Manuscript Details

Manuscript number JEB_2017_110

Title Between the Lines: Decipher the Firms' Fundamentals with Artificial Intelligience

Article type Research Paper

Abstract

Our paper studies whether language patterns elicited by a different level of stimulus during earnings conference calls are informative about firms' earnings. Executives endure various levels of pressure when presenting a scripted managerial discussion and spontaneous answering to analysts' scrutinized questions. Such stimulus induces evasive answers, incoherent responses, and disturbances in emotion and cognition. Our artificial intelligence based measures of language patterns, built upon deep learning and topic modeling, transform analysts' perceptions to FinTech solutions. We find that evasive and incoherent answers forecast firms' earnings and stock return. A trading strategy based on evasiveness yields a positive risk-adjusted return.

Keywords Earnings conference call; Asset Pricing; Deep Learning; Artificial Intelligence

Corresponding Author Yanzhen Chen

Corresponding Author's Institution The University of Texas at Austin

Order of Authors Yanzhen Chen, Huaxia Rui, Andrew Whinston

Submission Files Included in this PDF

File Name [File Type]

cl_title2.pdf [Cover Letter]

Highlights.pdf [Highlights]

jbe.pdf [Manuscript (without Author Details)]

Submission Files Not Included in this PDF

File Name [File Type]

jeb.zip [LaTeX Source File]

Cover Letter

Monday, June 15, 2017

Dear Professor Jagtiani, Professor Hunt, and Professor John,

I would like to submit the attached manuscript, "Between the Lines: Decipher the Firms' Fundamentals with Artificial Intelligence," for consideration for possible presentation during the FinTech conference in Philadelphia.

In our present paper, we study whether language patterns elicited by a different level of stimulus during earnings conference calls are informative about firms' earnings. Executives endure various levels of pressure when presenting a scripted managerial discussion and spontaneous answering to analysts' scrutinized questions. Such stimulus induces evasive answers, incoherent responses, and disturbances in emotion and cognition. Our artificial intelligence based measures of language patterns, built upon deep learning and topic modeling, transform analysts' perceptions to FinTech solutions.

We believe that this manuscript is appropriate for a presentation at your conference because of the following reasons. First, our study shed light upon automating analysts' interpretations using FinTech techniques. Second, our paper advances the literature by integrating different methodologies from artificial intelligence, asset pricing, and behavioral finance to study an emerging phenomenon of valuing qualitative information in disclosure and AI-driven trading from practitioners. It contributes to an emerging area of research and development at the intersection of big data analytics and finance.

This paper (or closely related research) has not been published or accepted for publication. It is not under consideration at another journal or conference.

Thank you for your consideration!

Sincerely,

Yanzhen Chen
Department of Information, Risk & Operations Management
McCombs School of Business
The University of Texas at Austin
Austin, TX 78712
yanzhen@utexas.edu

Between the Lines: Decipher the Firms' Fundamentals with Artificial Intelligience

Yanzhen Chen
University of Texas at Austin, Austin TX 78712, USA
yanzhen@utexas.edu
Huaxia Rui
University of Rochester, Rochester NY 14627, USA
huaxia.rui@simon.rochester.edu
Andrew. B. Whinston
University of Texas at Austin, Austin TX 78712, USA
abw@uts.cc.utexas.edu

Highlights

- Language patterns during earnings calls are informative about firms' earnings.
- Our approach transforms analysts' perceptions to FinTech solutions.
- Evasive answers forecast poor firm performances and lower stock returns.
- Our artificial intelligence incoherent measure predicts worse future earnings.
- A trading strategy based on evasiveness yields a positive risk-adjusted return.

Between the Lines:
Decipher the Firms' Fundamentals with Artificial Intelligence

Our paper studies whether language patterns elicited by a different level of stimulus during earnings conference calls are informative about firms' earnings. Executives endure various levels of pressure when presenting a scripted managerial discussion and spontaneous answering to analysts' scrutinized questions. Such stimulus induces evasive answers, incoherent responses, and disturbances in emotion and cognition. Our artificial intelligence based measures of language patterns, built upon deep learning and topic modeling, transform analysts' perceptions to FinTech solutions. We find that evasive and incoherent answers forecast firms' earnings and stock return. A trading strategy based on evasiveness yields a positive risk-adjusted return.

1. Introduction

The art of reading between the lines is as old as manipulated information.

– Serge Schmemann

Earnings conference calls have become a battlefield between analysts who are determined to obtain more information and executives who resolve to strategic disclosure or even information manipulation. We aim to infer the executives' mind from the behavioral expressions of their language and test how machine learning interpreters better inform investors. According to interpersonal deception theory (IDT) (Buller, Burgoon, Daly, and Wiemann (1994)), evasiveness, incoherence and emotional fluctuation raise concerns for deceptive intentions. Our paper operationalized this theory using artificial intelligence in speeches and conversations.

Behavioral finance, especially psychology-based asset pricing theory has been exploring executives creating overvaluation with selection and manipulation(Hirshleifer (2001)). Evasive answers from executives confuse investors with ambiguity, which is found to be distasted. One behavioral explanation is that investors associate ambiguity with potential hostile manipulation (Hirshleifer (2001)). Researchers also tried to identify dishonest behavior by examining 'speaking in two tongues' using aggregated opinions (Malmendier and Shanthikumar (2014)).

Recently there is an emerging interest in qualitative or behavioral leaks of pricing information in earnings calls from practitioners among financial industry. How open and forthright executives are when communicating with shareholders is one of the most important tactical behaviors being assessed by equity research teams. For example, a hedge-fund consulting firm, Business Intelligence Advisors (BIA) employs former CIA (Central Intelligence Agency) members to analyze language clues from earnings conference calls, interviews and other releases (Javers (2010)) By assessing whether executives answer questions forthrightly or try to divert the questions during earning conference calls, BIA has successfully predicted firms' profitability, such as UTStarcom.[1] In addition, the stock price is also

[1]After analyzing the conference call of UTStarcom of Aug 2, 2005, BIA rated UTStarcom's second quarter conference call as "medium high level of concern" and highlighted that the communication from its executives manifests " avoids providing information,..., indicates underlying concern". The BIA team, therefore, warned its client that their executives' behavior indicates they will release poor third-quarter results and highly double profitability in the fourth quarter. This BIA report had flagged the concerns with revenue recognition before the stock price shrunk to 2/3 of its original value on the next day after the third quarter earning announcement. The main reason that BIA analysts considered the UTStarcom's Executive Vice President and Chief Operating Officer, Michael Sophie as avoiding commenting on issues related to revenue recognition is that they found the following answer to be a "detour statement". During the Q&A session, Mike Ounjian, an analyst with Credit Suisse First Boston asked "Are there any issues related to recognizing revenues on these?". Sophie and the interim CFO tapped dance around subjects by responding "Yes,

responsive to definitive and direct answers in conference calls. On July 20, 2015, Jim Cramer from CNBC credited Google's share surge after its conference call to its new CFO Ruth Porat for being more down to earth in answering questions than her predecessors.[2] The Public Company Accounting Oversight Board (PCAOB) also has been emphasizing the importance of examining quarterly earnings calls as a means of detecting fraud (PCAOB (2008)). It also explicitly mandates that auditors should consider "observing or reading transcripts of earning calls," as a part of the procedure when assessing material misstatement.[3]

Although the above perceptions are made by experts, such as consultants, analysts, and accountants, advances in artificial intelligence and data analytics are reforming the way of making financial decisions (World Economic Forum 2015). Many of the biggest hedge funds are initiating AI-driven systems that can better trade and foresee market trends than humans. For example, Goldman Sachs is investing deeply in artificial intelligence. Starting from Feb 24, 2017, Goldman Sachs launches an investment trust in Japan that uses natural language processing to process news reports, analysts report and make decisions accordingly. As the co-head of Two Sigma, a prominent hedge fund, comments: "It's very hard for someone using traditional methods to juggle all the information of the global economy in their head (. . .). Eventually the time will come that no human investment manager will be able to beat the computer".

Despite the wisdom from industry, there is little empirical evidence by academic literature that documented to how to observe language directly during the earnings calls and the prevalence of FinTech in investment. To fill this void, we utilize the interactive nature of conference earning calls to analyze the company disclosure directly. We chose to analyze interactions which are the Q&A session of an earnings conference call because the demand for information is well defined and observable. Our paper directly measures incomplete disclosure using executives' language during earning calls and advances the measure to a refined semantic level compared to state-of-the-art measures which use the numbers of times that executives declined to answer a question (Hollander, Pronk, and Roelofsen (2010)). With the protection of the safe harbor, investors and analysts explicitly inquire executives to give more insight of the firms' fundamentals than in the legal or financial setting. The optimal disclosure policies according to the persuasion game or

with the backlog, the vast majority of the wireless backlog is clearly PAS (an acronym for one of the company's products, Personal Access System). I think you saw the announcement at the end of June where we announced on the PAS infrastructure orders in China. And again, it's just the timing of deployment and achieving final acceptance, we've also got some CDMA (an acronym for a type of mobile phone standard) to a lesser extent in the backlog. ... But Q3 is clearly a little more handset-oriented than we would typically run.". More examples of evaluations of openness and forthrightness in handling questions conducted by BIA analysts, please refer to LAING (2006)

[2] http://www.cnbc.com/2015/07/20/cramer-ruth-porat-key-to-googles-success.html

[3] PCAOB-Auditing Standard Release No. 2010-004 August 5, 2010

signaling game theories typically feature disclosure-withhold intervals and the threshold. However, human language carries rich and subtle meanings rather than a discrete choice of disclosing or suppressing.

