

QUANTITATIVE



Contents

Hanging on every word: Exploring word embeddings for financial sentiment analysis	1
Executive Summary	2
Introduction	5
An overview of word embeddings	5
Building a sentiment model based on word embeddings	11
Results for sentiment classification	13
Profiting from sentiment analysis	16
Conclusion	26
References	27
Appendix: Regional Results	27

Strategists

Macquarie Securities (Australia) Limited



Chanel Stuart-Findlay +61 2 8232 0811
chanel.stuart-findlay@macquarie.com



John Conomos, CFA +61 2 8232 5157
john.conomos@macquarie.com



Jeremy Lamplough +61 2 8232 1060
jeremy.lamplough@macquarie.com



Vivian Chua +61 2 8232 1731
vivian.chua@macquarie.com

Macquarie Capital Limited



Alvin Chao +852 3922 1108
alvin.chao@macquarie.com



Tracy Chow +852 3922 4285
tracy.chow@macquarie.com

This publication has been prepared by Macquarie's Quantitative Strategy team and is not a product of the Macquarie Research Department.

Quantamentals

Hanging on every word: Exploring word embeddings for financial sentiment analysis

Key points

- ▶ The Macquarie Quant team introduces a new model to measure the sentiment of earnings calls
- ▶ This model leverages the latest advances in word embeddings to improve our measurement of sentiment
- ▶ A trading strategy based on sentiment shows market-beating returns for up to 4 weeks after a call

Alternative data – today's innovation, tomorrow's requirement

With the amount of data available to investors growing exponentially it is becoming increasingly important for fund managers to have structured processes for interpreting data to maintain their edge. With a large proportion of this data being text-based, it is not surprising that Natural Language Processing (NLP) is the area where we are seeing the fastest adoption in the alternative data space.

A new way to look at global analyst conference calls

When companies report their results, in addition to the hard information announced, a wealth of soft information is provided in the form of earnings calls and broker reports that can help investors predict returns. In this report we leverage the latest advances in NLP to show how investors can unveil important information from the sentiment conveyed in conference call transcripts.

Machine learning sentiment model for earnings call transcripts

Our proposed method builds a sentiment model based on word embeddings to create lists of sentiment-charged words to improve upon existing sentiment dictionaries. Using this method, we ran a horserace between the most popular open source word embedding models, Word2Vec (Google), GloVe (Stanford) and fastText (Facebook) to determine which of these are most suited to measuring sentiment in a financial context.

Key Findings

- 1) Regardless of the measurement method, sentiment predicts returns beyond the immediate price reaction related to the earnings surprise.
- 2) The Q&A section, and more specifically the questions asked, is critical. Strategies derived purely from the sentiment of questions are more profitable than any other part of the call and take longer to decay.
- 3) There is value in using word embedding-based sentiment dictionaries with fastText from Facebook showing the most promise.

How can investors use our results?

Our findings can be used by fundamental and quant investors to better understand earnings results. Fundamental investors can isolate stocks with the most bullish or bearish tone or unexplained changes in sentiment that might require deeper analysis. It can also guide investors to find a more suitable entry point for a given stock. For quant investors, our research shows that systematic strategies can be built around the analysis of sentiment within earnings calls.

Executive Summary

Broad world of NLP

***Word embeddings
has revolutionised
the field of NLP***

The field of Natural Language Processing had its “Big Bang” moment in 2013, when Google open-sourced its word embeddings model. Since then the development of machine learning (ML) models based on them led to significant improvements in accuracy on a wide variety of language tasks. Every day by choosing typing suggestions, translating text or selecting search recommendations we depend on ML models that were built using word embeddings.

At the same time, word embeddings have been mostly absent from the academic literature related to analysis of financial disclosures, which has remained dependent on relatively simple methods that count words from hand-crafted lists. Our work attempts to bridge the gap between NLP and financial textual research and presents a simple yet powerful method that leverages word embeddings to enhance existing bag-of-words methods.

Enhancements to our previous research

***We build on our
previous research
and examine the
sentiment of earning
conference calls***

In our previous research, we made use of a dictionary-based sentiment method based on human-defined word lists. Using the latest advances in word embeddings and machine learning techniques, we train an ML model to learn the likelihood of a word having positive or negative sentiment and then use these results to create our own sentiment dictionary to improve upon and extend existing sentiment dictionaries. We then go on to show how one can profit from analysing the sentiment of earnings calls.

***We calculate the
sentiment of
earnings calls using
the latest advances
in NLP***

We have now also obtained access to a more extensive transcript database, which allows us to analyse the various components of conference calls in more detail and compare results on a global and regional basis. By evaluating the relationship between the different sections of the conference calls and subsequent returns, we can determine which parts of the calls have the biggest impact on the outcome.

Key Findings

Word embeddings improve dictionaries but existing methods already quite comprehensive

***fastText created by
Facebook is a cut
above the rest***

We perform our analysis using the word embedding models that are currently most popular - this includes Word2Vec (Google), gloVe (Stanford) and fastText (Facebook). We train a machine learning model to predict the likelihood of a word having positive or negative sentiment and create our own custom dictionaries based on these predictions. We then compare the results from using our word embedding dictionaries with the Loughran-McDonald dictionary as benchmark. When it comes to the word embedding models, we find that fastText provides the most compelling results, followed by Word2Vec and then gloVe. The Loughran-McDonald dictionary, however, sets a high benchmark and we do not find the word embedding models to outperform under all circumstances.

The importance of asking questions

***The questions from
analysts provide the
most useful data***

Firms are of course not oblivious to the increasing scrutiny of earnings calls. According to NIRI research¹, more than 15 percent of companies now prerecord formal earnings-call-comments prior to hosting a live question-and-answer session and it is not uncommon to hear about CEO's being trained on how to present their results. With calls becoming more rehearsed, we therefore need to be smarter about the way in which we analyse them.

To determine whether certain sections of the call are more informative than others, we split call transcripts into 5 sections, namely the full document, the management discussion, the question and answer session as well as the questions and answers separately.

We find that the management discussion is consistently more positive than the Q&A session and that a trading strategy focussed on analysing the sentiment of only Questions is the most profitable.

Results outside of the US market

We perform our analysis on a global developed market universe and then also examine the results for North America, Europe, Asia Pacific and Australia separately. We find that, in the US, the signal decays faster than in other regions and reverts strongly within 3 months. This reversal is most likely

¹ <https://www.niri.org/professional-development/annual-conference/2019-annual-conference/program/sessions/lunch-and-learn-corbin-advisors>

due to the quarterly reporting frequency of American companies. The profitability of the strategy is significantly higher in regions outside of the US and has a slower decay.

The relationship between Sentiment and Earnings Surprises

We also examine the relationship between our sentiment score and earnings surprises to determine whether our model adds value in addition to surprises. Our analysis shows that there is only a mildly positive relationship between our sentiment score and earnings surprises that varies over time and our backtest results indicate that a traditional post-earnings announcement drift (PEAD) strategy can be improved by incorporating information on sentiment.

How can Macquarie help investors?

Our tools allow us to process a large amount of text, measure the tone of the conversation or search for key topics.

Our sentiment screens provide investors with an objective measure of sentiment post an earnings call

Investors can make use of our sentiment signals in a variety of ways. The soft information embedded in tone can be used as a confirmatory signal to assess the quality of reported numbers. Investors can also filter for stocks whose reported numbers miss expectations but have a positive outlook. Our analysis shows that these stocks are more likely to rebound as their operational performance improves over the coming quarters. Similarly, we show that stocks which beat expectations but have a negative outlook are more likely to underperform over the coming quarters.

In Fig 1, we show an example of our daily screen that investors can use to evaluate the sentiment of a stock over time and in comparison to others.

Fig 1 Example of our earnings call sentiment screen



Source: Macquarie Quantitative Strategy, November 2019

In Fig 2, we show an example of one of our sentiment summary screens that can be used to evaluate the change in sentiment on a sector or country level.

Our sector sentiment screen allows one to see how sentiment changes on a sector basis

Fig 2 Example of sector sentiment summary screen

GICS	2017 Q1	2017 Q2	2017 Q3	2017 Q4	2018 Q1	2018 Q2	2018 Q3	2018 Q4	2019 Q1	2019 Q2	2019 Q3
Health Care	8.5%	2.7%	(4.0%)	(0.4%)	10.0%	(1.5%)	6.6%	(7.1%)	2.1%	(5.1%)	9.1%
Financials	18.7%	8.5%	3.3%	(24.2%)	29.3%	(3.3%)	2.7%	(14.6%)	(0.2%)	(0.5%)	6.2%
Communication Services	(1.3%)	(0.5%)	4.4%	(9.8%)	14.9%	(6.8%)	(1.3%)	3.5%	8.0%	(14.2%)	4.4%
Consumer Staples	(1.9%)	(5.1%)	5.3%	4.5%	8.0%	(6.1%)	0.8%	(14.6%)	19.8%	(4.8%)	(3.1%)
Information Technology	7.2%	1.4%	0.2%	0.8%	5.8%	(3.2%)	5.6%	(7.2%)	(10.9%)	2.7%	(5.7%)
Energy	15.9%	(5.2%)	2.0%	5.0%	1.9%	(4.7%)	2.2%	(8.9%)	(3.9%)	3.3%	(7.2%)
Real Estate	2.4%	0.4%	5.4%	(12.5%)	16.4%	(4.5%)	(1.6%)	(5.2%)	(3.8%)	9.6%	(7.6%)
Materials	2.6%	4.3%	(9.7%)	(4.7%)	10.2%	(7.9%)	4.1%	(12.0%)	0.1%	(3.9%)	(9.2%)
Consumer Discretionary	0.4%	11.5%	(5.2%)	1.6%	(1.4%)	4.3%	(6.2%)	(3.0%)	0.1%	0.6%	(9.2%)
Utilities	(4.4%)	(0.7%)	23.4%	(24.0%)	11.4%	4.0%	1.3%	0.9%	3.3%	6.2%	(9.5%)
Industrials	4.9%	26.5%	(10.8%)	(6.5%)	10.1%	(3.1%)	3.6%	0.2%	(3.6%)	(3.4%)	(11.0%)

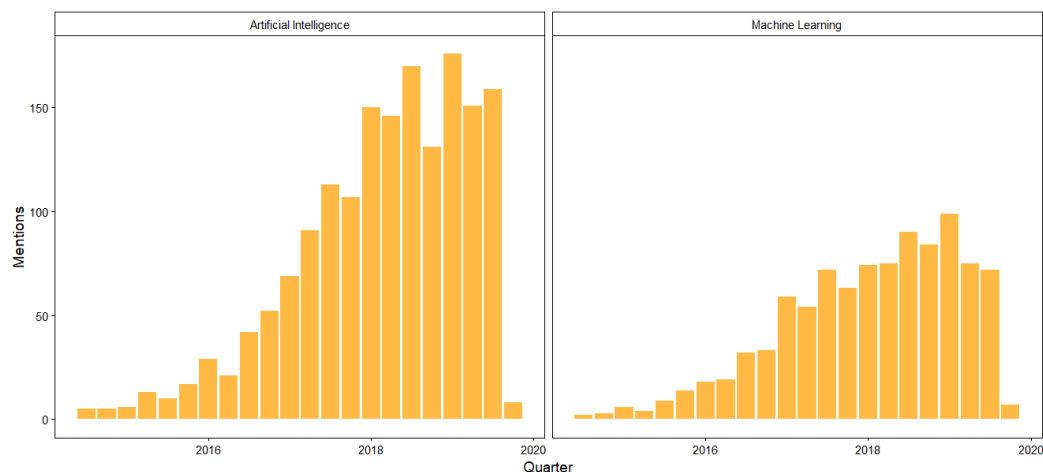
* This example shows the quarter-on-quarter change in sentiment per sector for a global universe

Source: Macquarie Quantitative Strategy, November 2019

Lastly, our text scraping capabilities allow us to efficiently scan through all transcripts to look for trends in the use of key words or phrases or find companies that used specific words during the call. As an example, Fig 3 shows how the use of the phrases 'artificial intelligence' and 'machine learning' has increased over recent years.

Our text scraping capabilities allow us to search for trends in the use of key words and phrases

Fig 3 Example of text scraping: number of companies that mentioned specific key words



Source: Macquarie Quantitative Strategy, November 2019

Introduction

*Previous Macquarie
quant papers using
textual analysis*

*There are many
different approaches
to measuring
sentiment, ranging
from simple rule-
based NLP
processes to
machine learning*

*We aim to use word
embeddings to
expand existing
sentiment
dictionaries*

*Word embeddings
has been one of the
most exciting
advances in NLP in
recent years*

The Macquarie quant team has published extensively on the use of text mining techniques in finance. We have looked at a range of different data sources and approaches, ranging from evaluating how data from RavenPack, a provider of real-time news sentiment, can add value to an investment process (Extra! Extra! Read all about it!, 2011), analysing 10-K reports for complexity (Camouflaged in Complexity, 2013), measuring the sentiment of our own analyst team's reports (How are you really feeling?, 2014), measuring quarterly conference calls for US companies (Positively Persuasive, 2013) and measuring earnings press releases of Russell 3000 companies (A surprising tone, 2014). In our most recent report on this topic (I just called to say I am bullish, 2015), we measured the sentiment of conference calls on a global universe using a Naïve Bayes approach.

There are several approaches to building sentiment models. The most widely adopted models have addressed the problem from a natural language processing (NLP) view while more recent models make use of machine learning. The most common NLP approach is to use a dictionary of positive and negative words, commonly called the 'bag of words' approach. With this approach, one only takes individual words into account and assign each a specific sentiment score which can be looked up in a sentiment dictionary. The weighted sentiment of all words is then calculated to arrive at a sentiment score for the document. Our team analysed a range of different sentiment dictionaries in the past and found that the Loughran-McDonald dictionary, which was specifically developed for financial applications, to be the most effective.

Another approach, based on machine learning, first requires someone to manually classify documents or sentences. Based on the classification, an ML model can then be developed that can learn from the human classification and make automated predictions of classification based on this. For example, in our report 'I just called to say I am bullish', our team manually classified 10,000 sentences and then used that to train a Naïve Bayes model to predict the sentiment of out-of-sample sentences.

