An Exploratory Study of Stock Price Movements from Earnings Calls

Sourav Medya[1], Mohammad Rasoolinejad[1], Yang Yang[2], Brian Uzzi[1]
[1]Kellogg School of Management & Northwestern Institute on Complex Systems, Northwestern University
[2]Syracuse University
{sourav.medya,mohammad.rasoolinejad,uzzi}@kellogg.northwestern.edu
从文本软数据 预测股价走势，效果超过交易硬数据 yyang87@syr.edu

ABSTRACT

Financial market analysis has focused primarily on extracting signals from accounting, stock price, and other numerical "hard" data reported in P&L statements or earnings per share reports. Yet, it is well-known that decision-makers routinely use "soft" text-based documents that interpret the hard data they narrate. Recent advances in computational methods for analyzing unstructured and soft text-based data at scale offer possibilities for understanding financial market behavior that could improve investments and market equity. A critical and ubiquitous form of soft data are earnings calls. Earnings calls are periodic (often quarterly) statements usually by CEOs who attempt to influence investors' expectations of a company's past and future performance. Here, we study the statistical relationship between earnings calls, company sales, stock performance, and analysts' recommendations. Our study covers a decade of observations with approximately 100,000 transcripts of earnings calls from 6,300 public companies from January 2010 to December 2019. In this study, we report three novel findings. First, the buy, sell and hold recommendations from professional analysts made prior to the earnings have low correlation with stock price movements after the earnings call. Second, using our graph neural network based method that processes the semantic features of earnings calls, we reliably and accurately predict stock price movements in five major areas of the economy. Third, the semantic features of transcripts are more predictive of stock price movements than sales and earnings per share, i.e., traditional hard data in most of the cases.

KEYWORDS

Earnings call, stock price movement, natural language processing

ACM Reference Format:
Sourav Medya[1], Mohammad Rasoolinejad[1], Yang Yang[2], Brian Uzzi[1]. 2018. An Exploratory Study of Stock Price Movements from Earnings Calls. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/1122445.1122456

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Earnings calls are the discussion sessions that happen after the public companies release their financial data for the reporting period which is usually a quarter of the year or a fiscal year. Analysts, investors, and journalists are often present during the earnings calls and the recordings of these can be accessed through corresponding company websites.

In this session, usually the CEO and/or other representatives from the management of the company present their financial achievements during the last quarter and give guidelines for the next. The management usually discusses details about important company information such as growth, risks, purchases, liabilities, lawsuits, share buybacks, increase/decrease in dividends, any change in executive teams and future goals. The session usually consists of management discussion and analysis (MD&A section) followed by a question-answering session involving the audience and the investors.

The earnings call is a major event as the stock price market reflects higher level of volatility and trading volume prior, during and after earnings calls [9]. Such volatility can result in bad investments, missing profit opportunities and huge losses for the investors. The large amount of available data on transcripts and stock market prices give us an opportunity to predict the directions of stock price movements more accurately. Moreover, there are other factors such as sentiment on social media and news that can also affect the stock market and have been studied in the literature [14, 36]. However, in this paper, we focus on associating the stock price movements from the transcripts of the earnings calls [28, 33].

Figure 1 shows an example of the stock price changes for the company Tesla, Inc. in 2019. The price movement after the occurrence of the earnings call can be positive or negative. As the figure shows daily effect, we define the label of the transcripts based on one day effect in the stock price (Def. 1 and Def. 2). While the price change is important, it is also meaningful to compare the change with respect to the broader market behavior. To achieve that we compare the rate of the stock price changes with the corresponding index (Table 1) over the next five business days. The labels of the transcripts are decided based on the majority of out-performances or under-performances with respect to the index value (Def. 3).

In this paper, we perform an extensive study on the power of the earnings calls to predict stock price movements. We formally define the stock price movement as three different classification problems. First, we establish that the analysts' ratings prior to the earnings call are not indicative of the stock price movement prediction while the emotions in the transcripts play a significant role. This motivates us to design more rigorous methods based on semantics of the transcripts. Using a graph neural network based method, we show

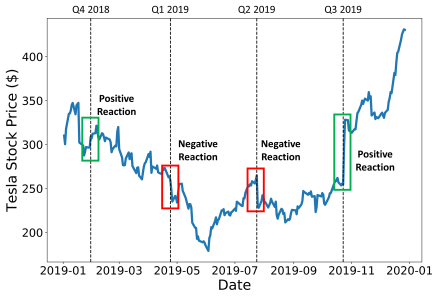


Figure 1: Reactions on stock price movements around occurrences of earnings calls from Tesla, Inc. during 2019.

that the earnings calls can predict upward or downward movements of the stock prices more accurately than alternative signals. Furthermore, while the differences between actual earnings and sales with the estimated ones impact the stock price movements, these have lower predictive power than the earnings call.