Language is versatile and valuable in behavioral finance research. It conveys nuances and behavioral hints permeated by the speaker's mind. Both are hard to be substituted by numbers. Analyzing language has several advantages in enhancing investment decisions. First, language reveals sentiment, which is beyond a statement of facts. Second, language demonstrates connections and aspects from any possible angels. They may be intrinsically qualitative and hard to be represented using even large numbers of variables. Lastly, numbers or aggregated opinions are usually processed, that are not sufficient to reflect the infinite variety of the underlying phenomena. For example, the recommendation from analysts can result from all market related factors, without going through the actual text, one is hard to tell whether the recommendation is logical. Our paper dive into the words of executives so as to 'read' executives' minds.

Specifically, we operationalized the three major cues of deception detection from the perspectives of what executives say and how they deliver the ideas. They are evasiveness in handling questions, incoherent responses and emotional & cognitive inconsistencies. Evasiveness to avoid material facts is among the three common ways of deceptions (Buller et al. (1994)). Deceptive statements are less coherent (Hauch, Blandón-Gitlin, Masip, and Sporer (2015)). Incoherence in thoughts helps executives to skirt issues by changing the subject. Lastly, according to deception cue theories,lying is usually more cognitively effortful, as a result invokes emotional fluctuation. We detect the inconsistencies in emotion and cognitive processes between management discussion (MD) and answers to analysts on the same topic.

We propose a novel evasiveness measure that uses a text mining technique of topic modeling (Blei, Ng, and Jordan (2003); Griffiths and Steyvers (2004)). Our evasive measure is based on the unstructured text of questions and their corresponding answers. This automatic system, the core of which is a Latent Dirichlet Allocation (LDA) algorithm, enables Bayesian-learning of underlying topics from the text. Based on the learned topics, topic modeling represents and summarizes the semantic of text using a probabilistic distribution over a set of underlying topics. By treating the questions and answers in one conversation as a pair of documents, the distance between topic distribution representations of the question and answer texts captures whether the answers are to the point. Our evasiveness measure leverages the expert knowledge of the industry and analysts by learning topics from their speeches and conversations and captures nuanced evasiveness using a continuous measure. We also discover the latent topics that are in the executives mind at the same time, which are well interpretable. This measure is another step forward in gauging the informativeness of answers, which is both subjective and expensive to quantify manually otherwise.

Complementary to content, we measure how executives deliver their ideas-the logical flow behind an answer. To measure coherence in a conference call, we utilize the deep learning technique: Skip-thought Models (Kiros, Zhu, Salakhutdinov, Zemel, Urtasun, Torralba, and Fidler (2015)). The motivation of using this deep learning based model is Tri-fold. Firstly, coherence in the context of earnings calls requires a deep understanding of the industry, finance, accounting, and so on. All this highly specialized domain knowledge is embedded in the managerial discussions and Q&As, which are difficult to hand-code by researchers reliably. Rather than engineer a set of ad-hoc and error-prone important features, deep learning models guide the model to find the best features using the information inherent in the data without extensive domain knowledge. Secondly, the Skip-thought model represents sentences' semantic and syntactic properties by vectors and provide insights of the hidden thoughts transition. Lastly, deep learning models enable us to automate the measures and analyze large-scale textual data. In this way, we avoid going through expert labeling, which is costly and may lack consistency among different raters.

Another aspect of delivery from the executives is emotional&cognitive inconsistency. Research in psychology defines the state of psychological arousal and discomfort as cognitive dissonance, which is caused by taking actions that deviate from their beliefs (Graham (2007); Mazar, Amir, and Ariely (2008)). Individuals who engage in deceptiveness or concealing the truth will experience cognitive dissonance which increases to the extent to the incompleteness and distortion of the information.

The Q&A sessions in earnings conference calls provide us with a strong external stimulus to elicit prominent cognitive dissonance. Executives are scrutinized by analysts using hard-to-predict questions and need to react in real-time in the Q&A session. On the contrary, the MD is drafted and polished by a specific support team and rehearsed by executives. In this way, executives can strategically disclose, as they see fit. The information environment settings of MD and Q&A sessions are significantly different, but the same group of executives participates. This fact enables us to contrast emotional and cognitive discrepancies considering both firm performance and individual fixed factors with the language patterns in MD as a psychophysiology baseline. The inconsistencies in context-free language patterns resulting from the spontaneity of answers compared to scripted managerial discussions during conference calls may be informative of a firm's performance.

We explore the formulation of emotion and cognitive processes which manifest in a speech (or an answer) as an assorted emotion and cognitive levels aroused by different topics. Combining matrix decomposition with LASSO technique, we learned the sparse representation of emotional and cognitive reactions. Based on the learned emotional and cognitive response to each topic, the inconsistencies of emotion and cognitive processes measure the changes specific to a given executive on a certain topic.

Our main result is that evasiveness and coherence measures forecast worse next quarter firm performance. A trading strategy based on evasiveness measures yields significantly positive excess return and raw return. The profitability only disappears when the transaction cost exceeds 0.25%.

This paper makes several contributions. It initiates the empirical study of incomplete disclosure of interactive communication. Different from the existing works, our paper focuses on real-time interactions rather than one-way communication such as firm released announcements and media coverage. Measures based on one-way communication is vulnerable to confound "no news" with less disclosure. The strong assumption that investors or analysts demand disclosures at a certain time and will punish the firms that fail to provide such information makes the measures less clean (Healy and Palepu (2001)). Our study advances the literature also by integrating different methodologies from artificial intelligence, asset pricing, and behavioral finance to study an emerging phenomenon of valuing qualitative information in disclosure and AI-driven trading from practitioners. Especially, behavioral finance investigates some financial phenomena that can be plausibly understood with some agents who are not fully rational. The field has two building blocks: limits to arbitrage, which argues prices may remain in a non-equilibrium state for protracted periods of time due to restrictions of funds; and psychology, which catalogs the kinds of deviations from full rationality. We introduce natural language processing techniques such as deep learning and topic modeling to asset pricing area for the first time to capture the psychological trails of deviation and test the limits of arbitrage based on them. The paper has clear industrial implications as it contributes to an emerging area of research and development at the intersection of big data analytics and finance, which is better summarized by a new area as FinTech. The economic value associated with language cues and asset price can be exploited meaningfully for a prediction on real-time and form a basis for a smart finance analysis platform. Our big data approach demonstrates the potential of extracting meaningful information from publicly available, unstructured data, through large-scale computation and transform the information into investment decision support system.

2. Literature Review

Our study is related to various streams of existing literature on disclosure communication games and textual analysis of corporate disclosures. Our innovative measures of evasiveness and coherence use topic modeling and deep learning. In addition, the rich literature on psychophysiology and psychology provide behavioral support for constructing our measures. We next review the relevant literature in these domains.

In the economics literature, firms' disclosure policy is usually described using three types of communication games: persuasion games, costless signaling games, and signal-jamming models.

In persuasion games, senders with information endowment can either report truthfully or withhold information in order to influence the behavior of a receiver(Shin (1994)). Milgrom (1981) showed if disclosure is costless and the sender is certain about receivers' response, complete revelation should occur in equilibrium. To explain the strategic disclosure, a substantial literature considers variants of the Grossman and Milgrom models. By relaxing the condition of the unraveling results from Milgrom (1981), Jovanovic (1982) derived the results that firms disclose information when it is sufficiently favorable. Other works find when receivers are of different levels of sophistication, firms may optimally withhold information (Fishman and Hagerty (2003)). The optimal disclosure policies in persuasion games feature disclosure and non-disclosure intervals. In our research question, information revealed in earnings calls largely engaged in forward-looking statements and forecasts, which is hard to verify and does not associate with a direct cost. Therefore, information manipulation is another choice besides disclose and non-disclose.

When the legal punishment of information manipulation is muted, especially when the senders report is hard to verify or protect by the safe harbor statement, there are no costs of misreporting. In these models ('cheap-talk' models), sender may report independently from their private signal (Stocken (2000)), defined as 'babbling equilibrium'. More realistically, the signal-jamming model features when a sender manipulates the report with cost, which often increases to the extent of information distortion. In these models, the sender has private information and if his/her incentives are common knowledge, there exists an equilibrium in which the sender dissembles the report while receivers rationally discount the report (Stein (1989); Narayanan (1985)). However, when senders privately observe their incentives, they may report less informatively without cost (Fischer and Verrecchia (2000)) In equilibrium, the report function weakly increases with the private signal of the sender. However, our research refined the decision of disclosure to be a less obvious and offensive way of responding to information inquiries beyond sending one-way signals.

Our paper empirically tests the predictions of the above theoretical works by examining directly the disclosure language. We build a series of measures to quantify disclosure in text from the perspectives of semantic, logical and psychological content.