In recent years, the field of NLP has evolved in leaps and bounds, with the most exciting advancements being around word embeddings. With this method, words are mapped to numeric vectors with similar words having similar vectors. Using deep learning models with word embeddings has proven to be very successful in sentiment prediction, especially when it comes to classifying small documents such as tweets, movie reviews or online comments.

In this analysis, we focus our efforts on analysing the sentiment of conference call transcripts.² These documents are typically lengthy, infrequent, and their release coincides with the publication of a large amount of hard data. For that reason, we believe that deep learning models, which would typically use total returns post the event to label documents and require large amounts of data to train, are more suitable for analysing sentiment based on news rather than call transcripts and filings. We therefore focus on a simpler method that combines techniques from traditional NLP while incorporating the latest advances in word embeddings. More specifically, we use word embeddings to extend traditional sentiment dictionaries, which can then be used to easily extend traditional bag-of-words approaches.

An overview of word embeddings

What are word embeddings?

There is no doubt that word embedding models have revolutionized Natural Language Processing. While the idea of numerical word representations that can capture relationships between words dates back to the 1990's (S. Deerwester, 1990), its popularity skyrocketed over recent years. Behind the recent success of word embeddings is a combination of factors – companies such as Google and Facebook making open-source libraries and pre-trained datasets available, the emergence of large open source textual datasets such as Common Crawl³, the introduction of new computationally efficient unsupervised language models that could handle these large datasets and a significant decrease in cost of computational power.

² We would also like to extend our thanks to our ex-colleague, Jakub Kolodziej, for his work on this project.

³ Common Crawl (<http://commoncrawl.org/>) is a free and open repository of 7 years of web crawl data.

**Word embedding
models map a set of
words to a vector of
numbers**

But what exactly are word embeddings? Put simply, word embedding is a collective term for models that learned to map a set of words or phrases in a vocabulary to vectors of numerical values.

Prior to word embeddings, most natural language processing (NLP) models treated each word as a single unit unrelated to other words. This is often called the “bag-of-words” approach. To translate words into numbers that a predictive model can use, one would create vectors that are the same size as the number of words in your vocabulary and represent them with a 1 where the word exists and a 0 everywhere else. This is also known as one hot encoding and is the standard process for converting categorical variables into a format that could be used as an input to ML algorithms. As an example, the one hot encoding for the words ‘Microsoft’, ‘Alphabet’ and ‘Facebook’ might look something like Fig 4.

Fig 4 Traditional representations of words that treat every word as unique and unrelated is sparse and high-dimensional (one hot encoding)

word	Microsoft	Alphabet	Facebook
...
Google	0	1	0
...
Facebook	0	0	1
...
Microsoft	1	0	0

270 000+ words

Source: Macquarie Quantitative Strategy, November 2019

With this approach, if we wanted to search our text for paragraphs that refer to technology companies, we would need to manually create a list of technology companies that we can search on and keep this list updated.

**Word embeddings
allows us to learn
about the meaning of
words**

Word embeddings, on the other hand, are based on the hypothesis that words that occur in the same contexts tend to have similar meanings. Therefore, if we can learn which words are most likely to occur together, we can learn a lot about their meaning. If we look at our words ‘Microsoft’, ‘Alphabet’ and ‘Facebook’ again, it is obvious that they would often occur in the same sentence as words such as ‘technology’, ‘company’ and ‘computer’. It is also likely that the word ‘Zuckerberg’ occurs more often in the same sentence as the word ‘Facebook’ than it would with the word ‘Microsoft’. By building a model based on how often words co-occur with others one would be able to conclude that all three words are related to technology companies and that there is a relationship between ‘Facebook’ and ‘Zuckerberg’. In this case, words are no longer represented as dummy variables, but rather as a numeric vector where each point in the vector captures a dimension of the word’s meaning. An example of what this could look like is shown in Fig 5.

Fig 5 Representations of words obtained from the word embedding models are denser and low-dimensional

Feature	Microsoft	Alphabet	Facebook
technology	0.90	0.88	0.92
company	0.98	0.97	0.99
social media	0.30	0.70	0.99
computer	0.70	0.55	0.35
Zuckerberg	0.40	0.48	0.96
Gates	0.96	0.60	0.40
Page	0.40	0.90	0.25

50 – 300 features

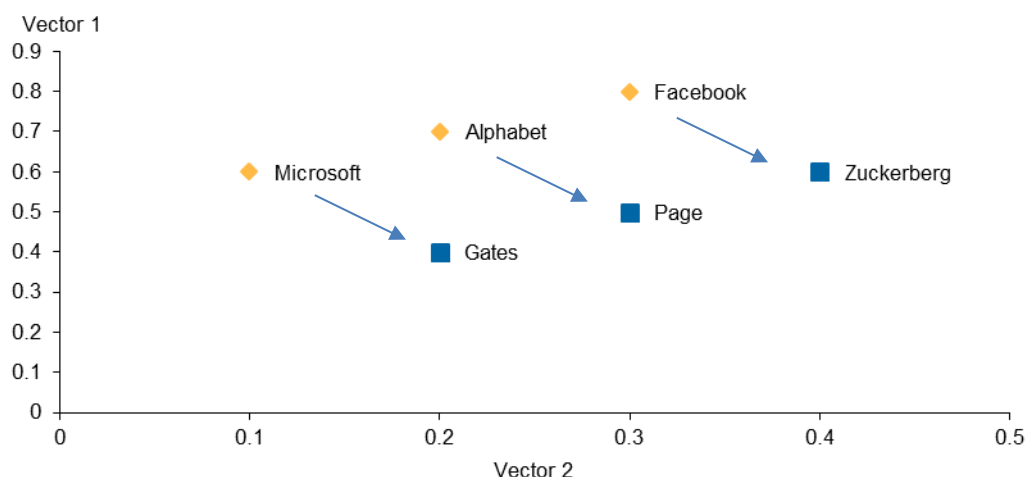
Source: Macquarie Quantitative Strategy, November 2019

What this means is that, if we search for words with a high correlation to both ‘technology’ and ‘company’, the words ‘Microsoft’, ‘Alphabet’ and ‘Facebook’ are highly likely to occur in this list.

An interesting benefit of this representation is that one can also perform certain mathematical operations on these word vectors. As a theoretical example, one would be able to learn that 'Zuckerberg' and 'Facebook' are related to each other in a similar way as 'Gates' is to 'Microsoft' as represented by the word vectors in Fig 6.

Fig 6 Example of words as vectors in geometric space

The relationships between words can be examined by performing mathematical operations between word vectors



Source: Macquarie Quantitative Strategy, November 2019

Word embeddings solve the curse of dimensionality

If we had to store the similarity of all words with all others, the data requirements would be extreme. For example, the Oxford English Dictionary has around 270,000 main headwords⁴, which means that we would require a mind-boggling number of parameters – over 73 billion – to represent the relationship between all these words. For that reason, an important feature of word embedding models is that they also reduce the dimensionality of the problem at hand.

Word embeddings can be equated with the factor loadings of a statistical risk model

Equity investors face the same problem when calculating covariance matrices and can relate the idea of word embeddings to the statistical risk models frequently used in finance to overcome the dimensionality problem. Such models are commonly based on principal component decomposition of the covariance matrix and consist of two components: a set of statistical factors and a matrix of factor scores for each asset.

Word embeddings are equivalent to the second element - vectors of factor exposures in the statistical risk model. In the same way as each stock gets represented with a set of factor exposures, word embedding models map words into real-valued multi-dimensional feature vectors. As stocks with similar risk factor exposures tend to have highly correlated returns, words that have feature vectors close to each other in the vector space such as 'good' and 'excellent', 'earnings' and 'net income' or 'profit' and 'revenue' tend to have similar meaning, despite not sounding anything like each other.

Representing words as numbers allows us to measure the relationship between them

Word embedding models can scan a large body of text and try to represent words in a limited number of dimensions so that words which frequently occur in each other's context have vectors that are close to each other in the vector space. As each word is summarised with a vector of up to 300 dimensions, we can now model the relationship between sentiment and word vector rather than words themselves. This change lowers the dimensionality of inputs in our estimation problem from 270,000 to 300! As we can expect that similar words e.g. 'profit' and 'revenue' have similar vector representations, we can expect that if our model learns something about sentiment from observing features of word 'profit' in the training set, it should be able to transfer this knowledge to the word 'revenue', even if 'revenue' does not occur in the sample used for fitting the model.

What are the benefits of word embeddings?

In addition to significantly reducing the dimensionality of our data, word embeddings offer several other advantages beyond the traditional approach:

⁴ <https://public.oed.com/how-to-use-the-oed/glossary/>

- **It reduces the reliance on incomprehensive hand-picked word lists:** The number of sentiment annotated examples in textual datasets is usually relatively small compared to the size of the vocabulary (reading and annotating text takes time and money). For example, the Loughran-McDonald sentiment dictionary contains only about 2700 words, while the English language consist of hundreds of thousands. These word lists are also subjective and since they are updated from time-to-time, it makes out-of-sample performance hard to measure.
- **Models can be trained on different bodies of text to account for domain specific connotations:** Sentiment is domain specific, for example words such as 'cost' or 'expense' are classified as having negative sentiment by the general sentiment Harvard Dictionary, but Loughran and McDonald (2011) show that many of these words are typically not considered negative in financial contexts.
- **Its use can be extended to topic modelling and classification:** Word embeddings are based on an unsupervised data-driven process that captures average relationships between words existing in the body of text. This means that word embeddings can be used for other tasks such as topic modelling or sentiment classification for which only a limited number of annotated examples are available. In this way, our model implicitly 'borrows' predictive power from the average statistical relationships captured by word vectors and should allow us to generalize well to the out-of-sample text.
- **It can easily be extended to different languages:** Currently we can obtain pre-trained word vectors for 157 languages. With pre-trained word vectors we are able to build textual models without being proficient in a given language. For instance, if we have a dataset of corporate news in Korean and want to model how text affect stock performance immediately after the publication, we can do it without knowing a single word of Korean and be relatively confident that our model will be able to perform well out-of-sample.

How are word embedding models trained?

Now that we have a high-level understanding of what word embeddings are and why they are so useful, we look at how they are created.

There are multiple open source libraries available to train word embeddings

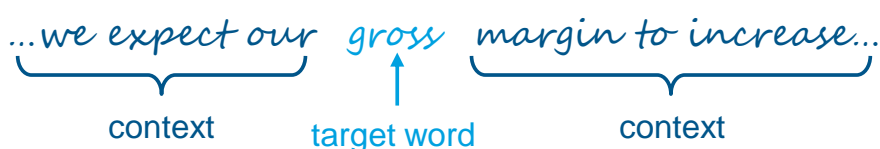
The good news is that, when using word embeddings for text classification, one has several algorithms and accompanying open source libraries to choose from. Currently, the most popular embedding models are Word2Vec, GloVe and FastText. We provide a high-level overview of each of these methods and one of the aims of this report is to determine which of these methods are the most suited to measuring sentiment in finance. Readers not interested in the technical details can skip to the results on page 16.

Word2Vec

Words that occur in the same contexts tend to have similar meaning

Word embeddings were first made popular by Google when they introduced their patented approach named Word2Vec (Tomas Mikolov, 2013). This algorithm, as well as most other word embedding algorithms, are based on the Distributional Hypothesis. This hypothesis states that words that occur in the same contexts tend to have similar meanings. What this means is that, if we can learn which words are most likely to occur together, we can learn a lot about their meaning.

Fig 7 Example of the target and context words in a sentence



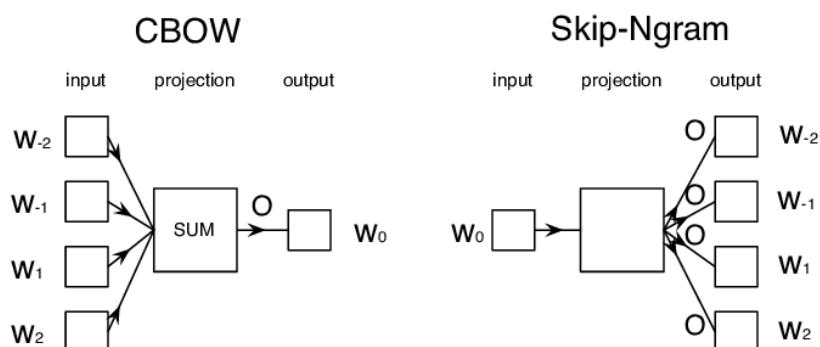
Word2Vec uses neural networks to predict which word is most likely given its context

Source: Macquarie Quantitative Strategy, November 2019

In Fig 7 we show an example to highlight the difference between a target word and context words. The word context is defined as words that occur in the short window before and after a given target word. A word embedding model tries to obtain a numeric vector representation for target word w_i and context word w_j so that some given similarity measure $f(w_i, w_j)$ gives large values for words that frequently occur in each other's context and close to 0 for words that do not co-occur often. If, for example, the word 'Facebook' regularly occurs in the same sentence as the word 'social media', the model will learn that there is a strong relationship between them.

Word2Vec provides two types of model architectures to create word embeddings, CBOW and Skip-gram

Fig 8 Training algorithms: CBOW vs Skip-gram



Source: Macquarie Quantitative Strategy, November 2019, www.researchgate.net

The Word2Vec method is based on a predictive approach whereby neural networks are used to calculate word embeddings based on the context of a word. There are two approaches that can be used to implement this idea. Firstly, there is the continuous bag of words (CBOW) approach. In this approach, the network tries to predict which word is most likely given its context. Words that are equally likely to appear can be interpreted as having a shared dimension. If we can replace 'profit' with 'revenue' in a sentence, this approach predicts a similar probability for both, even though they might not ever appear within the same sentence. Therefore, we infer that the meaning of these words is similar on at least one level.

The second approach is skip-gram. The idea is very similar, but the network works the other way around. Rather than trying to predict the target word from the context words, it aims to predict the context words from the target word. The difference between these two approaches are visualised in Fig 8.

