Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls

Katherine A. Keith
College of Information and Computer Sciences
University of Massachusetts Amherst
kkeith@cs.umass.edu

Amanda Stent
Bloomberg LP
astent@bloomberg.net

Abstract

Every fiscal quarter, companies hold *earnings calls* in which company executives respond to questions from analysts. After these calls, analysts often change their *price target recommendations*, which are used in equity research reports to help investors make decisions. In this paper, we examine analysts' decision making behavior as it pertains to the language content of earnings calls. We identify a set of 20 pragmatic features of analysts' questions which we correlate with analysts' pre-call investor recommendations. We also analyze the degree to which semantic and pragmatic features from an earnings call complement market data in predicting analysts' post-call changes in price targets. Our results show that earnings calls are moderately predictive of analysts' decisions even though these decisions are influenced by a number of other factors including private communication with company executives and market conditions. A breakdown of model errors indicates disparate performance on calls from different market sectors.

1 Introduction

Financial analysts are key sell-side players in finance who are employed to analyze, interpret, and disseminate financial information (Brown et al., 2015). For the firms they cover, financial analysts regularly release *recommendations* to buy, hold, or sell the company's stock, and stock *price targets*. Financial analysts' forecasts are of value to investors (Givoly and Lakonishok, 1980) and may be better surrogates for market expectations than forecasts generated by time-series models (Fried and Givoly, 1982).

Analysts' decisions are influenced by market conditions and private communications[1], so it is impossible to exactly reconstruct their decision making process. However, signals of analysts' decision making may be obtained by analyzing *earnings calls*—quarterly live conference calls in which company executives present prepared remarks (the *presentation* section) and then selected financial analysts ask questions (the *question-answer* section). Previous work has shown that earnings calls disclose more information than company filings alone (Frankel et al., 1999) and influence investor sentiment in the short term (Bowen et al., 2002). However, recently company executives and investors have questioned their value (Koller and Darr, 2017; Melloy, 2018).

Earnings calls are extremely complex, naturally-occurring examples of discourse that are interesting to study from the perspective of computational linguistics (see Figure 1). In this work, we examine analysts' decision making in the context of earnings calls in two ways:

- **Correlating analysts' question pragmatics with their pre-call judgements:** With domain experts, we select a set of 20 pragmatic and discourse features which we extract from the questions of earnings calls. Then we correlate these with analysts' pre-call judgments and find *bullish* analysts tend to be called on earlier in calls, and ask questions that are more positive, more concrete, and less about the past (§4).
- **Predicting changes in analysts' post-call forecasts:** We use the pragmatic features, along with representations of the semantic content of earnings calls, to predict changes in analysts' post-call price targets. Since analysts have a deep understanding of market factors influencing a company's performance and have access to private information, our null hypothesis is

[1] Brown et al. (2015) find over half of the 365 analysts they surveyed have five or more direct contacts per year with the CEO or CFO of companies they follow.

Brian Nowak, Analyst: Thanks for taking my questions. One on YouTube, I guess. Could you just talk to some of the qualitative drivers that are really bringing more advertising dollars on to YouTube? And then I think last quarter you had mentioned the top 100 advertiser spending was up 60% year-on-year on YouTube, wondering, if you could update us on that? And the second one on search, it sounds like mobile is accelerating. Where are you now in the mobile versus desktop monetization gap? And, Sundar, how do you think about that long-term? Do you see mobile being higher, reaching equilibrium? How do you see that trending?

Sundar Pichai, CEO: On the YouTube one. Look, I mean, the shift to video is a profound medium shift and especially in the context of mobile, you know and obviously users are following that. You're seeing it in YouTube as well as elsewhere in mobile. And so, advertisers are being increasingly conscious. They're being very, very responsive. So, we're seeing great traction there and we'll continue to see that. They are moving more off their traditional budgets to YouTube and that's where we are getting traction. On mobile search, to me, increasingly we see we already announced that over 50% of our searches are on mobile. Mobile gives us very unique opportunities in terms of better understanding users and over time, as we use things like machine learning, I think we can make great strides. So, my long-term view on this is, it is as-compelling or in fact even better than desktop, but it will take us time to get there. We're going to be focused till we get there.

Figure 1: Earnings calls are extremely complex examples of naturally-occurring discourse. In this example question-answer pair from a Google earnings call on October 27, 2016, the analyst asks **six distinct questions** in a single turn. Because the interaction originates as speech, there are **discourse markers and hedging.** The analyst and executive discuss **concrete entities and performance statistics** and **past, present and future** performance.

that earnings calls are not predictive of forecast changes. However, our best model gives a reduction of 25% in relative accuracy error over a majority class baseline (twice the reduction of a model using market data alone), suggesting there is signal in the noise. We also conduct pairwise comparisons of modeling features including: semantic vs. pragmatic features, Q&A-only vs. whole-call data, and whole-document vs. turn-level models (§5).

2 Related work

NLP is used extensively for financial applications (Tetlock et al., 2008; Kogan et al., 2009; Leidner and Schilder, 2010; Loughran and McDonald, 2011; Wang et al., 2013; Ding et al., 2014; Peng and Jiang, 2016; Li and Shah, 2017; Rekabsaz et al., 2017). Earnings calls, in particular, are shown to be predictive of investor sentiment in the short-term, including of increased stock volatility and trading volume levels (Frankel et al., 1999), decreased forecast error and forecast dispersion (Bowen et al., 2002), and increased absolute returns for intra-day trading (Cohen et al., 2012). Although most prior work on earnings calls treat each call as a single document, Matsumoto et al. (2011) find that the *question-answer* portion of the earnings call is more informative (in terms of intra-day absolute returns) than the *presentation* portion, and Cohen et al. (2012) show firms "cast" earnings calls by disproportionately calling on bullish analysts.