Our paper joins the emerging topic of textual analysis in understanding firm disclosure which attracts a lot of attention. The main disclosure characteristics that attract attention from the researchers are amount of information and sentiment. The most commonly used proxy for amount of information is the length (Lee (2012); Miller (2010)), although in many works, researchers also associate length with complexity. Researchers often measure sentiment of disclosure using dictionary software such as Diction and LIWC. In general, these works document that pessimism /optimism in news story and earnings announcements predict future return (Tetlock (2007); Tetlock, Saar-Tsechansky, and Macskassy (2008); Davis, Piger, and Sedor (2012); Brockman and Cicon (2013)) and soft information explains more of the announcement's effect on earnings forecasts (Brockman and Cicon (2013)).

The methodology of existing literature on textual analysis of firm disclosure varies from dictionary or statistical approaches. Tetlock et al. (2008) performed one of the earliest sentiment analysis on firm news and found that the negative tone of news forecasts worse future earnings. However, later researchers pointed out the caveat of measuring sentiment using the frequencies of a pre-defined positive/negative word list since context and domain knowledge plays a significant role in determining sentiment, which is largely ignored in the sentiment analysis approach (Li (2010)). For example, when responding to an estimation of future revenue of a product with steel as the main raw material, an executive said, "I personally believe that this is a very conservative estimate and especially now that metal price decreases a little bit in China.". This sentence will be labeled as having 2 negative words ("conservative" and "decrease") and no positive words, although it bears positive view about the future. Therefore, lacking specific prior knowledge of the firm and financial practice, most dictionaries are built for psychology which limits the application of the dictionary approach.

Our data-driven, analytics-based measure of language improve on granularity and provide a comprehensive measure of text on the level of semantic, logical and emotional of the disclosure content. The measure of evasiveness and incoherence leverage the domain knowledge embedded in the MD and conversations during earnings conference calls and are built upon context. The big data approach based on unstructured textual data automates the vision of analysts and can scale up to a real-time investment decision support platform.

When observing executives languages, prominent theories on deceptive discourse from communication and psychology literature reveal to what audience should be alert. Three aspects of language are claimed to draw special attention from the listeners based on these theories. Vagueness is one of the strategic behaviors used when trying to deceive others, according to the interpersonal deception theory and management obfuscation hypothesis (Buller et al. (1994)). High quality of cognitive operations embedded in the statements is another signal of fraud, according to the reality monitoring method (Johnson and Raye (1981)). Reality monitoring method predicts deceivers to substitute details and contextual reference with cognitive expressions and lying requires more cognitive effort. Lastly, negative affect warns the receiver of the genuineness of a statement, according to the Four-Factor Model of deception (Zuckerman, DePaulo, and Rosenthal (1981)) . Therefore, our content and content-free analysis of languages, which covers evasiveness, incoherence, emotion and cognition reflected in the conversation, reveal the new direction of detecting financial fraud and concealment of adverse news in big data era.

3. Methodologies and Constructions of Key Variables

Our key three variables are based on statistical text analysis of the conversations during the earnings calls. The content analysis reveals the rationale behind the overall equity analysts' recommendations from the behavioral perspectives of executives.

To measure the informativeness of an answer, we perform topic level analysis of the question-answer pairs. The probabilistic topic model enables us to discover the abstract topics and weights inherent in questions and their corresponding answers. The "topics" produced by topic modeling techniques are clusters of words. A question has a mix of topics and so does an answer. An evasive answer either skirting a question or avoiding the important and dwell on the trivial. The former is reflected in distinct components of latent topics in questions and answers. The later can be detected by discrepancies of topic weighs between the question and its answer.

Our coherence measure builds upon a statistical language model, which is a probability distribution over sequences of words and sentences. The goal is to evaluate the probability of a sentence. Statistical language models rely on the Chain Rule of probability to compute it, that is $P(w^1, w^2, ..., w^m) = \prod_{i=1}^{m} P(w^i | w^1, ..., w^{i-1})$. Usually, these models apply the Markov assumption are called n-gram models, in which we assume that the probability of observing the i^{th} word w^i can be approximated by the probability of observing it with a shorter context history. Only the recent history matters. Therefore, the probability of observing a sentence is approximated as $P(w^1, w^2, ..., w^m) = \prod_{i=1}^{m} P(w^i | w^{i-(n-1)}, ..., w^{i-1})$.

To perform the statistical modeling of textual data, researchers need to first encode the text into vectors. There are two different ways to achieve this goal: discrete approach and continuous space representation models. These vectorized words, (aka. word embeddings) are more desirable when the vectors contains information from multiple features of a word. For example, in Word2Vec, which is a most commonly used continuous space representation model, vectors represent relative position between words.

The pioneer work on n-gram language model used discrete approach, which codes every unique words using one dimension of an input vector (Chen and Goodman (1996)). For example, Green and Jame (2013) applies a trigram model on an english letter level to calculate how fluent and recognizable a company's symbol is and its impact of investment decisions and firm value. However, these discrete models face two major limitations. First, the possible n-grams increases exponentially with the size of the vocabulary, causing a curse of dimensionality. Second, it ignores semantic similarity between words. For example, 'meeting' and 'conference' carry similar meanings. However, in discrete approach, each is indicated by a different dimension of the input vector and the similarity is no longer preserved in the input representation.

Continuous space language models overcome the above two disadvantages using neural networks. It projects the word indices onto a continuous space so that the vector representation of the word has real valued elements rather than 0-1 entries. Neural networks represent words in a distributed way, which is a many-to-many relationship between words and neurons. Neural networks perform two tasks: project words onto a continuous space, and calculate the sentences' probability. The distributed representations encoded in the hidden layers of neural networks, represent the words (also called word embeddings). Therefore, each word is mapped onto an n-dimensional vector according to the size of the hidden layer. These representations model semantic relations between words and alleviates the curse of dimensionality. The most commonly used model for extracting word embeddings which convert texts to low dimensional vectors and then passed onto other neural probabilistic language models for sentence probability evaluation is word2vec model.

Our incoherence measure takes word embeddings from the word2vec model and process them using a special deep learning language model: skip-thought model (Kiros et al. (2015)). This model operates on the sentence level and represents sentences onto continuous space which enables us to quantify how latent thoughts underlying the language develop along with sentences. In section 3.2, we will discuss the application of deep learning models in analyzing earnings call interactions.

The inconsistencies in emotion and cognition are measured by the differences of emotional and cognitive levels demonstrated in MD presentations and answers. Since emotion and cognition reactions are independent of context, we adopt a bag-of-words approach. Using a well defined dictionary of emotion and cognition, drawn upon psychological research (Pennebaker, Francis, and Booth (2001)), we first measure the emotion and cognition reactions demonstrated in the MD presentation and answers. We trace these reactions back to sparse emotional and cognitive responses towards latent topics. The anomaly changes in reaction to the same topic warns analysts that the speaker maybe experiencing cognitive dissonance caused by deceptive intention.

In this section, we discuss the construction of the measures for behavioral leaks from their language which signals underlying negative information.

3.1. Evasiveness of Answers

Empirical evidence shows managers try to obfuscate information with more irrelevant information as a substitute for a clear answer (Larcker and Zakolyukina (2012)). Interpersonal deception theory models how people handle perceived deception at the conscious or subconscious level in dyadic and dialogic communication. IDT assumes that communication is a dynamic process in which information exchange is influenced by personal goals and the meaning of the interaction as it unfolds (Buller et al. (1994)). The sender's disclosure choices are affected by the reaction from the receiver, and vice versa. Commonly used ways to deceive receivers include equivocation and concealment.

Therefore, our evasiveness measures capture how answers are related to the questions in terms of intrinsic topics. We utilize topic modeling techniques. Topic modeling is used to discover latent "topics" for each document from a collection of documents. Its unsupervised learning nature enables the model to automatically generate the measure without requiring labor efforts in reading and labeling the evasiveness of each question-answer pair for training.

Formally, we treat questions and answers in the same round as a pair of documents[4].

3.1.1. *Construction of Evasiveness Measure*

Using a specific topic model: LDA topic model, we (i) discover different topics, where each topic consists of a bag of keywords, and (ii) identify the topic weights of each document. The key idea is that a specific document aims to convey information in a small number of topics, and the words observed in that document are the realizations of those topics. Thus, we can discover hidden topics by observing many documents.

We apply the LDA topic modeling algorithm to more than 0.25 million aggregated question-answer pairs, covering the S&P 500 firms from 2010 to Jun-2015. The algorithm produces K topics (we specify K=30 in our paper.[5]), where each topic is represented by a bag of key words. LDA outputs topic distributions for these documents, which are probability values assigned to each discovered topic for a given document. As a result, topic modeling generates topic representation of each question (or answer) for question. The topic representation of j-th round Q&A discussion for the earnings call of company i at time t is represented by a topic distribution $T^{Q}_{i,j,t}$, *and* $T^{A}_{i,j,t}$ respectively.

We interpret the discovered topic representation of questions and their answers as the focuses of the questions and answers. The higher a weight is assigned to a topic in the topic proportion, the more relevant this issue is to the point made by the speaker. Finally, we calculate the evasiveness $e_{i,j,t}$ of the j-th round Q&A discussion as the cosine similarity of the two corresponding topic distributions $T^{Q}_{i,j,t}$ and $T^{A}_{i,j,t}$, which can be written as follows: similarity of their focuses:

$$e_{i,j,t} = \frac{T^{Q}_{i,j,t} \cdot T^{A}_{i,j,t}}{||T^{Q}_{i,j,t}|| \cdot ||T^{A}_{i,j,t}||}$$

We report the topics discovered by topic modeling in Table 1. We can see that the topics are interpretable. The three topics correspond to operations, markets, health care issues, accounting and energy production issues.

