Conversation Analytics: Can Machines Read between the Lines in Real-Time Strategic Conversations?

Yanzhen Chen
HKUST
imyanzhen@ust.hk

Huaxia Rui
The University of Rochester
huaxia.rui@simon.rochester.edu

Andrew Whinston
The University of Texas at Austin
abw@uts.cc.utexas.edu

Abstract

Strategic conversations involve one party with an informational advantage and the other with an interest in the information. This paper proposes machine-learning based measures to quantify the degrees of evasiveness and incoherence of the informed party during real-time strategic conversations. The specific empirical context is the questions and answers (Q&A) part of earnings conference calls during which managers endure high pressure as they face analysts' scrutinizing questions. Being reluctant to disclose adverse information, managers may resort to evasive answers and sometimes respond less coherently due to increased cognitive load. Using data from the earnings calls of the S&P 500 companies from 2006 to 2018, we show that the proposed measures predict worse next-quarter earnings. Moreover, the stock market perceives incoherence as a negative signal. This paper contributes methodologically by developing two novel machine-powered measures to automatically evaluate behavioral cues during real-time strategic conversations. The proposed analytical tools are particularly beneficial to resource-constrained and informationally disadvantaged parties such as retail investors who may not be able to effectively trade on signals buried deep in unstructured conversational data.

Keywords: conversation analytics, topic modeling, deep learning, conference calls.

1. Introduction

We humans are good at, and often proud of our ability to listen between the lines — that we can infer information implicitly conveyed and sometimes inadvertently revealed by a speaker. Being able to discern subtleties in conversations is such a unique human skill that we generally consider it as an important aspect of intelligence that machines can hardly emulate. This, however, is changing, as we witness a new technology revolution.

With the rapid advances in machine learning methods and computing power over the past few decades, the performance of computer programs has increased so dramatically that we now routinely refer to them as artificial intelligence, or AI. Not surprisingly, companies in almost all industries are racing to augment their human intelligence with machine intelligence. Indeed, if we think of human intuitions and experiences as algorithms embodied in biological rather than artificial neurons, it is only logical that AI algorithms, equipped with faster computation and more data, may eventually surpass where our intuitions and experiences can take us.

The broad context of this paper is real-time strategic conversations where one party has an informational advantage while the less informed party typically has an interest in such information. Our objective is to measure the degrees of evasiveness and incoherence during such real-time conversations using machine learning algorithms. We believe these two measures are particularly meaningful in such contexts because of the following three ingredients: the incentives of the two parties are not aligned; the informed party usually has certain flexibilities in whether, to what extent, and how to reveal information; and the informed party often needs to improvise in answering unexpected questions due to the real-time nature of the conversation. As social psychologists (Goffman 1959) have long recognized, there are two types of expressions during social interactions: expressions given and expressions given off. Expressions given are the verbal or non-verbal signals we intend for others to receive while expressions given off are those we do not intend for others to receive. Traditionally, the less informed party relies on their shrewdness and acumen to detect expression given off. This paper is based on the premise that the performance of modern machine learning algorithms has reached the tipping point of being able to detect expression given off and may sometimes even do so better than humans.

The specific context for our empirical evaluation is earnings conference calls where managers of a public company discuss the financial results of a reporting period. Unlike the management discussion part of an earnings conference call which is typically scripted and well-prepared, the questions and answers (Q&A) part is conversational and impromptu. Managers need to improvise in answering questions posed to them, which are often difficult to predict. They must also do so under

URI: https://hdl.handle.net/10125/102830
978-0-9981331-6-4

time constraints without real-time support from their staff members. Such a "conversational dance" between managers and analysts is an example of real-time strategic conversation. Moreover, once we measure the degrees of evasiveness and incoherence of such a conversation, we can study their implications using objective quantities such as stock market reactions and next-quarter earnings.

Using the earnings conference calls of the S&P 500 companies from January 2006 to December 2018, we find that both evasiveness and incoherence forecast worse next-quarter earnings, and incoherence also predicts worse next-day abnormal stock return. These results demonstrate the value of our proposed measures and also provide clear policy implications for regulators to mitigate information processing inequalities between institutional and retail investors which otherwise may "deter stock market participation and impede economic growth" (Lev, 1988, Blankespoor et al. 2020, page 21).

2. Conversation analytics: literature foundation

Conversation analysis focuses on what speakers do to deliver and interpret information in oral interaction (Liddicoat 2021). Using recorded service conversations or live comments on video streaming platforms respectively, Li et al. (2020) and Zhang et al. (2020) are the only studies exploring the interaction or responses between content providers and receivers. Though its application in business research is still in its infancy, conversation analytics can potentially help manage customer service conversations to improve satisfaction (Li et al. 2020) and predict sales (Zhang et al. 2020). Lacking reliable tools to analyze strategic conversations limits the development of this line of research.