Glove

One year after Google released Word2Vec, Stanford published GloVe (Jeffrey Pennington, 2014). Word2Vec learns embeddings by relating target words to their context. However, it ignores whether some context words appear more often than others. For Word2Vec, a frequent co-occurrence of words creates more training examples, but it carries no additional information.

In contrast, GloVe stresses that the frequency of co-occurrences is vital information and should not be "wasted" as additional training examples. Instead, GloVe builds word embeddings in a way that a combination of word vectors relates directly to the probability of these words' co-occurrence in the text.

This is done by evaluating an equation that calculates the probability of a target word appearing in the context of another word or words.

Once all these probabilities are calculated and stored in a matrix, a technique from linear algebra called singular value decomposition (SVD) is performed to reduce the dimensions of the matrix.

The most important difference between GloVe and Word2Vec is that GloVe is not a predictive model in the same sense that Word2Vec is. Instead, its embeddings can be interpreted as a statistical summary of the text with low dimensionality that reflects co-occurrences.

fastText

FastText is a word embedding library created by Facebook's AI Research lab. Facebook makes available pretrained models for 157 languages.

fastText is an extension of Word2Vec, but while Word2Vec treats each word as an individual unit, fastText treats each word as a combination of characters. For example, the word 'profitable' will be the sum of the vectors or n-grams of 'profitable', 'profi', 'profit' etc.

There are three main advantages to the fastText approach. First, generalization is possible if new words have the same characters as known ones. For example, if the words 'profit' and 'profitability' were in our training set, our model will be able to understand the word 'profitable' the first time that it occurs in the text, despite never seeing it before. Second, less training data is needed since much more information can be extracted from each piece of text. Lastly, the approach is computationally significantly more efficient.

GloVe is based on the simple intuition that the ratio at which words occur close to each other allows one to learn about their meaning

fastText breaks a word down into its components, which extends well to unseen data

Academic evidence suggests fastText's bag-of-ngrams approach provides comparable performance at a much faster speed than models that take word order into account

We examine multiple sets of pre-trained word embeddings for their suitability to financial applications

While this approach is technically still a 'bag-of-words' approach, the creators found that their approach provides comparable performance to approaches that explicitly take word order into account, while being substantially more computationally efficient (Arman Joulin, 2016). They therefore describe it as a 'bag of n-grams' approach. Unlike word vectors from Word2Vec, fastText word features can also be average together to form good sentence representations.




Pre-trained vs Self-trained Word Embeddings

Since all the above libraries are available as open source tools, one has the option of using them to train your own word embeddings on your own proprietary text library. While there are clear benefits to training word embeddings on the most relevant texts, the main problem with this is the size of the required datasets to effectively train word embedding models. Further good news is that, along with making algorithms available via open source libraries, multiple sets of pre-trained word embeddings have also been released into open source and is simple to download⁵.

These embeddings mostly differ on the model used to train them, the body of text on which they were trained as well as the parameters used, such as the final dimensions and number of words. We use the three different types of word vectors described above in order to assess how different methods of training word embeddings affect sentiment predictions and if some embeddings are more suitable for financial applications than others.

For comparative purposes, we also trained our own word embeddings using the full set of call transcripts in the FactSet database using fastText. This provided us with 3 billion words of text existing of 60 0 000 unique words. We selected 300 dimensions and a window size of 5 to train our model.

Fig 9 Word embeddings summary

Creator	Word embedding	Creators	Date	Training model	Corpora	Corpora size	Dimension	Nr words	Label	Type
	word2vec	Mikolov et al.	2013	skip-gram	Google News	100 billion words	300	3 million words and phrases	Word2vec	Pre-trained
	Glove	Pennington et al	2014	GloVe	English Wikipedia and Gigaword	6 billion tokens	50, 100, 200 and 300	400 000 words	Glove	Pre-trained
					Common Crawl	42 billion tokens	300	1.9 million words	Glove Crawl	Pre-trained
	fastText	Mikolov et al.	2018	CBOW	Wikipedia from 2017 and news datasets from UMBC webbase, statmt.org and Gigaword	16 billion tokens	300	1 million words	FastText Wiki+News	Pre-trained
					Common Crawl	600 billions tokens	300	1.9 million words	FastText Crawl	Pre-trained
					All Factset Call Transcripts from 2007 to 2019	3 billion tokens	300	600 000 words	MQR FastText	Self-trained

Source: Macquarie Quantitative Strategy, November 2019

The implication of the above is that, while the idea behind word embeddings might sound complicated, by leveraging the open source libraries and data that are readily available, one can incorporate them into a process with a fraction of the original effort.

⁵ Word vectors can be downloaded from the following sites:

Glove: <https://nlp.stanford.edu/projects/glove/>, **Word2Vec:** <http://vectors.nlpl.eu/repository/>
fastText: <https://fasttext.cc/docs/en/english-vectors.html>

Building a sentiment model based on word embeddings

Now that we understand word embeddings, the next step is to look at how we can use them to measure the sentiment of financial text. In this section we describe the methods we employed in order to measure sentiment from each conference call transcript. We start with the bag-of-words approach that is frequently used in the financial literature. Then we present our proposed method of extending the sentiment dictionary using a linear classifier based on word embeddings.

Loughran-McDonald Bag-of-words approach

To calculate a bag-of-words sentiment score we follow a similar methodology to the one presented in [Quantamentals - I just called to say I'm bullish](#). This methodology requires two word lists: one for positive and one for negative terms. Words that are not included in either of these lists are ignored. We compute the sentiment score using a simple formula:

$$\text{Sentiment} = \frac{\text{Count of Positive Words} - \text{Count of Negative Words}}{\text{Count of All Words}}$$

In our previous report we examined a range of word lists and, consistent with Loughran & McDonald's (2011) argument that sentiment in financial domain differs from the sentiment in other contexts, our research shows that the Loughran-McDonald (**LM**) dictionary outperforms other general sentiment word lists such as Diction and LIWC in forecasting stock returns. Therefore, in this report we only consider the LM dictionary for our baseline model. In our previous research, we have included various advanced techniques to improve on this basic methodology, but for the purposes of this report, we are predominantly interested in understanding the benefits of extending the dictionary using word embeddings and we therefore keep these extensions for future research.

Extending the Loughran-McDonald dictionary using word embeddings

While word embedding models are powerful, one of the problems that they face when it comes to sentiment analysis is that antonyms often tend to occur in similar contexts. For example, the words 'upwards' and 'downwards' would tend to co-occur in sentences such as 'The next move in earnings are expected to be *upwards/downwards*'. This would mean that antonyms might have vectors that are very similar. To use word embeddings for sentiment analysis, we therefore need to train a model to determine whether a word is associated with positive or negative sentiment.

There are various approaches that one could follow to do this. The idea behind our model is to train a classifier on word vectors using a relatively short list of positive and negative examples, so that later the model can be used to assess sentiment of all words in the vocabulary. The implicit underlying assumption, which we will confirm in the first section of our Results, is that our model can learn from these examples about the concept of sentiment in a financial context.

In the sections below, we first discuss the data pre-processing steps that we apply to our text so that we can use it as an input into our model. Next we describe our sentiment model specification and model training. Finally, we show how we select the optimal classifier threshold for predicting stock returns.

Step 1: Data pre-processing

Tokenization

Tokenization refers to the process of breaking up a piece of text into its underlying pieces. This mostly involves separating words using spaces and punctuation, but also makes provisions for the handling of special characters, hyphenated words and plurals.

One of the important issues that we discovered when working with word embeddings is that different embedding models are implemented using different tokenizers and text pre-processing methods.

To understand the difference between text processing methods, we use an example of a short sentence:

"We'll achieve a growth rate of 5% year-over-year."

This sentence is split into 11 tokens in GloVe model:

("we", "ll", "achieve", "a", "growth", "rate", "of", "5", "%", "year-over-year", ".")

While in Word2Vec this sentence is represented with 10 different tokens:

Two lists of words containing positive and negative terms are required

The Loughran-McDonald dictionary is best suited to finance

We need to train our model to differentiate between words with similar embeddings but different sentiment

Word embeddings allow us to extend the Loughran-McDonald dictionary across the entire vocabulary

It is important to match tokenization with that of the word embedding model

(“We’ll”, “achieve”, “a”, “growth”, “rate”, “of”, “5”, “year”, “over”, “year”).

As authors of pre-trained word vectors do not provide much detail about how they process text loaded into their models, we spent a considerable amount of time to match as closely as possible our text processing pipeline to each word embedding set.

Normalisation of word vectors

Word vectors are normalised by subtracting the average word vectors from each vector average word

The unsupervised methods used for training word embeddings do not impose any explicit limit on the vector lengths. We find that norms of pre-trained word vectors vary significantly across words. This is an undesirable property for our sentiment modelling task, because the linear activation in our model might have very different magnitudes for different words. At the same time, we do not want to modify word vectors in a way that alters the linear relationships encoded in the embeddings vector space.

Therefore, we rescale each vector to the unit norm, but we avoid normalizing feature dimensions, which would inadvertently affect relations between words. This should not have a significantly detrimental impact on effectiveness of regularization in the model, because the variation in word vector norms is considerably larger than variation in feature dimension.

To normalize word vectors, first we calculate an average word vector and subtract it from each vector. Next we scale each vector by dividing it by its norm. We compute the average word vector as a word frequency weighted mean and this way it should represent the most common feature corresponding to stop words. This is in some sense similar idea to deducting the projection of the embedding on first eigenvector proposed by Arora et al. (2017). We find that frequency weighting improves results slightly on the sentiment prediction task relative to the equal weighting.

Step 2: Word sentiment model

Three word lists containing positive, negative and neutral terms are used to train the model

Once we obtain the normalized word vectors, we match all 2,709 words from the Loughran-McDonald (LM) dictionary plus an additional list of 252 neutral words created by us. We select this list by filtering most frequent words in the embeddings vocabularies that do not seem to have any polarity – mostly stop words (e.g. and, the, a etc), numerals and names of weekdays and months. We add this neutral word list, as we find in our experiments this improves the ability of the model to generalize to the entire vocabulary. An alternative option would be to remove stop words from the vocabulary.

A multinomial logistic regression model is used to discriminate between classes

We then train a multinomial logistic regression model on this classification to predict the likelihood of whether a word is associate with positive, negative or neutral sentiment.

Step 3: Classifier threshold selection

Thresholds are based on the correlation between sentiment scores and subsequent total returns

Once the sentiment models are trained, we need to select the thresholds above which we include words in the positive and negative lists of our enhanced sentiment dictionary. We use a data driven process to select these threshold levels. Each year we calculate rank correlation between sentiment scores calculated with different threshold levels for probability of positive and negative sentiment and 21 day returns after the event. We select the optimal threshold based on the highest rank correlation using an expanding window. We then apply this threshold in the backtest for one year forward.

As the levels of sentiment scores are changing with these thresholds over time, we recalculate the levels for inclusion to the top and bottom sentiment basket annually based on an expanding window.

Results for sentiment classification

Before we look at whether sentiment is predictive of future returns, we determine how successful our model is at using word embeddings to extend a sentiment diary.

We use pre-trained word embeddings to extend the LM sentiment dictionary

To do this, we use a subset of the sentiment scores provided by the LM dictionary to train our model. To test the predictive power of our model, we predict the sentiment of the words that we withheld from the sample and compare the predicted sentiment with that provided by Loughran-McDonald.

The first step in this process is to extract all the words contained in the Loughran-McDonald dictionary with its sentiment score and then combine this with the word embedding vector for each word. Fig 10 provides an example of what this dataset could look like. On the left, we have the LM dictionary containing a list of words with its sentiment score, and on the right, we have the word embedding as a numeric vector for each word. Our aim is to predict a sentiment score for words not contained in the LM dictionary.

Fig 10 Example dataset for predicting sentiment using word embeddings

Sentiment dictionary		Word embedding										
word	sentiment	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
...
proactive	1	0.07	(0.30)	(0.11)	(0.18)	(0.07)	(0.13)	0.26	(0.10)	0.04	0.14	0.01
problem	-1	0.15	(0.04)	0.04	0.01	(0.12)	0.20	0.18	(0.04)	0.15	(0.04)	0.06
problematic	-1	0.12	(0.14)	(0.04)	(0.00)	(0.09)	0.17	0.16	0.01	0.01	0.15	0.01
proficiency	1	(0.30)	(0.03)	(0.01)	(0.20)	(0.05)	(0.06)	0.17	(0.04)	0.14	0.12	0.23
profitable	1	0.13	(0.27)	0.08	0.08	(0.07)	(0.07)	(0.14)	(0.01)	0.16	0.21	0.07
progress	1	0.06	0.08	0.05	(0.06)	(0.11)	(0.11)	0.05	0.07	0.22	0.01	(0.02)
progressing	1	0.10	(0.20)	(0.06)	(0.08)	(0.15)	(0.12)	0.13	0.16	0.12	0.07	0.10
...

Source: Macquarie Quantitative Strategy, November 2019

As a sensibility check, we start by visually inspecting the top 10 most positive and negative words from the predictions of each word embedding model. The results are shown in Fig 11. It is good to note that the results all seem very sensible. It is also interesting to see that there is little overlap between the models.

Fig 11 Top 10 most positive and negative words for each word embedding model

Word2Vec		GloVe		fastText	
Positive	Negative	Positive	Negative	Positive	Negative
artistic	avoided	affordable	consequences	bright	blamed
dynamic	barring	cutting_edge	irresponsible	comfortable	blaming
expertise	blaming	elegant	leaks	consistent	fatal
keen	causing	fabulous	lied	flexible	imprisoned
performer	dumping	finest	prohibited	oriented	irrational
showcase	fearing	robust	purish	pioneer	jailed
skill	seizure	superb	scare	proud	risk
talent	stemming	terrific	troubling	remarkable	scared
talented	triggered	thrilled	unconstitutional	superb	trapped
talents	triggering	unique	withdrawn	talented	unconstructive

Source: Macquarie Quantitative Strategy, November 2019

Another interesting check is to look at word similarity. In the below table we show the words that are deemed to be most like 'profit' in each word embedding model. This clearly illustrates the difference between fastText and the other models. While Word2Vec and gloVe can only compare complete words, fastText breaks words up into smaller components, and is therefore much better at recognizing that the word 'profit' and its derivations are similar.

