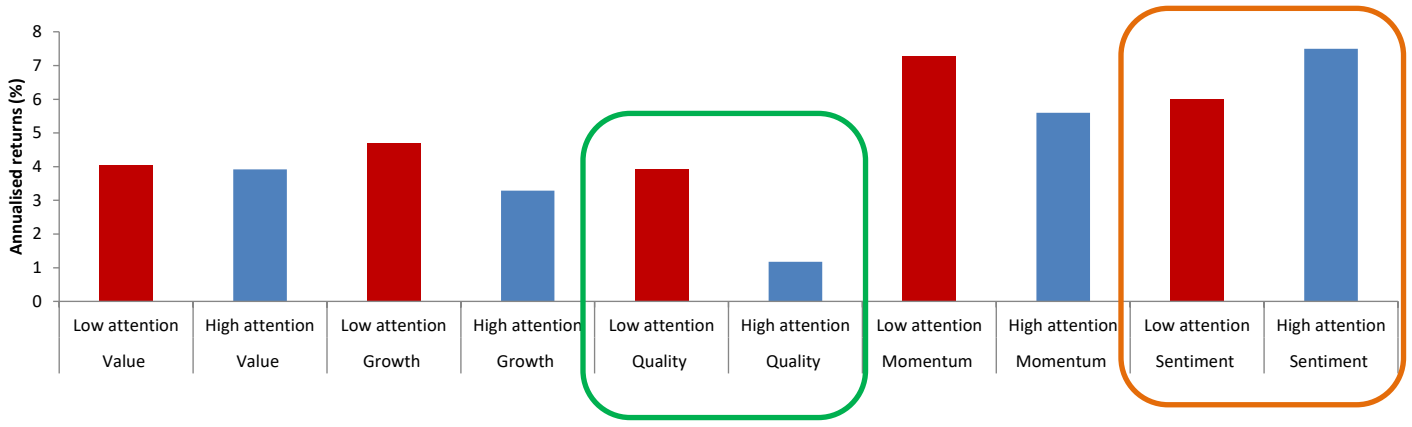


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

We also backtest the performance of common factors in two high/low attention universes based on the analyst participation in the conference call. A "high attention" universe means active analyst participation, while a "low attention" universe reflects few analyst participation. As expected, the performance of common stock-selection factors is much stronger for low attention universe particularly for Quality-type of factors (see Figure 32). The only exception is the Sentiment factor, which is again intuitive. Sentiment factors are derived from sell-side analyst expectations. If there is low participation/interest from the sell-side, sell-side sentiment will not be particularly reflective/accurate.

Figure 32 Common factors performance, in low-high attention universes

A) Long/short quintile portfolio annualized returns



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES.

TOPIC MODELLING

The third vast area of NLP research is on topic modelling, i.e., how to extract abstract "topics" that occur in a document. In a supervised learning approach, topics are typically pre-defined. For example, in a collection of M&A related news articles; topics can be manually defined as “completed acquisition”, “failed merger” and “spin-off announcement”, etc. However, in the case of earnings calls, the number of potentially meaningful topics is too large to pre-determine manually. In addition, the manual approach to topic modelling is never exhaustive and often biased, unless the sole focus of the study centers on a particular subject such as buy-backs or spin-offs.

In this section, we apply an unsupervised learning model. We train our NLP algorithms to systematically identify topics presented in a collection of texts, classify documents accordingly, and capture hidden patterns exhibited by the text corpus. The idea is to find a collection of words (“topics”) in large clusters of texts. These topics create a repeating pattern of co-occurring terms in the document collections.

BAG-OF-WORDS, N-GRAMS

Similar to sentiment analysis, a crude approach to topic modelling is tokenization. We perform a word frequency count on each of the call transcripts and use this vocabulary of words as the list of topics. Therefore, the number of topics is the same as the number of unique words in the corpus, which is often in the range of hundreds of thousands. We can then apply machine learning techniques to extract the most relevant words out of this large set. However, using thousands of dependent variables (“unigram topics”) with limited number of call transcripts is a classic trap for overfitting.

More importantly, single word (unigram) topics are not very meaningful. For example, the topic “margin” is not nearly as meaningful as “margin contraction”. Therefore, we further tokenize our text into bigrams (two words), trigrams (three words), and n-gram (i.e., n words) to extract more meaningful topics.

Let’s take another concrete example. For the sentence, 'We have reduced the gearing, winning new contracts and our growth outlook looks positive.” We first remove punctuations, special characters and stop words. As a reminder, stop words are commonly occurring words, such as “have”, “we” and “the” in the above sentence. Next, we tokenize the sentence into n-grams (unigrams, bigrams and trigrams).

- **Unigrams:** ('reduced'), ('gearing'), ('winning'), ('new'), ('contracts'), ('growth'), ('outlook'), ('looks'), ('positive')
- **Bigrams:** ('reduced', 'gearing'), ('gearing', 'winning'), ('winning', 'new'), ('new', 'contracts'), ('contracts', 'growth'), ('growth', 'outlook'), ('outlook', 'looks'), ('looks', 'positive')
- **Trigrams:** ('reduced', 'gearing', 'winning'), ('gearing', 'winning', 'new'), ('winning', 'new', 'contracts'), ('new', 'contracts', 'growth'), ('contracts', 'growth', 'outlook'), ('growth', 'outlook', 'looks'), ('outlook', 'looks', 'positive')

We can clearly see that the unigram topics do not make a lot of sense on a standalone basis. While some of the bigrams are more meaningful such as “reduced gearing” and “new contracts”. With trigrams, we are able to capture even more useful topics such as “winning new contracts” and

"outlook looks positive". In an ideal topic model, we wish to extract these informative topics and at the same time ignore other non-relevant ones.

We can do a frequency count of these n-gram topics across transcripts. Ignore the less commonly occurring topics and hope that the more frequently occurring topics are the more meaningful. Similar to stop words, however, extremely popular phrases with very high "document frequency" (defined as high occurrence across documents) might be trivial. Using bigrams or trigrams also effectively reduce the number of topics, as those meaningless combinations of words get ignored due to their low frequency counts.

An effective algorithm to refine weights for these topics is called Term Frequency-Inverse Document Frequency (TF-IDF). This is a statistical measure used to evaluate how important a word is to a document in a text collection (corpus). The importance increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

$$TF(t,d) = (\#\ of\ times\ term\ t\ appears\ in\ a\ document) / (Total\ \#\ of\ terms\ in\ the\ document)$$

$$IDF(t,d,D) = log(Total\ \#\ of\ documents / \#\ of\ documents\ with\ term\ t)$$

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

Another way to refine topics is through POS tagging, as elaborated in a previous section. For example, the occurrence of a noun word may be more relevant as a topic than a preposition or conjunction. Alternatively, those words conveying sentiment are probably more relevant. In the next few sections, we explore a few interesting machine learning techniques for topic modelling.

ABSTRACT, UNSUPERVISED LEARNING WITH VECTORIZATION

NLP algorithms conventionally treat words as discrete atomic entities. For instance, it may assign an "ID:225" to "oil" and an "ID:570" to "energy", without regard to the relationship between the two individual words. These encodings are arbitrary, and provide no useful information. Another problem with the above representation is that assigning a unique ID to each word leads to data-sparsity. We end up with thousands of unique words, which mean that we may not have enough data to train a useful machine learning model.

A sophisticated unsupervised learning model to help segregate documents into topics is document vectorization and word embedding. Given that most machine learning algorithms are more suited to numbers than text, we can restructure our textual information and represent it through a set of numbers. Such representation can be effective as long as readability is not important, as the algorithm is able to learn the relationship among words.

DISTRIBUTED REPRESENTATION WITH WORD2VEC (WORD EMBEDDING)

Vector representations of words or word embedding, can overcome some of the impediments with traditional representation. The model hypothesizes that the words appearing in the same contexts should share semantic meaning. Word embedding uses a distributed representation, where semantically similar words are mapped to nearby points. The model computes the statistics of how often some words co-occur with their neighboring words in a large text corpus, and then map these count-statistics down to a small, dense vector of continuous numbers for each word.

Given a collection of text documents, the Word2Vec algorithm first constructs a vocabulary from the training text data and then learns vector representation of words. In this representation, words with similar meanings appear in close clusters. Word2Vec was published by Google in 2013. It is a neural network implementation that learns vector representations of words.

Now, let's use a simple example to demonstrate the concept. We train the Word2Vec algorithm on the Q&A section of all call transcripts published in the year of 2015. Figure 33 shows the vector representation of a few words after the training process. Here we are trying to represent each word with a vector of five numbers. We see clear patterns where closely connected words have similar vector representation.

Our choice of five numbers is arbitrary and to make it human digestible. In our research, each word is represented with around 100 numbers. Think of these numeric vector representations as abstract topics. Earnings calls from thousands of companies should at least be talking about a few hundred topics.