The nascent field of non-cooperative/semi-cooperative dialogues in computer science and computational linguistics shed some light on analyzing conversation strategies. This literature explores conversations when an agent tries to manipulate her adversary (Lewis et al. 2017). Such a setting has been barely explored but is particularly relevant for gauging information proactively, persuading, negotiating against by hiding information (Cuayáhuitl et al., 2015) or learning and thus playing tactics (Zhou et al., 2019), and even deceiving other players (Lewis et al., 2017). A growing interest in the studies of strategic conversation also explores reasoning together with strategy generation (Lewis et al., 2017). These works learn conversation strategies from dialogues by mostly using (deep) reinforcement learnings with a reward function that provides positive feedback only when it meets the user's goals. Our paper is in the same spirit because we try to quantify two specific aspects of real-time strategic conversation. However, instead of assisting humans in non-cooperative conversations with algorithms, our approach focuses on human-augmented AI measures of strategic conversation, rather than AI-augmented human interactions. In our empirical setting, analysts are well-trained and experienced investors with insightful questions in mind. We aim to construct scalable methods to evaluate strategic conversations elicited by expert intelligence, including the aspects of evasiveness and incoherence.

The notion of evasiveness in conversation analytics is inspired both by the literature and by the conventional wisdom. An evasive answer occurs when the informed party uses shifting & refocusing tactics by, for example, briefly commenting on the thorny aspect but dwelling on the favorable parts, or replacing the original question with a related but different one. The strategy of providing an evasive answer, although intuitive, has its theoretical root in the classical work of Crawford and Sobel (1982) and a class of game-theoretical models called the persuasion games. In a persuasion game, the sender, who has an information advantage, can withhold information, but cannot lie because the receiver can verify any information reported. Under some mild conditions, Milgrom (1986) showed that the receiver's unique equilibrium strategy is the so-called "assume the worst" strategy in which the receiver infers that leads to the least favorable decision for the sender, conditional on the information available. This result confirms the intuition that an evasive strategy should be negatively interpreted by a rational receiver.

For our particular empirical context, such a strategy is often formulated as the management obfuscation hypothesis. The hypothesis argues that managers obfuscate information when firm performance is unsatisfactory so that the processing cost of adverse information would increase, which may delay or even prevent an adverse stock market reaction to the information. Because outright lying and silence during the Q&A of an earnings call are out of the question, answering in an evasive way that clouds, disguises, or even distorts private information could be a feasible option for some managers (Khalmetski et al. 2017). Indeed, there are empirical evidences suggesting that managers sometimes present less relevant information to avoid giving a direct answer (Larcker and Zakolyukina 2012). Naturally, analysts and investors are interested in evaluating behavioral or linguistic cues that suggest evasiveness. In fact, the value proposition of some companies is exactly their expertise to uncover deceptive behavior, including the detection of executives' evasive responses. For example, Business Intelligence Advisors (BIA), a hedge-fund consulting firm, hires former Central Intelligence Agency (CIA) employees to analyze language clues (Javers 2010).

Specialists at BIA try to gauge whether managers directly answer questions rather than dancing around the topics. An important type of behavioral cues they look for is "management replies larded with irrelevant specifics." For example, by analyzing how managers of UTStarcom diverted questions during one earnings call, BIA successfully predicted their profitability. In accounting practice, the Public Company Accounting Oversight Board (PCAOB) also has emphasized that auditors should consider observing or reading transcripts of earnings calls, as a part of the procedure when assessing material misstatement (PCAOB 2008). Besides revealing information about firm fundamentals, how definitive and direct managers answer questions during conference calls also seems to influence stock price movement. For example, 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.

To construct the evasiveness measure, we use the Latent Dirichlet Allocation (LDA) (Blei et al. 2003), an unsupervised algorithm that relies on a set of parametric assumptions and the co-occurrence patterns of words in different documents to uncover latent topics in each document. Using the LDA output, we summarize each question or answer using a topic vector where each element of the numerical vector indicates the proportion of the corresponding topic being covered in that document. Once we represent each document as a topic vector from the same topic space, we compare the topic vectors of a pair of question and answer to measure how "thematically matched" they are. The more similar an answer is to the corresponding question in terms of their topic coverage, the less evasive the answer is.

Our incoherence measure is motivated by theories and evidences from the psychology literature on deception. According to this literature, deception induces greater cognitive load out of a need to avoid contradicting former statements or facts that the observer may know about. As a result, deceptive accounts appear less coherent (Hauch et al. 2015). In our empirical context, managers who hide or distort some adverse information will experience a greater cognitive load which may lead to less coherent answers. To construct the incoherence measure, we quantify how smooth a manager's thoughts flow within the answer. This is inherently a challenging task, if possible at all, because thought is a very abstract construct. Luckily, recent advances in deep learning research allow us to encode the thought carried by a sentence using a numerical vector. In particular, we rely on a state-of-the-art deep learning model which is native to coherence in text tasks (Lan et al. 2019) to measure the smoothness of the thought flow. Our incoherent measure learns representations of conversations by maximizing the likelihood of sentences orders which is similar to Yarats and Lewis (2018).

3. Measure constructions

3.1. Evasiveness

Our measure of evasiveness is designed for conversational data and is based on statistical language modeling. It can be computed in real-time without human involvement, hence is scalable. Moreover, the measure is domain agnostic because the underlying topic space is not pre-defined.