Fig 12 Words most like 'profit' according to each embedding model

word2vec	glove	fastText
profits	profits	profits
EBITDAO	revenue	profitable
npat	gains	prifiting
CAG.N	earnings	business
revenue	profitable	profitability
extraordinaries	income	revenue

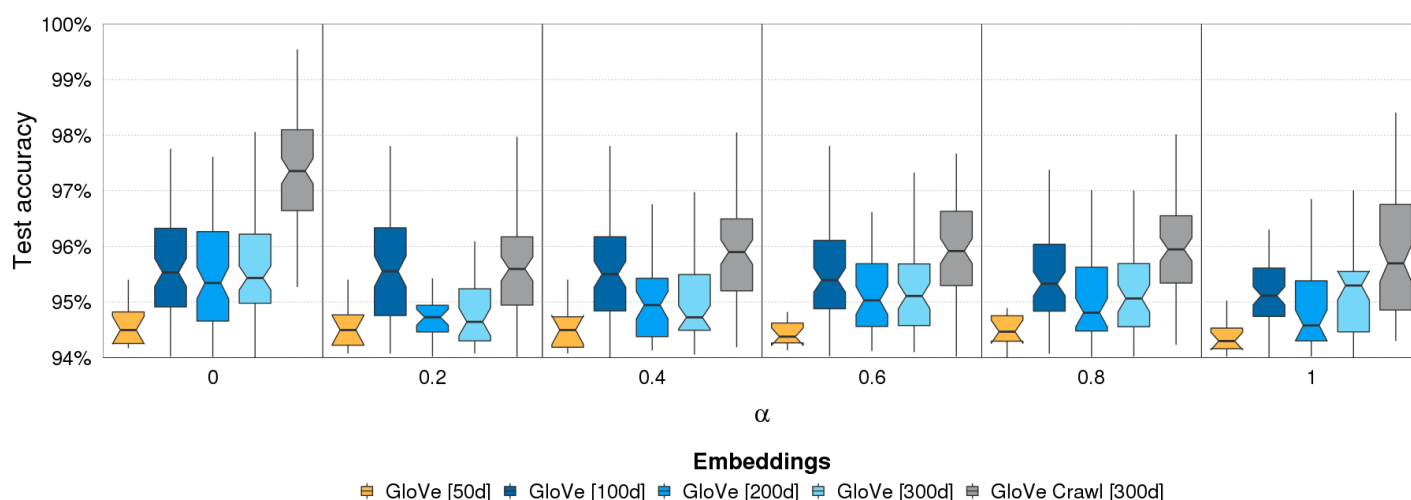
Source: Macquarie Quantitative Strategy, November 2019

All our word embeddings have very high accuracy in predicting sentiment

To assess and compare the performance of sentiment models based on different pre-trained word embeddings in a more objective and comprehensive manner, we perform a test where we randomly shuffle the words in our dictionary 100 times and split it into training, validation and test sets in 60:20:20 proportions. We then fit our model (a multinomial elastic net logistic regression) to predict the sentiment of each word based on its word embedding vector. Since the LM dictionary has substantially more negative than positive words, we use a weighted regression to ensure that the model doesn't overfit to negative words.

We report the test set accuracy across all our models in Fig 13 and Fig 14. We report results for different values of the hyperparameter α , which is used to tune an elastic net model⁶. Since GloVe is the only model where we have embeddings with different dimensions, we start our analysis by comparing all the GloVe models in Fig 13 and then compare GloVe with Word2Vec and fastText in Fig 14.

Fig 13 Test set accuracy comparison for the sentiment models based on GloVe word embeddings



Source: Macquarie Quantitative Strategy, November 2019

GloVe models with higher dimensions do better than models with lower dimensions

In Fig 13 we report test set accuracy across all GloVe models. Box plots in this chart show the distribution of accuracy from 100 resamples and includes the notch, which displays the 95% confidence interval for the median. All models achieve high out-of-sample accuracy that exceeds 94%. This suggests that our multinomial model can capture the notion of sentiment and generalize it to out-of-sample words. Based on these results we can make a few additional observations:

- Accuracy of our models seems to be relatively **insensitive to value of the hyperparameter α** indicating that the model is robust to the specification of the regularization distribution (L1 vs L2).

⁶ An elastic net regression includes both L1 (used in Lasso models) and L2 (used in Ridge models) penalties. The hyperparameter α determines the importance of the L1 and L2 penalty, where $\alpha = 0$ corresponds to Ridge and $\alpha = 1$ to Lasso.

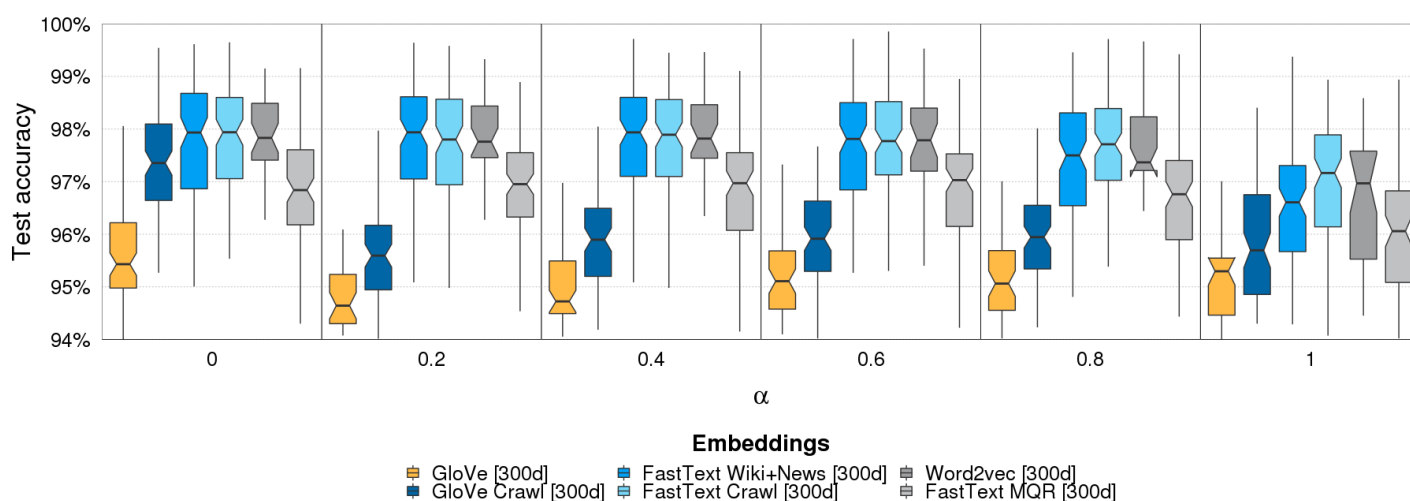
A larger corpus used in training results in better model performance

- Higher dimensionality of word embeddings leads to higher variance in forecasting accuracy; however, this increase in variance seems to be more than offset by the improvement in the median accuracy. All **higher dimensional vectors have statistically higher median accuracy** compared to 50-dimensional word vectors at 5% significance level.
- The word embeddings trained over the **larger corpus** (GloVe Crawl) seem to have **superior performance** over the word vectors trained on a smaller corpus with the same dimensionality (GloVe [300d]). The improvement in accuracy is around 1%.

In Fig 14 we compare test set accuracy for models based on different pre-trained embeddings – GloVe, Word2Vec and fastText. To make the comparison fair we only include 300-dimensional word vectors. To measure the importance of the size of the training corpus, we report results for the same models on all available corpora. Based on these results we can conclude that:

- Accuracy of the models is relatively **insensitive to value of the hyperparameter α** . However, it seems to be slightly lower for higher values of α suggesting that overweighing lasso regularization and promoting sparsity of parameters hurts performance.
- Models trained with **CBOW (fastText) and skip-gram (Word2Vec) models seem to outperform GloVe** models by around 2%. While corpus size is different for each set of embeddings, even GloVe word vectors trained on a very large body of text (Common Crawl) underperform other methods, suggesting that this difference in accuracy is driven by the model rather than the text used to train the model.
- As the number of parameters is the same for each model, **the variance in the forecasting accuracy seems to be comparable** across all models.
- The fastText model trained on Common Crawl and Wikipedia outperform the model trained on the FactSet transcripts by a considerable margin. We believe that this is due to two reasons. Firstly, the **size of the corpus**. While we trained our model on a respectable 3 billion tokens, this pales into comparison to the 16 billion used by Wikipedia and News dataset and the 600 billion used by Common Crawl. Secondly, we suspect that the fact that our training data consists of **spoken as opposed to written text** could also have something to do with it. Written datasets are typically created by a single person, drafted and edited and has relatively straightforward relationships between clauses. Spoken text on the other hand contains the dialogue from multiple people, contains more grammatical and spelling mistakes and has more intricate relationships between clauses (Marilisa Amoia, 2012).

Fig 14 Test set accuracy comparison for the sentiment models based on word embeddings trained with different models (GloVe, Word2Vec and fastText)



Source: Macquarie Quantitative Strategy, November 2019

Given the above results, we focus the rest of our analysis on word-embeddings trained on the largest available dictionary using vectors with 300 dimensions. This means that we use the Word2Vec model trained on GoogleNews and the gloVe and fastText models trained on Common Crawl.

Profiting from sentiment analysis

In this section we assess whether extending our sentiment dictionary using word embeddings can aid in the prediction of future returns. We start by providing an overview of the data source and software used in our analysis and then provide the results from our backtests.

Input Data and Processing

Data Source

In our analysis we use conference call transcripts provided by FactSet through OnDemand DocRetrieval API. Documents are available in XML format which allows us to split calls into sections (management discussion and Q&A) and paragraphs for each speaker. Additional metadata in XML transcripts allows identification of each speaker and their affiliation as well as separation of questions from answers and corporate participants from sell-side analysts.

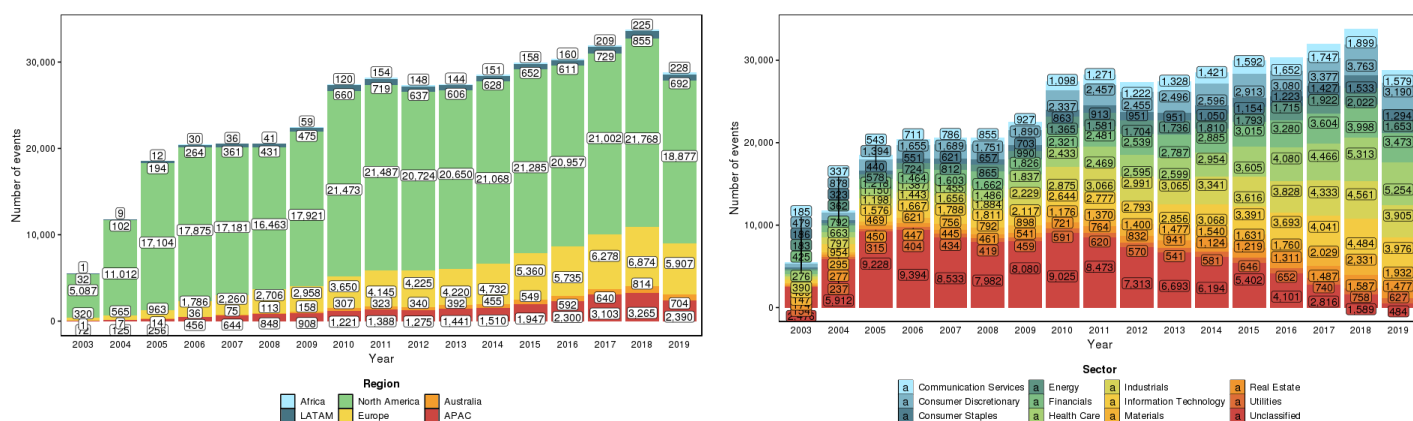
This structure in XML documents enables us to perform more granular analysis of the sentiment compared to our report [Quantamentals - I just called to say I'm bullish](#). We can now calculate sentiment within each document as well as separately for the management discussion (MD) and the Q&A. In addition, we also split the Q&A into only questions and only answers.

Our sample includes transcripts from March 2007 to September 2019. While FactSet provides transcripts for earnings calls from 1999, documents published before 2007 were uploaded retrospectively and made available through the API in 2007. We only consider documents that were uploaded by FactSet within 48 hours from the start of the call. This requirement should not be too restrictive, as FactSet targets publication of the first complete version of the transcript within three times the call's length, measured from the start of the call. Therefore, for a typical one-hour conference call a raw transcript should be available within two hours of the call's completion.

In Fig 15 we provide an overview of the number of events in the FactSet transcripts database. This includes events such as Earnings Calls, Presentations, Special Situation Calls, Analyst and Shareholder Meetings, Sales and Revenue Calls and Guidance Calls.

FactSet API provides data with a well-defined structure as well as additional metadata to refine our analysis

Fig 15 Summary of FactSet transcripts by region and sector

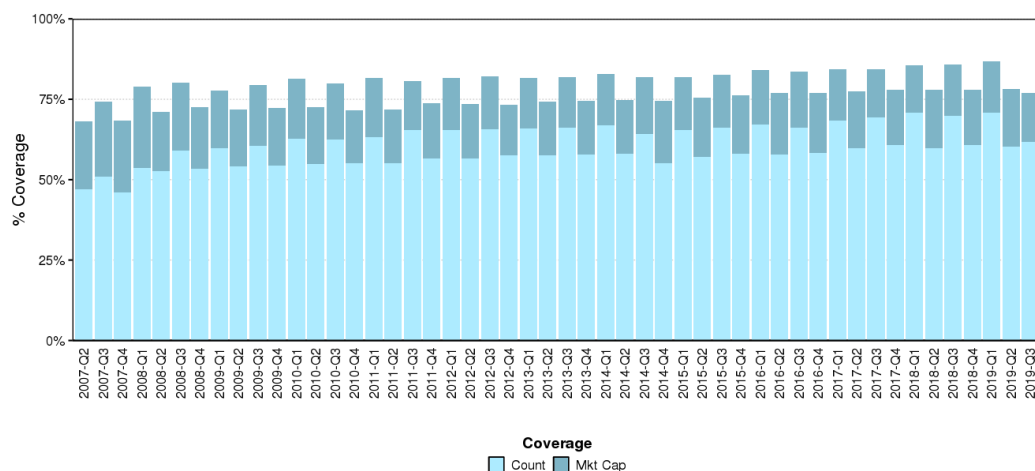


Source: Macquarie Quantitative Strategy, November 2019

We made use of all the text in the database from 2007 onwards for training our own fastText embeddings, but for backtest purposes, we restrict ourselves to only Earnings Calls. In Fig 16 we show the proportion of MSCI World covered by both the number of stocks and by market cap.