3.1.2. *Measure Validation*

To ensure the evasiveness measure captures the construct of interest, we need to observe and measure variation in the extent of evasiveness felt from speeches and conversations using the historical conference earnings call transcript

[4]We define a round of Q&A discussion as a series of questions (answers) by the same speaker before the operator connects to another questioner.

[5] Our statistical results are robust to K=50 and 80

Table 1: Sample of Latent Topics

Topics	**Most Freq Key Words**							
1	customers	capacity	production	business	fleet	shipment	fuel	new
2	europe	growth	market	china	brazil	international	u.s	asia
3	million	data	question	launch	clinical	study	fda	going
4	million	year	quarter	guidance	rate	tax	number	impact
5	projects	going	that's	rigs	coast	crude	gulf	oil

Only the top eight words are displayed due to space limit.

samples from individual firms. We divide documents into evasive and forthright groups, using documents with top and bottom 25% measures of evasiveness by topic modeling and human rating. We then compare the selected documents rated by subjects and the machine learning algorithm. Sixty seven undergraduate business major students from the University of Texas at Austin participated in the survey. Participants were 55.22% female, with a median age of 20.2 years.

The results are shown in Figure 1, our topic modeling based evasiveness measure performs better when identifying evasive documents with low evasiveness values. The fraction of documents overlapped among those with bottom 25% of evasiveness measured by LDA Topic Model, and the human rating is 82%. The overlap rate in documents selected as top 25% forthright is 69%. Both ratios are significantly higher than a random labeling, which will yield 25% overlap rate in expectation. Therefore, our topic model based evasiveness measure does capture useful information in quantifying a documents' evasiveness.

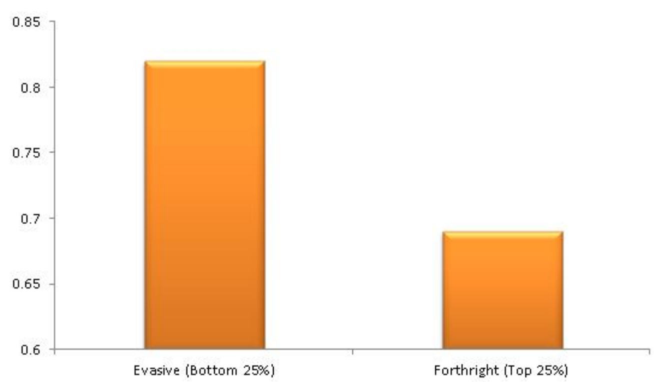


Fig. 1. Fractions of documents overlapped with different measures of evasiveness

3.2. *Incoherent Answers*

We also aim to replace analysts' judgments of coherence with FinTech techniques. As thoughts are linked by a chain of reasoning, a logical path of sorts, we measure incoherence by the way thoughts skip in a conversation. With the help of deep learning, we trained a 'skip-thought vector' model to represent thought in sentences in its relation to the context which leads to the artificial construction of common sense. Vector representations of sentences represent thoughts. These vectorized thoughts are especially useful in mastering natural, conversational language, and the ability to make leaps of logic.[6] The aim is to distill natural language down to vector space mathematics using a deep learning framework.

Deep learning is connecting different layers of neural networks to form an increasingly meaningful representation of the data. As shown in Figure 2, each node in the hidden layer is a neuron. The system begins by transforming words into one-hot vectors, which are a group of bits that are all zeros except the one unique bit to represent a certain word. After that, data represented by nodes in the input layer, travels through the connections between the neurons, which are of different strength, captured by weights. A neuron becomes activated through activation function when the signal is strong enough, which is shown by black nodes in Figure 2. The activation function we use is a logistic function. Therefore, in each layer in Figure 2, the neural network feeds a vector of inputs through multiple logistic regressions and generates a vector of outputs. However, here we do not need to know the predicted variables. Instead, the ultimate training goal that directs what the hidden variables should be, so as to predict the next layer precisely. By selectively activating neurons and strengthening the connections between neurons, the neural network identifies important features.

We use a deep learning based skip-thought method (Kiros et al. (2015)), which exploits rich semantic and contextual information between the lines, to represent sentences by thoughts. The model subjects to an encoder-decoder architecture (Sutskever, Vinyals, and Le (2014)). Both the encoder and decoder are deep learning networks. The encoder maps a sentence into a fixed-length vector representation. Deep learning helps us to avoid handcrafting features that are time-consuming, often over-specified and incomplete. The model feeds each encoded vector through a "narrow passage" that guides the network to abstract a small number of important features. This encoded vector is then decoded into an output by the decoder. The main idea is depicted in Figure 3.

[6]More discussions see Professor Geoffrey Hinton's talk at the Royal Society, May-2015

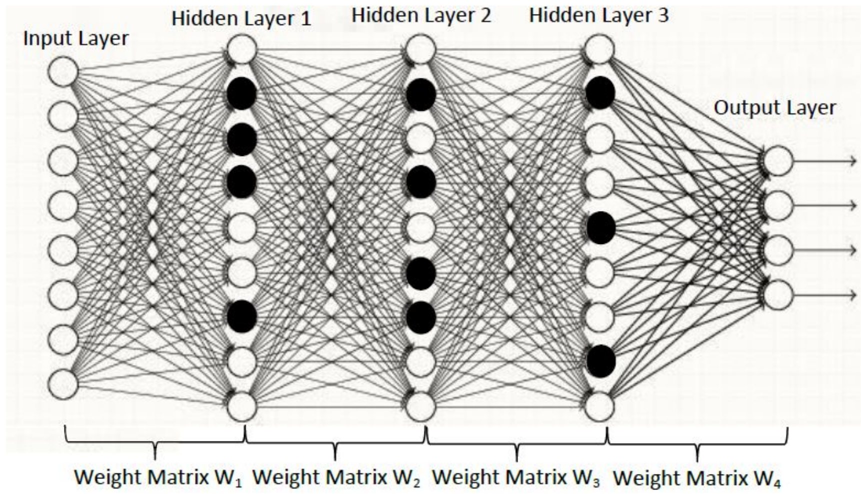


Fig. 2. A Neural Network with Three Hidden Layers
The solid black nodes denote the activated neurons which are important to improve prediction performance.

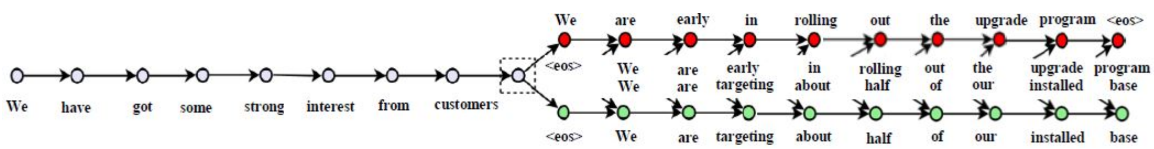


Fig. 3. Predict the Preceding sentence and the Next Sentence Using the Skip-thought Model

For an observed tuple (s_{i-1}, s_i, s_{i+1}) of consecutive sentences (in red, grey and green respectively), the Skip-thought model encodes the sentence s_i to reconstruct the previous and following sentences. In the example above, the input is a sentence triplet *We are early in rolling out the upgrade program. We have got some strong interest from customers. We are targeting about half of our installed base. This is a simplified conversation between the CEO and an analyst of 1-May-2013 on ADT Security Services. Unattached arrows are connected to the encoder output. The components of the same color share the same parameters. <eos> is the token for the end of a sentence.*

The architecture of the skip-thought model is shown in Figure 4. The encoder takes a word embedding layer to convert each word into a vector. Word Embedding is a map of a word to a vector. A common way of extracting word embedding is to pretrain a neural network such as word2vec[7] (Kiros et al. (2015)). Then the encoder will capture the patterns of the word sequence. The hidden layers of the encoder are fed as inputs into two decoders, for the prediction of preceding and following sentences respectively. Each decoder uses a deep learning neural network, which shares the same word embedding with the encoder.

When implementing the skip-thought model, the most widely used deep learning networks of encoder and decoder, are recurrent neural networks (RNN). Specifically, our encoder is RNN with gated recurrent unit (GRU) (Chung, Gulcehre, Cho, and Bengio (2014)) activations. For a decoder, we use RNN with a conditional GRU. There are two

[7] For more information about word2vec neural network, please refer to Mikolov, Chen, Corrado, and Dean (2013)

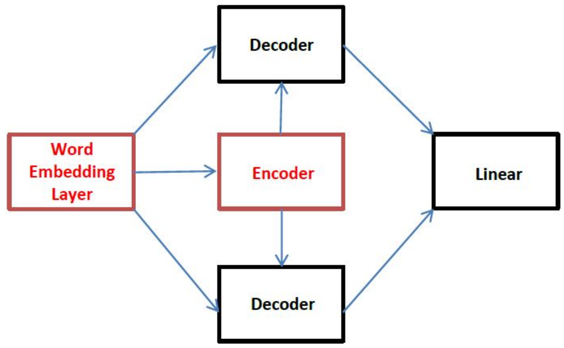


Fig. 4. The architecture of the Skip-thought Model

After learning the skip-thought model, the encoder and word embedding layer can serve as a generic feature extractor for new tasks, as is shown here using red boxes.

decoders for the preceding and following sentences respectively. Each decoder uses a different set of recurrent layers and shares the embedding layer. An RNN model operates on sequences and can simulate interactions of words in the compositional process. Each hidden layer is updated with information from two sources: current word embedding layer activations and the previous hidden layer's activations. RNNs have a "memory" which captures information about what has been calculated so far. Theoretically, RNNs can preserve information in arbitrarily long sequences, but in practice they are limited to only a few steps. To better capture long-term dependencies, we choose RNN-GRU. By using RNN-GRU, the model decides what to clear from memory so as to be efficient at capturing long-term dependencies (Chung et al. (2014)).