Before describing our approach, we first review how current business literature measure concepts that are related to our notion of evasiveness. The first related concept is informativeness which is often measured by the number of words, i.e., length (Lee 2012, Miller 2010, Ertugrul et al. 2017, Loughran McDonald 2014). We believe such a crude measure is too noisy and inaccurate to capture evasiveness because evasiveness answers can be long and direct answers can be short. Two more closely related concepts are readability and vagueness. Readability is often measured by word complexity (Biddle et al. 2009, Miller 2010, Lehavy et al. 2011, Lawrence 2013) and the presence of grammatical errors (Hwang Kim 2017), while vagueness is typically measured by the proportion of words from a pre-defined vagueness lexicon: a category of words that indicate uncertainty (Loughran McDonald 2013, Ertugrul et al. 2017, Dzieliński et al. 2017). The underlying motivation for this approach is that these function words (e.g., may, might, could) may reflect the attitude or mood of the manager. A big drawback of this approach, however, is that a person's use of these function words can also result from the need to be polite. According to one of the most influential politeness theories (Brown et al. 1987), speaking indirectly is a strategy to mitigate face threats directed to the listener during social interaction. Thus, the use of words from the vagueness lexicon may alternatively reflect the need to reduce face threats whenever a disagreement arises rather than indicating an evasive strategy.

Our approach is fundamentally different from the abovementioned approaches all of which are implemented at the lexical level. We aim to evaluate evasiveness directly at the semantic level[1] which is how

[1] By semantic level, we meant compositional semantics.

we humans detect evasiveness. Of course, the challenge is to use an algorithm to evaluate how the content of an answer "matches" that of the corresponding question. Intuitively, we need to represent a question and an answer as two topic vectors where each element of a vector indicates the proportion or weight of that topic covered by the question or answer. We can then measure how "relevant" an answer is in terms of properly touching on each topic in the corresponding question by comparing the similarity of their topic vectors.

While this semantic-level approach can be implemented using any algorithm that returns a topic vector for each document (i.e., a question or an answer), a natural choice is the LDA topic model, which allows us to represent each document as a dense topic vector once we specify the number of topics in the topic space. The LDA model has been widely applied in a variety of domains. In the IS literature, researchers have used LDA to compare content similarity between documents. For example, Shi et al. (2016) use LDA to represent each company's textual description on CrunchBase as a topic vector, and then use those topic vectors to compute a business proximity measure between each pair of companies. The measure is then validated using an application of mergers and acquisitions in the U.S. high technology industry. Chen et al. (2021) compare an executive's job description and the executive's tweets using LDA in order to gauge the job relevance of those tweets which is then used to construct a measure of social media personal branding. They then study whether social media personal branding improves a job candidate's labor market performance in the context of executive employment and compensation. Along this line of work, we also measure the relevance between two documents: a financial analyst's question and a manager's answer, in order to measure the degree of evasiveness in the manager's response.

Technically, the LDA model assumes a two-step document generative process using a Dirichlet distribution for topic proportion per document and another Dirichlet distribution for topic generation per corpus, where a topic is defined as a distribution over a fixed vocabulary and each document is assumed to cover a mixture of topics. For each document, LDA draws a topic proportion from the first Dirichlet distribution. Then, for each word of that document, it first draws a topic based on the realized topic proportion and draws the word based on the topic definition which, shared by all documents in the corpus, is drawn from the second Dirichlet distribution. Once we feed the algorithm with a collection of textual documents, the algorithm estimates all model parameters, including the topic vector for each document.

In our context of business conversation, suppose the topic vector of the j-th question document for the informed party i is $T_{i,j}^{Q}$, and the topic vector of the corresponding answer is $T_{i,j}^{A}$. We calculate $e_{i,j}$, the evasiveness of the j-th answer as the cosine distance of the two corresponding topic vectors of the question and answer, i.e.,

$$e_{i,j} \equiv 1 - \frac{T_{i,j}^{Q} \cdot T_{i,j}^{A}}{||T_{i,j}^{Q}|| \cdot ||T_{i,j}^{A}||}$$

The evasiveness of an entire conversation is then calculated by averaging $e_{i,j}$ over all questions. To reduce noise, we ignore trivial questions which are defined as those routine greetings or checking (e.g., a check for connections during conference calls). To overcome the limitation of LDA for short documents, we also apply the following heuristic rules: If the response from the informed party is short and definite, i.e., less than 5 words and containing words such as *yes*, *sure*, *correct*, *you bet*, *yeah*, *right*, *no*, we set its evasiveness measure to 0.

3.2. Incoherence

Linguistic research on text coherence and cohesion has exploited the connection of neighboring text for measurement (Hobbs, 1979; Halliday and Hasan, 1976; Grosz et al., 1995; Foltz et al. 1998). Recent advancements in natural language processing often train deep learning models by predicting words in adjacent sentences (Kiros et al. 2015; Devlin et al. 2019), predicting following sentences (Gan et al. 2017) and discourse markers (Jernite et al. 2017). Among these, BERT (Devlin et al. 2019) has stirred the machine learning community by presenting state-of-the-art results in a wide variety of NLP tasks.

Our incoherence measure is based on a variant of BERT called the lite BERT for self-supervised learning of language representations, or ALBERT which aims to address the "ineffectiveness of the next sentence prediction (NSP) loss proposed in the original BERT" (Yang et al., 2019; Liu et al., 2019). The ALBERT model introduces a loss for sentence-order prediction (SOP), which focuses on inter-sentence coherence to boost the performance of BERT in predicting the next sentence. Different from BERT which combines topic prediction and coherence prediction in constructing loss, ALBERT focuses primarily on coherence, which guides the model to "learn finer-grained distinction about discourse-level coherence properties" (Lan et al. 2019), resulting in a drastic improvement in capturing text ordering. More specifically, ALBERT models achieve this goal by using natural sequences and swapped ones as positive and negative examples rather than treating sentences from different documents as

negative examples.