Our main contributions can be summarized as follows:

- We study the problem of stock movement prediction from earnings calls and formulate three classification problems considering the daily and weekly effect of the earnings call on stock prices.
- We provide descriptive analyses on labels as well as sentiments for our created large dataset of approximately 100,000 transcripts.
- From the recommendations of analysts, we demonstrate that analysts' ratings available prior to the earnings calls do not have a significant impact on the stock price changes while the emotion traits available in the transcripts play a significant role in predicting stock price movements.
- Through experimental evaluation, we show that our proposed graph neural network based method can predict stock price movement from earnings call transcripts with relatively high quality.
- We show that the transcripts have more predictive power than actual and estimated values of sales and earnings per share. In particular, the semantic features of transcripts produce at least 10% higher average recall and 33% higher average precision in two of the fastest growing areas of the economy, "Technology" and "Services".

2 STOCK PRICE MOVEMENTS

We formulate the stock price movements as binary labels. Let $\mathbb{C} = \{C_1, C_2, \cdots, C_m\}$ be the set of m companies. We denote the closing stock price of the company c on the day d as S^c_d. As the stock prices vary during a day, we consistently use the closing price of the stock on a particular day throughout our analysis. For a company c, the set of t transcripts are denoted by $\mathbb{T}^c = \{T^c_{d_1}, T^c_{d_2}, \cdots, T^c_{d_t}\}$ where $T^c_{d_i}$ represents the earnings call transcript on day d_i. The value of t can vary depending on the history of earnings calls.

2.1 Daily Movement

Our aim is to measure the local impact of the earnings call transcripts on the corresponding stock values. Thus, we define stock price movements (positive or negative) based on the stock values on the day before and after the occurrence of the earnings call. More specifically, for a company c, if the earnings call happen on the day d, the stock values on the $d-1$ and $d+1$ are compared and an upward and downward movements get binary values of 1 and 0 respectively. The $d+1$ and $d-1$ days denote the next and previous business days of the day d. Formally, this value based stock movement label of a transcript can be defined as follows:

Definition 1. ***Value Based Label Function (VBL):*** *We define the label function $y_v(T^c_d) \in \{0, 1\}$ for a transcript T^c_d of a company c on the day d as follows :*

$$y_v(T^c_d) = \begin{cases} 1, & if\, S^c_{d+1} > S^c_{d-1} \\ 0, & otherwise \end{cases}$$

where $d+1$ and $d-1$ denote the next and previous business days respectively of the day d.

The previous definition does not account for the value of the change. We capture a significant amount of change (**shock**) both in upward and downward directions in another label function:

Definition 2. ***Shock Based Label Function (SBL):*** *We define the label function $y_s(T^c_d) \in \{0, 1\}$ for a transcript T^c_d of a company c on the day d as follows :*

$$y_s(T^c_d) = \begin{cases} 1, & if\, \frac{S^c_{d+1} - S^c_{d-1}}{S^c_{d-1}} \geq \tau \\ 0, & if\, \frac{S^c_{d-1} - S^c_{d+1}}{S^c_{d-1}} \geq \tau \end{cases}$$

where $d+1$ and $d-1$ denote the next and previous business days respectively of the day d; τ is a threshold (we set it as .05 i.e., 5%).

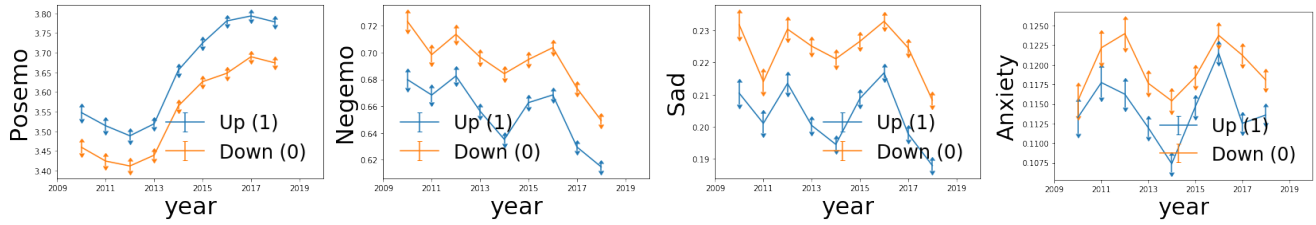
Our aim is to investigate the predictive power of the earnings call transcripts of the stock price movement. To do so, we learn a prediction function f, where the features constructed from the earning calls are given as input and the defined labels (y_v and y_s) act as output variables (Sec. 5).