Most prior applications of NLP to earnings calls use only shallow linguistic features and correlation analyses, specifically correlations between political bigrams and stock return volatility (Hassan et al., 2016); contrastive words and share prices (Palmon et al., 2016); and euphemisms and earnings surprise (Suslava, 2017). Other work analyzes earnings calls from a sociolinguistic perspective, including in terms of discourse connectives (Camiciottoli, 2010), indirect requests (Camiciottoli, 2009), unanswered questions (Hollander et al., 2010), persuasion (Crawford Camiciottoli, 2018) and deception (Larcker and Zakolyukina, 2011). Focusing on only the audio of earnings calls, Mayew and Venkatachalam (2012) extract managers' affective states using commercial speech software. In the work most similar to ours, Wang and Hua (2014) use named entities, part-of-speech tags, and probabilistic frame-semantic features in addition to unigrams and bigrams to correlate earnings calls with financial risk, which they defined as the volatility of stock prices in the week following the earnings call.

NLP-based corpus analyses of decision making are rare. Beňuš et al. (2014) analyze the impact of entrainment on Supreme Court justices' subsequent decisions. Multiple groups have examined the impact of various semantic and pragmatic features on modeling opinion change using reddit ChangeMyView discussions (e.g. (Hidey et al., 2017; Jo et al., 2018; Musi, 2018)), and there has been other work on opinion change using other web discussion data (e.g. (Tan et al., 2016; Habernal and Gurevych, 2016; Lukin et al., 2017)). Because many factors influence decision making behavior, the fact that any signal can be obtained

Earnings calls total (2010-2017)	12,285
Train (2010-2015)	9,770
Validation (2016)	1,066
Test (2017)	1,449
Unique companies	642
Total Q&A sets	573,550
Ave. Q&A sets per doc.	44.3
One call, ave. unique analysts speaking	10.9
One call, ave. analysts w/ price targets	9.6
Ave. num. of tokens per doc.	8,761
Ave. turn length (num. tokens), Q&A	62.7

Table 1: Data statistics for S&P 500 companies' earnings calls. A Q&A set consists of two or more turns, one containing an analyst's question(s) and the rest containing company representatives' answer(s).

from linguistic analyses of isolated language artifacts is scientifically interesting.

3 Data and pre-processing

Our data[2] consists of transcripts of 12,285 earnings calls held between January 1, 2010 and December 31, 2017. In order to control for analyst coverage effects (larger companies with a greater market share will typically be covered by more analysts), we include only calls from S&P 500 companies. We split the data by year into training, validation and testing sets (see Table 1).

The transcripts are XML files with metadata specifying speaker turn boundaries and the name of the speaker (or "Operator" for the call operator). In order to identify *speaker type* (analyst or company representative) we use the following heuristic: if the transcript explicitly includes the speaker type with the speaker name (e.g. "John Doe, Analyst"), we do exact string matching for ", Analyst"; else, we assume the names of speakers between the first and second operator turns (i.e. in the *presentation* section) are those of company representatives and all other speakers are analysts. We manually checked this heuristic on a few dozen documents and found it to have high precision.

We remove turns spoken by the operator as well as turns that have fewer than 10 tokens since manual analysis revealed the latter were largely acknowledgment and greeting turns (e.g. "Thank you for your time" and "You're welcome"). We also lexicalized named entities and represented them as a single token. We obtained tokenization, part of speech tagging, and dependency parsing via a proprietary NLP library[3].

4 Pragmatic correlations with analysts' pre-call judgments

We are interested in whether and how the forms of analysts' questions reflect their pre-call judgments about companies they cover. Analysts' questions are complex: a single turn may contain several questions (or answers). An example question-answer pair is shown in Figure 1.

We compute Pearson correlations between linguistic features indicating certainty, deception, emotion and outlook (§4.1) and the *type* of analyst (bullish, bearish, or neutral) asking the question. We use a mapping of analysts' recommendations to a 1-5 scale[4] where a 1 denotes "strong sell" and a 5 denotes "strong buy." We label each analyst according to their recommendation of the company before the earnings call:

- *bearish* if analysts give a company a 1 or 2,
- *neutral* if they give a 3, and
- *bullish* if they give a 4-5.

We have analyst recommendations for 160,816 total question turns and the distribution over analyst labels is 4.5% bearish, 35.7% neutral, and 59.7% bullish. Following other correlation work in NLP (Preoţiuc-Pietro et al., 2015; Holgate et al., 2018), we use Bonferroni correction to address the multiple comparisons problem.

4.1 Pragmatic lexical features

We extract 20 pragmatic features from each turn by gathering existing hand-crafted, linguistic lexicons for these concepts[5]. See Table 2 for statistics about the lexicons and Table 3 for examples.

Named entity counts and concreteness ratio. For each turn, we calculate the number of named entities in five coarse-grained groups constructed from the fine-grained entity types of OntoNotes[6]

[2] In Appendix A in supplemental material we provide the stock tickers for the calls in our data; the corpus can be re-assembled from multiple sources, such as https://seekingalpha.com/.