Figure 33 An example of Word2Vec representation, using 2015 call transcripts

Word/Vector	Number1	Number2	Number3	Number4	Number5
"oil"	-0.336	0.598	0.205	0.499	-0.488
"rig"	-0.321	0.416	0.753	-0.085	-0.387
"crude"	-0.428	0.586	0.181	0.440	-0.498
"technology"	-0.639	-0.221	-0.725	-0.110	0.066
"system"	-0.677	0.038	-0.716	-0.121	-0.113
"hardware"	-0.705	0.014	-0.644	0.273	-0.116
"services"	-0.693	0.089	-0.657	0.225	-0.173

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES.

Now that we have transformed words into machine readable numbers, we can easily evaluate the relationship between them. Figure 34 shows the correlation matrix among the same five words. The algorithm is able to find two clusters of words. For example, "crude", "oil" and "rig". Furthermore, within the cluster, "crude" and "oil" are often used together, while "oil" and "rig" form another sub-cluster. Similarly, "hardware", "services", "system", and "technology" are often used together.

Figure 34 Correlation matrix for the words using the Word2Vec representation

	Crude	Oil	Rig	Hardware	Services	System	Technology
Crude	100%	100%	67%	52%	58%	39%	3%
Oil	100%	100%	65%	50%	56%	36%	0%
Rig	67%	65%	100%	-17%	-12%	-19%	-48%
Hardware	52%	50%	-17%	100%	99%	91%	85%
Services	58%	56%	-12%	99%	100%	93%	81%
System	39%	36%	-19%	91%	93%	100%	90%
Technology	3%	0%	-48%	85%	81%	90%	100%

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES.

The Word2Vec model can reveal many interesting patterns. Given a set of words, it can identify which word is dissimilar from the others. For example, it is able to separate the word “analyst” from the set of words “consumers”, “guest”, “user”, “shopper”, and “retailer”.

Our trained model is now capable of distinguishing changes in meaning owning to a financial context. In Figure 35, we ask the trained model to give out those words most similar to the ones listed on the left side of the table. India and Japan are most frequent alongside China, followed by other countries. It can differentiate a cluster of countries from currencies.

Figure 35 Most similar words to the word mentioned on left using the Word2Vec algorithm

china	=	'india',	'japan',	'brazil',	'taiwan',	'europe',	korea'
usd	=	'rmb',	'rub',	'aud',	'mxn',	'brl',	zar'
retail	=	'wholesale',	'supermarket',	'foodservice',	'grocery',	channel',	ecommerce'
airline	=	'airlines',	'airplane',	'ife',	'lufthansa',	'lessor',	widebody'
bank	=	'banks',	'banking',	'telekom',	'banker',	'syndicate',	nonbank'
economy	=	'gdp',	unemployment'	'tourism',	'economic',	macro'	market'

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo’s QES.

Document to Vector

We can further improve the model with other variations. For example, we can average the word vectors across all words in a document to get an average vector representation of each document. A sophisticated TF-IDF weighting scheme can be added.

Instead of Word2Vec representation, we can conduct a Document2Vec representation. A Document2Vec model should be more effective than the bag-of-words or Word2Vec approaches, since vector averaging and clustering on word vectors lose the word order, whereas Document2Vec representation preserves it.

DEEP LEARNING VIA CONVOLUTIONAL NEURAL NETWORKS (CNN)

Whichever way we summarize the numerical representation of words, we are bound to lose some useful information. First, the Word2Vec representation adds one layer of dimensionality, which needs to be reduced. Then another machine learning model needs to be trained for predictive analysis on the summarized vector space.

An alternative approach is to combine these analytics all at once. In recent years, deep learning, in particular CNN (Convolutional Neural Networks) has been successfully applied to image processing, speech recognition, and even the ancient GO game. In image processing, CNN models honor the high-dimensional spatial structure in images, whilst are still robust to the position and orientation of learned objects in the scene. The same principles can be used on textual information to preserve the sequence of words in a document. The same characteristics that make the CNN model effective at recognizing objects in images can potentially be used to learn structure in sentences and documents.

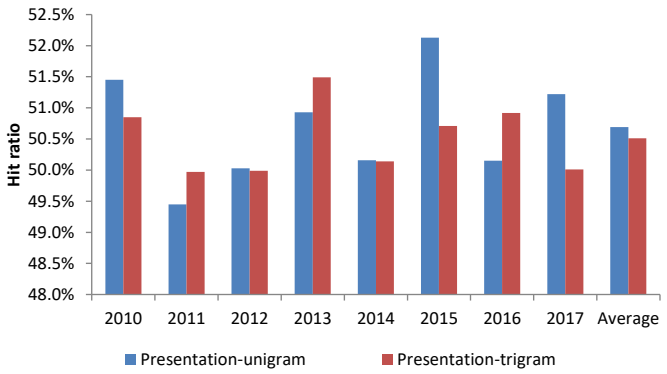
We apply the CNN model on earnings calls and iterate on various sizes of Word2Vec representations. We use the forward three-month stock return to classify outperformers/underperformers as the binary dependent variable. The performance of the trained model is relatively weak (see Figure 36). The hit rates range between 50% to 51%, for both unigram

and trigram based models, with the presentation section being slightly more informative. The disappointing results may be due to the relative short history of training data and sparsely distributed information in the text.

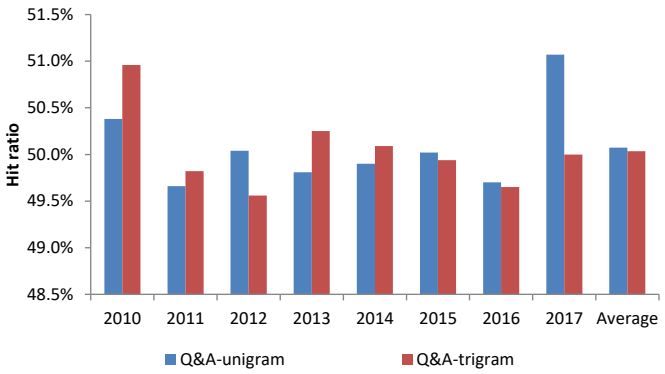
These results are not unusual; Liang [2016] apply multiple cross-validated machine learning models on earnings call transcripts but find that the transcripts alone are not informative enough to predict stock price changes. Instead of training a CNN model from scratch we could have also initialized the word vectors with the Word2Vec model trained in the previous section. However, given the extremely heavy computational cost of deep learning models, we judiciously move on and explore other NLP algorithms.

Figure 36 Accuracy of CNN models, trained on earnings call transcript data

A) Hit rate, the presentation section



B) Hit rate, the Q&A section



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

LATENT DIRICHLET ALLOCATION

For any machine learning algorithm to handle text based documents effectively, we need to first make the text machine readable. We have discussed various approaches to this. The bag-of-words approach, for example, creates a unique identifier for each word (unigrams) or phrase (bigrams, trigrams). Next, we explore a more abstract Word2Vec approach that can translate continuous words to vector representation.

In a more traditional approach, the idea behind topic modelling is to extract the salient features from the document while at the same time preserving its readability. Topics are represented as a collection of phrases, which makes the process intuitive. Essentially, the goal is to convert documents from bag-of-words to bag-of-topics. The above approach to topic modelling involves a lot of subjective processes. Should we choose bigrams or trigrams? Should we remove low frequency words or the words which are too common?

Latent Dirichlet Allocation (LDA) is an algorithm that can systematically optimize along these dimensions. LDA assumes documents are related to a set of topics and these topics relate to a set of words. It is a matrix decomposition technique on the document-term matrix (DTM). In the vector space, a collection of documents can be represented as DTM, which is a matrix of word frequency count by document. LDA decomposes this DTM into two lower dimensional matrices – DTM1 and

DTM2. DTM1 is a document-topics matrix (frequency of topics by documents) and DTM2 is a topic-terms matrix (frequency of words by topics).

The LDA algorithm first randomly assigns words to topics. Then it optimizes based on the proportion of words in document d that are currently assigned to topic t, and the proportion of assignments to topic t over all documents D that come from the word w.

A Base LDA Model with Words Pruned using Frequency Count

Let's start with a simple example. We train the LDA algorithm on the Q&A section of the call transcript database from the year of 2015. There are 25,000 transcripts with about 100,000 unique numbers of words. To reduce dimensionality, we remove all terms appearing in less than 5% of the documents (too sparse/uncommon) and those in more than 50% of documents (too trivial). This shrinks the number of unique words from 100,000 to 2,000. The two hyper-parameters can be further tuned and optimized. If the criteria are too restrictive, we may eliminate some important terms; while if they are too loose, we end up with too many topics.

Figure 37 shows the top 10 topics picked by the LDA algorithm. Each line is a topic with three words and the weights for each term related to that topic. To make it intuitive for our readers, we manually assign a meaningful topic name to each collection of words. For example, we call first topic "BioPharma", since words with highest weights for this group are "patient", "data" and "study".