Sector-wise comparison: In the experiments (Sec. 5.2), we perform classification tasks based on the aforementioned labels for the companies that are in the same sector. Each index is a measure showing the performance of companies in terms of stock values in the corresponding sector. The indices for different sectors are given in Table 1. Note that we compute the index value based labels using the corresponding sector index.

2.2 Weekly & Normalized Movement

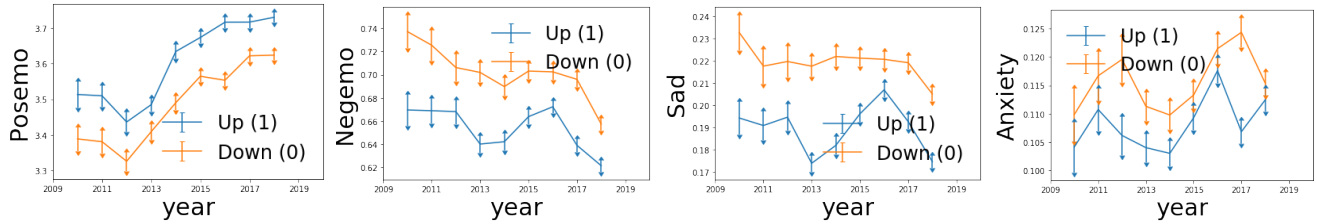
The previous definitions of the labels for transcripts do not capture the stock movement compared to the broader market such as companies in the same sector. We further use index (see Table 1) values to make a comparison and define positive and negative movements accordingly. Additionally, the previous labels consider stock values for only one day after the earnings call happens. We extend this notion of locality to one week (i.e., five business days) and compare it with the index values for next five business days.

Let I_d denote the index value for the day d. We compare the rates of increase in the stock prices and index values. Let an earnings call (T^c_d) occur for a company c on the day d. The rate of increase in the value is computed based on the previous day. So, the rate of increase (R) on the day $d+1$ for the stock value would be, $R(S^c_{d+1}) = \frac{S^c_{d+1} - S^c_d}{S^c_d}$. The stock movement is upward on the day $d+1$ if $R(S^c_{d+1}) > R(I_{d+1})$ and we denote this upward movement by 1. Otherwise, we call it downward movement and denote by 0. From this definition we construct a 5-dimensional vector $\mathbb{L}(T^c_d)$ containing binary values



(a) Positive (b) Negative (c) Sad (d) Anxiety

Figure 2: Value Based Label (y_v): The average sentiment scores in the transcripts over the years for both labels (0 and 1). As expected, while the positive sentiment is higher for upward labels over the years; negative, sad and anxiety sentiments are higher for transcripts downward labels.



(a) Positive (b) Negative (c) Sad (d) Anxiety

Figure 3: Shock Based Label (y_s): The average sentiment scores in the transcripts over the years for both labels (0 and 1). As expected, consistent with the case in value based label, the positive sentiment higher for transcripts with upward labels; negative, sad and anxiety sentiments are higher for transcripts with downward labels.

for the next five business days after the day d. We define a parameter k that controls the output of label based on k zeros and ones in $\mathbb{L}(T_d^c)$.

Definition 3. ***Index Based Label Function (IBL):*** *We define the label function $y_{I,k}(T_d^c) \in \{0, 1\}$ for a transcript T_d^c of a company c on the day d as follows :*

$$y_{I,k}(T_d^c) = \begin{cases} 1, & if \mathbb{L}(T_d^c) \text{ has} \geq k \text{ 1s} \\ 0, & if \mathbb{L}(T_d^c) \text{ has} \geq k \text{ 0s} \end{cases}$$

Similarly, as in Sec. 2.1, we learn a prediction function where the features constructed from the earning calls are given as input and the defined label ($y_{I,k}$) acts as an output variable. Note that we consider k to be between 3 and 5 as we define the final label based on majority of zeros or ones in $\mathbb{L}$.

While the daily (value and shock) labels are relevant, they are miopoic in terms of investments and have high implied volatility [9]. The weekly index based label (IBL) reflects a company's performance over a week (less miopic and less volatile). This also helps to show whether the earnings calls are a predictor of the stock price movement for a longer duration.