Fig 16 Percentage of MSCI World covered per quarter by number of stocks and market cap

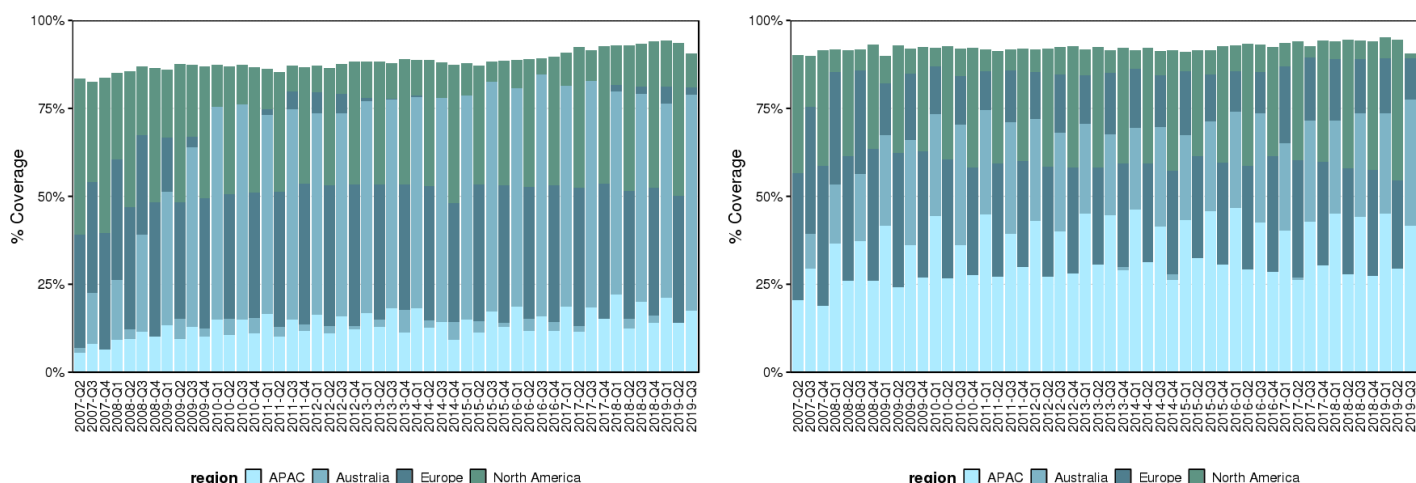
FactSet provides transcripts for more than 80% of MSCI World stocks by market cap



Source: Macquarie Quantitative Strategy, November 2019

We see that in terms of coverage, we have roughly 87% by market cap and 70% by count over recent quarters. The sawtooth pattern can be explained by differences in earnings announcement frequency. Most large American companies hold analyst calls quarterly while other regions might only do so on a semi-annual or even an annual basis.

In Fig 17 we break down the coverage by region using overlapping bars. We note that the coverage for North America is consistently high (above 95% for the most recent quarters) by both market cap and number of stocks, averaging about 690 calls per quarter. European coverage is currently at about 88% (440 calls per quarter), Australian coverage at 75%, while coverage for Asia Pacific is the lowest at 50% by market cap and less than 30% by number of stocks (145 calls per peak quarter).

Fig 17 Percentage of MSCI World covered per quarter and per region by count (left) and market cap (right)

Source: Macquarie Quantitative Strategy, November 2019

Software

Similar to our work on using AI for stock selection in [Self-driving portfolios: The artificial intelligence approach to picking stocks](#) we use RStudio Server deployed on a large AWS EC2 instance in combination with the machine learning library H2O. More details about the software itself can be found in that report. Here we described how we carry out tasks related specifically to this report.

We use R packages for most text processing tasks - parsing XML documents and most of the word tokenization. For Penn Treebank tokenizer we use the CoreNLP package (Manning et al. 2014).

H2O is used for fitting our ML models. Given that the problems are computationally expensive, we rely heavily on parallel computing using the doParallel package in R.

To train our own word embeddings, we make use of fastText. While there are two R packages available that provide an R interface to fastText, we find training models via R painstakingly slow. We therefore recommend accessing fastText from the Linux command prompt. It took us 5 hours to train the model across a total of 3 billion words on a 34-core machine, which is a once-off job. We then use the R package, fasttext, to work with the generated word embeddings.

Stock universe and stock returns

We use a broad developed market global universe (MSCI World) for our analysis and consider only events for stocks which were constituents of the index as of the time of the event. We calculate post event returns for each call over 5 and 21 days after. We assume that the stock could be traded at close on the day after the transcript upload date if upload happened before or during the market hours or on the second day after if the upload happened after the market close. This ensures that we only consider trades which an investor could realistically enter.

Event returns are calculated as active returns against the MSCI World sector total return index based on the stock's GICS sector.

For our regional results, we assign each stock to a region and report the results for that subset. To ensure enough breadth, we restrict our regional analysis to North America, Europe and Asia Pacific.

For Australian clients, we also run a separate analysis using the ASX 200 as our universe, since MSCI World only contains a small number of Australian stocks.

Sentiment Score Calculation

To measure the sentiment of a document or a section within a document, we obtain the probability of each word being positive, negative or neutral within the text. To determine the optimal level for classifying a word as either positive or negative, we perform cross-validation to find the threshold that maximises the historical IC between the sentiment score and post-event returns. This means that a word will be classified as positive only if its probability of being positive is above the optimal positive threshold and classified as negative if the probability of it being negative is below the optimal negative threshold. These thresholds are updated on an annual basis using an expanding window of out-of-sample data.

To calculate the overall sentiment for the document, we then calculate a weighted average sentiment score. In unreported results, we examined two weighting schemes – equally-weighted and word-frequency weighted and opted to use word-frequency weighted.

Event Study: Is there a statistically significant difference in performance based on sentiment?

To compare whether our hypothesis that stocks with positive sentiment has returns that differs from stocks with negative sentiment, we perform an event study. We calculate the excess return of stocks with the best and worst sentiment, defined as those in the top and bottom tercile after ranking all the available observations by sentiment. To avoid look-ahead bias, we only use information that was available at that point in time, which means that our thresholds change over time and our baskets are therefore not necessarily equal in size. To determine whether the difference in mean between the return series are statistically significant, we therefore perform a two-tailed t-test.

The results from this analysis are summarised in Fig 18 where we depict the t-stat of the spread between the top and bottom tercile baskets based on the sentiment measured using each of our embedding models. We show the results for both 1 week and 1 month after the call. We also include the results from the Loughran-McDonald dictionary (LM) as our benchmark.

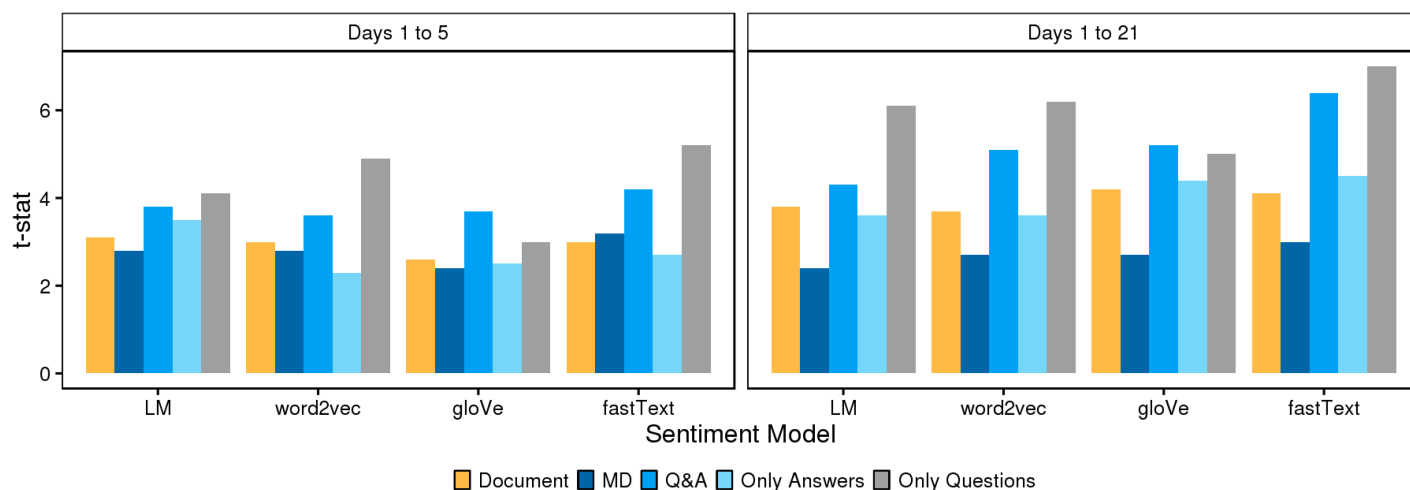
We perform our analysis on a broad developed market global universe

We are conservative in our assumptions of when trades can be entered

Words are classified as positive/negative if the probability of being a polarising word is above a certain threshold

We perform an event study over different horizons

Fig 18 T-stat of return drift of long-short baskets conditioned on sentiment post conference call



Source: Macquarie Quantitative Strategy, November 2019

We make several observations from these results:

- Regardless of which model we use, the spreads between high and low sentiment baskets are **significantly positive** for horizons of 5 and 21 days post the event.
- Most of the return drift occurs in the first week but returns **continue to drift** upwards for the month post the results. After that, returns start to mean revert.
- The Management Discussion is almost always **less informative** than the Q&A section.
- Within the Q&A section, there is a very **clear benefit to focussing on the sentiment of the Questions section** of a call. For all models and horizons, this provides the best outcome.
- The Loughran-McDonald dictionary provides a **hard benchmark** to beat and despite its simplicity, provides good results.
- On average we find **that fastText provides the best results** of all the models, followed closely by Word2Vec, while our gloVe model tends to underperform the benchmark.

In Fig 19 we provide a detailed breakdown of the results for 5 and 21 days after the call.

Fig 19 Return drift post conference call conditioned on sentiment

5 days after the call

Section	Basket			High-Low Spread		
	Low	Medium	High	Spread	t-stat	p-value
LM						
Document	-9.8	-4.1	9.2	19.0	3.1	0.002
MD	-7.7	-0.6	6.4	14.1	2.8	0.005
Q&A	-10.4	-1.0	8.9	19.3	3.8	0.000
Only Answers	-9.3	0.8	7.7	17.0	3.5	0.000
Only Questions	-11.9	-5.5	12.6	24.5	4.1	0.000
Word2vec						
Document	-4.8	6.1	10.9	15.8	3.0	0.002
MD	-2.0	2.8	12.0	14.0	2.8	0.006
Q&A	-6.7	6.3	12.0	18.7	3.6	0.000
Only Answers	-3.4	9.1	8.3	11.7	2.3	0.023
Only Questions	-10.6	7.2	14.3	25.0	4.9	0.000
gloVe						
Document	-2.7	5.9	10.2	12.9	2.6	0.009
MD	1.2	-0.1	12.8	11.6	2.4	0.018
Q&A	-6.8	6.0	12.3	19.2	3.7	0.000
Only Answers	-3.2	7.6	9.5	12.7	2.5	0.011
Only Questions	-4.9	7.3	10.4	15.3	3.0	0.003
fastText						
Document	-5.5	8.4	9.9	15.3	3.0	0.003
MD	-3.0	3.0	12.9	15.9	3.2	0.001
Q&A	-10.2	11.1	11.2	21.4	4.2	0.000
Only Answers	-3.2	6.5	10.4	13.6	2.7	0.007
Only Questions	-11.9	9.5	14.4	26.2	5.2	0.000

21 days after the call

Section	Basket			High-Low Spread		
	Low	Medium	High	Spread	t-stat	p-value
LM						
Document	-5.2	20.5	41.7	46.9	3.8	0.000
MD	5.8	24.3	36.2	30.4	2.4	0.015
Q&A	-0.9	18.8	41.4	42.3	4.3	0.000
Only Answers	-0.3	21.3	40.3	40.6	3.6	0.000
Only Questions	-9.4	23.6	48.4	57.8	6.1	0.000
Word2vec						
Document	15.1	35.0	48.7	33.6	3.7	0.000
MD	21.2	35.4	45.3	24.2	2.7	0.007
Q&A	4.9	38.0	52.4	47.5	5.1	0.000
Only Answers	15.7	35.5	48.3	32.6	3.6	0.000
Only Questions	-0.2	39.5	55.9	56.1	6.2	0.000
gloVe						
Document	15.1	32.0	52.4	37.2	4.2	0.000
MD	27.3	24.4	51.1	23.8	2.7	0.006
Q&A	6.4	35.1	53.9	47.5	5.2	0.000
Only Answers	10.0	39.8	49.2	39.3	4.4	0.000
Only Questions	11.9	27.7	56.9	45.0	5.0	0.000
fastText						
Document	14.7	33.4	51.2	36.5	4.1	0.000
MD	24.4	26.2	51.0	26.7	3.0	0.003
Q&A	-2.2	41.9	55.5	57.7	6.4	0.000
Only Answers	11.2	36.6	51.2	39.9	4.5	0.000
Only Questions	0.2	33.9	62.0	61.7	7.0	0.000

Source: Macquarie Quantitative Strategy, November 2019

This allows us to make the following additional observations:

- The performance differential between our word embedding models and the LM benchmark **increases with the time horizon**. It is hard to know what the reason for this would be, but it is possible that our extended model picks up more nuances in the conversation, which takes longer to be disseminated into the market.
- The return difference between the LM model and fastText seems to be correlated to the **size of the underlying text**. For the whole document, LM does a good job and is harder to beat, but the smaller the section, the better fastText fares in comparison. This is particularly evident in the Question section, which could sometimes be very short. This could be because LM consists of a relatively small proportion of words and one therefore needs a larger amount of text to accurately read sentiment.
- Given the difference in basket sizes, a t-test is the most robust measure to compare strategies, but the average spread in returns does give us an indication of the possible profitability of the strategy. We note that, depending on strategy and document section, the **return spread varies between 12 – 26 bp over 5 days and 23 – 62 bp over 21 days**.