[3] Bloomberg's libnlp

[4] Qualitative analyst rating labels vary from firm to firm. For example, some firms use the standard "buy, hold, sell" labels while others might use different labels such as "outperform, peer perform, underperform." We use ratings from a proprietary financial database that have been manually normalized to 1-5 scale.

[5] Appendix B in supplemental material gives details about the sources of our lexicons.

[6] Version 5, https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf Section 2.6.

No.	Pragmatic Lexicon	Examples	Source	Num. terms
10	Positive sentiment, financial	booming, efficient, outperform	LM	354
10	Positive sentiment, general-purpose	perfection, enthrall, phenomenal	T	2,507
11	Negative sentiment, financial	accidents, recession, stagnant	LM	2,353
11	Negative sentiment, general-purpose	cheater, devastate, loathsome	T	3,692
12	Hedging, unigrams	basically, generally, sometimes	PH	79
12	Hedging, multi-word	a little, kind of, more or less	PH	39
13	Weak Modal	appears, could, possibly	LM	27
13	Moderate Modal	likely, probably, usually	LM	14
13	Strong Modal	always, clearly, undoubtedly	LM	19
14	Uncertain	assume, deviate, turbulence	LM	297
15	Constraining	bounded, earmark, indebted	LM	184
16	Litigious	adjudicate, breach, felony, lawful	LM	903

Table 2: Detailed examples and the number of words for lexicons used as pragmatic features. LM is (Loughran and McDonald, 2011), PH is (Prokofieva and Hirschberg, 2014) and T is (Taboada et al., 2011). Feature numbers (No.) correspond to the text description in §4.1.

No.	Pragmatic Feat.	Example	Score
6	Concreteness	Yes. Andrew for the_quarter the total inter-company sales for the_first_quarter was roughly 4.6_million and about 600,000 was related to medical, it was 4_million via DSS.	0.29
10	Positive sentiment	Good morning, gentlemen. Nice job on the rebound quarter.	0.33
11	Negative sentiment	And this is a slightly delicate question. With some of the terrible events that have been happening, what is this duty or potential liability or cost of insurance?	0.15
12	Hedging	It may vary Michael. So, some might be much better than that, but then you got some of that – that's not as much right. So, all-in, yeah.	0.22

Table 3: Pragmatic features as highlighted tokens. Note, named entities are lexicalized (e.g. "4.6_million"). Feature numbers (No.) correspond to the text description in §4.1.

(Hovy et al., 2006): (1) events, (2) numbers, (3) organizations/locations, (4) persons, and (5) products. We also calculate (6) a *concreteness ratio*: the number of named entities in the turn divided by the total number of tokens in the turn.

Predicate-based temporal orientation. Temporal orientation is the emphasis individuals place on the past, present, or future. Previous work has shown correlations between "future intent" extracted from query logs and financial market volume volatility (Hasanuzzaman et al., 2016). We determine the temporal orientation of every predicate in a turn. We extract OpenIE predicates via a re-implementation of PredPatt (White et al., 2016). For each predicate, we look at its Penn Treebank part-of-speech tag and use a heuristic[7] to determine if it is "past," "present," or "future." : We calculate the number of (7) "past" oriented predicates, (8) "present" oriented predicates and (9) "future" oriented predicates in each turn.

Sentiment. We calculate the ratio of (10) positive sentiment terms and (11) negative sentiment terms to the number of tokens in each turn. We use the financial sentiment lexicons developed by Loughran and McDonald (2011) from fourteen years of 10-Ks. We supplement these with a general-purpose sentiment dictionary (Taboada et al., 2011), to account for the relative informality of earnings calls.

Hedging. We calculate (12) the ratio of hedges to tokens in each turn. Hedges are lexical choices by which a speaker indicates a lack of commitment to the content of their speech (Prince et al., 1982). We use the single- and multi-word hedge lexicons from (Prokofieva and Hirschberg, 2014).

Other lexicon-based features. We compute the ratios of (13) modal, (14) uncertain, (15) con-

[7] If the part-of-speech tag for the predicate is VBD or VBN the temporal orientation is "past"; otherwise if it is VB, VBG, VBP, or VBZ it is "present" unless the predicate has a dependent of the form *will, 'll, shall* or *wo* indicating "future", *are is, am,* or *are* indicating "present", or *was* or *were* indicating "past".

No.	Feature	Pearson's r	p-value
1	Named entities event	0.0041	0.0999
2	Named entities number	0.0064	0.0099
3*	Named entities org.	0.0185	$< 1e^{-4}$
4*	Named entities person	0.0247	$< 1e^{-4}$
5	Named entities product	0.0022	0.3777
6*	Concreteness ratio	0.0115	$< 1e^{-4}$
7*	Num past preds	−0.0086	0.0006
8	Num present preds	0.0052	0.0378
9	Num future preds	0.0033	0.1914
10*	Sentiment positive	0.0162	$< 1e^{-4}$
11*	Sentiment negative	−0.0104	$< 1e^{-4}$
12	Hedging	0.0017	0.5019
13	Modal	0.0075	0.0028
14	Uncertainty	0.0055	0.0287
15	Constraining	0.0005	0.8399
16	Litigiousness	−0.0072	0.0037
17*	Turn order	−0.1034	$< 1e^{-4}$
18	Num. tokens	0.0050	0.0459
19	Num predicates	0.0011	0.6692
20	Num sents.	0.0043	0.0854

Table 4: Results from Pearson correlations of pragmatic lexical features from §4.1 and prior-to-call labels of analysts, (*bearish*, *neutral*, or *bullish*). Statistical significance after Bonferroni correction is marked by (*) for $p < 0.0025$. Total 160,816 question turns.

straining, and (16) litigious terms in each turn using the respective lexicons from Loughran and McDonald (2011). In each case, we compute the ratio of terms in the category to the number of tokens in the turn.