3 DATA

Our study includes three datasets— earnings call transcripts, earnings per share (EPS) and sales estimates, and recommendations from analysts. We have collected transcripts that are available online[1]. The other datasets are from Zack investment[2] dataset obtained via Wharton Research Data Services (WRDS)[3]. We have collected full texts of 152,483 transcripts of which 97,478 have complete data; the remaining transcripts were unusable due to missing data. The earnings calls for these transcripts happen between January of 2010 to December of 2019. The dataset has all the relevant information such as the name of the company, ticker symbol, transcript release date, industrial sector, and quarter of the earnings. These transcripts are from 6,300 different companies that belong to one of nine predefined sectors as shown in Table 1.

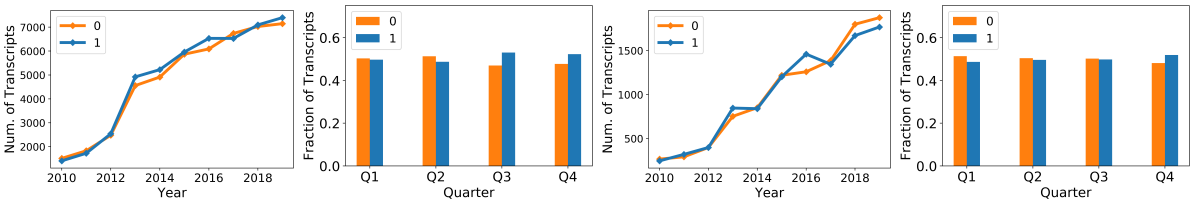
Sector	Ref. Index	Percentage
Services (Service)	IYC	18.5
Technology (Tech)	XLK	18.3
Financial (Fin)	XLF	18.1
Healthcare(Health)	XLV	12.6
Basic Materials (Mat)	XLB	11.3
Consumer (Con)	XLY	9.2
Industrial (Ind)	XLI	8.8
Utilities (Util)	XLU	3.0
Not Specified	SP500	0.78

Table 1: Percentage of transcripts in different company sectors: the sector of a company is determined from the reference index (Ref. Index).

3.1 Descriptive Analysis with Sentiments

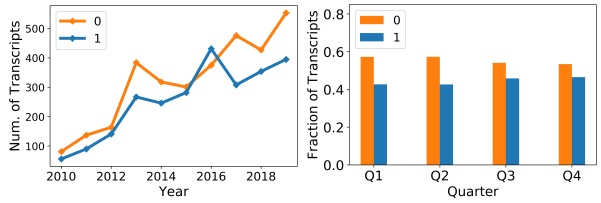
We use the Linguistic Inquiry and Word Count (LIWC)[4] dictionary to identify the sentiment traits in the transcripts. We include positive and negative emotions, anxiety, and sadness as affect traits among others. Figs. 2 and 3 show the year-wise sentiments in the transcripts for y_v (value label) and y_s (shock label) respectively (the results with the index label ($Y_{I,5}$) are similar and omitted due to space). We make two interesting observations: (1) While the positive sentiment is higher for upward labels over the years; negative, sad

[1] https://seekingalpha.com/
[2] https://www.zacks.com/
[3] https://wrds-www.wharton.upenn.edu/pages/about/data-vendors/zacks/
[4] https://www.kovcomp.co.uk/wordstat/LIWC.html



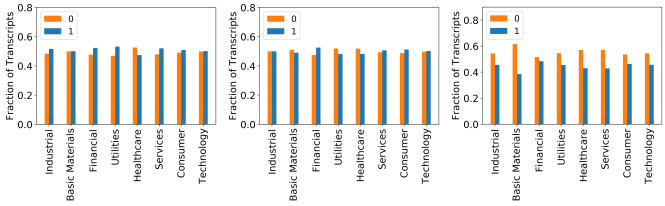
(a) Year-wise for y_v (b) Quarter-wise for y_v (c) Year-wise for y_s (d) Quarter-wise for y_s

Figure 4: The (a) number of transcripts over the years and (b) fraction of transcripts in different quarters with both labels (orange for 0 and blue for 1) for the value based label (y_v). The (c) number of transcripts over the years and (d) fraction of transcripts in different quarters for the shock based label (y_s).



(a) Year-wise for $y_{I,5}$ (b) Quarter-wise for $y_{I,5}$

Figure 5: The (a) number of transcripts over the years and (b) fraction of transcripts in different quarters for the index based label ($y_{I,k}$ when $k = 5$).



(a) y_v (b) y_s (c) $y_{I,5}$

Figure 6: The fraction of transcripts with both labels (0 and 1) in different sectors when the labels are (a) value based (y_v) (b) shock based label (y_s) and (b) index based ($y_{I,5}$).

and anxiety sentiments are higher for downward labels. (2) Over the years, the positive emotion has been increased and the negative and sad emotions have been decreased indicating that the earnings calls are being more enthusiastic over the years. We perform a detailed analysis of the role of sentiments in the stock price movement prediction in Sections 4.2 and 8.5.