Figure 37 Top 10 LDA topics using earnings calls in the year of 2015

```
"BioPharma"      = (0, '0.033*"patient" + 0.022*"data" + 0.017*"study"')
"Retail"         = (1, '0.034*"store" + 0.027*"brand" + 0.014*"category"')
"Geo-politics"   = (2, '0.021*"china" + 0.020*"europe" + 0.015*"currency"')
"Oil and Gas"    = (3, '0.023*"project" + 0.016*"oil" + 0.014*"gas"')
"Banking"        = (4, '0.020*"loan" + 0.018*"bank" + 0.013*"credit"')
"Financials"     = (5, '0.022*"asset" + 0.011*"portfolio" + 0.011*"acquisition"')
"Not meaningful" = (6, '0.012*"eur" + 0.009*"ph" + 0.008*"please"')
"Software"       = (7, '0.014*"technology" + 0.010*"system" + 0.008*"team"')
"Not meaningful" = (8, '0.009*"fourth" + 0.006*"mix" + 0.006*"expense"')
"IT"             = (9, '0.019*"service" + 0.014*"data" + 0.011*"platform"')
```

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES.

As we can see from the above list, the LDA algorithm is able to group a similar set of words together to create a relevant topic. The first topic ("patient", "data", "study") relates mostly to BioPharma companies; the second topic ("store", "brand", "category") is about the retail sector; and the third topic ("china", "europe" and "currency") relates to geopolitical risks. In this example, the algorithm has done a good job in identifying relevant topics and representing them through a combination of related key words.

One unique feature and for our purpose, a problem, is that the topics are predominantly sector specific themes. As shown in Figure 37, seven of the 10 topics are related to a certain sector (i.e., "BioTech", "Retail", "Oil and Gas", "Banking", "Financials", "Software", and "IT"). Since we already have an accurate industry classification system based on the GICS for each company, this additional information is not useful. More importantly, our purpose is to predict residual returns above and beyond risk factors such as sectors/industries.

It is extremely difficult to isolate topics that are not sector specific, since most conversations in the financial context are dominated by sector related issues and discussions. We have attempted to run multiple iterations of the LDA model, removing sector-related topics at each iteration. However, this process is not only extremely slow, but prone to overfitting. We have also tried to train a separate model for each GICS sector, but then we end up with mostly industry related topics.

Removing Sector-Specific Topics using the Herfindahl Index

We develop a more efficient approach. First, we do a term frequency count within GICS sector. Then compute a concentration score using the Herfindahl Index. Lastly, we remove the words with high sector concentration (i.e., words with the Herfindahl Index greater than 25%).

Interestingly, not only do we remove most sector specific terms, but also those phrases that are infrequent, misspelt, non-English language, and names of persons.

Although unigram models perform well, we expect them to be more correlated to other NLP signals we have discussed in the previous sections. In addition to single word, we also perform the same procedure using bigram and trigrams topics. Most trigram terms are infrequent and trivial triplets of most common words in call transcripts. Therefore, we stick to bigrams in this section.

A Logistic Regression Model

Now, we want to understand what bigram topics (identified with our LDA model, filtered by the Herfindahl Index) are predictive of future stock returns. Rather than a more complicated CNN model presented in the previous section, we decide to go for a simple logistic regression model.

The dependent variable in our logit model is a binary variable – outperformers versus underperformers. The outperformance/underperformance is defined by sector neutral forward three month returns. Stocks in the top (bottom) tercile are classified as outperformers (underperformers). The logit model is used to determine what bigram topics are predictive of future winners/losers.

We expect the topics with positive tone to have positive coefficients in the logit regression, and vice-versa, negative coefficients for phrases with negative sentiment. Figure 38 shows the list of topics with the most positive coefficients in our logit model. Interestingly, most topics are intuitively related to good news, e.g., “market share gain”, “profit margin improvement”, “buyback programs”, and “earnings growth”; and therefore are linked to potential future share price outperformance.

Figure 38 LDA base model bigram topics with most positive coefficients in the logit regression

```
“market share”    = ("share_gain", "market_growth", "gain_market")
“profit margin”   = ("profit_margin", "margin_increased", "margin_improved")
“buyback”         = ("repurchase_program", "repurchased_numericss", "stock_repurchase")
“interest”        = ("interest_income", "interest_margin", "net_charge")
“margin improvement”  = ("margin_expansion", "margin_improvement", "numericss_eps")
“Not meaningful” = ("rate_increase", "numericss_member", "cost_trend")
“sentiment”       = ("good_year", "good_news", "numericss_good")
“prices”          = ("commodity_price", "lower_price", "low_price")
“growth”          = ("profitable_growth", "earnings_growth", "drive_growth")

Note: We use “numericss” to represent numbers.
```

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo’s QES.

New Initiatives or Management Pet Projects

Figure 39 shows the list of topics with most negative coefficients in the logit regression. Interestingly, "strategic initiatives" and "new platform" are linked to poor future performance. These new initiatives seem to be mostly management pet projects which do not add significant shareholder value. Cliché words without concrete meanings such as "next generation", "strategic plan" are also penalized by the market. Company executives are not economists; therefore, when they speak of "prevailing market condition" or "challenging market", they do so to blame poor performance to something outside of their control. Management is typically more interested in expanding business than cost control. Therefore, when "cash cost" and "cost per" are heavily used, they are more likely to be associated with underperformance. The topics specifically related to power and retail sectors are probably not surprising, given the tough environment faced by these two sectors in recent years.

One surprising observation is that, unlike the topics with positive coefficients, there are few negative words in the topics with negative coefficients. In the Loughran and McDonald [2011] dictionary, there are around 2300 negative words, but only 350 positive ones. Wouldn't we expect more negative words? In the sentiment analysis section (e.g., Figure 15), company management is twice more likely to use positive than negative words in the earnings calls. It is probably not surprising that top management can always find something positive to say in an investor's call, even if financial performance is not all that impressive.

Figure 39 LDA base model bigram topics with most negative coefficients in the logit regression

```
"market condition"   = ("market_environment", "low_level", "challenging_market")
"cost"           = ("cash_cost", "cost_per", "per_numericss")
"mixed"          = ("value_proposition", "technology_solution", "solution_business")
"new development"= ("new_platform", "new_way", "million_people")
"new technology" = ("new_technology", "next_generation", "numericss_technology")
"retail sector"  = ("retail_sale", "retail_business", "numericss_retail")
"new initiative" = ("strategic_initiative", "growth_initiative", "strategic_plan")
"sales"          = ("comparable_sale", "sale_increase", "total_sale")
"power sector"   = ("power_plant", "power_generation", "numericss_power")
```

Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES.

We can further refine our algorithm by forcing the bigrams to have at least one word with a sentiment/tone or a noun.

Specifying the Final LDA Topic Model and the Logit Regression on the LDA Topics

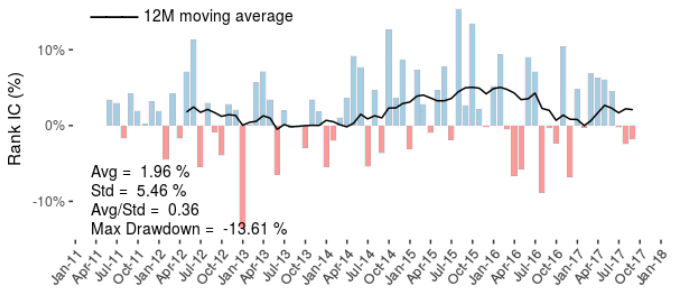
The examples in the previous section are based on in-sample analysis, i.e., we use all call transcript data to fit our logistic regression models. The goal is to understand the intuition behind the LDA (with Herfindahl filter) and logit models, rather than real-time prediction.

To perform a pure out-of-sample backtest, we train a separate LDA/Herfindahl/Logit model for each year from 2009 to 2017, using an expanding window. For example, we use all transcripts in 2009 to train the logit model for January 2010. We require a minimum of one year's worth of data for our first backtest; therefore, our results start in January 2011.

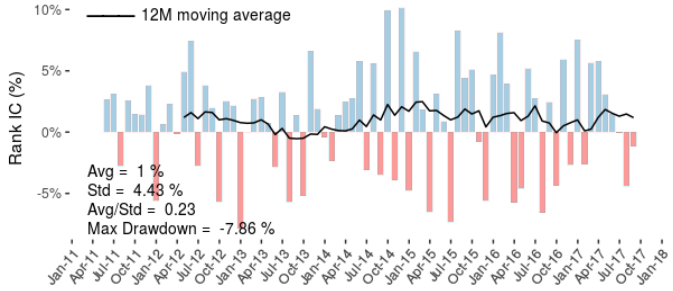
Figure 40 shows the performance (as measured by Rank IC) of the LDA/Herfindahl/Logit models, using the information from the presentation and the Q&A section, respectively. This particular model seems to be more powerful to extract useful information content from the presentation section than from the Q&A's.