Because ALBERT specializes in sentence order prediction with a focus on inter-sentence coherence (Lan et al., 2019), we build our incoherence measure based on this algorithm. Our use of the BERT framework is native because the training of BERT is based on a combined loss function of the Masked Language Modeling (MLM) and the Next Sentence Prediction (NSP), which captures coherence and cohesion on word and sentence level respectively. In other words, measuring incoherence is a native task for the BERT-family models. Consequently, this fact not only improves the interpretability of our measure of incoherence, but also offers some face validity at the algorithmic design level, which also differentiates our use of the BERT network structure from other applications of BERT.

Furthermore, we modify the loss using human perceptions rather than crude and artificially constructed negative examples, and take the last layer hidden state of the first token of the sequence (the CLS token) and further process it by a linear layer followed by a softmax activation function. The linear layer weights are learned from the next sentence prediction (classification) objective during fine-tuning. We define incoherence as one minus the output probability of being coherent. Thus, we extend ALBERT to integrate human intelligence with inter-sentence coherence.

4. Measure validations

4.1. Evasiveness

We validate the proposed evasiveness measure by comparing our algorithm-based evasiveness measure and human perception.

We compare whether and to what extent humans agree with answers considered evasive or direct by our measure. We recruited 30 undergraduate business major students from a large U.S. university and asked them to rate the degree of evasiveness of each answer on a scale from 0 to 9 for 335 pairs of questions and answers (161 conversations) from our data. Again we examine both linear and rank-order correlations between our measure and human perceptions. We found that in both cases the correlation is reasonable and statistically significant (Pearson correlation=0.166 with p = 0.036 and spearman correlation= 0.178 with p=0.024).

4.2. Incoherence

To validate our incoherence measure, we compare incoherence measured by our method and incoherence perceived by humans using conference calls. We invite 16 business major undergraduate students from a public university in the U.S. to label 347 earnings call responses. Table 1 reports a significant positive correlation between student rating and our measure.

Table 1. Comparing human perceived incoherence and the incoherence measures using earnings calls sample

Domain	Earnings Call
Total data num	347
Train pos num	50
Train neg num	50
Batch size	64
Epoch num	8
Test num	247
Corr Coef	0.411***

Note. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The same notations apply to all tables in this paper.

In summary, we find that deep learning based incoherence measure captures the human perception of incoherence well in various domains.

5. Application: deciphering conference calls

In this section, we analyze conference calls using our proposed measures. Conference calls are held in conjunction with earnings announcements as a form of voluntary disclosure. There is typically a management presentation part during which managers share their interpretations of company performance and provide any additional information, followed by a Q&A session during which analysts may question those interpretations and request additional information. We use our proposed measures to evaluate the evasiveness and incoherence of the Q&A session and investigate whether such information predicts next-day abnormal stock return and next-quarter earnings surprises.

5.1. Data and variables

We obtained transcripts of all earnings conference calls held by S&P 500 companies between January 2006 and Dec 2018 from the Factset database. To construct our key measures, we first organize the transcript of each conference call into document pairs where each pair consists of one document containing a question raised by an analyst and another document containing the corresponding answer given by a manager. Our evasiveness measure is based on how matched the two documents are on the topic level.

Although our results are robust to different numbers of topics (e.g., 50), we set it to 30 for two reasons. First, this model best describes the text we observed according to the model perplexity. Second, the topics learned seem meaningful based on human reading. With a very large

number of topics, keywords often overlap significantly across topics, making them less interpretable. For incoherence, we apply the fine-tuned ALBERT model to managers' responses.

Our two dependent variables are the next-day abnormal stock return and next-quarter earnings surprise. We obtained stock returns data from the CRSP database, and analysts' forecast data from the I/B/E/S database. We also obtain firm balance sheet information from the Compustat database. After removing earnings calls with missing data, we obtain our final sample of 6,812 earnings calls.

5.2. Abnormal return

We first investigate how the stock market responds to managers' evasiveness and incoherence during the earnings conference call. To do so, we predict the next-day stock abnormal return using these measures along with many control variables. We measure abnormal return based on the three-factor model (Fama French 1993). Like other papers in this literature, we do not claim strong causality in the sense of randomized experiments or quasi-experiments.

To control for the earnings shock, we include the earnings surprise of the quarter just ended. We also control for firm characteristics and risk factors that are often considered in the literature. More specifically, we include: (1) the abnormal return of the earnings call day ($FFCAR_{0,0}$), the prior day ($FFCAR_{-1,-1}$) and the previous two days ($FFCAR_{-2,-2}$) to control for short-term stock returns; (2) the abnormal return of the previous month ($FFCAR_{-30,-3}$) to control for medium-term stock returns; (3) the momentum effect from the past year; (4) the firm size and book-to-market ratio to control for the priced factors (Fama and French 1992); and (5) the trading volume.

Due to the conversational nature of our data, we also account for question complexity using the logarithm of the number of topics with positive probabilities in a question. Moreover, we take the content information and sentiment into consideration using the log of total words delivered by executives and its percentage of negative sentiment measured by Loughran and McDonald (2011).