Encoder. Let $x_i^{(t)}$ be the word embedding for the t-th word in the sentence i. At each step the encoder produces a hidden state $h_i^{(t)}$ that represents the memory of the network. It captures information in all the previous steps. The output at each step is calculated based on the memory only at time t.Therefore, the hidden state of the last word of the sentence represents the full sentence. Consequently, the encoder captures both syntactic and semantic properties. The encoding is an iterated process of the following equations (dropping the subscript i):

$$
\begin{aligned}
&Reset\ Gate : r^{(t)} = \sigma(W_r x^{(t)} + U_r h^{(t-1)}) \\
&Update\ Gate : z^{(t)} = \sigma(W_z x^{(t)} + U_z h^{(t-1)}) \\
&New\ memory\ : \bar{h}^{(t)} = tanh(W x^{(t)} + U(r^{(t)} \circ h^{(t-1)})) \\
&Hidden\ state\ : h^{(t)} = (1 - z^{(t)}) \circ h^{(t-1)} + z^{(t)} \circ \bar{h}^{(t-1)},
\end{aligned}
\tag{1}
$$

where W_r, W_z , W and U are weighting matrices that need to be learned. We use $\circ$ to denote a Hadamard product, which is an element-wise multiplication.

The intuition of each equation can be summarized in Figure 5. The new memory generation equation captures how to consolidate a new input word embedding $x^{(t)}$ with the past hidden state $h^{(t-1)}$. In this equation, the model combines the two sources of information to summarize the new word in light of the contextual past as the $\bar{h}^{(t)}$. In the reset gate, a reset signal $r^{(t)}$ determines how important $h^{(t-1)}$ is to summarize $\bar{h}^{(t)}$. In other words, the reset gate can ignore or diminish the past hidden state should the $h^{(t-1)}$ be less relevant to the new memory. The update gate determines how $h^{(t-1)}$ should be incorporated in the next hidden state. The next hidden state will be the same as $h^{(t-1)}$ when the update signal $z^{(t)} \approx 1$. It will reflect mostly the new memory $\bar{h}^{(t)}$ when the update signal $z^{(t)} \approx 0$. With the opinion from update gate, the hidden state $h^{(t)}$ is generated using the past hidden state $h^{(t-1)}$ and the new memory $\bar{h}^{(t)}$.

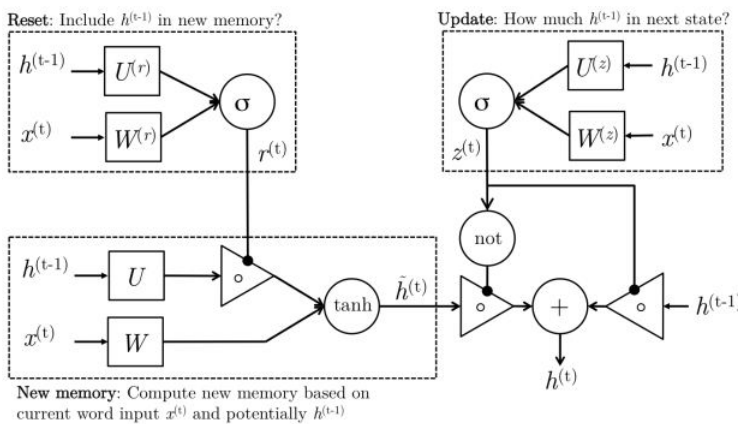


Fig. 5. The Detailed Internals of The Encoder Network

The final hidden state vector represents the thought that the sentence expresses. After learning skip-thought vectors, we fixed the model and used the embedding layer and encoder as a generic feature extractor.

Decoder. The decoder is another RNN but with a conditional GRU. It conditions on the encoder output h_i. The equations are the same as the encoder except that there are bias matrices for update gate, reset gate and hidden state. Different decoders have different parameters for predicting adjacent sentences.

Our coherence measure captures natural reasoning. We measure it by evaluating how well the sentence fits into its context, given the thoughts in a sentence. Therefore, the optimal value of the objective function after we trained and froze the model serves our purpose. Given a tuple (s_{i-1}, s_i, s_{i+1}), the objective optimized the sum of the log-likelihood for the preceding and following words conditioned on the representation of the current sentence:

$$\sum_i (\sum_{t_{i+1}} logP(w^t_{i+1}|w^{<t}_{i+1},h_i) + \sum_{t_{i-1}} logP(w^t_{i-1}|w^{<t}_{i-1},h_i)).$$

We evaluated the coherence of every executive in a given earnings conference and we feed the training model with all the Q&A conversations with people in the same industry.

3.3. Inconsistencies in Emotion and Cognitive Processes

Behavioral finance researchers have been trying to detect strategic information distortion via inconsistencies. Malmendier and Shanthikumar (2014) examined analysts' deceptive behavior by the difference between recommendation and forecast optimism. This novel metric proxies analysts' ingenuine behavior but are also open to alternative explanations. Using only the summary of the text restricts the strength of the proxy since analysts may have nonstrategic and valid reasoning for this with-in analyst discrepancy in recommendations and earnings forecasts explained in the text but was not taken into consideration. In this paper, we advances inconsistencies detection by looking into the original textual evidence and reconstructing the deep structure of emotion and cognitive reactions.

The Four-Factor Model of deception Zuckerman et al. (1981); Zuckerman, Driver, Siegman, and Feldstein (1985) describes the four processes or factors that influence deceivers' behaviors as: (1) Attempted Control is their cue leakage, such as body moments, (2) psychological and physical arousal, such as pupil dilation and blink rate (3) negative affect evoked (4) extra cognitive effort. Since the first two can only be detected with video data, while so far a large fraction of earnings calls are still carried out by teleconference, our context-free measure focuses on emotion and cognitive prospects.

Analogy to polygraph test, the audience needs baseline observation to capture the abnormal changes in emotion and cognitive efforts. In our research we chose the levels exemplified in the MD presentation as our baseline to construct our inconsistencies measure.

Our emotional and cognitive levels conveyed via firm i's presentation ($P^{pres}_{i,t}$), is calculated based on a dictionary approach. We use the software program Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al. (2001)), which provides different words category dictionaries based on the bulk of research on psychology and disclosure. This text analysis software program includes 32 pre-specified word categories related to psychological constructs (e.g., affect, cognition, biological processes). We calculate the degree to which executives use different categories of words, to quantify psychological properties on a certain dimension.

Executives' emotional and cognitive reactions vary from topic to topic according to how the firm is doing in that certain aspect and the high/low external stimulus they endure when making the responses during the MD presentation/Q&A Session. We interpret the observed psychological reactions in the language of executives via a deep structure of two layers: What topics are in the executive's mind; and how they feel towards each topic. In other words, we decompose psychological responses in each document (the presentation or answers by speaker s during an earnings call) to topic representation of the document ($T_{s,t}^{pres}$ and $T_{s,t}^{A}$), which is the output of Topic Modeling, and psychological reaction towards topics ($L_{s,t}^{pres}$and $L_{s,t}^{A}$). That is,

$$P_{s,t}^{pres} \approx L_{s,t}^{pres} \cdot T_{s,t}^{pres}$$

$$P_{s,t}^{A} \approx L_{s,t}^{A} \cdot T_{s,t}^{A},$$

where $L_{s,t}^{pres}$ and $L_{s,t}^{A}$ are matrices with each column representing the psychological responses to each topic in presentation and the answer from executives respectively.

In order to train the latent emotional reaction to each topic under two different scenarios, and generate a distinguishable underlying emotional reaction to each topic, we leverage LASSO regression. Our approach consists of considering sparse approximations

To calculate the topic psychological response matrices $L_{s,t}^{pres}$and $L_{s,t}^{A}$, we apply sparse regression:

$$\min_{L_{s,t}^{pres}} \sum_{s} \left(\frac{1}{2} \parallel P_{s,t}^{pres} - L_{s,t}^{pres} \cdot T_{i,j,t}^{pres} \parallel_2^2 + \lambda_s \parallel L_{s,t}^{pres} \parallel_1 \right)$$

$$\min_{L_{s,t}^{A}} \sum_{s} \left(\frac{1}{2} \parallel P_{s,t}^{A} - L_{i,t}^{A} \cdot T_{s,t}^{A} \parallel_2^2 + \lambda_s \parallel L_{s,t}^{A} \parallel_1 \right),$$

where each λ_s is a tuning parameter that can be learned from the data. Additionally, our empirical results are robust to ridge regression, which is calculated as replacing the norm $\|\cdot\|_1$ by $\|\cdot\|_2^2$ in the above optimization problem.