Regional Results

In the previous section we looked at the results for a global developed market universe. We now split that universe into three regions to determine whether there are any material differences in stock behaviour between regions. We limit our focus to North America, Europe and Asia Pac to ensure that we have a broad enough sample to analyse and then also provide separate results for Australia. We summarise the results in Fig 20 and provide the detailed breakdowns in the Appendix.

North America

We split global results into three regions – North America, Europe and Asia Pacific

Starting with North America, we see that the results are substantially poorer than that for a global universe. We note that most of the drift occurs in the first 5 days and that it is almost exhausted by 21 days. This is in line with anecdotal evidence that a large proportion of calls for American stocks are now pre-recorded and that live calls are becoming increasingly rehearsed. It could also be a function of strategies based on call sentiment becoming more mainstream in the US, thereby resulting in reduced alpha.

To highlight a main difference between North America and other regions, we also extend the results to 63 days after the call. This indicates that a reversal occurs over this horizon, with the Question section being the only notable exception. This reversal is not apparent in other regions, bar Australia. We believe that this could partly be explained by the fact that American companies typically release results quarterly, which is not the case for most other regions. This would mean that, by 63 days, a new set of results would have been released, making the previous signal obsolete.

Our results suggest that North American investors can improve the results of this strategy by focussing only on the sentiment of the Questions section of a call. Global investors should be cognisant of the fact that this strategy requires shorter holding periods in American stocks than in other regions.

Europe

Looking at Europe, we see that returns to these strategies are considerably more profitable than in North America, showing positive returns for all strategies up to 1 month after the event. We also do not note the same level of reversal after 3 months. Once again, in most cases, a strategy based on fastText provides the best results.

Asia Pacific

Results in Asia Pacific are more mixed, with models based on Word2Vec and fastText showing positive returns, while our gloVe model is not effective. Once again, fastText appears to be the most effective model. We would have expected the results in Asia Pacific to be better than that for Europe, which suggests that one should consider country-neutralising the strategy.

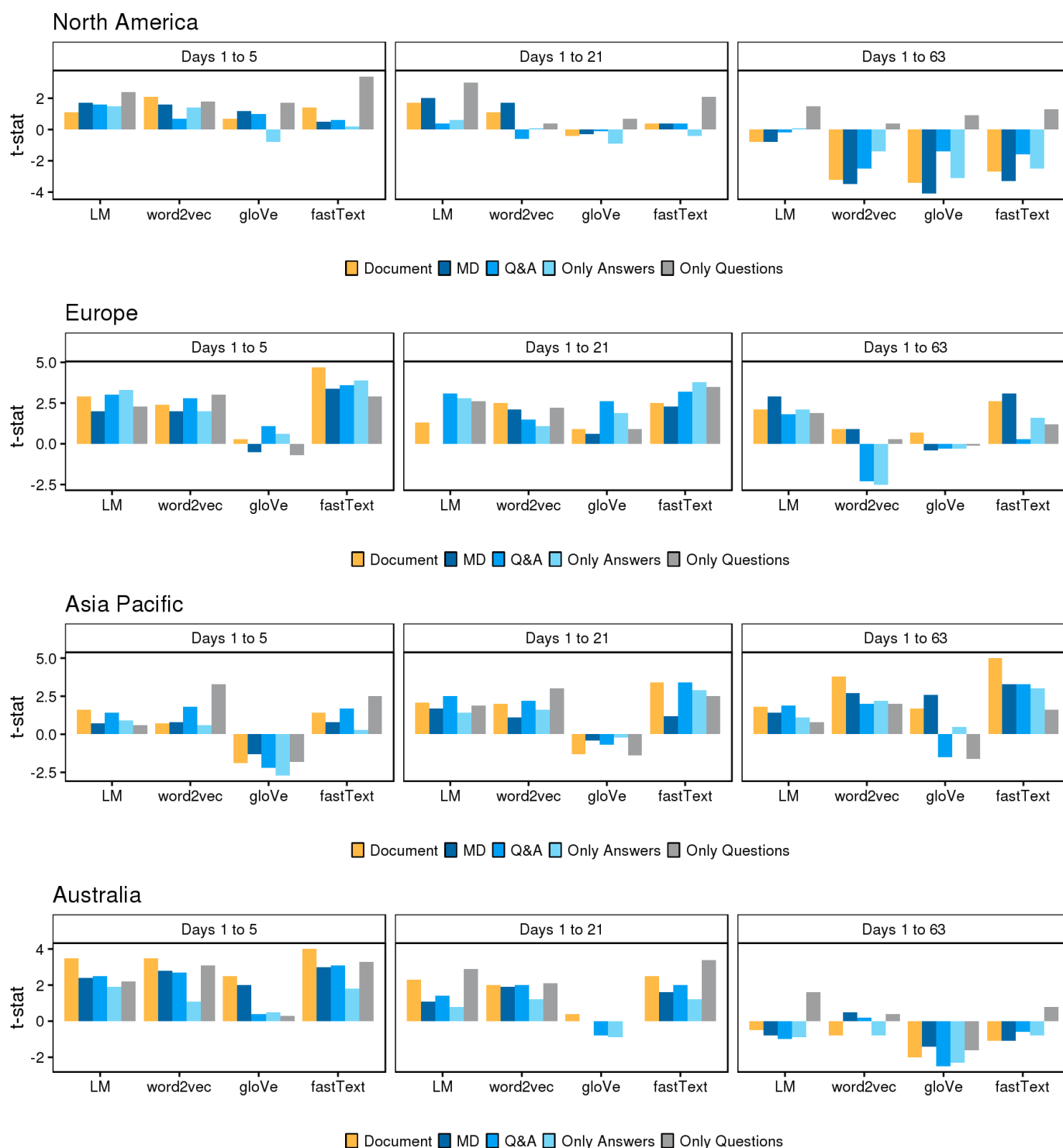
Interestingly, when comparing the results of Europe and Asia Pacific against that of North America, we note that our observation that the best results are generally achieved by focussing on the Question section is not as consistently true. It seems that this observation is predominantly driven by the US. This could be a function of lesser scrutiny on conference calls, which results in less pre-recordings and preparations and suggests that in those case, a better result could be obtained from analysing the full transcript.

Australia

We also examine results for Australia (ASX 200)

We see strong results from sentiment models in Australia for 5 days and 21 days after the event, where after the returns start to decay. Once again, fastText shows the most promising results.

We do not find that the Question section is more predictive than the full document over 5 days but do see this play out over longer time periods.

Fig 20 Return drift post conference call conditioned on sentiment for different regions

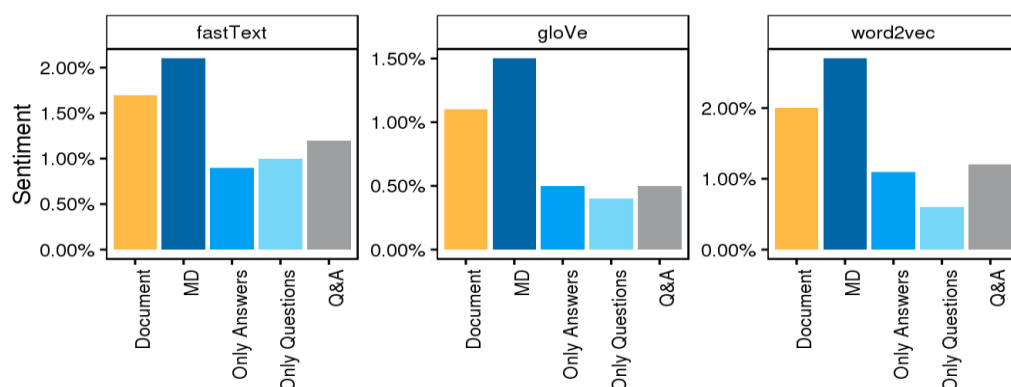
Source: Macquarie Quantitative Strategy, November 2019

Why is the Question section more informative?

To get further insight into the reason why the Question section of the call is the most informative, we show the average sentiment per section across all earnings calls in Fig 21 and split into region in Fig 22. This shows us that, regardless of which model or region we look at, the Management Discussion is consistently more positive than the Q&A that follows. This strengthens the argument that Management Discussions are prepared and rehearsed and that by focussing on the Q&A, and more specifically, the Questions section, investors can obtain a more accurate reflection of sentiment around the earnings call.

Fig 21 Average sentiment per document section for each embedding model

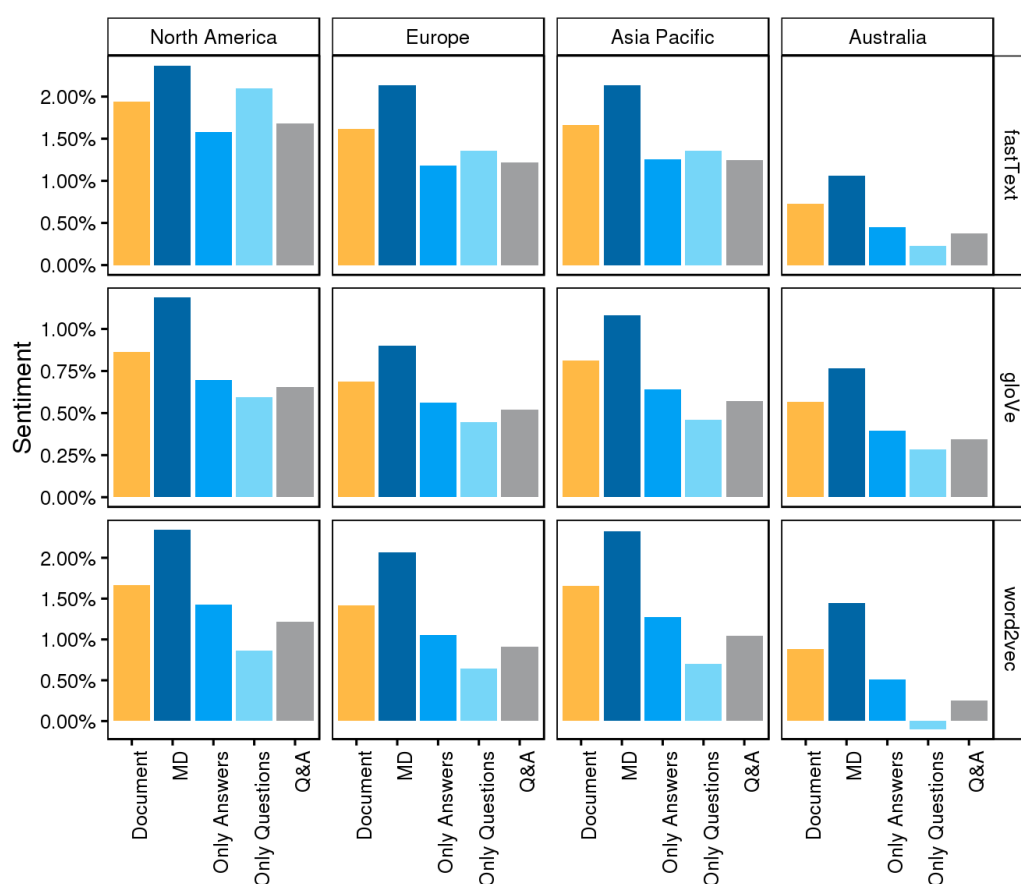
The management discussion is consistently more positive than the Q&A



Source: Macquarie Quantitative Strategy, November 2019

Fig 22 Average sentiment per document section for each embedding model and region

This is the case in all regions



Source: Macquarie Quantitative Strategy, November 2019

Using Sentiment to differentiate Earnings Surprises

In our first set of results we formed baskets purely based on sentiment. An obvious question is whether the above results aren't more easily captured by simply measuring the Earnings Surprise post a result. Put otherwise, if we know the Earnings Surprise of a stock, does analysing the sentiment provide any incremental value?

To measure the earnings surprise, we make use of EPS surprise information from I/B/E/S. This is calculated as follows:

$$Surprise_{i,t}^{EPS} = \frac{EPS_{i,t} - E(EPS_{i,t})}{|E(EPS_{i,t})|}$$

where:

$EPS_{i,t}$ is actual EPS for stock i at time t ,

$E(EPS_{i,t})$ is a mean I/B/E/S consensus estimate one day before the earnings announcement for stock i at time t .

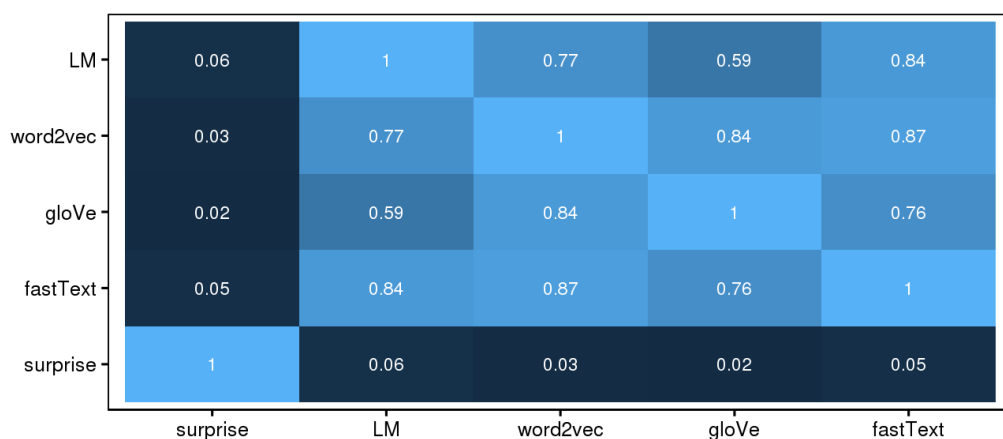
We limit our analysis to stocks for which we can match EPS surprise information from I/B/E/S with earnings call transcripts. This allows us to assess the impact of sentiment after controlling for the post-earnings-announcement drift.