3.2 Descriptive Analysis on Label Distributions

We show descriptive analysis of the data on the distribution of the labels. Figs. 4 and 5 represent the number and fraction of the transcripts in both classes (0 and 1) over the years as well as in specific quarters as the earnings calls happen quarterly from a company. Unsurprisingly, the number of transcripts increase (Figs. 4a, 4c, and 5a) over the years for all label functions as the number of company increases. After aggregating the transcripts over different quarters, we find there is no significant difference in number of transcripts between two classes (0 and 1) for the value and shock based labels (Figs. 6a and 6b). However, this difference is more prominent in the index based label where class 0 is larger (Fig. 6c).

Variables	Model 1	Model 2	Model 3	Model 4	Model 5
Positive Emotion		0.26***	0.26***	0.26***	0.29***
		(0.012)	(0.013)	(0.013)	(0.018)
Negative Emotion		-0.64***	-0.33***	-0.34***	-0.32***
		(0.033)	(0.057)	(0.057)	(0.077)
Anxiety Score			-0.26*	-0.24*	-0.19
			(0.11)	(0.11)	(0.16)
Anger Score			0.17	0.19	0.19
			(0.12)	(0.12)	(0.16)
Sad Score			-0.73***	-0.75***	-1.22***
			(0.091)	(0.092)	(0.13)
Certainty Score			-0.0097	0.029	0.061
			(0.032)	(0.036)	(0.050)
Cognitive Score				-0.019**	-0.034**
				(0.0075)	(0.011)
Insight Score				-0.020	-0.038
				(0.021)	(0.029)
Causation Score				-0.024	0.021
				(0.021)	(0.029)
Discrepancy Score				0.062*	0.027
				(0.029)	(0.042)
Actual Sales					8.33e-05***
					(2.36e-05)
Estimated Sales					-8.71e-05***
					(2.42e-05)
Actual EPS					0.0056
					(0.0041)
Estimated EPS					-0.0059
					(0.0043)
Sector	YES	YES	YES	YES	YES
Year	YES	YES	YES	YES	YES
Observations	93,682	93,165	93,165	93,165	51,306
BIC	129,921	128,040	127,968	127,994	70,465

Table 2: Logistic Regression of Value Based Label Function (y_v) on sentiment features. Standard errors are in parentheses. Significance levels marked with stars: *** p<0.001, ** p<0.01, * p<0.05. YES denotes inclusion of the variable. Bayesian Information Criterion (BIC) [6] is used for model selection and lower BIC is preferred.

4 PILOT REGRESSION ANALYSIS

In this section, we run several regression analyses to explore the association between text narratives and corporate stock market performance that are defined by y_v, y_s and $y_{I,k}$ (see details in Section 2).

4.1 Effect of Sentiments

Our objective is to validate whether information in the text of the earnings call transcripts have predictive power after controlling for several frequently used performance indicators of corporate's performance, such as earnings per share (EPS) and sales (see details in Section 8.1).

Here, we use y_v, y_s and $y_{I,k}$ as our dependent variables in the regression (logistic). Our key independent variables are quantitative measures (e.g., sentiments) extracted from the transcripts. We quantify sentiments using LIWC dictionary. There are two categories

of them. First, we have major metrics, which quantify the levels of positive and negative emotion along with more detailed emotion information, such as anxiety, anger, and sadness. Second, we also measure personality features in the texts, such as how certain the language is, how many insight words are used, and how many cognitive processes words are used in the speech among others.

In Table 2, we present the estimates (coefficients, standard errors, significance) of the variables in five regression models where y_v is the dependent variable along with *year* and *sector* as fixed effects. We have several interesting observations. First, we find that positive emotion and negative emotion are significantly predictive of stock market performance in terms of y_v (model 2 in Table 2). Second, both of them remain significant after controlling for corporate's actual sale, estimated sale, actual EPS, and estimated EPS. This implies that the sentiment information extracted from texts can provide useful signal for stock price movements other than frequently used indicators (e.g., sales, EPS). Thirdly, the signs of positive sentiment and negative sentiment are in the opposite directions. This means that earnings call transcripts with positive emotion are predicting stock price increase. On the other hand, the amount of negative emotion in the transcripts correlates with stock price decrease. Similar results are found for sadness score and cognitive process words. In Appendix (Table 8), we present the same analyses for the shock based label y_s as dependent variable. All results are consistent with our above observations.

To summarize, this pilot analysis using metrics from texts indicate a connection between transcript narratives and stock market performance. This leads us to design more rigorous NLP methods based on semantics (Sec. 5).