We further detect the topic-wise psychological reaction differences between MD presentation and responses to analysts. Due to real time responding pressure and lack of assistance, executives face strong external stimulus in Q&A

session. As a result, they demonstrate different emotional reactions compared to that in the Q&A session. Consistent with the four-factor model of deception, concealing usually requires more cognitive energy to formulate the deceptive speech, make responses consistent and therefore arouse emotional fluctuation. We measure the inconsistencies in emotion and cognition using the distance between topic psychological reaction matrices $L_{s,t}^{pres}$ and $L_{s,t}^{A}$:

$$DP_{s,t} = \parallel L_{s,t}^{A} - L_{s,t}^{pres} \parallel_2 .$$

This measure of emotional discrepancy alleviates concerns about unobserved speaker characteristics by differencing it out. It is a clean measure to detect psychological inconsistencies on the same topics.

Specifically, consistent with the finding in (Zuckerman et al. (1981, 1985)), we incorporate emotions of anger, anxiety, sadness, affection. In the same spirit of cognitive factors in deception, we consider uncertain and causal.

4. Data Description

Our sample include 6554 earnings conference call transcripts from the S&P 500 Companies' between 2011 to June-22 2015. The unique number of participants is 20320 including executives and analysts.

To construct our measure of evasiveness, we analyzed the transcripts on the level of questions (answers) by specific speakers per round in each conference. We often observed that analysts raised questions and were not satisfied with the answers at once. They often demanded more details or asked another add-on question, which was highly relevant to the answers they got from executives. Therefore, we aggregate the add-on questions (all the complementary answers) in the same round as an analyzed unit. The Q&A round is defined as different conversations separated by operators that connect one questioner at one time.

Different from the way we processed data to construct evasiveness measure, our observation unit when computing the coherence measure and inconsistencies are on the level of speakers per earnings call. We evaluate each speaker's coherent level using all his/her responses during the earnings call. The inconsistencies measure is furthermore based on the speakers' language , who both gave an MD presentation and responses to questions later.

The analysts' forecast data is from I/B/E/S database. We extract returns data from the CRSP database; and the balance-sheet information of the firms are from Compustat database.

5. Listening Between the Lines: Predict Earnings

In this section, we investigate whether language patterns detected during earnings conference calls provide extra information about firms' fundamentals. In order to test whether evasiveness, incoherence, and emotional & cognitive inconsistencies convey novel information, we use two measures of earning news as dependent variables: standardized unexpected earnings (SUE) and standardized analysts' forecast errors (SAFE).

Following Chan, Jegadeesh, and Lakonishok (1996),we compute firm i 's standardized unexpected earnings at time t as a seasonal random walk model:

$$SUE_{i,t} = \frac{E_{i,t} - E_{i,t-4}}{\sigma_{i,t}}$$

where $E_{i,t}$ is the quarterly earnings per share just announced before the earnings conference call, $E_{i,t-4}$ is the earnings per share last year of the same quarter, $\sigma_{i,t}$ is the standard deviation of unexpected earnings, and $E_{i,t} - E_{i,t-4}$ is for the past eight quarters. We drop the firms with missing earnings data for the most recent 10 quarters and assume zero trends for firms with fewer than 3 years of earnings data.

Another measure of earnings surprise is standardized analysts' forecast errors (SAFE). I included control variables based on a firm's lagged earnings, size, book-to-market ratio, trading volume, three measures of recent stock returns, analysts' earnings forecast revisions, and analysts' forecast dispersion. SAFE is calculated as the median equity analysts' forecast errors adjusted by earnings volatility. We use the most recent forecast error prior to three days before the earnings call.

Our control variables include lagged earnings, size (log(Market Equity)), logarithm of book-to-market ratio and trading volume following Fama and French (1992), all of the above are evaluated at the end of the preceding year. The trading volume is the log of annual shares traded adjusted by outstanding shares at the end of the previous year. We also control stock returns and analysts' forecasts.

Our three control variables for a firm's recent returns are based on an earnings announcement event study methodology. Using benchmark returns based on Fama and French (1993) three-factor model. We denote the earnings call date as t and the next quarterly earnings announcement date as day 0. We include two control variables for a firm's recent returns, the cumulative abnormal return from the [t, -3] trading day window ($\text{FFCAR}_{t,-3}$) and the abnormal return on day -2 ($\text{FFCAR}_{-2,-2}$). Our third control variable is related to the Jegadeesh and Titman (1993) return momentum effect, which is based on firms' relative returns over the previous calendar year, prior to the date of earnings call, FFAlpha $_{t-222,t-1}$. It is the estimated intercept from the event study regression; It measures the in-sample cumulative abnormal return over the previous calendar year.

We also controlled the information from analysts' forecasts, both in perspectives of revision and dispersion. For analysts' forecast revisions, we use a three-month moving average of past changes in earnings forecast:

$$REV_{i,t} = \sum_{j=0}^{3} \frac{f_{i,t-j} - f_{i,t-j-1}}{p_{i,t-j-1}}$$

where $f_{i,t}$ is the median analyst's quarterly forecast and $p_{i,t-j-1}$ is the prior month's stock price. We calculate forecast dispersion as the standard deviation of analysts' most recent quarterly earnings forecasts prior to the next earnings announcement and scaled by earnings volatility ($\sigma_{i,t}$).

We estimate the predictive ability of language patterns on earnings using polled ordinary least squares regressions, and standard errors clustered by quarter and industry division based on SIC codes.

We report the estimates of the prediction power of future firm performance using two earnings measures in Table 2. Evasiveness measures consistently predict a lower firm performance, with both SUE and SAFE as the measure of earnings. It is positively and statistically significant at the 95% level. The economic magnitude is: the conditional expectation of SUE will be 0.0363×4=0.1452 standard deviation lower when evasiveness measures decrease from two standard deviations above its mean to two standard deviations below. Similarly, the predicted next quarter earning will be 0.1484 standard deviations lower as coherence measures move from two standard deviations below to above the mean. Emotional & Cognitive inconsistencies suggest a lower future earning in the specification of predicting SUE, however, it is not significant when we use SAFE to measure the future earnings.

Among the control variables, we find lagged earnings and return momentum effect are strong predictors for the future earnings.In sum, after controlling for recent stock returns, analyst forecast and revisions and other financial indicators, we find that language clues provides extra information in predicting future firm performance.

Table 2: Listening Between the Lines: Predict Earnings

This table shows estimates of the predictive ability of the language patterns to predict next quarter earnings (SUE and SAFE) using OLS regressions with firm fixed effect. For each measure of earnings, we display two specifications: only financial and analysts' forecast predictors, and with extra information from language. We clustered standard errors by calendar quarter and industry division.

	SUE		SAFE	
	(1)	(2)	(3)	(4)
Lag(SUE)	0.0296**	0.0302**		
	(0.0143)	(0.1430)		
Lag(SAFE)			0.0229**	0.0253**
			(0.0104)	(0.1148)
Forecast dispersion	-0.0568	-0.0551	-0.1886	-0.1875
	(0.0344)	(0.0347)	(0.1703)	(0.1708)
Forecast Revisions	0.0498**	0.0526	0.0139	0.0147
	(0.0231)	(0.0253)	(0.0160)	(0.0159)
log(Market Equity)	-0.2513**	-0.2526**	-0.1362	-0.1414
	(0.1024)	(0.0327)	(0.0911)	(0.0920)
log(Book/Market)	-0.2549**	-0.2646***	-0.1092	-0.1187
	(0.1005)	(0.1008)	(0.0874)	(0.0886)
log(Share Turnover)	0.0447	0.0471	-0.1390***	-0.1363***
	(0.0441)	(0.0440)	(0.0499)	(0.0501)
$\alpha_{t-222,t-1}$	0.1125***	0.1111***	0.0523***	0.0521***
	(0.0244)	(0.0245)	(0.0187)	(0.0185)
$FFCAR_{t,-3}$	0.0314	0.0304	0.0289*	0.0285*
	(0.0221)	(0.0223)	(0.0173)	(0.0170)
$FFCAR_{-2,-2}$	0.0082	0.0057	0.0222	0.0213
	(0.0262)	(0.0262)	(0.0148)	(0.0147)
Evasiveness		0.0363**		0.0373**
		(0.0168)		(0.0175)
Coherence		0.0371**		0.0178*
		(0.0273)		(0.0096)
Inconsistencies		-0.0290*		-0.0142
		(0.0178)		(0.0200)
Constant	-0.0021	-0.0025	0.0064	0.0064
	(0.0211)	(0.0212)	(0.0151)	(0.0154)
Observations	4,343	4,343	3,684	3,684
Clusters	148	148	132	132
Firm Fixed Effect	Yes	Yes	Yes	Yes
R-squared	0.1087	0.1483	0.2002	0.2605

Robust standard errors in parentheses.
** p<0.01, ** p<0.05, * p<0.1

6. Listening Between the Lines: Predict Stock Returns

Since evasiveness and incoherence measures predict firms fundamentals, we further investigate whether investors immediately respond fully to the information embedded in language

We first assess whether language clues predict stock returns by examining the contemporaneous stock price changes. We measure market return reactions using both raw and abnormal returns. The raw return we use is the next day return following the earnings call. While the abnormal returns is measured as the abnormal return on day 1 using the benchmark of Fama and French (1993) three-factor model. In this section we define the earnings call date as the event date.