Table 2 reports the estimation results of four different specifications. The results show that incoherence consistently predicts lower stock returns at least the 5% level of statistical significance. Comparing the results from column 1 and 2, in terms of economic magnitude, the two language cues boost the adjusted R^2 by 20.78% when we predict the next-day abnormal return ($FFCAR_{1,1}$). An increase of incoherence by one standard deviation lowers the abnormal stock return by about 0.05 standard deviation. Clearly, managers' failure to answer coherently during the earnings calls is negatively perceived by investors. Investors understand managers' informational advantage and interpret incoherence as negative signals. However, evasive answers did not trigger a statistically significant decrease in the next-day abnormal return. These findings remain robust if we include year and/or firm fixed effects, as we report in columns 3 and 4.

As for control variables, we find, not surprisingly, that the just-announced earnings surprise strongly explains the abnormal return. Since the earnings announcement is the most informative event in many cases, we expected and find that the same-day abnormal return $FFCAR_{0,0}$ is insignificant after controlling for earnings surprises. The long-term momentum demonstrates a reversal. We also find that the total information content and negative sentiment to be informative in predicting returns. Not surprisingly, quite a few control variables are insignificant because, in a highly efficient stock market, few variables would contain information that can predict stock returns.

Table 2. Predict stock returns

DV: Risk-Adj Returns ($FFCAR_{1,1}$)	(1)	(2)	(3)	(4)
Incoherence		-0.0492***	-0.0468***	-0.0397**
		(0.0137)	(0.0138)	(0.0200)
Evasiveness		-0.0116	-0.0124	-0.0053
		(0.0128)	(0.0128)	(0.0140)
$FFCAR_{0,0}$	-0.0104	-0.0111	-0.0119	-0.0175
	(0.0133)	(0.0133)	(0.0133)	(0.0136)
$FFCAR_{-1,-1}$	0.0044	0.0046	0.0037	0.0070
	(0.0137)	(0.0137)	(0.0137)	(0.0140)
$FFCAR_{-2,-2}$	0.0150	0.0148	0.0149	0.0165
	(0.0134)	(0.0133)	(0.0134)	(0.0136)
$FFCAR_{-30,-3}$	-0.0078	-0.0104	-0.0144	-0.0256*
	(0.0143)	(0.0144)	(0.0146)	(0.0150)
$FF\alpha$	-0.0355**	-0.0379***	-0.0391***	-0.0545***
	(0.0141)	(0.0141)	(0.0148)	(0.0160)
Earnings Surprises	0.1365***	0.1346***	0.1383***	0.1673***
	(0.0225)	(0.0225)	(0.0225)	(0.0240)
log(Market Equity)	0.0018	-0.0029	-0.0006	-0.2069***
	(0.0160)	(0.0160)	(0.0161)	(0.0597)
log(Book /Market)	-0.0067	-0.0007	0.0005	0.0159
	(0.0132)	(0.0133)	(0.0134)	(0.0561)
log(Share Turnover)	0.0254	0.0203	0.0315*	0.0415

	(0.0156)	(0.0157)	(0.0164)	(0.0329)
log(total words)	0.0640***	0.0598**	0.0499**	0.0775**
	(0.0243)	(0.0243)	(0.0254)	(0.0315)
Sentiment	-0.0490***	-0.0466***	-0.0460***	-0.0775***
	(0.0164)	(0.0164)	(0.0166)	(0.0218)
Question Complexity	-0.0078	-0.0122	-0.0136	-0.0147
	(0.0128)	(0.0129)	(0.0131)	(0.0144)
Year FE			Yes	Yes
Firm FE				Yes
Observations	6,808	6,808	6,808	6,807
Adjusted R-squared	0.0077	0.0093	0.0094	0.0071

5.3. Earnings surprises

To understand whether trading on evasiveness and incoherence detected during earnings conference calls is well-grounded in firm fundamentals or merely results from market overreaction, we further investigate whether evasiveness and incoherence actually predict worse next-quarter earnings surprises. To do so, we follow the accounting literature and assess firm fundamentals using the standardized earnings surprise based on analysts' forecast errors (AFE) which is calculated as the median of the forecast errors by all equity analysts of a firm's quarterly earnings using each equity analyst's most recent forecast. Following the literature, we adjusted AFE by the volatility of seasonal changes in earnings which are calculated using seasonal changes in earnings in the past up to 18 quarters.

We include as control variables the lagged dependent variable, firm size measured by the logarithm of its market equity, and the logarithm of its book-to-market ratio, the last two of which are evaluated at the end of the preceding year. We control for the trading volume, which is the logarithm of annual shares traded adjusted by outstanding shares at the end of the previous year.

To remove the predictive power from past returns, we include three control variables for a firm's recent returns which are calculated from an earnings announcement event study using the benchmark returns based on the three-factor model (Fama and French 1993). In particular, denoting the earnings call date as day t and the next quarterly earnings announcement date as day 0, we include the cumulative abnormal return during the trading window [-30, -3] ($FFCAR_{-30,-3}$) and the abnormal return on day -2 ($FFCAR_{-2,-2}$). These two variables capture the return information for the past 29 trading days prior to the next earnings announcement, which should have incorporated the most recent information about firm fundamentals. To control for the firm's return momentum (Jegadeesh and Titman 1993) over the previous year before the earnings call, we include the control variable $FF\alpha$, which is the estimated intercept from the event study regression. It measures the in-sample cumulative abnormal return of the previous year.