4.2 Sentiment of Transcripts and Analysts' Recommendations

We explore the effect of recommendations from the analysts in this section for stock price movement prediction. Section 4.2.2 shows that while analysts' recommendations prior to earnings calls are not significant, the sentiment traits of the transcripts are important.

探索分析师的建议与我们提出的基于价值（价值）的标签之间的关联

4.2.1 Settings. Analysts play a role in affecting the perceptions of the investors about the future prospects of a business [1, 3, 12] and thus their recommendations can affect the stock price. We aim to explore the association between the analysts' recommendations and our proposed value based (y_v) label (the results on index ($y_{I,k}$) based label are in Sec. 8.2.1 of the Appendix). The recommendations are captured in a categorical variable with five possible values: strong buy, moderate buy, hold, moderate sell, and strong sell. The number of recommendations ranges from 10 to 500 for each company.

To measure the extent of the associations between the recommendations and the earnings calls, the dates of occurrences are important. The analysts' recommendations are not always available on the dates of the earnings calls. Thus, we create variables correspond to the "majority recommendation" on a company within time t prior to the date of the earnings call. For instance, if t is 1 month, a company's earnings calls date is 2015-09-20, 10 recommendations are made since 2015-08-20, and six of them (or the majority) are marked as "hold"; the value of the recommendation variable corresponding to this earnings call will be "hold". In our analysis we set t as one month and denote this Monthly Analysts' Recommendation (MAR) variable as MAR_{1m}. Additionally, we create another variable to measure the "local after-affect". In particular, we aggregate the recommendations from the analysts within 5 days after the earnings calls date and assign the values with majority recommendations as before. We denote it as MAR_{5d}.

Variables	Model 1	Model 2	Model 3	Model 4	Model 5
positive	0.27***	0.27***	0.26***	0.27***	0.34***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
negative	-0.43***	-0.43***	-0.42***	-0.42**	-0.63***
	(0.12)	(0.12)	(0.13)	(0.13)	(0.19)
anxiety	0.35	0.35	0.28	0.28	0.07
	(0.23)	(0.23)	(0.24)	(0.24)	(0.35)
anger	0.56*	0.54*	0.57*	0.56*	0.62
	(0.24)	(0.24)	(0.25)	(0.25)	(0.36)
sad	-0.66***	-0.58**	-0.63**	-0.63**	-0.57
	(0.19)	(0.19)	(0.20)	(0.20)	(0.29)
certain	-0.01	-0.02	-0.01	-0.00	0.12
	(0.07)	(0.07)	(0.07)	(0.07)	(0.11)
MAR_{1m}		YES	YES	YES	YES
Sector			YES	YES	YES
Year				YES	YES
MAR_{5d}					YES
Observations	18,289	18,289	18,289	18,289	8,994
BIC	25185	25201	25271	25346	12474

Table 3: Regression Results: Value Based Label Function (y_v) on Analysts' Recommendations within 1 month (MAR_{1m}) prior and 5 days (MAR_{5d}) after the earnings call date and sentiment of the transcripts. Standard errors are shown in parentheses for the coefficients. Significance levels marked with stars: *** p<0.001, ** p<0.01, * p<0.05. YES denotes the inclusion of the variable.

4.2.2 Effect of Sentiments and Recommendations. We attempt to understand the associations among analysts recommendations, sentiment of the earnings call transcripts and our defined labels (classes). We use p-value [35] to evaluate the significance of a variable. Bayesian Information Criterion (BIC) [6] is used for the performance measure in model selection and the model with the lowest BIC is preferred.

The key independent variables are analysts' recommendations and sentiment of earnings call transcripts. For analysts' recommendations, the variables MAR_{1m} and MAR_{5d} describe analysts' ratings one month prior to and five days after the day of earnings calls respectively. There are six variables related to sentiments of transcripts: *positive*, *negative*, *anxiety*, *anger*, *sad*, and *certain*.

Results: Table 3 shows the regression results for y_v. Model 1 only takes six sentiment variables as independent ones. In Model 2, we add analysts recommendations (MAR_{1m}) as another independent variable. Similarly, we add fixed effect of the areas of the companies, fixed effect of the year of the earnings calls, and analysts' recommendations after Earnings call (MAR_{5d}) in Model 3, Model 4 and Model 5 respectively. The values in the same row indicate the coefficients of the variables with significance levels (p-values).