As a first step, we calculate the average cross-sectional correlation between our fastText Sentiment score and the IBES Earnings Surprise for all MSCI World stocks where we have scores for both metrics. The results are shown in Fig 23. This shows that there is only a mildly positive correlation between Earnings Surprises and Sentiment. Comparing the results between Sentiment models is also interesting. We see that all the sentiment models have a high correlation to each other as well as to LM. It is not surprising to see that fastText and Word2Vec have a high correlation, given that they are based on similar underlying methods, while their correlation to gloVe, which is based on an entirely different methodology, is lower.

Results are compared with traditional earnings surprises

The correlation between Earnings Surprise and Sentiment is only marginally positive

Fig 23 Correlation between Earnings Surprise and Sentiment

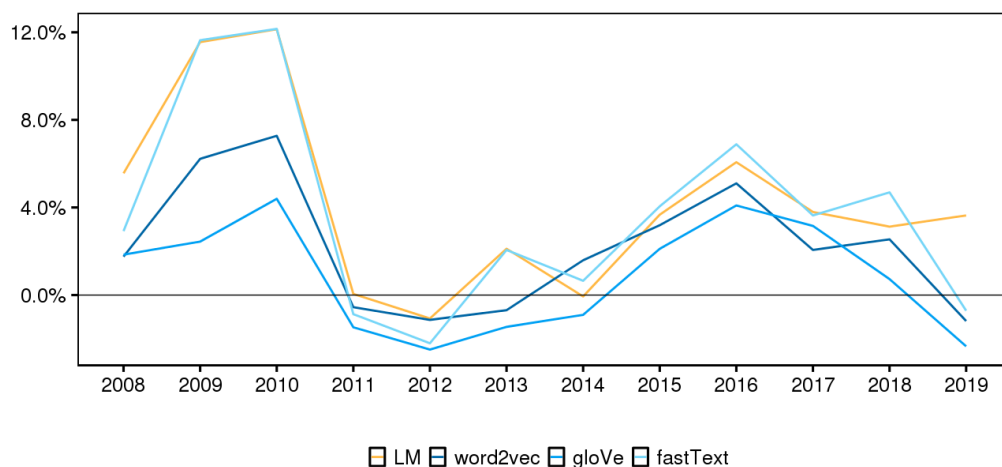


Source: Macquarie Quantitative Strategy, November 2019

To determine how this relationship has evolved over time, we display the annual correlation between Earnings Surprise and each of our sentiment models below. We note that this relationship varies over time, but always tends to be low. This suggests that Sentiment is not just a proxy for Earnings Surprise.

Fig 24 Correlation between Earnings Surprise and Sentiment per annum

*The relationship
between Earnings
Surprise and
Sentiment varies
over time*



Source: Macquarie Quantitative Strategy, November 2019

*We perform a double
sort between
Earnings Surprise
and Sentiment*

To determine whether there are profits to be made from combining information on Earnings Surprises and Sentiment, we repeat our event study. To control for the effect of earnings surprise on tone levels, we perform a double sort where we first form tercile baskets based on Earnings Surprise and then further divide baskets on Sentiment within each tercile. We use our fastText model to determine sentiment. By comparing the excess return of each portfolio, we can assess whether, for example, an earnings beat followed by a bullish conference call has significantly different implications for returns compared to an earnings beat with negative sentiment.

These results are reported in Fig 25. Looking, for example, at the results for 21 days after the event, we see that stocks that had both a positive surprise and positive sentiment had an average excess return of 78.2 bps, while stocks that had a positive surprise with negative sentiment only returned 23.7 bps. Once again, we find that sentiment measured from the Questions section provides the biggest differentiation in results, with stocks with both positive Surprise and Sentiment drifting 93 bps, while stocks with both a negative surprise and sentiment lost 29.4 bps in the month after the call. This provides further evidence that sentiment does indeed provide additional information on top of the Earnings Surprise.

Fig 25 Return drift 21 days after event for portfolios sorted on Earnings Surprise and fastText Sentiment for MSCI World

Sentiment	Document			MD			Q&A		
	High	Medium	Low	High	Medium	Low	High	Medium	Low
	78.2	53.6	13	74.7	65.3	5.2	86.3	46.2	12.1
	71.6	44.8	19.1	48.4	30.2	36.8	59.3	53.1	60.6
	5.4	32.5	23.7	43.5	35.6	12	9.6	34.8	-13.6
Only Answers	Only Questions								
	High	Medium	Low	High	Medium	Low	High	Medium	Low
	89.8	51.9	13.5	93.2	50.6	44.9			
	47.7	40	26.4	16.1	51.5	23.3			
Only Questions	High	Medium	Low	High	Medium	Low			
	19.4	43.5	32.8	42.7	29.1	-29.4			

Source: Macquarie Quantitative Strategy, November 2019

This highlights again the benefit of having a quantitative toolkit for measuring the sentiment of conference calls. With more than a hundred global calls scheduled per week during peak reporting season, having an objective measure of sentiment as well as earnings surprises allows portfolio managers and analysts to have a broad overview of results without having to listen to every call.

In addition, sentiment can be measured as soon as the transcript is received, which is typically within a few hours after the call, while quant funds typically need to wait for overnight feeds to provide the fundamental data released to accompany the results.

Conclusion

In this report we introduced a simple, but effective way of extending a dictionary-based sentiment model using word embeddings.

We show that, regardless of which method is used, the soft information available in conference calls does predict future stock returns. This is likely because the hard information released during results season is backwards looking, while investors will trade based on the expectation of future results. If the sentiment of a call is positive, this bodes well for future returns, despite the most recent results potentially being disappointing.

We find that even a simple strategy based on the Loughran-McDonald's dictionary is predictive of future stock returns. By extending this dictionary using fastText, we show that incremental improvements can be achieved. This improvement is most evident when measuring the sentiment of the Questions section of the call.

Further work

Now that we have determined that there is value in using word embeddings for financial sentiment analysis, there are various avenues for future research:

- In this report, we purposefully restricted ourselves to using unigrams (single words) and a simple model. Word embedding models can, however, easily be extended to bigrams (multiple words), which should improve results further. There are also various simple-to-implement NLP rules such as adjustments for negation that can be incorporated.
- Our previous research suggests that one can improve results further by analysing both the level of sentiment as well as changes in sentiment.
- Rather than the unsupervised learning used in this report, one can use supervised deep learning trained on forward returns using a large body of real-time text.
- Since we now have access to a database that allows us to recognise speakers' names and positions, we can also look at the following topics:
 - The impact of participants on calls' tone and informativeness – Cicon (2014) finds that CEO participation inhibits information discovery, in contrast number of analysts active on the call contributes positively
 - Measures of tone controlled for manager-specific tone – Davis et al. (2014) find that part of the abnormal tone of conference call has a significant manager-specific component.
 - The impact of analysts' tone and participation on earnings forecast accuracy – Mayew et al. (2013) find that analysts who participated actively in conference call issue more accurate forecasts

Once again, the differential based on Questions is the biggest

References

- Arman Joulin, E. G. (2016). Bag of Tricks for Efficient Text Classification. *EACL*.
- Conomos, J. (2011). *Extra! Extra! Read all about it!* Australia: Macquarie Quantitative Strategy.
- David Bew, C. R. (2018). Modelling Analysts' Recommendations via Bayesian Machine Learning. *The Journal of Financial Data Science Winter 2019*, 1 (1) 75-98.
- Guida, G. C. (2019). Training trees on tails with applications to portfolio choice. Available at SSRN: <https://ssrn.com/abstract=3403009>.
- Guida, T. a. (2019). Ensemble learning applied to quant equity: gradient boosting in a multi-factor framework. In *Big Data and Machine Learning in Quantitative Investment*. New York: Wiley Finance Series.
- Jeffrey Pennington, R. S. (2014). *GloVe: Global Vectors for Word Representation*.
- Jones, K. C. (2019). Machine Learning for Stock Selection. *Financial Analysts Journal*.
- Leon Chen, Z. D. (Spring 2015). Implementing Black-Litterman using an Equivalend Formula and Equity Analyst Target Prices. *The Journal of Investing*.
- M. Dixon, D. K. (2017). Classification-based Financial Markets Prediction using Deep Neural Networks. *Algorithmic Finance*.
- Macquarie Quantitative Strategy. (2019). *Self-driving portfolios: The Artificial Intelligence Approach to Picking Stocks*.
- Marilisa Amoia, K. K.-K. (2012). Coreference in Spoken vs. Written Texts: a Corpus-based Analysis. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).
- Prado, M. L. (2018). *Advances in Financial Machine Learning*. John Wiley & Sons, Inc., Hoboken, New Jersey.
- Research, M. Q. (2013). *Positively Persuasive*.
- S. Deerwester, S. D. (1990). Indexing By Latent Semantic Analysis. *Journal of the Association for Information Science and Technology*, 391-407.
- Shihao Gu, B. K. (2018). Empirical Asset Pricing via Machine Learning. Available on SSRN at: <http://ssrn.com/abstract=3159577>.
- Silver, N. (2012). *The Signal and the Noise: Why so many predictions fail - but some don't*. Penguin Group.
- Strategy, M. Q. (2013). *Camouflaged in Complexity*.
- strategy, M. Q. (2014). *A surprising tone*.
- Strategy, M. Q. (2014). *How are you really feeling?*
- Strategy, M. Q. (2015). *I just called to say I am bullish*.
- Tomas Mikolov, K. C. (2013). Efficient estimation of word representations in vector space. *Proceedings of Workshop at International Conference on Learning Representations (ICLR 2013)*.
- Womack, L. B. (2004). Analysts, Industries, and Price Momentum. *Journal of Financial and Quantitative Analysis*.
- Y. Bengio, R. D. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 1137 - 1155.
- Zhi Da, E. S. (2011). Relative valuation and analyst target price forecasts. *Journal of Financial Markets*, 161 - 192.

Appendix: Regional Results

North America

Fig 26 Return drift post conference call conditioned on sentiment

5 days after the call

Section	Basket			High-Low Spread		
	Low	Medium	High	Spread	t-stat	p-value
LM						
Document	3.2	5.2	12.3	9.1	1.1	0.270
MD	-0.6	9.5	12.5	13.0	1.7	0.083
Q&A	-0.3	5.7	13.8	14.1	1.6	0.103
Only Answers	1.8	4.3	14.1	12.3	1.5	0.143
Only Questions	-0.8	0.7	18.4	19.2	2.4	0.018
Word2vec						
Document	1.7	8.5	12.0	10.3	2.1	0.040
MD	4.5	6.8	12.1	7.7	1.6	0.108
Q&A	5.5	10.7	9.4	3.8	0.7	0.461
Only Answers	1.6	14.9	8.5	6.9	1.4	0.169
Only Questions	3.1	9.1	11.8	8.6	1.8	0.067
gloVe						
Document	7.4	6.7	10.8	3.4	0.7	0.460
MD	8.7	0.8	14.3	5.6	1.2	0.217
Q&A	6.6	7.7	11.1	4.5	1.0	0.328
Only Answers	12.5	6.6	9.0	-3.5	-0.8	0.431
Only Questions	1.6	14.1	9.1	7.5	1.7	0.092
fastText						
Document	4.6	7.7	11.4	6.7	1.4	0.175
MD	6.5	10.1	9.0	2.5	0.5	0.595
Q&A	6.6	9.8	9.5	3.0	0.6	0.571
Only Answers	8.8	7.9	9.8	1.1	0.2	0.834
Only Questions	-1.8	8.2	14.7	16.5	3.4	0.001

21 days after the call

Section	Basket			High-Low Spread		
	Low	Medium	High	Spread	t-stat	p-value
LM						
Document	34.2	46.1	60.1	25.9	1.7	0.081
MD	33.7	50.0	60.6	26.9	2.0	0.046
Q&A	48.3	46.6	55.2	6.9	0.4	0.660
Only Answers	48.0	43.8	57.0	9.0	0.6	0.548
Only Questions	26.5	41.5	69.2	42.7	3.0	0.003
Word2vec						
Document	44.6	61.6	54.4	9.8	1.1	0.260
MD	44.5	56.0	58.7	14.2	1.7	0.088
Q&A	61.2	53.4	55.6	-5.6	-0.6	0.538
Only Answers	55.4	54.4	56.6	1.2	0.1	0.893
Only Questions	62.7	35.9	66.1	3.3	0.4	0.689
gloVe						
Document	56.7	55.0	53.7	-3.1	-0.4	0.699
MD	59.9	46.5	57.9	-2.0	-0.3	0.800
Q&A	58.2	52.2	57.6	-0.6	-0.1	0.940
Only Answers	59.7	57.1	52.4	-7.3	-0.9	0.353
Only Questions	51.2	57.6	56.3	5.1	0.7	0.515
fastText						
Document	54.7	49.5	58.4	3.7	0.4	0.668
MD	57.2	45.0	60.5	3.3	0.4	0.681
Q&A	55.1	52.4	58.6	3.4	0.4	0.709
Only Answers	59.5	52.4	56.3	-3.2	-0.4	0.721
Only Questions	46.3	49.9	64.2	18.0	2.1	0.034