To remove the predictive power from analysts' forecast dispersion and forecast revision before the next earnings announcement, we control for both variables. To construct forecast revision, we sum the median of the scaled moving changes in earnings forecast within the past three months: $REV_{i,t} = \sum_{j=0}^{2} (f_{i,t-j} - f_{i,t-j-1})/p_{i,t-j}$ where $f_{i,t}$ is the median analyst's quarterly forecast of firm i in month t. The monthly revision is scaled by the stock price, $p_{i,t-j}$. We calculate forecast dispersion as the standard deviation of the most recent earnings forecasts before the next earnings announcement, scaled by the volatility of seasonal changes in earnings.

We estimate a pooled OLS model with and without year fixed effects. In accordance with earnings surprise literature (Mayew and Venkatachalam 2012; Kelley and Tetlock 2013; Chen 2014), we did not include firm fixed effects since the firm trend in earnings is already adjusted. Indeed, since the earnings surprise measures analysts' forecast error, any persistent under- or over-estimate of the earnings for a firm over years cannot exist in equilibrium.

Table 3 reports the estimation results. We find that both incoherence and evasiveness predict unexpected lower future earnings. In terms of magnitude, the conditional expectation of SAFE will be 0.0172 × 4=0.0688 standard deviation lower when the incoherence measure increases from two standard deviations below its mean to two standard deviations above. Similarly, the predicted next-quarter earing surprise will be 0.0724 standard deviations lower as the evasiveness measure moves from two standard deviations below to two standard deviations above the mean. Among the control variables, we find that lagged earnings surprise, return momentum, and recent returns are strong predictors for future earnings surprises. We also find analyst forecasts dispersion to be informative in predicting earnings surprises in both specifications.

To summarize, after controlling for factors commonly suggested in the literature that predict earnings, we find that evasiveness and incoherence still contain additional information on a firm's future earnings surprise.

Table 3. Predict earnings surprises

	(1)	(2)
Incoherence	-0.0172**	-0.0198**

	(0.0087)	(0.0088)
Evasiveness	-0.0181**	-0.0175**
	(0.0081)	(0.0081)
Earnings Surprises (lagged)	0.2178***	0.2128***
	(0.0143)	(0.0143)
Forecast Dispersion	0.1507***	0.1518***
	(0.0081)	(0.0082)
Forecast Revision	0.0084	0.0133
	(0.0139)	(0.0141)
log(Market Equity)	0.0172*	0.0145
	(0.0102)	(0.0102)
log(Book/Market)	-0.0158*	-0.0145*
	(0.0085)	(0.0086)
log(Share Turnover)	0.0087	-0.0016
	(0.0100)	(0.0104)
$FF\alpha$	0.0452***	0.0519***
	(0.0083)	(0.0086)
$FFCAR_{-30,-3}$	0.0351***	0.0382***
	(0.0085)	(0.0085)
$FFCAR_{-2,-2}$	0.0235***	0.0223***
	(0.0078)	(0.0078)
log(total words)	-0.0003	0.0151
	(0.0154)	(0.0161)
Sentiment	-0.0100	-0.0141
	(0.0104)	(0.0105)
Question Complexity	0.0113	0.0084
	(0.0082)	(0.0083)
Year FE		Yes
Observations	6,812	6,812
Adjusted R-squared	0.1092	0.1129

Combining results in Table 2 and 3, we find evidence that, both evasiveness and incoherence reveals additional information about firm profitability, but only incoherence is currently well captured by investors.

6. Conclusion

In this paper, we proposed two machine learning based measures to quantify evasiveness and incoherence in real-time strategic conversations. We validated these measures in different contexts, and demonstrated their business value through a concrete business application where we analyze managers' responses to questions during earnings conference calls. We showed that both measures provide additional information about a firm's earnings in the following quarter, and the incoherence measure also predicts a lower next-day stock abnormal return after the firm's earnings call.

This paper contributes methodologically to the literature by proposing and evaluating a novel measure of evasiveness based on semantics which differs from the lexicon-based vagueness measure previously used, and by introducing the measure of incoherence which is new to that literature. The paper also contributes to the emerging field of FinTech by demonstrating how financially valuable information can be extracted from conversations between managers and analysts during conference calls, a type of data mostly overlooked by FinTech algorithms. Finally, this paper also pioneers the development of combining machine learning with asset pricing, a direction potentially with a high impact given the rapid advances of AI.

This research has important practical implications. From the perspective of stock traders and investors, our proposed measures of evasiveness and incoherence can be incorporated to form profitable trading and investment strategies. In particular, retail investors can benefit from such information which used to be accessible almost exclusively by institutional investors with in-house information acquisition and professional financial expertise. Therefore, technology can be "the ultimate empowerment of the individual" (Pettit and Jaroslovsky 2001). More broadly, the idea of using machine learning to analyze the "conversational dance" between a party with more information and another with an interest in such information, could be particularly fruitful. In this cat-and-mouse game of information seeking, machines may ultimately win, leaving only one viable option for managers: be honest and be forthright. Finally, from the perspective of financial analysts, our results highlight the evolving role of equity analysts in probing every subtle detail of earnings conference calls. Machine learning algorithms are encroaching on many territories that are traditionally considered uniquely operated by human intelligence. In this looming new age of AI, we believe analysts who are open to and can harness new technologies to augment their abilities are likely to survive and thrive.

7. References

Adamopoulos, P., Ghose, A. and Todri, V. (2018) The impact of user personality traits on word of mouth: Text-mining social media platforms. *Information Systems Research*, 29(3): 612-640.