The results on the sentiments are consistent with the ones in Sec. 4.1. Moreover, adding prior recommendations analysts, industry information and year do not improve the model's fitness as the BIC values in Model 2, 3 and 4 are larger (smaller means better) than that of Model 1. However, Model 5 has a much lower BIC value. The reasons could be the followings: the sample size (N) is different; and Model 5 includes posterior analysts' recommendations MAR_{5d}, which intuitively should have better explanation power of the dependent variable y_v.

Implications: There are two major implications of these results.

- The ratings from the analysts, that are made prior to the earnings call do not have significant impact on the stock price movement (SPM) prediction after the earnings call. These ratings are usually categorical variables and thus, these might not translate directly to the actions that investors might take and affect SPM. However, sentiments can be easily visible or interpretable to the investors through special adjective or words.
- As analysts are domain experts, they usually have a global viewpoint on the growth of the company. However, our defined binary classification tasks capture a localized trend after the occurrence of the earnings call. Therefore, global viewpoints that are made prior to it, might not be useful to capture such local effects.

5 EARNINGS CALLS AND STOCK PRICE MOVEMENT PREDICTION

Our goal is to learn prediction functions where the features are from the earnings call transcripts with value based (y_v), shock based (y_s) and index based $y_{I,k}$ labels. We begin with descriptions of our approaches to solve these classification problems. We show the main results in Sec. 5.2 and a comparison against a baseline consisting of "hard" data such as earnings per share (EPS) and sales in Sec. 5.3.

5.1 Our Method

To test the predictive power of the content in the earnings call, we use their semantics to solve the aforementioned problems. Our main method StockGNN is a combination of a graph neural network (GNN) architecture [18] and a semantic feature generator by the well-known Doc2Vec method [21].

The StockGNN Method: GNN based methods have been popular tools in several natural language processing and related tasks such as event detection [8, 15, 19, 37]. Inspired by the method in [37], our architecture (Figure 7) is based on a Gated GNN [24] and have four following components:

1) Graph generation: We first build a graph $G = (V, E)$ from each document. Each unique word in that document becomes a node and the words appear in its neighbourhood (or context) become its neighbors. The neighbourhood size becomes a hyper-parameter.

2) Gated GNN: After building the graph, a Gated GNN is applied to learn the node (word) embeddings. We start with initial (given or constructed features) embeddings ($z \in \mathbb{R}^{|V| \times d}$, a vector of size d) of the nodes or words. The gated GNN produces an embedding of node x based on aggregation of its own features and its neighbours' features. To achieve higher order (k-hop neighbours) interactions, this process is repeated k times. More specifically, the Gated GNN [24] interactions are captured by the following equations:

$$x^k = \sigma(\mathbf{W}_x a^k + \mathbf{U}_x z^{k-1} + b_x) \text{ where } a^k = \mathbf{A} z^{k-1} \mathbf{W}_a$$
$$r^k = \sigma(\mathbf{W}_r a^k + \mathbf{U}_r z^{k-1} + b_r)$$
$$\hat{z}^k = tanh(\mathbf{W}_z a^k + \mathbf{U}_z (r^k \otimes z^{k-1}) + b_z)$$
$$z^k = \hat{z}^k + z^{k-1} \otimes (1 - x^k)$$

Here, $\otimes$ is element-wise multiplication, σ is the sigmoid function, z^k is the final embedding after k-hop interactions, $\mathbf{A}$ is the adjacency matrix of the graph, $\mathbf{W}, \mathbf{U}$ and b are trainable weights (parameters) and biases. Furthermore, x and r are update and reset gate respectively and they are used to determine the degree of information

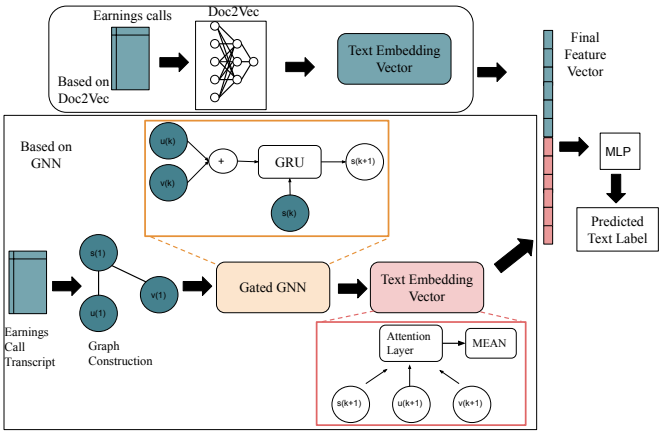


Figure 7: The architecture of StockGNN: It combines traditional context based Doc2Vec embeddings with GNN based embeddings that capture text level word interactions.

from the neighbours and the current node into consideration. As the words in the same document form a graph, Gated GNN produces embeddings based on both sentence (neighbours are created from context window) and document level interactions.