Figure 40 LDA/Herfindahl/Logit model, Rank IC, Russell 3000

A) Rank IC for presentation section



B) Rank IC for Q&A section

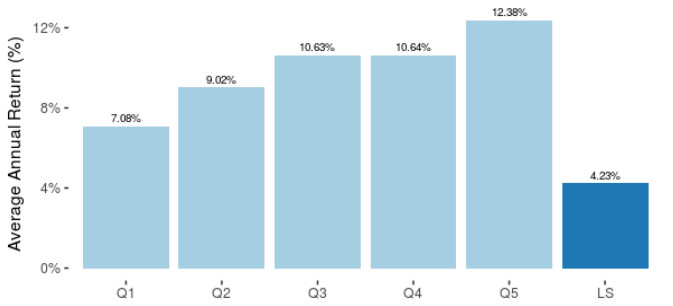


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

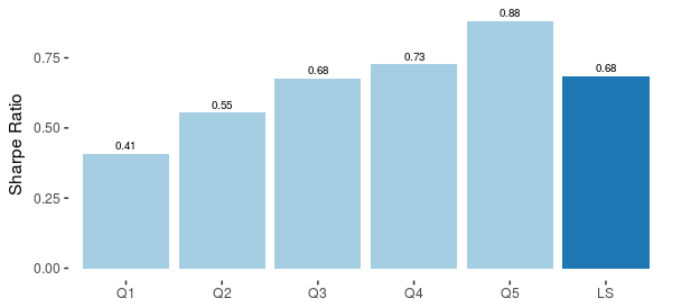
Figure 41 shows the performance of the long/short quintile portfolio based on the LDA/Herfendahl/Logit model trained on the presentation section. Performance is comparable to some of the strongest sentiment signals presented in the previous sections. However, the LDA/Herfendahl/Logit model is also among the most computationally intensive.

Figure 41 LDA/Herfindahl/Logit model, quintile portfolio performance, Russell 3000

A) Annualized returns



B) Sharpe ratio



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

BUILDING A COMPOSITE MODEL OF CALL TRANSCRIPTS

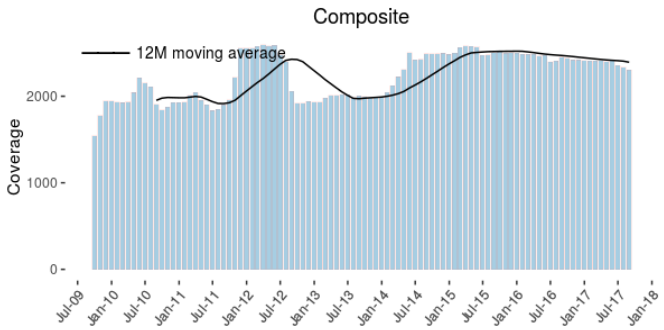
As you can see from our analysis thus far, there is a suite of interesting NLP and machine learning techniques that can be applied to the S&P Capital IQ's Transcript database. In the final section of this paper, we would like to conclude with a composite model – SMEC or Systematic Mining of Earnings Calls. The historical out-of-sample data for the SMEC is available upon request.

SYSTEMATIC MINING OF EARNINGS CALLS (SMEC COMPOSITE)

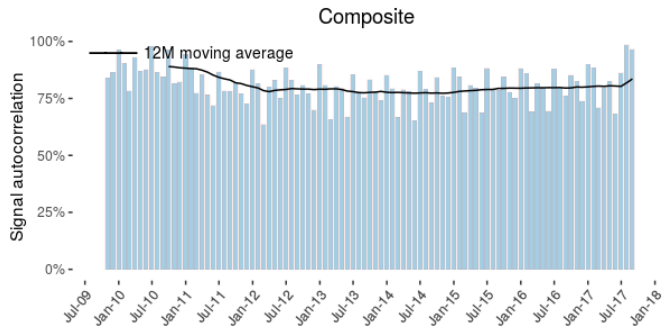
In the end, after showcasing many sophisticated text mining techniques, we would like to keep the final composite model relatively straightforward. The SMEC model equally weights all relevant text mining factors presented in this paper. In this section, the performance in the US equity market (Russell 3000 universe) is shown. Results for other international markets are presented in the next section. The SMEC model covers over 2000 stocks in the US (see Figure 42 A), with modest turnover[7] (see Figure 42 B).

Figure 42 SMEC composite coverage, Russell 3000

A) Coverage



B) Signal Autocorrelation



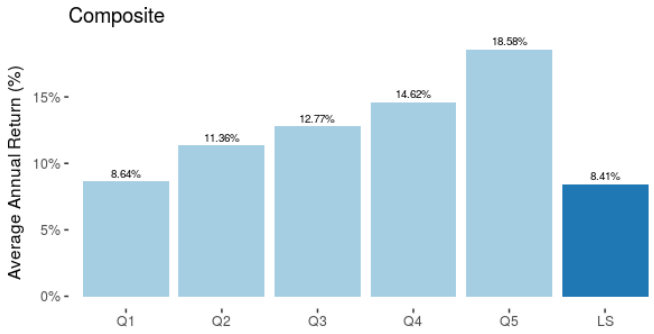
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

Equally weighted quintile portfolios show a monotonic return pattern (see Figure 43 A and B). A long/short portfolio delivers an annual Sharpe ratio close to 1.5x. Performance has been relative consistent over time (see Figure 43 C), with a forecasting horizon beyond a year (see Figure 43 D).

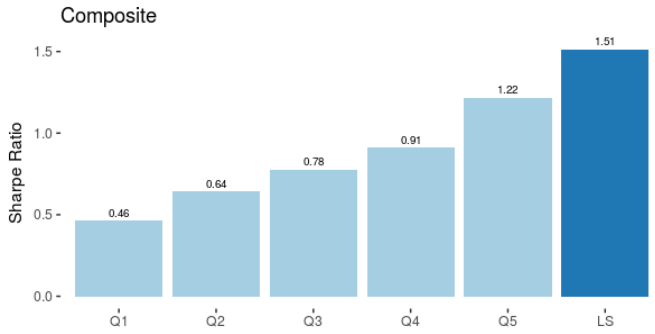
[7] We use the monthly signal autocorrelation as a proxy for turnover. The autocorrelation is always between +100% (in this case, a constant signal and therefore, no turnover) and -100% (in this case, a complete turnover of 100%).

Figure 43 SMEC composite quintile portfolio performance, Russell 3000 sector neutral universe

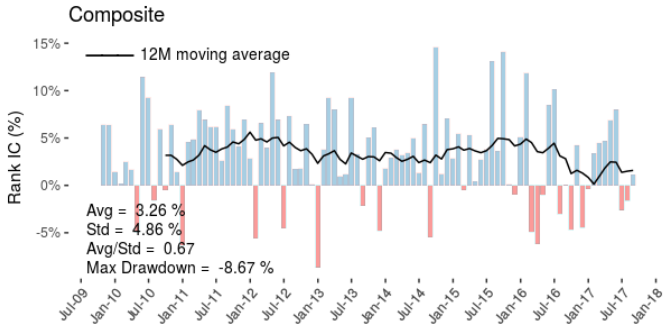
A) Annualized returns



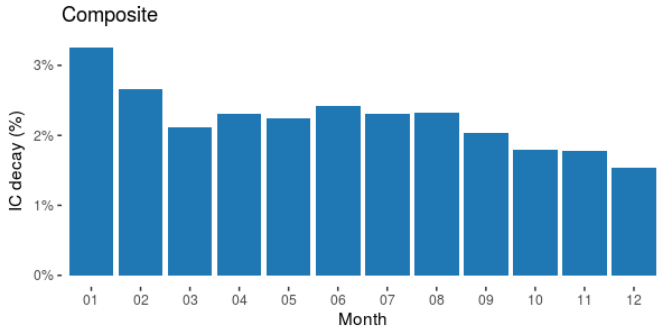
B) Sharpe ratio



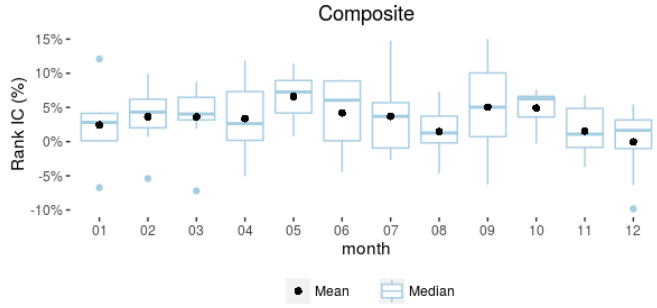
C) Rank IC



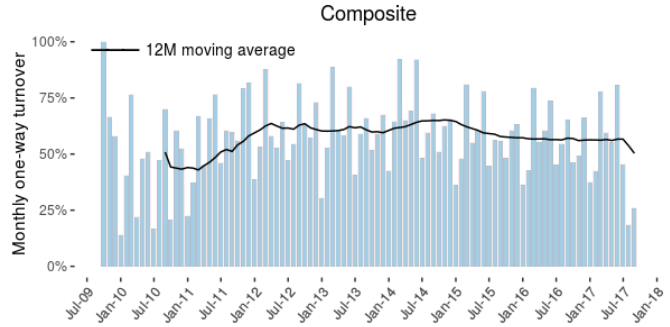
D) IC decay



E) IC seasonality



F) Portfolio turnover



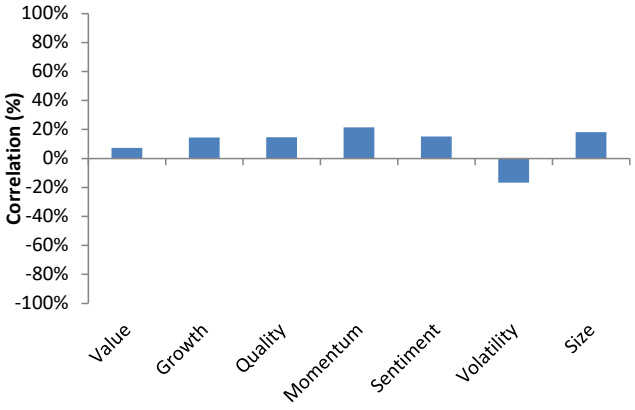
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