In each regression, we control the predictors that have proved to have significant explanatory power for future stock returns (Fama and French (1993); Daniel, Grinblatt, Titman, and Wermers (1997)). These control variables include firm characteristics and risk factors. Specifically, for returns we control the short term (abnormal return on the earnings call date and two days before, $FFCAR_{0,0}$, $FFCAR_{-1,-1}$, $FFCAR_{-2,-2}$), middle-term (the abnormal return one month before $FFCAR_{-30,-3}$) and long-term (the momentum effect for the past year $\alpha_{-252,-31}$). For past earnings, we control the recent earnings announcement (SUE). Finally, to take the priced factors Fama and French (1992) into consideration, we control firm size, book-to-market ration, and trading volume. The calculation of these control variables are the same as in Section5.

Using two different specifications, we test whether language patterns provide incremental information beyond the predictors of returns, earnings, and priced risk factors. We report the results in Table 3. Besides using next-day abnormal return, we also use next-day raw return as a dependent variable to ensure that our results are robust to the benchmark process. We compute robust standard errors. In this way we account for the correlations between firms' stock returns within the same quarter and industry division.

Table 3 shows in both specifications, evasiveness measures robustly predict lower stock returns. The estimates of evasiveness are positively significant in column (2) and (4). The economic magnitude for the coefficient is that the next-day abnormal return ($FFCAR_{1,1}$) will drop 2.87% when evasiveness measures decreases by one standard deviation. This empirical evidence reveals that executives' avoidance of giving relevant information is perceived as bad news and punished by the market. Coherence measure reveals the tendency of lowering next day raw close-to-close return at the significance level of 10%. In a highly efficient market, we do not expect many control variables to have strong predictive power of returns. This coincides with the few significant control variables in Table 3.

Since evasiveness measure provides significant incremental explanatory power for the next-day return, we investigate what the economic magnitude is when caused by the market's sluggish reaction to evasive executives. We compute the differences in cumulative abnormal return between evasive and forthright companies with Fama-French three-factor model as the benchmark for expected returns. The evasive (forthright) companies are those with evasive-

Table 3: Listening Between the Lines: Predict Stock Return

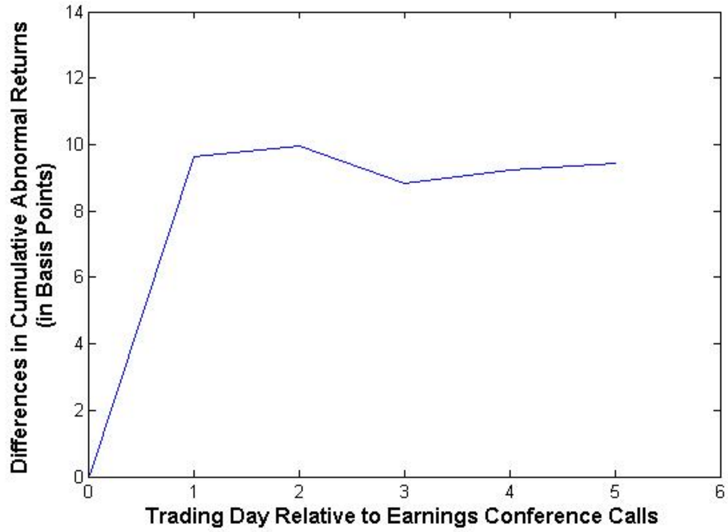
This table shows estimates of the predictive ability of the language patterns to predict on the following day ($Ret_{+1,+1}$ and $FFCAR_{+1,+1}$) using OLS regressions with firm fixed effect. For each measure of earnings, we display two specifications: only financial and recent returns, and with extra information from language. We clustered standard errors by calendar quarter and industry division.

	$Return_{1,1}$		$FFCAR_{1,1}$	
	(1)	(2)	(3)	(4)
$FFCAR_{0,0}$	0.5303**	0.5442*	0.1148**	0.1068*
	(0.2543)	(0.3268)	(0.0563)	(0.5793)
$FFCAR_{-1,-1}$	0.0109	0.0116	0.0252	0.0261
	(0.0246)	(0.0247)	(0.0248)	(0.0254)
$FFCAR_{-2,-2}$	0.0105	0.0099	0.0192	0.0185
	(0.0204)	(0.0204)	(0.0203)	(0.0203)
$FFCAR_{-30,-3}$	0.0288	0.0290	0.0209	0.0213
	(0.0228)	(0.0228)	(0.0225)	(0.0229)
$\alpha_{-252,-31}$	-0.0195	-0.0202	0.0783	0.0781
	(0.0244)	(0.0244)	(0.0572)	(0.0482)
SUE	0.0146	0.0141	0.0230	0.0202
	(0.0203)	(0.0202)	(0.0186)	(0.0180)
log(Market Equity)	0.0029	0.0034	-0.2245	-0.2337
	(0.0157)	(0.0158)	(0.1430)	(0.1420)
log(Book/Market)	-0.0180	-0.0211	0.1178	0.1014
	(0.0239)	(0.0246)	(0.1191)	(0.1189)
log(Share Turnover)	0.0220	0.0220	-0.0954**	-0.1027**
	(0.0304)	(0.0304)	(0.0413)	(0.0403)
Evasiveness		0.0246**		0.0287**
		(0.0113)		(0.0134)
Coherence		0.0051*		0.0391
		(0.0029)		(0.0281)
Inconsistencies		-0.0148		-0.0137
		(0.0135)		(0.0166)
Constant	0.0008	0.0009	0.0000	-0.0048
	(0.0241)	(0.0241)	(0.0179)	(0.0175)
Observations	4,265	4,265	4,265	4,215
Clusters	136	136	136	136
Firm Fixed Effect	Yes	Yes	Yes	Yes
R-squared	0.1184	0.0038	0.1203	0.1195

Robust standard errors in parentheses.
** p<0.01, ** p<0.05, * p<0.1

ness measures lower (higher) than the median of evasiveness among the same industry division in the same quarter. Figure 6 pictures the difference in cumulative abnormal return between evasive and forthright companies. We can see that the return premium for forthright companies reached to its fullest one day after the earnings call and stayed relatively stable afterwards.

Fig. 6. Firms' valuations after earnings call.



7. Evasiveness Based Trading Strategy

The lingering differences in the cumulative abnormal returns between evasive and forthright firms hinders a potential trading strategy based on evasiveness measure. In this section, we explore this possibility and the boundaries of profiting from such a strategy.

7.1. Evasiveness Trading Strategy Returns

Our Evasiveness based trading strategy is: For every date that has more than seven earnings call happen, we form two equal-weighted portfolios based on the way executives handle questions during the earnings call. Our long portfolio consists of all the companies with evasiveness measure that are greater than the median of the same industry division in the same quarter. The rest of the companies are in the short portfolio. We hold both portfolio for one trading day and rebalance at the end of the next trading day.

Table 4 shows the daily risk-adjust return using three different risk adjustments for the contemporaneous market, size, book-to-market, and momentum factors. They are CAPM Model, Fama-French three-factor model and Carhart four-factor model (Carhart (1997)). We report the α and factor loadings using time series regression of the long-short evasiveness based portfolios. The standard errors of the coefficients are computed using the White (1980) heterskedasticity-consistent covariance matrix.

Table 4: Sensitivity of Evasiveness-Based Trading Returns to Trading Cost Assumptions

This table shows the daily risk-adjusted returns (α) from an evasiveness-based trading strategy. The first regressions uses CAPM Model to adjust risk. The second regression use the Fama and French (1992) three-factor model to adjust the trading strategy returns for the impact of contemporaneous market (Market), size (SMB), and book-to-market (HML) factors. The last regressions use the Carhart (1997) four-factor model to account for the momentum factor (UMD). This table reports the alpha and loadings from the time-series regression of the long-short evasiveness-based portfolio returns on each of the four factors. We exclude the rare days in which there are less than 7 firms holding conference calls. We compute all coefficient standard errors using the White (1980) heteroskedasticity-consistent covariance matrix. The robust standard deviations are in parentheses.

	CAPM	**Three-Factor**	**Four-Factor**
Alpha	0.2719**	0.2631**	0.2657**
	(0.1134)	(0.1134)	(0.1164)
Market	-0.0001	0.0004*	0.0003
	(0.0011)	(0.0002)	(0.0011)
SMB		-0.0022	-0.0021
		(0.0028)	(0.0028)
HML		-0.0020**	-0.0018*
		(0.0009)	(0.0010)
UMD			0.0003
			(0.0018)
Earnings Call Days	158	158	158
Adjusted R^2	0.0002	0.0049	0.0052

*,**,and *** denote statistical significance at the 10%, 5% and 1% level, respectively.

Consistent with Table 3, Table 4 confirms that an evasiveness-based trading strategy would earn substantial risk-adjusted returns under the idea trading condition of no transaction cost and zero price impact. The excess return is 0.27% per day adjusted by the Fama-French three-factor model. The excess return is positively significant with all three risk adjustment models. This confirms that our evasiveness based trading strategy yields excess return robustly.