Other pragmatic features. We also calculate (17) the turn order, (18) the number of tokens, (19) the number of predicates, and (20) the number of sentences in each turn.

4.2 Interpretation of correlation results.

Full results for the pragmatic correlation analysis are given in Table 7. For a number of features the correlations are not statistically significant. However, we expand upon the statistically significant results for negative (−) and positive (+) correlations with the bullishness of an analyst:

- *(+) Bullishness and turn order.* This suggests bullish analysts tend to be called on earlier in the call and bearish and neutral analysts tend to be called on later in the call which confirms the conclusion of Cohen et al. (2012).
- *(+) Bullishness and positive sentiment.* Bullish analysts tend to ask more positive (less negative) questions and the reverse is true for neutral/bearish analysts. Intuitively, this makes sense since bullish analysts are more favorable towards the firm and thus probably cast the firm in a positive light.
- *(+) Bullishness and entities.* Here we find that bullish analysts are slightly more concrete in their questions towards the company and tend to ask more about organizations and people.
- *(−) Bullishness and past predicates.* This suggests bearish and neutral analysts tend to talk about the past more.

These correlations could be used by journalists and investors to flag questions that follow atypical patterns for a particular analyst.

5 Predicting changes in analysts' post-call forecasts

We are interested in what earnings-call related information is indicative of analysts' subsequent decisions to change (or not change) their *price target* after an earnings call. A *price target* is a projected future price level of asset; for example, an analyst may give a stock that is currently trading at $50 a six-month price-target of $90 if they believe the stock will perform better in the future.

We design experiments to answer the following **research questions**: (1) Is the text of earnings calls predictive of analysts' changes in price targets from before to after the call? This is an open research question since analysts may change their price targets at any time and consider external information (e.g. current events or private conversations with company executives); (2) If the text is predictive, is the text more predictive than market-based features such as the company's stock price, volatility, and earnings? (3) If the text is predictive, what linguistic aspects (e.g. pragmatic vs. semantic) are more predictive and with which feature representations? (4) Is the *question-answer* portion of the call more predictive than the *presentation* portion? (5) Does a turn-based model of the call provide more signal than "single document" representations?

5.1 Representing analysts' forecast changes

We model analysts' changes in forecasts as both a regression task and a 3-class classification task because different formulations may be of interest to various stakeholders[8].

[8]For instance, investors may care more about small changes in forecast price targets whereas journalists may care more about relative changes (e.g. whether an earnings call will move analysts' price targets up or down).

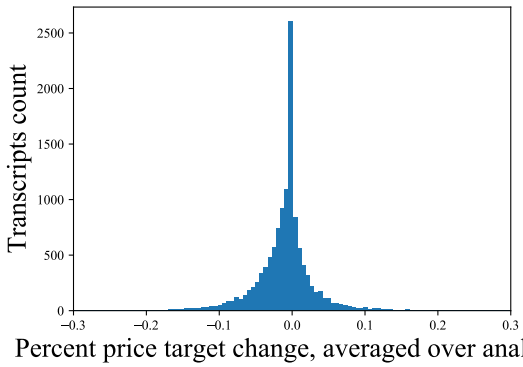

Figure 1: Distribution across the entire corpus of prediction y-values, percentage price change in analyst price targets.

Dataset	−1	0	1
Train	33.3%	38.3%	28.4%
Validation	29.2%	30.5%	40.3%
Test	33.6%	38.7%	27.7%

Table 5: Percentage of examples in each class $(-1, 0, 1)$ for the training, validation, and test sets.

Regression. For each earnings call in our corpus, $i \in \mathcal{D}$, and each analyst in the set of analysts covering that call, $j \in J_i$, let b_j be the price target of analyst j before the call and let a_j be the price target after the call[9]. Then the *average percent change in analysts' price targets* is

$$y_i = \frac{1}{|J_i|} \sum_{j \in J_i} \frac{a_j - b_j}{b_j}. \quad (1)$$

See Figure 1 for the distribution of y_i.
Since analysts can report price targets at any time, we set cut-off points for a_j and b_j to be 3 months before and 14 days after the earnings call date respectively (a majority of analysts who change their price targets do so within two weeks after a call).

Classification. We create three (roughly equal) classes (*negative, neutral,* and *positive* change) by binning the y_i values calculated in the equation above into thirds. For each earnings call i, $c_i = -1$ if $y_i < -0.0167$, $c_i = 0$ if $-0.0167 \leq y_i \leq 0.0$, and $c_i = 1$ if $0 < y_i$. Table 5 shows the class breakdown for each split of the data.

[9]Because the company holding the earnings call chooses which analysts to call on for questions, our data includes analyst ratings and recommendations from analysts who do *not* ask a question in a call. Also, because individual analysts' recommendations may be sold to different vendors, we do not have analyst ratings and recommendations for all analysts who ask questions in our data.

5.2 Features

We compare models with market-based, pragmatic, and semantic features.