FACTOR EXPOSURE

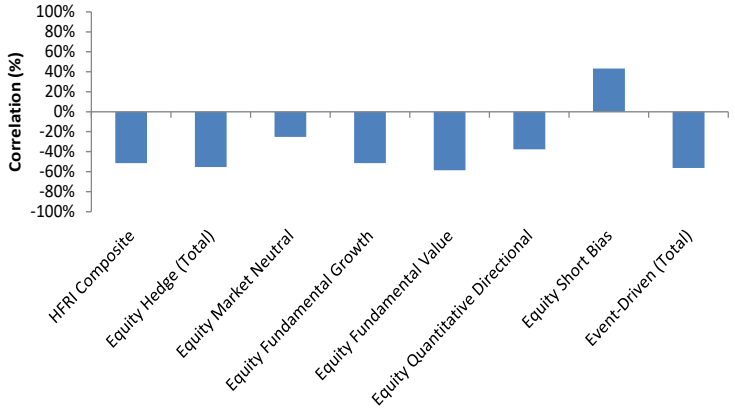
SMEC model is uncorrelated and orthogonal to most common quantitative and fundamental strategies. As shown in Figure 44, the SMEC model has marginal correlation with common factors. Moreover, our SMEC model is negatively correlated to most hedge fund styles (proxied by HFRI Indices), meaning the SMEC strategy should provide great diversification benefit to a hedge fund allocation.

Figure 44 Correlation of SMEC composite with common factors and HFRI Indices

A) Correlation with common factors



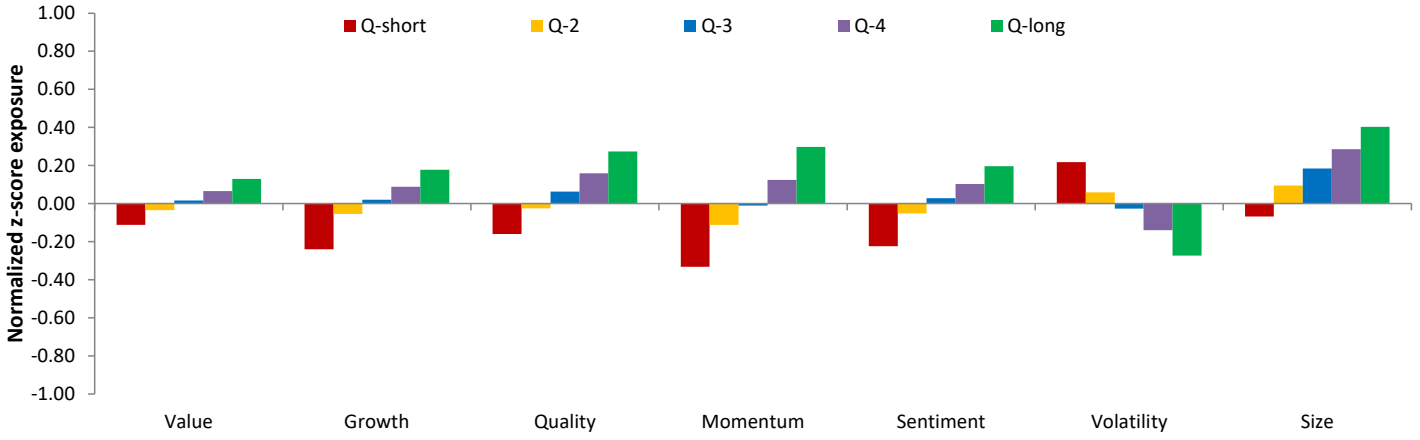
B) Correlation with HFRI Indices



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

As shown in Figure 45, the SMEC signal is positive tilted towards cheap valuation, high growth, high quality, positive price momentum, low volatility and large market cap. Overall, the exposures to common styles are relatively weak.

Figure 45 Factor exposure of SMEC composite



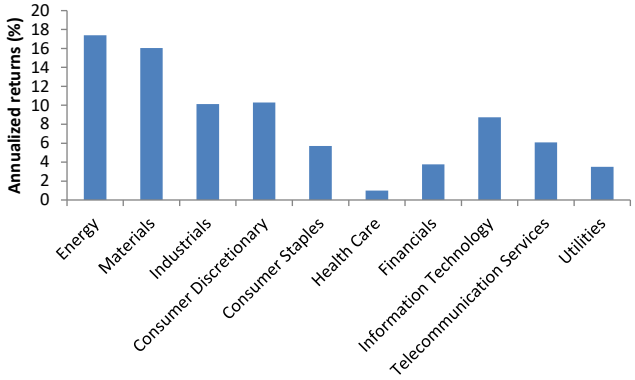
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES.

PERFORMANCE BY GICS SECTORS

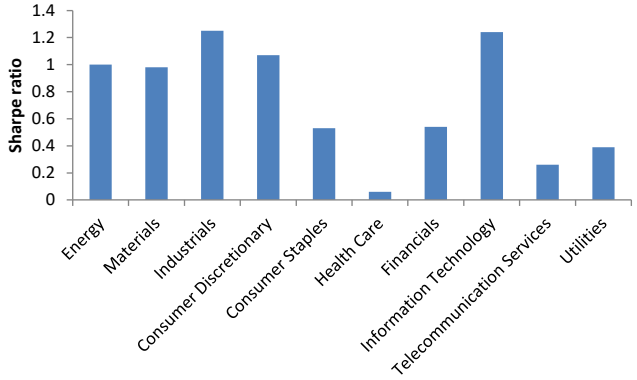
Not all sectors are created equal. As shown in Figure 46, the SMEC model performs much better in the cyclical sectors (e.g., energy, materials, industrials, consumer discretionary, info tech) than the defensive sectors (e.g., health care, telecom services, and utilities). Interestingly, the SMEC model demonstrates to have reasonable predictive power in the financial sector.

Figure 46 SMEC composite model, quintile portfolio performance, within each GICS sectors

A) Annualized returns



B) Sharpe ratio



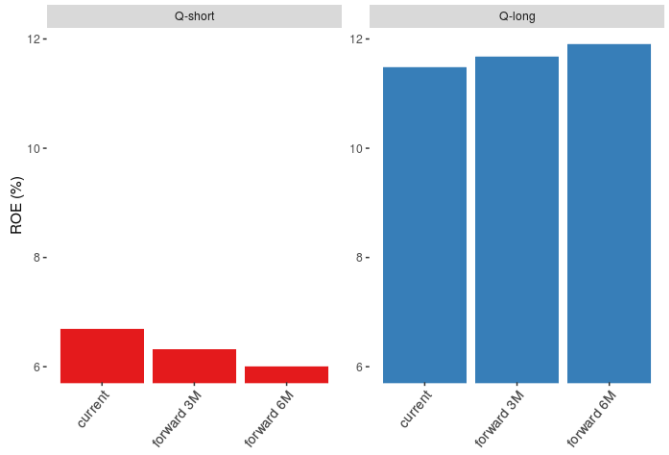
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

PROGRESSION OF KEY FUNDAMENTALS

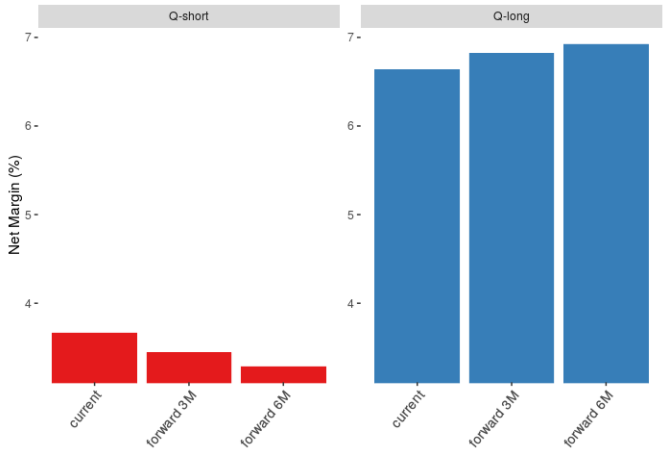
The language used in the earnings calls not only predict future stock returns, but also future financial and operational performance. Companies with positive SMEC ratings (in the top quintile) are more profitable, represented by higher ROE and net profit margin than firms in the bottom quintile (see Figure 47). More importantly, better rated companies see their profitability continue to improve in the subsequent quarters.