Bai, X., Marsden, J.R., Ross Jr, W.T. and Wang, G. (2020). A Note on the Impact of Daily Deals on Local Retailers' Online Reputation: Mediation Effects of the Consumer Experience. *Information Systems Research*, 31(4): 1132-1143.

Barzilay R, Lapata M (2008) Modeling local coherence: An entity-based approach. *Computational Linguistics* 34(1):1–34.

Blankespoor, E., Dehaan, E., Wertz, J. and Zhu, C., 2019. Why

do individual investors disregard accounting information? The roles of information awareness and acquisition costs. *Journal of Accounting Research*, 57(1): 53-84.

Biddle GC, Hilary G, Verdi RS (2009) How does financial reporting quality relate to investment efficiency? Journal of accounting and economics 48(2-3):112–131.

Blankespoor, E., deHaan, E. and Marinovic, I., 2020. Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics*, 70(2-3), p.101344.

Brahma, A., Goldberg, D.M., Zaman, N. and Aloiso, M. (2021). Automated mortgage origination delay detection from textual conversations. *Decision Support Systems*, 140, :113433.

Chen, H., De, P., Hu, Y.J. and Hwang, B.H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5):1367-1403.

Chan LK, Jegadeesh N, Lakonishok J (1996) Momentum strategies. *The Journal of Finance* 51(5):1681–1713.

Chen, J., Lin, S.T. and Durrett, G., (2019). Multi-hop question answering via reasoning chains. arXiv preprint arXiv:1910.02610.

Chen, H., Xiao, K., Sun, J. and Wu, S. (2017). A double-layer neural network framework for high-frequency forecasting. *ACM Transactions on Management Information Systems*, 7(4):1-17.

Chen, Y., Rui, H. and Whinston, A., 2021. Tweet To The Top? Social Media Personal Branding And Career Outcomes. *MIS Quarterly*, 45(2).

Cho K, Van Merriënboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

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

Crossley, S. and D. McNamara (2011). Text Coherence and Judgments of Essay Quality: Models of Quality and Coherence. *Cognitive Science* 33:1236-1247.

Cuayáhuitl, H., Keizer, S. and Lemon, O., 2015. Strategic dialogue management via deep reinforcement learning. arXiv preprint arXiv:1511.08099.

De Amicis, C., Falconieri, S. and Tastan, M. (2020). Sentiment analysis and gender differences in earnings conference calls. *Journal of Corporate Finance*, p.101809.

Dong B, Li EX, Ramesh K, Shen M (2015) Priority dissemination of public disclosures. *The Accounting Review* 90(6):2235–2266.

Dzieliński M, Wagner AF, Zeckhauser RJ (2017) Straight talkers and vague talkers: The effects of managerial style in earnings conference calls. Technical report, National Bureau of Economic Research.

Ertugrul M, Lei J, Qiu J, Wan C (2017) Annual report readability, tone ambiguity, and the cost of borrowing. *Journal of Financial and Quantitative Analysis* 52(2):811–836.

Fama EF, French KR (1992) The cross-section of expected stock returns. *the Journal of Finance* 47(2):427–465.

Fama EF, French KR (1993) Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33(1):3–56.

Fishman MJ, Hagerty KM (2003) Mandatory versus voluntary disclosure in markets with informed and uninformed customers. *Journal of Law, Economics, and organization* 19(1):45–63.

Gao, M., and Huang, J. K. 2020. "Informing the Market: The Effect of Modern Information Technologies on Information Production," *Review of Financial Studies* (33:4), pp. 1367-1411.

Georgila, K. and Traum, D., 2011. Reinforcement learning of argumentation dialogue policies in negotiation. In Twelfth Annual Conference of the International Speech Communication Association.

Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proceedings of the National academy of Sciences* 101(suppl 1):5228–5235.

Goffman E (1959) The presentation of self in everyday life. Anchor.

Hauch V, Blandón-Gitlin I, Masip J, Sporer SL (2015) Are computers effective lie detectors? a meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review* 19(4):307–342.

Hill F, Cho K, Korhonen A (2016) Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483* .

Hwang BH, Kim HH (2017) It pays to write well. *Journal of Financial Economics* 124(2):373–394.

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(1):65–91.

Jennings R, Starks L (1985) Information content and the speed of stock price adjustment. *Journal of Accounting Research* 336–350.

Karamanis N, Poesio M, Mellish C, Oberlander J (2004) Evaluating centering-based metrics of coherence for text structuring using a reliably annotated corpus. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 391 (Association for Computational Linguistics).

Kelley, E.K. and Tetlock, P.C., (2013). How wise are crowds? Insights from retail orders and stock returns. *The Journal of Finance*, 68(3):1229-1265.

Khalmetski K, Rockenbach B, Werner P (2017) Evasive lying in strategic communication. *Journal of Public Economics* 156:59–72.

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

Ko, W.J., Dalton, C., Simmons, M., Fisher, E., Durrett, G. and Li, J.J., 2021. Discourse Comprehension: A Question Answering Framework to Represent Sentence Connections. arXiv preprint arXiv:2111.00701.

Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S. and Prendinger, H. (2018) Deep learning for affective computing: Text-based emotion recognition in decision support. Decision Support Systems, 115:24-35.

Kraus, M. and Feuerriegel, S. (2017). Decision support from

financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, 104:38-48.