5.2.1 Market features

For each company and call in our dataset, we obtain 10 market features for the trading day prior to the call date: open price, high price, low price, volume of shares, 30-day volatility, 10-day volatility, price/earnings ratio, relative price/earnings ratio, EBIT yield, and earnings yield[10]. We impute missing values for these features using the mean value of features in the training data[11]. We scale features to have zero mean and unit variance.

5.2.2 Semantic features

Doc2Vec. We use the *paragraph vector* algorithm proposed by (Le and Mikolov, 2014) to obtain 300-dimensional document embeddings. Depending on the model, we train doc2vec embeddings over whole calls, question-answer sections only, and individual turns. Using the `Gensim`[12] implementation (Řehůřek and Sojka, 2010), we train *doc2vec* models for 50 epochs and ignore words that occur less than 10 times in the respective training corpus.

Bag-of-words. We lowercase tokens, augment them with their parts of speech, and then limit the vocabulary to the top 100K content words[13] in the training data. Depending on the model, we calculate bag-of-words feature vectors over the whole document, over the *Q&A* section, and over each turn separately.

5.2.3 Pragmatic features

We combine the 20 pragmatic features described in Section 4.1 into a single feature vector. These features are only used in our turn-level models.

5.3 Models

We use several models to predict changes in analysts' forecasts.

[10]See Appendix B in supplemental material for detailed definitions of these finance terms.

[11]There are missing values for less than 1% of the data. The missing values are mainly due to company acquisitions and changing of company names.

[12]Version 3.6.0

[13]UD Part of speech tags ADJ, ADV, ADV, AUX, INTJ, NOUN, PRON, PROPN, VERB.

		Regression Task				Classification Task			
Feature type	Feature	Model	MSE	R^2	% err.	Model	Acc.	F1	% err.
Baselines	Random (ave. 10 seeds)	–	0.32987	-199.9	–	–	0.340	0.338	–
	Training mean	–	0.00165	$-1e^{-5}$	0.0	–	–	–	–
	Predict 0	–	0.00177	-0.072	–	–	–	–	–
	Predict majority class	–	–	–	–	–	0.387	0.186	0.0
Market	Market	RR	0.00160	0.0478	3.0	LR	0.435	0.408	12.4
Semantic	Bag-of-words	RR-WD	0.00140	0.1500	15.2	LR-WD	**0.482**	**0.475**	**24.8**
		RR-Q&A	0.00165	-0.0043	0.0	LR-Q&A	0.388	0.189	0.3
	doc2vec	RR-WD	**0.00137**	**0.1718**	**17.0**	LR-WD	0.479	0.468	23.8
		RR-Q&A	0.00165	-0.0031	0.0	LR-Q&A	0.385	0.220	0.5
		LSTM	0.00155	0.0598	6.1	LSTM	0.442	0.400	14.2
Pragmatic	Pragmatic lexicons	LSTM	0.00164	-0.0020	0.6	LSTM	0.415	0.368	7.2
Fusion	doc2vec + prag	LSTM	0.00155	0.0573	6.1	LSTM	0.461	0.460	19.1
Ensemble	doc2vec + prag + market	Ens.	0.00154	0.0619	6.7	Ens.	0.460	0.461	18.9

Table 6: Test-set regression and classification results. Models are ridge regression (RR), long short-term memory networks (LSTM), logistic regression (LR), and ensemble (Ens.). *WD* denotes whole-document models, while *Q&A* denotes Q&A-only models. Evaluation metrics are mean squared error (MSE), the coefficient of determination (R^2), accuracy (Acc.), and macro-level F1. For regression, percent error reduction (% err.) is from the MSE of the baseline of predicting the training mean; for classification, it is from the accuracy of predicting the majority class.

5.3.1 Whole-document models

Ridge regression[14]**.** For regression, we use ridge regression[15] which has a loss function that is the linear least squares function and is regularized with an L2-norm. To tune hyperparameters, we perform a five-fold cross-validation grid search over the regularization strength[16]. We evaluate on mean squared error (MSE) and the coefficient of determination (R^2) scores.

Logistic regression[17]**.** For classification, we train logistic regression with a L2 penalty[18] and we tune C, the inverse regularization constant, via a grid search and 5-fold cross validation on the training set. We evaluate validation and test sets using accuracy and macro F1 scores.

5.3.2 Q&A-only models

In order to compare the relative influence of the *presentation* versus *question-answer* sections of the earnings calls, we remove the *presentation* portion of each call and only predict on the $Q\&A$ portion[19]. Except for this difference, Q&A-only models are identical to whole-document models.

5.3.3 Turn-by-turn models

LSTM for regression. We model transcripts as a sequence of turns using long-short term memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997). Let $x_t \in \mathbb{R}^k$ be the input vector at time t for embedding dimension k, and let L be the total length of the sequence. Each x_t is fed into the LSTM in order and produces a corresponding output vector h_t. Then the final output vector is passed through a linear layer $y = w^y h_L + b^y$ for output $y \in \mathbb{R}$ with $w^y \in R^k$. For a given mini-batch b, L_b is fixed as the maximum number of turns among all documents and the sequences for the other documents in the mini-batch are padded. The network is trained with mean squared error (MSE) loss.

LSTM for classification. The LSTM architecture for classification is similar to that used for regression except that there is an additional softmax layer after the final linear layer. This network is trained with cross-entropy loss.