Figure 47 ROE/net profit margin based on SMEC composite, in the Russell 3000 universe

A) ROE current and next two quarters



B) Net profit margin current and next two quarters



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

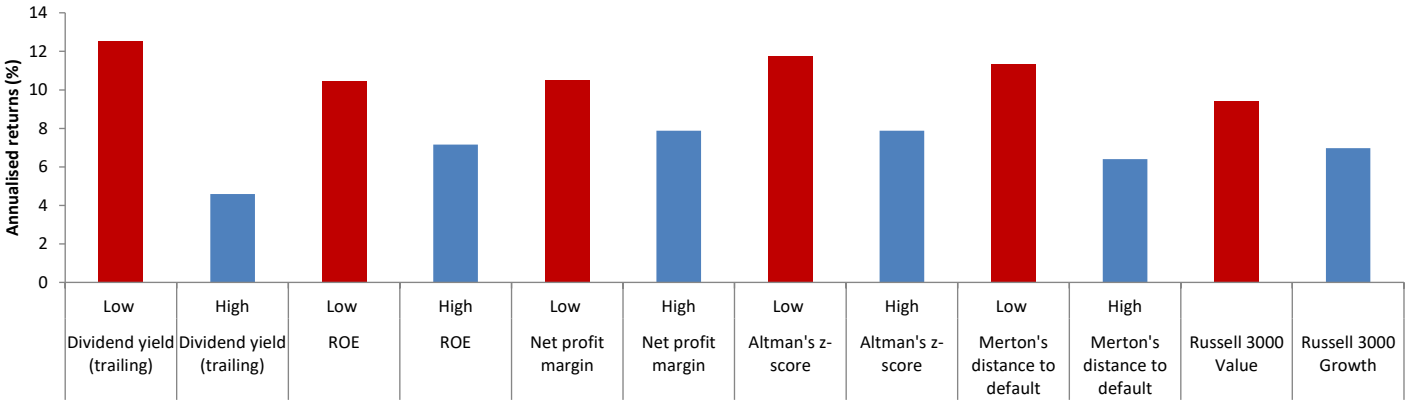
PERFORMANCE WITHIN LOW-HIGH QUALITY AND VALUE-GROWTH PORTFOLIOS

Price, et al [2011] notice that the Q&A section of call transcripts has incremental explanatory power of future stock performance, which is primarily concentrated in firms that do not pay dividends. For companies that choose to reinvest rather than distribute earnings to shareholders, investors face greater level of cash flow uncertainty. Therefore, the informational content from management presentation is likely to generate more scrutiny from investors.

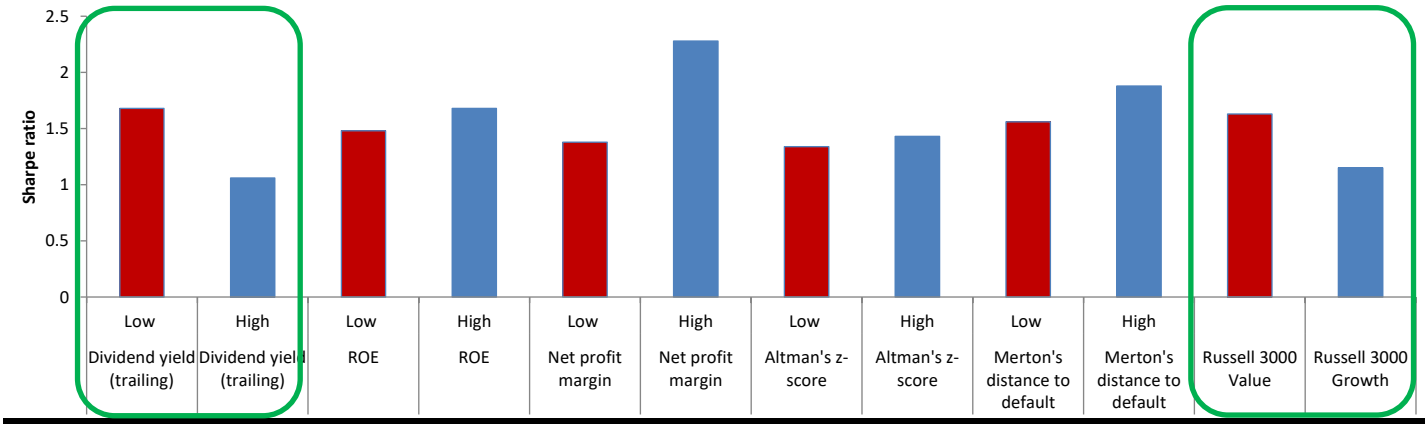
Motivated by Price, et al [2011], we backtest our SMEC model separately for the low and high quality universes. We use a variety of factors to define quality – dividend yield, ROE, net margin, Altman's z-score and Merton's distance to default. The SMEC model produces higher returns in the low quality universe, but not necessarily higher Sharpe ratios. Consistent with Price, et al [2011], we also find that performance of the SMEC model is particularly strong for companies that do not distribute or pay little dividends (see Figure 48). Sharpe ratio is 50% higher for companies with low dividend yield. We also find that the performance of the SMEC model is much stronger in the Russell 3000 Value universe relative to Russell 3000 Growth universe (see Figure 48).

Figure 48 SMEC composite L/S model performance, in low-high quality stock and value/growth universes

A) Long/short quintile portfolio annualized returns



B) Long/short quintile portfolio Sharpe ratio



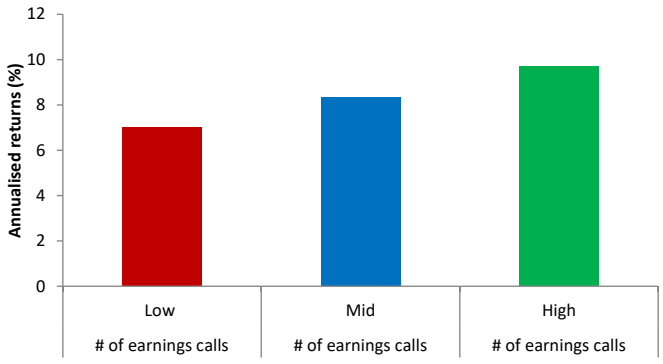
Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES.

LIMITED ATTENTION DURING HIGH NEWS DAYS

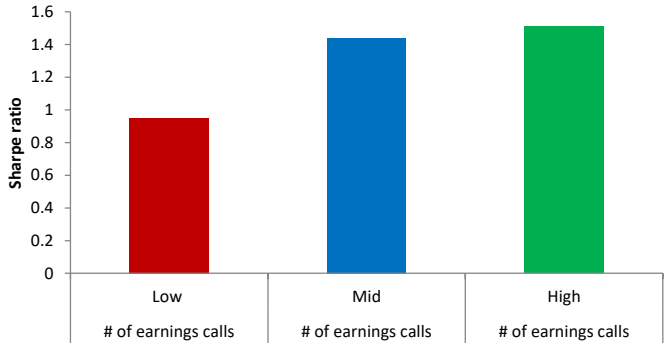
During earnings season, there could be nearly 400 conference calls taking place on the same day. When people perform multiple tasks at the same time, performance tends to suffer. Limited attention, therefore, is likely to affect how investors process information and how market reacts to news (see Hirshleifer, et al [2009]). To test this hypothesis, we divide our investment universe into two subsets, based on number of earnings calls released on a given day. It is interesting to note that performance of our call transcript based SMEC model is much stronger for earnings call released on busy/high news days than the ones occurred on low news days (see Figure 49 A and B). We also find that the SMEC model performs better among companies beating consensus EPS estimates than those meeting/missing EPS expectations (see Figure 49 C and D).