7.2. *Sensitivity of Evasiveness Trading Strategy Returns*

Since our trading strategy profits in a frictionless market, we further examine the sensitivity of trading return based on evasiveness strategyto trading cost.[8] Although, recently the rapid development in FinTech makes trading with lower nominal transaction cost or even zero cost possible (Examples of such investing apps include Robinhood, Openfolio,FeeX, etc), according to microstructure theory, invisible transaction cost still exists. Such implicit trading costs include price impact and costs related to the bid-ask spread (Anand, Irvine, Puckett, and Venkataraman (2012); Frazzini, Israel, and Moskowitz (2012)). We calculate the excess return with a different amount of round-trip

[8]Note: In this subsection, we calculate the annual return as if one could trade according to the evasiveness based strategy for every trading day. The result is for illustration purpose. For those trading days that do not have enough earnings calls, investors cannot trade according to this strategy.

transaction costs, from 0 to 35 basis points.

Table 5 shows that our evasiveness based trading strategy remains profitable for transaction costs lower than 25 basis points. The annualized raw return without transaction cost is 66.08%. Note that this is based on the assumption that investors can trade every day in a year (252 trading days). However, the trading strategy is only applicable when there are companies holding earnings calls. The unique number of days that companies hold earnings calls varies per year.

Table 5: Sensitivity of Evasiveness-Based Trading Returns to Trading Cost Assumptions

This table shows the evasiveness-based trading strategy's profitability given different levels of transaction costs . We compute the trading strategy returns for 8 alternative assumptions about a trader's round-trip transaction costs: 0, 5, 10, 15 . . . or 35 basis points (bps, 0.01%) per round-trip trade. The abnormal annualized cumulative evasiveness-based returns for each assumption appear below. The risk-adjustment are based on the CAPM model, Fama-French three-factor model and Carhart (1997) model.

$Trading$ $Costs(bps)$	$AbnormalAnnualized$ $Returns : CAPM$	$AbnormalAnnualized$ $Returns : FF3$	$AbnormalAnnualized$ $Returns : C4$	$RawAnnualized$ $Returns$
0	0.6851	0.6629	0.6696	0.6608
5	0.5588	0.5366	0.5433	0.5246
10	0.4324	0.4103	0.4170	0.3905
15	0.3061	0.2839	0.2906	0.2585
20	0.1797	0.1576	0.1643	0.1286
25	0.0534	0.0313	0.0380	0.0006
30	-0.0729	-0.0950	-0.0884	-0.1254
35	-0.1993	-0.2214	-0.2147	-0.2495

Note: The results are annualized return in the case that investors can continuous trade for 252 days in a year.

8. Conclusion

Our paper is the first to propose a big data approach so as to evaluate firms' disclosure strategies and their market responses. Our results show that both evasiveness and incoherence provide incremental information to capture opaque aspects of the firm. This information is beyond quantitative measures of the firm and analysts' forecasts. We also find the evasiveness measure to be a significant determinant of a company's next-day return. The market's one-day delay in absorbing this information detected sheds light on a trading strategy accordingly.

Besides analysis from the investors' point of view, our results reveal that the market is reasonably efficient. Although the qualitative information has been almost left out from the asset pricing literature, the market is able to adjust within one trading day. This finding also highlighted the role of equity analysts in forming the information disclosure. The slight under-reaction of the market motivates analysts to probe the underlying truth by raising challenging questions.

Lastly, our research methodology which integrates data science into asset pricing provides a generic frame work for future language processing. For example, forecast policy changes from a president's press interview. Businessmen would also predict the interest rate by analyzing the Federal Reserve Chairman's language. For all these language mining tasks, our paper provides a reliable, consistent and economical way to extract soft information embedded in language.

References

Anand, A., Irvine, P., Puckett, A., Venkataraman, K., 2012. Performance of institutional trading desks: An analysis of persistence in trading costs. Review of Financial Studies 25, 557–598.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. Journal of machine Learning research 3, 993–1022.

Brockman, P., Cicon, J., 2013. The information content of management earnings forecasts: an analysis of hard versus soft information. Journal of financial research 36, 147–174.

Buller, D. B., Burgoon, J. K., Daly, J., Wiemann, J., 1994. Deception: Strategic and nonstrategic communication. Strategic interpersonal communication pp. 191–223.

Carhart, M. M., 1997. On persistence in mutual fund performance. The Journal of finance 52, 57–82.

Chan, L. K., Jegadeesh, N., Lakonishok, J., 1996. Momentum strategies. The Journal of Finance 51, 1681–1713.

Chen, S. F., Goodman, J., 1996. An empirical study of smoothing techniques for language modeling. In: *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp. 310–318.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 .

Daniel, K., Grinblatt, M., Titman, S., Wermers, R., 1997. Measuring mutual fund performance with characteristic-based benchmarks. The Journal of finance 52, 1035–1058.

Davis, A. K., Piger, J. M., Sedor, L. M., 2012. Beyond the numbers: Measuring the information content of earnings press release language. Contemporary Accounting Research 29, 845–868.

Fama, E. F., French, K. R., 1992. The cross-section of expected stock returns. the Journal of Finance 47, 427–465.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of financial economics 33, 3–56.

Fischer, P. E., Verrecchia, R. E., 2000. Reporting bias. The Accounting Review 75, 229–245.

Fishman, M. J., Hagerty, K. M., 2003. Mandatory versus voluntary disclosure in markets with informed and uninformed customers. Journal of Law, Economics, and organization 19, 45–63.

Frazzini, A., Israel, R., Moskowitz, T. J., 2012. Trading costs of asset pricing anomalies .

Graham, R., 2007. Theory of cognitive dissonance as it pertains to morality. Journal of Scientific Psychology pp. 20–23.

Green, T. C., Jame, R., 2013. Company name fluency, investor recognition, and firm value. Journal of Financial Economics 109, 813–834.

Griffiths, T. L., Steyvers, M., 2004. Finding scientific topics. Proceedings of the National academy of Sciences 101, 5228–5235.

Hauch, V., Blandón-Gitlin, I., Masip, J., Sporer, S. L., 2015. Are computers effective lie detectors? a meta-analysis of linguistic cues to deception. Personality and Social Psychology Review 19, 307–342.

Healy, P. M., Palepu, K. G., 2001. Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. Journal of accounting and economics 31, 405–440.

Hirshleifer, D., 2001. Investor psychology and asset pricing. The Journal of Finance 56, 1533–1597.

Hollander, S., Pronk, M., Roelofsen, E., 2010. Does silence speak? an empirical analysis of disclosure choices during conference calls. Journal of Accounting Research 48, 531–563.

Javers, E., 2010. Cia moonlights in corporate world. Politico. com .

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. The Journal of finance 48, 65–91.

Johnson, M. K., Raye, C. L., 1981. Reality monitoring. Psychological review 88, 67.

Jovanovic, B., 1982. Truthful disclosure of information. The Bell Journal of Economics pp. 36–44.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Skip-thought vectors. In: *Advances in neural information processing systems*, pp. 3294–3302.

LAING, J. R., 2006. Is your ceo lying. Barron's, June 26, 2006.

Larcker, D. F., Zakolyukina, A. A., 2012. Detecting deceptive discussions in conference calls. Journal of Accounting Research 50, 495–540.

Lee, Y.-J., 2012. The effect of quarterly report readability on information efficiency of stock prices. Contemporary Accounting Research 29, 1137–1170.

Li, F., 2010. Textual analysis of corporate disclosures: A survey of the literature. Journal of accounting literature 29, 143.

Malmendier, U., Shanthikumar, D., 2014. Do security analysts speak in two tongues? Review of Financial Studies 27, 1287–1322.

Mazar, N., Amir, O., Ariely, D., 2008. The dishonesty of honest people: A theory of self-concept maintenance. Journal of marketing research 45, 633–644.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .

Milgrom, P. R., 1981. Good news and bad news: Representation theorems and applications. The Bell Journal of Economics pp. 380–391.

Miller, B. P., 2010. The effects of reporting complexity on small and large investor trading. The Accounting Review 85, 2107–2143.

Narayanan, M., 1985. Managerial incentives for short-term results. The Journal of Finance 40, 1469–1484.

PCAOB, 2008. Proposed auditing standards related to the auditor's assessment of and response to risk. PCAOB release No. 2008-005.PCAOB Washington, DC .

Pennebaker, J. W., Francis, M. E., Booth, R. J., 2001. Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates 71, 2001.

Shin, H. S., 1994. News management and the value of firms. The RAND Journal of Economics pp. 58–71.

Stein, J. C., 1989. Efficient capital markets, inefficient firms: A model of myopic corporate behavior. The Quarterly Journal of Economics pp. 655–669.

Stocken, P. C., 2000. Credibility of voluntary disclosure. The RAND Journal of Economics pp. 359–374.

Sutskever, I., Vinyals, O., Le, Q. V., 2014. Sequence to sequence learning with neural networks. In: *Advances in neural information processing systems*, pp. 3104–3112.

Tetlock, P. C., 2007. Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance 62, 1139–1168.

Tetlock, P. C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. The Journal of Finance 63, 1437–1467.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: Journal of the Econometric Society pp. 817–838.

Zuckerman, M., DePaulo, B. M., Rosenthal, R., 1981. Verbal and nonverbal communication of deception. Advances in experimental social psychology 14, 1–59.

Zuckerman, M., Driver, R. E., Siegman, A. W., Feldstein, S., 1985. Telling lies: Verbal and nonverbal correlates of deception. Multichannel integrations of nonverbal behavior pp. 129–147.