Both LSTMs are trained via a grid search over the following hyperparameters: learning rate, hid-

[14]We also tried Kernel ridge regression with a Gaussian (RBF) kernel, which gave similar results. See Appendix C for more details.

[15]Implemented with `scikit-learn`.

[16]α in scikit-learn for values 10^{-3} to 10^8 by logarithmic scale.

[17]We also tried support vector machines; see Appendix C.

[18]Implemented with `sklearn`.

[19]Of the 12,285 documents, there were 246 that only contained the *presentation* section and did not have the *question-and-answer* section. In the Q&A modeling we completely remove these documents.

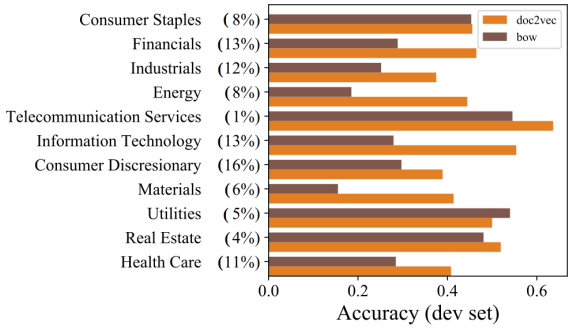

Figure 2: Per-industry breakdown of errors on the validation set for *doc2vec* (overall dev acc. 44.6%) and *bag-of-words* (bow) (overall dev acc. 30.4%) models. Y-axis denotes the 11 GICS industries and their percentage of documents across the entire corpus.

den dimension, batch size, number of layers, and L2-penalty (a.k.a. weight decay). The networks are written in Pytorch[20] and optimized with Adam (Kingma and Ba, 2014) .

5.3.4 Fusion and ensembling

Early fusion. We use early fusion (Atrey et al., 2010) to combine semantic and pragmatic feature vectors at every turn and feed these into a LSTM.

Ensembling via stacking. We use "stacked generalization" (Wolpert, 1992) (a.k.a. "stacking") to combine fusion and market-based models. For regression, we take the output values from the fusion and market-based models as features into a ridge regression model. For classification, we take the three-dimensional probability vector output from the fusion and market-based models and concatenate these as features into a logistic regression model. In both cases, hyperparameters are tuned on validation data.

5.3.5 Baselines.

We compare against several baselines: (1) random, drawing a random variable from a Gaussian centered at the mean of the training data, (2) predicting the mean change in forecast across all documents in the training set (regression), and (3) predicting 0, the majority class (classification).

5.4 Results.

See Table 6 for full results. We address our original **research questions** from the beginning of §5.

(1) **Predictiveness.** We find earnings calls are moderately predictive of changes in analysts' forecasts, with an almost 25% relative error reduction in classification accuracy from the baseline of predicting the majority class. While the accuracy of our best model may seem modest, for this task, analysts' decisions can be influenced by many external factors outside of the text itself and our ability to find any signal among the noise may be interesting to financial experts.

(2) **Text vs. market.** Semantic features are more predictive of changes in analysts' price targets than market features (a 24.8% error reduction over baseline for bag-of-words and a 23.8% reduction for doc2vec, vs. a 12.4% error reduction for market features).

(3) **Semantic vs. pragmatic.** Semantic features (doc2vec and bag-of-words) are more predictive than pragmatic features. This suggests the semantic content of the earnings call is important in how analysts make decisions to change their price targets.

(4) **Q&A-only vs. whole-doc.** Contrary to Matsumoto et al. (2011) who find the *question-answer* portions of earnings calls to be most informative, we find the Q&A-only models are much less predictive for doc2vec (accuracy 0.479 vs. 0.385) and bag-of-words (accuracy 0.482 vs. 0.388) models.

(5) **Whole-doc vs turn-level.** Whole-document models are more predictive than turn-level models (the best LSTM model achieves 19.1% error reduction over baseline, vs. 24.8% for the best whole-doc model). We hypothesize that turn-level models might capture more signal if they incorporate speaker metadata, e.g. the role of the speaker or the analysts' pre-calls judgment for the company. Although whole-document models are more predictive, turn-level analyses of analysts' behavior may be more useful to alerting stakeholders to predictive signals in real-time (e.g. an important analyst analyst question mid-way through a live earnings call) since financial markets can vary significantly in short time periods.

Breakdown of results by industry. We analyze errors on the validation data by segmenting earnings calls by each company's Global Industry Classification Standard (GICS) sector[21]. See Figure 2 for the breakdown results. Notably, the bag-of-words model performs almost 2.5 times worse on earnings calls from the *Materials* sector versus the *Utilities* and *Telecommunication Services* sectors. This suggests industry-specific models may

[20] https://pytorch.org/

[21] See https://www.msci.com/gics. There are 11 broad industry sectors.

be important in future work.

6 Conclusions and future work

In this work we (a) correlate pragmatic features of analysts' questions with the pre-call judgment of the questioner, (b) explore the influence of market, semantic and pragmatic features of earnings calls on analysts' subsequent decisions. We show that bullish analysts are more likely to ask slightly more positive and concrete questions, talk less about the past, and be called on earlier in a call. We also demonstrate earnings calls are moderately predictive of changes in analysts' forecasts.