Figure 49 SMEC composite L/S performance on high/low news days

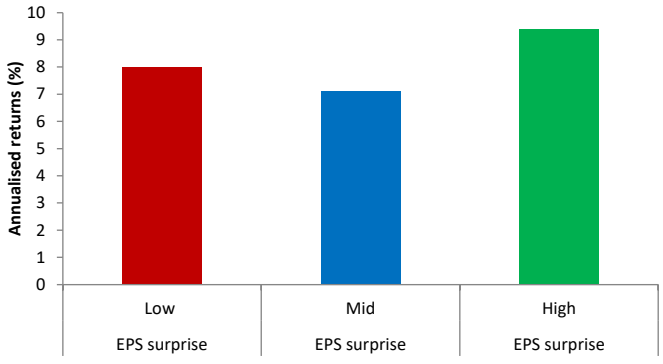
A) Annualized returns



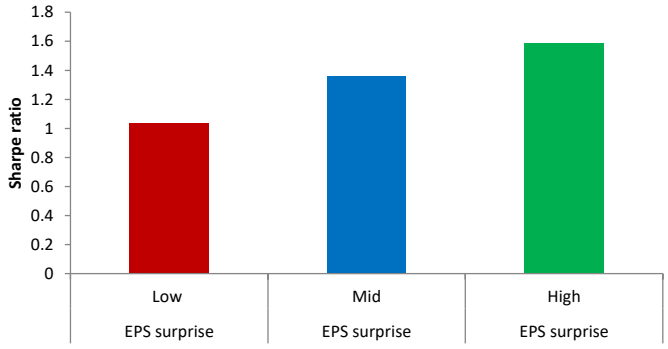
B) Sharpe ratio



C) Annualized returns



D) Sharpe ratio

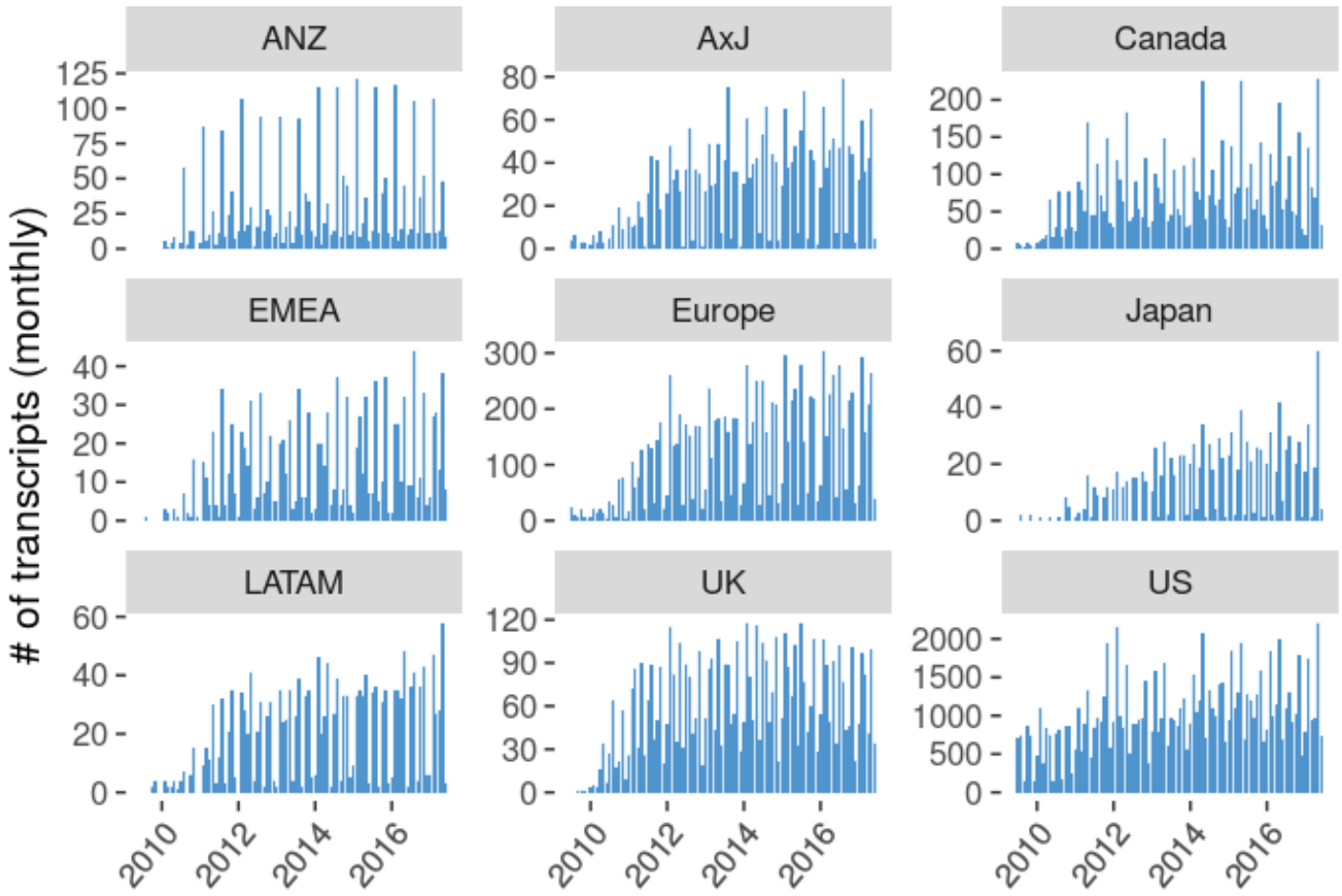


Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

SMEC GLOBAL

Outside the US, coverage is reasonable, especially for those countries where English is widely spoken, such as Canada, UK, developed Europe, and Australia (see Figure 50). Coverage has also been improving, with nearly 1,200 stocks currently in the call transcript database (see Figure 50 A).

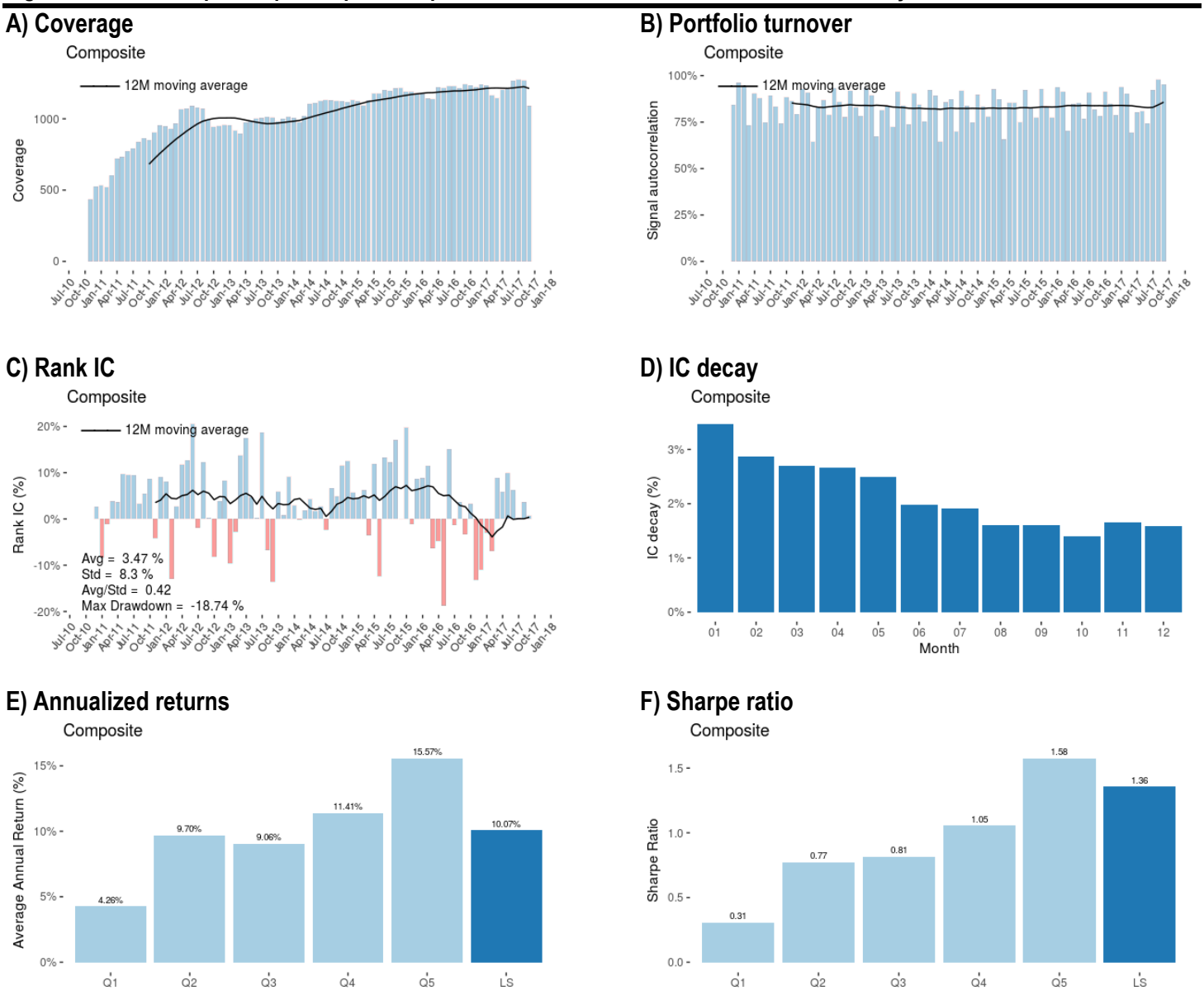
Figure 50 Monthly coverage of call transcripts data across the world



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

We combine earnings calls from all countries outside the US into a Global ex US universe. We have coverage for nearly 1200 stocks. The SMEC model performance in the international context remains strong. A long/short portfolio based on the model delivers an annual Sharpe ratio of almost 1.5x (see Figure 51). Information decay of the signal is marginally faster than in the US, possibly because the coverage outside of the US is tilted toward large cap companies that are closely monitored by investors.