Kraus, M. and Feuerriegel, S (2019). Forecasting remaining useful life: Interpretable deep learning approach via variational Bayesian inferences. *Decision Support Systems*, 125:113100.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R. (2019) Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

Lee YJ (2012) The effect of quarterly report readability on information efficiency of stock prices. Contemporary Accounting Research 29(4):1137–1170.

Lev, B., 1988. Toward a theory of equitable and efficient accounting policy. *Account. Rev*. 1-22.

Lewis, M., Yarats, D., Dauphin, Y.N., Parikh, D. and Batra, D., 2017. Deal or no deal? end-to-end learning for negotiation dialogues. arXiv preprint arXiv:1706.05125.

Li EX, Ramesh K, Shen M, Wu JS (2015) Do analyst stock recommendations piggyback on recent corporate news? an analysis of regular-hour and after-hours revisions. *Journal of Accounting Research* 53(4):821–861.

Li, X. T., Zhu, H. W., and Zuo, L. 2021. "Reporting Technologies and Textual Readability: Evidence from the Xbrl Mandate," *Information Systems Research* (32:3), pp. 1025-1042.

Li, W., Chen, H. and Nunamaker Jr, J.F. (2016). Identifying and profiling key sellers in cyber carding community: AZSecure text mining system. *Journal of Management Information Systems*, 33(4):1059-1086.

Li, Y., Packard, G. and Berger, J., 2020. Conversational Dynamics: When Does Employee Language Matter?.*Working Paper*.

Liddicoat, A.J., 2021. An introduction to conversation analysis. Bloomsbury Publishing.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Loughran, T. and McDonald, B., (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1): 35-65.

Loughran T, McDonald B (2013) Ipo first-day returns, offer price revisions, volatility, and form s-1 language. *Journal of Financial Economics* 109(2):307–326.

Mahmoudi, N., Docherty, P. and Moscato, P., (2018) Deep neural networks understand investors better. *Decision Support Systems*, 112: 23-34.

Mayew, W.J. and Venkatachalam, M. (2012) The power of voice: Managerial affective states and future firm performance. *The Journal of Finance*, 67(1):1-43.

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

Miller BP (2010) The effects of reporting complexity on small and large investor trading. The Accounting Review 85(6):2107–2143.

Mousavi, R., Raghu, T.S. and Frey, K. (2020). Harnessing Artificial Intelligence to Improve the Quality of Answers in Online Question-answering Health Forums. *Journal of Management Information Systems*, 37(4):1073-1098.

Myatt, D.P., Wallace, C., 2012. Endogenous information acquisition in coordination games. *Rev. Econ. Stud*. 79 (1), 340-374.

Patell JM, Wolfson MA (1984) The intraday speed of adjustment of stock prices to earnings and dividend announcements. *Journal of Financial Economics* 13(2):223–252.

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

Pettit, D., and Jaroslovsky, R. 2001. *The Wall Street Journal Online's Guide to Online Investing: How to Make the Most of the Internet in a Bull or Bear Market*. Three Rivers Press (CA).

Plüss B, Piwek P (2016) Measuring non-cooperation in dialogue 1925–1936.

Poria, S., Majumder, N., Mihalcea, R. and Hovy, E., 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. IEEE Access, 7, pp.100943-100953.

Shi, Z., Lee, G.M. and Whinston, A.B., (2016). Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence. *MIS quarterly*, 40(4).

Shim, J. and Arkin, R.C., 2013, October. A taxonomy of robot deception and its benefits in HRI. In 2013 IEEE international conference on systems, man, and cybernetics (pp. 2328-2335). IEEE.

Stein, K. 2018. "Remarks at SEC Speaks: Increasing Product Complexity: What's at Stake?", from https://www.sec.gov/news/speech/stein-sec-speaks-increasing-product-complexity

Tofangchi, S., Hanelt, A., and Li, S. (2019). Advancing Recommendations on Two-Sided Platforms : A Machine Learning Approach to Context-Aware Profiling, in International Conference on Information Systems, ICIS 2019.

Vo, N.N., He, X., Liu, S. and Xu, G. (2019). Deep learning for decision making and the optimization of socially responsible investments and portfolio. *Decision Support Systems*, 124:113097.

Wang, P., Li, J. and Hou, J., (2021). S2SAN: A sentence-to-sentence attention network for sentiment analysis of online reviews. Decision Support Systems, p.113603.

Xu, W., Wang, T., Chen, R. and Zhao, J.L., (2021). Prediction of initial coin offering success based on team knowledge and expert evaluation. *Decision Support Systems*, p.113574.

Yarats, D. and Lewis, M., 2018, July. Hierarchical text generation and planning for strategic dialogue. In International Conference on Machine Learning (pp. 5591-5599). PMLR.

You, Y., Li, J., Hseu, J., Song, X., Demmel, J. and Hsieh, C.J. (2019) Reducing BERT pre-training time from 3 days to 76 minutes. arXiv preprint arXiv:1904.00962.

Zhang, Q., Wang, W. and Chen, Y., 2020. Frontiers: In-consumption social listening with moment-to-moment unstructured data: The case of movie appreciation and live comments. Marketing Science, 39(2), pp.285-295.

Zhou, Y., He, H., Black, A.W. and Tsvetkov, Y., 2019. A dynamic strategy coach for effective negotiation. arXiv preprint arXiv:1909.13426.