Promising directions for future research include examining additional features and feature representations: pragmatic features such as formality (Pavlick and Tetreault, 2016) or politeness (Danescu-Niculescu-Mizil et al., 2013); acoustic-prosodic features from earnings call audio; more sophisticated semantic representations such as claims (Lim et al., 2016), automatically induced entity-relation graphs (Bansal et al., 2017) or question-answer motifs (Zhang et al., 2017) (these representations are non-trivial to construct because a single turn may contain many questions or answers); or even discourse structures. The models used in this work aim to be just complex enough to determine whether useful signals exist for this task; future modeling work could include training a complete end-to-end system such as a hierarchical attention network (Yang et al., 2016), or building industry-specific models.

Acknowledgments

We thank Sz-Rung Shiang, Christian Nikolay, Clay Elzroth, David Rosenberg, and Daniel Preotiuc-Pietro for guidance early on in this work. We also thank Abe Handler, members of the UMass NLP reading group, and anonymous reviewers for their valuable feedback. This work was partially supported by NSF IIS-1814955.

References

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379.

Trapit Bansal, Arvind Neelakantan, and Andrew McCallum. 2017. Relnet: End-to-end modeling of entities & relations. *arXiv preprint arXiv:1706.07179*.

Štefan Beňuš, Agustín Gravano, Rivka Levitan, Sarah Ita Levitan, Laura Willson, and Julia Hirschberg. 2014. Entrainment, dominance and alliance in Supreme Court hearings. *Knowledge-Based Systems*, 71:3–14.

Robert M Bowen, Angela K Davis, and Dawn A Matsumoto. 2002. Do conference calls affect analysts' forecasts? *The Accounting Review*, 77(2):285–316.

Lawrence D Brown, Andrew C Call, Michael B Clement, and Nathan Y Sharp. 2015. Inside the black box of sell-side financial analysts. *Journal of Accounting Research*, 53(1):1–47.

Belinda Crawford Camiciottoli. 2009. "Just wondering if you could comment on that": Indirect requests for information in corporate earnings calls. *Text & Talk-An Interdisciplinary Journal of Language, Discourse & Communication Studies*, 29(6):661–681.

Belinda Crawford Camiciottoli. 2010. Discourse connectives in genres of financial disclosure: Earnings presentations vs. earnings releases. *Journal of Pragmatics*, 42(3):650–663.

Lauren Cohen, Dong Lou, and Christopher Malloy. 2012. Casting conference calls. *Available at SSRN*.

Belinda Crawford Camiciottoli. 2018. Persuasion in earnings calls: A diachronic pragmalinguistic analysis. *International Journal of Business Communication*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Richard Frankel, Marilyn Johnson, and Douglas J Skinner. 1999. An empirical examination of conference calls as a voluntary disclosure medium. *Journal of Accounting Research*, 37(1):133–150.

Dov Fried and Dan Givoly. 1982. Financial analysts' forecasts of earnings: A better surrogate for market expectations. *Journal of Accounting and Economics*, 4(2):85–107.

Dan Givoly and Josef Lakonishok. 1980. Financial analysts' forecasts of earnings: Their value to investors. *Journal of Banking & Finance*, 4(3):221–233.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Mohammed Hasanuzzaman, Wai Leung Sze, Mahammad Parvez Salim, and Gaël Dias. 2016. Collective future orientation and stock markets. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*.

Tarek A Hassan, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. 2016. Aggregate and idiosyncratic political risk: Measurement and effects. *Available at SSRN*.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Why swear? analyzing and inferring the intentions of vulgar expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Stephan Hollander, Maarten Pronk, and Erik Roelofsen. 2010. Does silence speak? an empirical analysis of disclosure choices during conference calls. *Journal of Accounting Research*, 48(3):531–563.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Yohan Jo, Shivani Poddar, Byungsoo Jeon, Qinlan Shen, Carolyn P Rosé, and Graham Neubig. 2018. Attentive interaction model: Modeling changes in view in argumentation. *arXiv preprint arXiv:1804.00065*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Tim Koller and Rebecca Darr. 2017. Earnings calls are a waste of time and 3 other ways to fight the fast money. *MarketWatch*.

David F Larcker and Anastasia A Zakolyukina. 2011. Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, 50(2):495–540.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Jochen L Leidner and Frank Schilder. 2010. Hunting for the black swan: risk mining from text. In *Proceedings of the ACL 2010 System Demonstrations*.

Quanzhi Li and Sameena Shah. 2017. Learning stock market sentiment lexicon and sentiment-oriented word vector from stocktwits. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

Wee-Yong Lim, Mong-Li Lee, and Wynne Hsu. 2016. Claimfinder: A framework for identifying claims in microblogs. In *# Microposts*, pages 13–20.

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

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Dawn Matsumoto, Maarten Pronk, and Erik Roelofsen. 2011. What makes conference calls useful? the information content of managers' presentations and analysts' discussion sessions. *The Accounting Review*, 86(4):1383–1414.

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

John Melloy. 2018. Here are highlights of Elon Musk's strange Tesla earnings call: 'They're killing me'. *CNBC*.

Elena Musi. 2018. How did you change my view? a corpus-based study of concessions argumentative role. *Discourse Studies*, 20(2):270–288.

Dan Palmon, Ke Xu, and Ari Yezegel. 2016. What does 'but' really mean?–evidence from managers' answers to analysts' questions during conference calls. *Available at SSRN*.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association of Computational Linguistics*.

Yangtuo Peng and Hui Jiang. 2016. Leverage financial news to predict stock price movements using word embeddings and deep neural networks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop*

on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality.

Ellen F Prince, Joel Frader, Charles Bosk, et al. 1982. On hedging in physician-physician discourse. *Linguistics and the Professions*, 8(1):83–97.