Figure 51 SMEC composite quintile portfolio performance, in the S&P BMI Global ex US country neutral universe



Sources: Bloomberg Finance LLP, FTSE Russell, S&P Capital IQ, Thomson Reuters, SEC, Wolfe Research Luo's QES

BIBLIOGRAPHY

Ahmad, N. and Zinzalian, A. [2010]. "Predicting Stock Volatility from Quarterly Earnings Calls and Transcript Summaries using Text Regression", Stanford NLP

Antweiler, W. and Frank, M. [2002]. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards", *Journal of Finance*, Vol 59, Issue 3, pp. 1259-1294

Bradley, M.M. and Lang, P.J. [1999]. "Affective Norms for English Words (ANEW)". The Center for Research in Psychophysiology, University of Florida

Call, A., Sharp, N. and Shohfi, T. [2017]. "Implications of Buy-Side Analysts' Participation in Public Earnings Conference Calls", SSRN Working Paper

Chall, J.S., and Dale, E. [1995]. *Readability Revisited*, ISBN 151290087

Dale, E., and Chall, J. [1948]. "A Formula for Predicting Readability", *Educational Research Bulletin*, 27: 11-20

Das, S. and Chen, M. [2001]. "Yahoo! For Amazon: Sentiment Parsing from Small Talk on the Web", SSRN Working Paper

Goldberg, L.R. [1993]. "The Structure of Phenotypic Personality Traits", *American Psychologist*, 48, 26-34

Graham, J.R., Campbell, R.H., and Puri, M. [2013]. "Managerial Attitudes and Corporate Actions", *Journal of Financial Economics*, 109, 103-121

Gunning, R. [1952]. *The Technique of Clear Writing*, McGraw-Hill

Hirshleifer, D., Lim, S.S., and Teoh, S.H. [2009]. "Driven to Distraction: Extraneous Events and Underreaction to Earnings News", *Journal of Finance*, 60, pp. 2289-2325.

Huang, A., Lehavy, R., Zang, A. Y., and Zheng, R. [2014]. "A Thematic Analysis of Analyst Information Discovery and Information Interpretation Roles", SSRN Working Paper

Hutto, C.J. and Gilbert, E. E. [2014]. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text", Eighth International Conference on Weblogs and Social Media (ICWSM-14)

Jussa, J., Luo, Y., Wang, S., and Rohal, G. [2017a]. "Space – The Next Alpha Frontier", Wolfe Research Luo's QES, May 17, 2017

Jussa, J., Alvarez, M., Luo, Y., Wang, S., and Rohal, G. [2017b]. "Introducing D^2", Wolfe Research Luo's QES, November 2, 2017

Jussa, J., Luo, Y., Alvarez, M., Wang, S., and Rohal, G. [2017c]. "Proceeds from our First Global Quantitative and Macro Investment Conference", *Journal of Quantitative, Economics, and Strategy*, November 17, 2017

Liang, D. [2016]. "Predicting Stock Price Changes with Earnings Call Transcripts", University of North Carolina at Chapel Hill

Li, F. [2006]. "Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports?", SSRN Working Paper

Loughran, T., and McDonald, B. [2011]. "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks", SSRN Working Paper

Loughran, T., and McDonald, B. [2013]. "IPO First-Day Returns, Offer Price Revisions, Volatility, and Form S-1 Language", SSRN Working Paper

Luo, Y., Jussa, J., and Wang, S. [2016]. *Journal of Quantitative, Economics, and Strategy*, Wolfe Research Luo's QES, December 12, 2016

Luo, Y., Jussa, J., and Wang, S. [2017a]. "The Big and The Small Sides of Big Data", Wolfe Research Luo's QES, February 8, 2017

Luo, Y., Jussa, J., and Wang, S. [2017b]. "Signal Research and Multifactor Models", Wolfe Research Luo's QES, February 16, 2017

Luo, Y., Jussa, J., and Wang, S. [2017c]. "Style Rotation, Machine Learning, and The Quantum LEAP" Wolfe Research Luo's QES, February 24, 2017

Malmendier, U., and Tate, G. [2008]. "Who Makes Acquisitions? CEO Overconfidence and the Market's Reaction", *Journal of Financial Economics*, 89, 20-42

Malmendier, U., Tate, G., and Tate, G. [2005]. "CEO Overconfidence and Corporate Investment", *Journal of Finance*, 60, 2661-2700

Price, S.M., Doran, J., Peterson, D. and Bliss, B. [2011]. "Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone", SSRN Working Paper

Rohal. G., Luo, Y., Jussa, J., and Wang, S. [2017]. "Text Mining Unstructured Corporate Filing Data" Wolfe Research Luo's QES, April 20, 2017

Senter, R.J.S. [1967]. "Automated Readability Index", Wright-Patterson Air Force Base: iii. AMRL-TR-6620

Tetlock, P., Saar-Tsechansky, M. and Macskassy, S. [2007]. "More Than Words: Quantifying Language to Measure Firms' Fundamentals", SSRN Working Paper

Wang, S., Luo, Y., Jussa, J., and Rohal, G. [2017a]. "Crowdsourcing Earnings and Revenue Estimates", Wolfe Research Luo's QES, April 24, 2017

Wang, S., Luo, Y., Jussa, J., Alvarez, M., and Rohal, G. [2017b]. "Machine Learning Takeovers", Wolfe Research Luo's QES, September 12, 2017

Wang, S., Luo, Y., Alvarez, M., Jussa, J., Rohal, G. [2017c]. "The Silk Road to China", Wolfe Research Luo's QES, December 1, 2017

Zhao, F. [2017]. "Natural Language Processing – Part I: Primer", S&P Capital IQ

DISCLOSURE SECTION

Analyst Certification:
The analyst of Wolfe Research primarily responsible for this research report whose name appears first on the front page of this research report hereby certifies that (i) the recommendations and opinions expressed in this research report accurately reflect the research analysts' personal views about the subject securities or issuers and (ii) no part of the research analysts' compensation was, is or will be directly or indirectly related to the specific recommendations or views contained in this report.

Other Disclosures:
Wolfe Research, LLC does not assign ratings of Buy, Hold or Sell to the stocks it covers. Outperform, Peer Perform and Underperform are not the respective equivalents of Buy, Hold and Sell but represent relative weightings as defined above. To satisfy regulatory requirements, Outperform has been designated to correspond with Buy, Peer Perform has been designated to correspond with Hold and Underperform has been designated to correspond with Sell.

Wolfe Research Securities and Wolfe Research, LLC have adopted the use of Wolfe Research as brand names. Wolfe Research Securities, a member of FINRA (www.finra.org) is the broker-dealer affiliate of Wolfe Research, LLC and is responsible for the contents of this material. Any analysts publishing these reports are dually employed by Wolfe Research, LLC and Wolfe Research Securities.

The content of this report is to be used solely for informational purposes and should not be regarded as an offer, or a solicitation of an offer, to buy or sell a security, financial instrument or service discussed herein. Opinions in this communication constitute the current judgment of the author as of the date and time of this report and are subject to change without notice. Information herein is believed to be reliable but Wolfe Research and its affiliates, including but not limited to Wolfe Research Securities, makes no representation that it is complete or accurate. The information provided in this communication is not designed to replace a recipient's own decision-making processes for assessing a proposed transaction or investment involving a financial instrument discussed herein. Recipients are encouraged to seek financial advice from their financial advisor regarding the appropriateness of investing in a security or financial instrument referred to in this report and should understand that statements regarding the future performance of the financial instruments or the securities referenced herein may not be realized. Past performance is not indicative of future results. This report is not intended for distribution to, or use by, any person or entity in any location where such distribution or use would be contrary to applicable law, or which would subject Wolfe Research, LLC or any affiliate to any registration requirement within such location. For additional important disclosures, please see www.wolferesearch.com/disclosures.

The views expressed in Wolfe Research, LLC research reports with regards to sectors and/or specific companies may from time to time be inconsistent with the views implied by inclusion of those sectors and companies in other Wolfe Research, LLC analysts' research reports and modeling screens. Wolfe Research communicates with clients across a variety of mediums of the clients' choosing including emails, voice blasts and electronic publication to our proprietary website.

This report is limited for the sole use of clients of Wolfe Research. Authorized users have received an encryption decoder which legislates and monitors the access to Wolfe Research, LLC content. Any distribution of the content produced by Wolfe Research, LLC will violate the understanding of the terms of our relationship.