Source: Macquarie Quantitative Strategy, November 2019

Europe

Fig 27 Return drift post conference call conditioned on sentiment

5 days after the call

Section	Basket			High-Low Spread		
	Low	Medium	High	Spread	t-stat	p-value
LM						
Document	-20.0	0.5	19.1	39.2	2.9	0.004
MD	-16.3	-7.1	11.9	28.2	2.0	0.051
Q&A	-19.3	5.0	23.4	42.7	3.0	0.003
Only Answers	-18.6	-1.1	26.9	45.5	3.3	0.001
Only Questions	-15.4	1.6	18.0	33.4	2.3	0.021
Word2vec						
Document	-2.0	-12.6	18.8	20.8	2.4	0.016
MD	-2.6	-10.6	14.3	16.8	2.0	0.048
Q&A	-13.8	8.8	11.5	25.3	2.8	0.004
Only Answers	-7.9	0.4	10.4	18.3	2.0	0.041
Only Questions	-16.3	6.9	10.2	26.4	3.0	0.003
gloVe						
Document	-0.4	-3.1	2.5	2.9	0.3	0.747
MD	1.1	-1.8	-2.9	-4.0	-0.5	0.650
Q&A	-9.3	10.0	0.4	9.7	1.1	0.265
Only Answers	-5.6	4.7	-0.6	5.0	0.6	0.562
Only Questions	5.6	-10.0	-0.2	-5.8	-0.7	0.503
fastText						
Document	-6.9	-22.1	33.5	40.4	4.7	0.000
MD	-19.8	4.7	9.9	29.7	3.4	0.001
Q&A	-16.0	5.6	15.1	31.2	3.6	0.000
Only Answers	-18.3	6.6	15.1	33.4	3.9	0.000
Only Questions	-12.2	1.8	13.8	26.0	2.9	0.004

21 days after the call

Section	Basket			High-Low Spread		
	Low	Medium	High	Spread	t-stat	p-value
LM						
Document	-32.1	-22.1	-0.1	32.0	1.3	0.193
MD	-16.1	-26.2	-16.7	-0.7	0.0	0.979
Q&A	-53.8	0.9	27.6	81.5	3.1	0.002
Only Answers	-39.1	-29.2	30.6	69.6	2.8	0.006
Only Questions	-52.8	3.9	17.1	69.9	2.6	0.008
Word2vec						
Document	-14.3	-22.5	23.5	37.8	2.5	0.013
MD	-22.5	-4.7	8.6	31.1	2.1	0.037
Q&A	-18.3	-9.1	4.7	23.0	1.5	0.145
Only Answers	-17.8	-6.7	-0.1	17.7	1.1	0.260
Only Questions	-38.2	17.0	-3.1	35.2	2.2	0.025
gloVe						
Document	-15.3	0.9	-1.1	14.2	0.9	0.378
MD	-6.7	-19.9	2.0	8.7	0.6	0.576
Q&A	-35.4	12.6	5.3	40.7	2.6	0.008
Only Answers	-29.7	7.1	-0.5	29.1	1.9	0.058
Only Questions	-4.1	-33.4	9.7	13.8	0.9	0.369
fastText						
Document	-9.5	-36.4	27.6	37.1	2.5	0.014
MD	-34.5	4.6	0.6	35.1	2.3	0.023
Q&A	-31.4	-5.1	17.4	48.8	3.2	0.002
Only Answers	-36.0	-5.9	22.4	58.4	3.8	0.000
Only Questions	-24.4	-20.3	31.5	55.9	3.5	0.000

Source: Macquarie Quantitative Strategy, November 2019

Asia Pacific

Fig 28 Return drift post conference call conditioned on sentiment

5 days after the call

Section	Basket			High-Low Spread		
	Low	Medium	High	Spread	t-stat	p-value
LM						
Document	-71.2	-95.5	-4.0	67.2	1.6	0.115
MD	-76.3	-48.8	-46.8	29.5	0.7	0.490
Q&A	-68.3	-100.6	-9.2	59.0	1.4	0.168
Only Answers	-69.2	-84.2	-30.3	38.9	0.9	0.357
Only Questions	-67.5	-60.7	-39.1	28.3	0.6	0.539
Word2vec						
Document	-69.9	-54.1	-53.6	16.3	0.7	0.495
MD	-71.4	-52.0	-53.1	18.4	0.8	0.428
Q&A	-79.6	-78.1	-38.6	41.0	1.8	0.078
Only Answers	-60.6	-89.5	-46.8	13.8	0.6	0.575
Only Questions	-100.6	-63.5	-22.7	77.9	3.3	0.001
gloVe						
Document	-31.2	-81.2	-73.6	-42.5	-1.9	0.061
MD	-39.0	-80.1	-70.1	-31.1	-1.3	0.186
Q&A	-34.6	-70.1	-88.9	-54.3	-2.2	0.027
Only Answers	-32.6	-58.4	-97.7	-65.0	-2.7	0.007
Only Questions	-37.1	-74.0	-83.7	-46.6	-1.8	0.072
fastText						
Document	-75.5	-54.9	-42.6	32.9	1.4	0.162
MD	-69.5	-56.5	-50.4	19.1	0.8	0.418
Q&A	-77.2	-77.6	-35.6	41.5	1.7	0.091
Only Answers	-69.4	-61.0	-60.8	8.6	0.3	0.730
Only Questions	-79.9	-84.1	-19.8	60.1	2.5	0.014

21 days after the call

Section	Basket			High-Low Spread		
	Low	Medium	High	Spread	t-stat	p-value
LM						
Document	-178.8	-184.1	-25.5	153.3	2.1	0.034
MD	-211.3	-90.7	-85.8	125.5	1.7	0.090
Q&A	-126.0	-306.0	53.0	179.0	2.5	0.012
Only Answers	-123.4	-217.3	-25.4	98.0	1.4	0.165
Only Questions	-109.2	-274.7	39.1	148.3	1.9	0.053
Word2vec						
Document	-159.1	-105.9	-74.9	84.1	2.0	0.041
MD	-149.6	-82.9	-104.0	45.6	1.1	0.267
Q&A	-156.5	-104.2	-71.5	85.1	2.2	0.028
Only Answers	-133.9	-115.4	-73.2	60.8	1.6	0.119
Only Questions	-152.1	-122.8	-35.0	117.1	3.0	0.003
gloVe						
Document	-108.9	-67.7	-158.0	-49.1	-1.3	0.207
MD	-124.4	-71.3	-139.2	-14.7	-0.4	0.714
Q&A	-94.4	-121.7	-123.6	-29.3	-0.7	0.480
Only Answers	-110.2	-94.5	-116.5	-6.3	-0.2	0.877
Only Questions	-84.2	-84.6	-140.6	-56.5	-1.4	0.174
fastText						
Document	-164.3	-136.1	-29.9	134.4	3.4	0.001
MD	-138.4	-107.7	-92.1	46.3	1.2	0.243
Q&A	-175.5	-106.5	-33.0	142.5	3.4	0.001
Only Answers	-150.7	-132.2	-33.2	117.4	2.9	0.004
Only Questions	-146.3	-90.8	-47.2	99.1	2.5	0.012

Source: Macquarie Quantitative Strategy, November 2019

Australia

Fig 29 Return drift post conference call conditioned on sentiment (Australia)

5 days after the call

Section	Basket			High-Low Spread		
	Low	Medium	High	Spread	t-stat	p-value
LM						
Document	-39.8	-18.1	26.0	65.8	3.5	0.001
MD	-33.5	-30.6	22.6	56.1	2.4	0.019
Q&A	-34.6	13.6	12.2	46.8	2.5	0.011
Only Answers	-12.6	6.4	9.3	21.9	1.9	0.054
Only Questions	-21.9	-10.1	45.5	67.4	2.2	0.030
Word2vec						
Document	-45.5	6.7	51.1	96.5	3.5	0.001
MD	-45.2	31.3	33.4	78.6	2.8	0.005
Q&A	-44.6	39.7	33.1	77.7	2.7	0.006
Only Answers	-8.0	26.5	23.0	31.0	1.1	0.271
Only Questions	-20.3	-14.5	63.4	83.6	3.1	0.002
gloVe						
Document	-25.1	4.3	43.9	69.0	2.5	0.011
MD	-19.3	11.5	34.5	53.9	2.0	0.048
Q&A	12.8	7.9	25.1	12.4	0.4	0.659
Only Answers	1.2	32.6	14.5	13.3	0.5	0.629
Only Questions	16.8	4.6	26.1	9.3	0.3	0.744
fastText						
Document	-58.2	13.8	54.2	112.4	4.0	0.000
MD	-45.5	21.0	41.9	87.4	3.0	0.002
Q&A	-29.8	1.6	59.1	88.9	3.1	0.002
Only Answers	-12.3	15.0	37.0	49.3	1.8	0.072
Only Questions	-32.0	-5.0	58.7	90.7	3.3	0.001

21 days after the call

Section	Basket			High-Low Spread		
	Low	Medium	High	Spread	t-stat	p-value
LM						
Document	-91.1	-98.3	24.0	115.1	2.3	0.020
MD	-54.8	-78.7	3.4	58.2	1.1	0.262
Q&A	-63.8	-94.7	8.0	71.8	1.4	0.150
Only Answers	-47.1	-59.3	-6.4	40.7	0.8	0.397
Only Questions	-100.5	-35.1	3.7	104.2	2.9	0.004
Word2vec						
Document	-33.1	28.9	62.3	95.4	2.0	0.049
MD	-37.4	41.7	54.7	92.1	1.9	0.058
Q&A	-50.9	59.3	48.2	99.1	2.0	0.046
Only Answers	-11.7	37.3	46.6	58.3	1.2	0.239
Only Questions	-31.9	29.0	67.7	99.6	2.1	0.034
gloVe						
Document	16.0	30.5	35.6	19.6	0.4	0.675
MD	44.0	-6.2	45.4	1.5	0.0	0.976
Q&A	51.0	25.2	14.7	-36.3	-0.8	0.451
Only Answers	37.5	61.9	-4.1	-41.6	-0.9	0.373
Only Questions	55.4	-17.2	54.5	-0.9	0.0	0.985
fastText						
Document	-59.7	48.5	63.6	123.3	2.5	0.012
MD	-39.2	65.6	39.0	78.3	1.6	0.114
Q&A	-32.3	31.8	63.6	95.8	2.0	0.047
Only Answers	-14.4	45.9	44.1	58.5	1.2	0.217
Only Questions	-72.4	26.7	89.3	161.7	3.4	0.001

Source: Macquarie Quantitative Strategy, November 2019

Important information:

This publication represents the views of the Sales and Trading Quant Strategy Department and/or Sales and Trading Desk strategists of Macquarie. It is not a product of Macquarie Research and the view of Quant Strategists may differ from the views of Macquarie Research and other divisions at Macquarie. Macquarie has policies in place to promote the independence of Macquarie Research and to manage conflicts of interest, including policies relating to dealing ahead of the dissemination of Macquarie Research. These policies do not apply to the views of the Quant Strategy contained in this report.

Quant Strategy Disclosure

The name "Macquarie" refers to Macquarie Group Limited and its worldwide affiliates and subsidiaries (the Macquarie Group). This information is provided on a confidential basis and may not be reproduced, distributed or transmitted in whole or in part without the prior written consent of Macquarie.

This publication has been prepared by Macquarie Sales and Trading personnel and is not a product of the Macquarie Research Department. Any views or opinions expressed are the views of the author and the Macquarie Sales and/or Trading desk from which it originates ('the Authors') and those views may differ from those of the Macquarie Research Department. Prior to distribution of this publication, information contained herein may be shared with Macquarie Trading desks who are not subject to prohibitions on trading prior to the dissemination of this publication. The views are not independent or objective of the interests of the Authors and other Macquarie Sales and/or Trading desks that trade as principal in the financial instruments mentioned within and who may be compensated in part based on trading activity. The views do not include and are not intended as trading ideas or recommendations specifically tailored for the needs of any particular investor.

This communication is provided for information purposes only, is subject to change without notice and is not binding. Any prices or quotations in the information provided are indicative only, are subject to change without notice and may not be used or relied on for any purpose, including valuation purposes. This communication is not a solicitation to buy or sell any product, or to engage in, or refrain from engaging in, any transaction, except to the extent covered by the CFTC Rules (see Important Derivatives Disclosure). Nor does it constitute investment research, a research report or a personal or other recommendation. Nothing in the information provided should be construed as legal, financial, accounting, tax or other advice.

Important Derivatives Disclosure: This material constitutes a solicitation for entering into a derivatives transaction only for the purposes of, and to the extent it would otherwise be subject to, U.S. Commodity Futures Trading Commission Regulations §§ 1.71 and 23.605 promulgated under the U.S. Commodity Exchange Act (the "CFTC Rules"). Futures options and derivatives products are not suitable for all investors and trading in these instruments involves substantial risk of loss.

Macquarie is a global provider of banking, financial advisory, investment and funds management services. As such, Macquarie may act in various roles including as provider of corporate finance, underwriter or dealer, holder of principal positions, broker, lender or adviser and may receive fees, brokerage or commissions for acting in those capacities. In addition, Macquarie and associated personnel may at any time buy, sell or hold interests in financial instruments referred to in this information either on behalf of clients or as principal. Therefore, this information should not be relied upon as either independent or objective from the interests of Macquarie and associated personnel which may conflict with your interests.

To the extent permitted by law, Macquarie accepts no responsibility or liability (in negligence or otherwise) for loss or damage resulting from the use of or relating to any error in the information provided. This information has been prepared in good faith and is based on information obtained from sources believed to be reliable, however, Macquarie is not responsible for information stated to be obtained from third party sources. Any modelling, scenario analysis, past or simulated past performance (including back-testing) contained in this information is no indication as to future performance.

Canada: This report has been prepared by a Quant Strategist in the Institutional Sales Group based upon general comments by research analysts or other sources we believe to be reputable. This report should not be construed to be a research report, nor is this article investment advice and should not be relied on as such. The author of this report is NOT a Macquarie research analyst, salesperson or investment advisor. The views expressed herein are those of the author alone and are not necessarily those of Macquarie. Please refer to this [link](#) to our research reports for the most current research on any of the names discussed herein. Those Research Reports contain important disclosures regarding our research ratings and Capital Advisory relationships.

The financial products and/or services referred to in this information may not be eligible for sale in all jurisdictions. This information is directed at institutional clients who have professional experience as defined by applicable law and/or regulation in the relevant jurisdiction. It is not for retail clients and it is not for distribution into any jurisdiction where this information is not permitted. For important country-specific disclosures regarding information from Macquarie Sales and Trading, please click on the region relevant to you at: www.macquarie.com/salesandtradingdisclaimer.

© Macquarie Group