Anna Prokofieva and Julia Hirschberg. 2014. Hedging and speaker commitment. In *Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data*.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Alexander Dür, Linda Andersson, and Allan Hanbury. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based ir models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kate Suslava. 2017. 'Stiff business headwinds and unchartered economic waters': The use of euphemisms in earnings conference calls. *Available at SSRN*.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the International World Wide Web Conference (WWW)*.

Paul C Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467.

Chuan-Ju Wang, Ming-Feng Tsai, Tse Liu, and Chin-Ting Chang. 2013. Financial sentiment analysis for risk prediction. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*.

William Yang Wang and Zhenhao Hua. 2014. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

David H Wolpert. 1992. Stacked generalization. *Neural Networks*, 5(2):241–259.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. 2017. Asking too much? the rhetorical role of questions in political discourse. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

A Calls Used in This Work

See https://kakeith.github.io/attach/acl2019_supplement.pdf for the list of earnings calls used in this work, i.e. all earnings call transcripts available to us for every company that was in the S&P 500 on the date of the call, between 2010 and 2017 inclusive. The overall number of S&P 500 companies in our data (642) is greater than 500 because we look at company inclusion in the S&P 500 index *daily*; companies regularly enter and leave this index.

B Additional Details Regarding Definitions and Sources of Features

B.1 Market features

The **relative price/earnings ratio** is a stock's price/earnings ratio relative to the price/earnings ratio of a relevant index, in this case the S&P 500.
The **EBIT yield** is equivalent to the (trailing 12-month operating income per share / last price) *100.
The **earnings yield** is equivalent to the (trailing 12-month earnings per share before extraordinary items)/ last price) *100.

B.2 Pragmatic lexicons

B.2.1 OntoNotes five-coarse grained groups

For the pragmatic entity features, we construct five coarse-grained groups from the fine-grained entity types of OntoNotes[22] (Hovy et al., 2006): (1) events (OntoNotes' EVENT); (2) numbers (OntoNotes' ORDINAL, MONEY, PERCENT, CARDINAL, TIME, DATE, QUANTITY); (3) organization/locations (OntoNotes' LOC, NORP, FACILITY, GPE, LOCATION, ORGANIZATION); (4) persons (OntoNotes' PERSON); and (5) products (OntoNotes' PRODUCT).

[22] Version 5, https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf Section 2.6

		Regression Task				Classification Task			
Feature type	Features	Models	MSE	R^2	% err.	Models	Acc.	F1	% err.
Market	Market	GK	0.00163	0.0117	1.2	SVM	0.423	0.379	9.3
Semantic	Bag-of-words	GK	0.00152	0.0765	7.9	SVM	–	–	–
	doc2vec	GK	0.00140	0.1513	15.2	SVM	0.476	0.455	23.0

Table 7: Results on the test set for additional models. Comparable to Table 6 in the main document.

B.2.2 Sentiment

As a financial sentiment lexicon, We used the positive and negative word lists from: https://sraf.nd.edu/textual-analysis/resources/[23] (Loughran and McDonald, 2011), as retrieved in July 2018.

As a general sentiment lexicon, we used the SO-CAL dictionary from: https://github.com/sfu-discourse-lab/SO-CAL/tree/master/Resources/dictionaries/English (Taboada et al., 2011), as retrieved in July 2018.

If a unigram appears in opposite categories for the general and financial sentiment lexicons, we defaulted to the sentiment given by the financial sentiment lexicon. There were 14 instances of terms defined as positive in SO-CAL and negative in Loughran-McDonald: *unpredictably, conviction, correction, force, seriousness, toleration, missteps, overcome, condone, tolerate, exonerate, upset, challenging, unpredictable*.

We also deleted *question* and *questions* from the negative Loughran-McDonald list since these were abundant in the question-answer portions of earnings calls.

B.2.3 Hedging

We used the unigram and ngram hedging dictionaries from https://github.com/aproko/hedge_nn (Prokofieva and Hirschberg, 2014), as retrieved in July 2018.

B.2.4 Uncertainty, Litigiousness, Modal, Constraining

We used the word lists from https://sraf.nd.edu/textual-analysis/resources/[24] (Loughran and McDonald, 2011), as retrieved in July 2018.

C Other modeling experiments

For the prediction task in §5, in addition to ridge regression and logistic regression, we also experimented with Gaussian kernel ridge regression and support vector machines but found they performed worse or similarly. See Table 7 for the full results.

C.1 Gaussian kernel ridge regression.

Kernel ridge regression combines ridge regression with the kernel trick and we implement the model with `sklearn`. We use a Gaussian (RBF) kernel. To tune hyperparameters, we perform a five-fold cross-validation grid search over the regularization strength, α, and the inverse of the radius of influence of samples selected by the model as support vectors, γ.

C.2 SVC with RBF kernel.

We also train a support vector classifier (SVC) with an RBF kernel and we implement the model with `sklearn`. We tune the hyperparameters "C" (the penalty parameter of the error term) and gamma (free parameter of the Gaussian radial basis function). The SVM trained on the bag-of-words features ran out of memory, even on a machine with a large amount of RAM.

[23] Archived at https://web.archive.org/web/20181203160914/https://sraf.nd.edu/textual-analysis/resources/

[24] Archived at https://web.archive.org/web/20181203160914/https://sraf.nd.edu/textual-analysis/resources/