3) Embedding vector: The classification label is associated with the document, i.e., the corresponding graph. So the node embeddings learned via the Gated GNN are further aggregated by a function and generate final embeddings (z_G) for the entire graph or document. The aggregation function is as follows:

$$z_G = \frac{1}{|V|} \sum_{v \in V} z_v \text{ where } z_v = \sigma(MLP(z_v^k)) \otimes tanh(MLP(z_v^k))$$

where z_v is embedding for the node v and MLP is Multilayer Perceptron. Afterwards the embeddings are fed into a softmax layer and trained by a cross-entropy function.

4) Combining embeddings via Doc2Vec and final classification: We further generate unsupervised embeddings via the well-known Doc2Vec method [21] and concatenate with the (supervised) embeddings from the previous step. Thus, we exploit having both unsupervised embeddings based on content similarity of transcripts from Doc2Vec and sophisticated supervised ones from Gated GNN that produces embeddings based on both sentence and document level interactions. These final embeddings are then fed into a MLP to generate the final classification.

The main advantages of StockGNN are threefold: (i) easy to implement, (ii) captures both sentence and document level interactions, (iii) can be generalized with other GNN architectures. In experiments, we show that StockGNN also predicts relatively accurate daily and weekly stock price movements with all labels.

Other methods: Our aim is to show the predictive power of the semantic features constructed from the transcripts as well as to measure the accuracy of the proposed StockGNN method compared to baselines. To achieve that we propose several baselines which have two phases. First, we generate unsupervised low-dimensional embeddings via the well-known Doc2Vec method [21]. Intuitively, similar documents will be embedded closer in that low-dimensional space. Second, these embeddings are further fed into a classifier for the final classification or label prediction. We have used three different classifiers: Support Vector Machine (SVM), Logistic Regression

Measures	Accuracy					Avg. Precision					Avg. Recall				
Methods	Fin	Health	Mat	Service	Tech	Fin	Health	Mat	Service	Tech	Fin	Health	Mat	Service	Tech
DEsvm	.624	.549	.74	.647	.527	.633	.533	.661	.58	.529	.632	.526	.605	**.592**	.525
DElogreg	.595	.524	.632	.630	.534	.609	.516	.584	.57	.542	.606	.516	.610	.56	.545
DEmlp	**.666**	.540	.724	.595	.565	**.664**	.524	.615	.564	.538	**.665**	.520	.575	.568	.539
StockGNN	.582	**.60**	**.761**	**.65**	**.583**	.631	**.60**	**.69**	**.581**	**.554**	.61	**.578**	**.614**	.55	**.553**

Table 4: Index Based Label ($y_{I,5}$) Results: The accuracy, average precision and recall produced by different methods in five major sectors or areas. The maximum standard deviations in the results for DEsvm, DElogreg, DEmlp, and StockGNN are .002, .002, .008 and .003 respectively. The best performances are shown in bold. Our main method StockGNN outperforms the baselines in most of the cases. A base rate (label all data as the label of the larger class) will produce an average recall of .5.

Measures	Accuracy					Avg. Precision					Avg. Recall				
Methods	Fin	Health	Mat	Service	Tech	Fin	Health	Mat	Service	Tech	Fin	Health	Mat	Service	Tech
DEsvm	.563	.603	.52	**.57**	.57	**.62**	.61	.52	.57	.58	.563	**.61**	.52	**.57**	.58
DElogreg	.551	.59	.52	.56	.58	.555	.59	.516	.56	.58	.555	.59	.51	.56	.58
DEmlp	.592	.57	.53	.54	.60	.587	.57	.52	.54	.60	.585	.57	.52	.54	.60
StockGNN	**.60**	**.61**	**.62**	.568	**.61**	.60	**.61**	**.62**	**.57**	**.61**	**.602**	.604	**.62**	**.57**	**.613**

Table 5: Shock Based Label (y_s) Results: The accuracy, average precision and recall produced by different methods. The maximum standard deviations in the results for DEsvm, DElogreg, DEmlp, and StockGNN are .003, .001, .007 and .004 respectively. The best performances are shown in bold. Our main method StockGNN outperforms the baselines in most of the cases.

Measures	Accuracy					Avg. Precision					Avg. Recall				
Methods	Fin	Health	Mat	Service	Tech	Fin	Health	Mat	Service	Tech	Fin	Health	Mat	Service	Tech
DEsvm	.544	.582	.554	.567	.597	.54	.577	.555	.568	.597	.54	.579	.555	.567	.598
DElogreg	.55	.584	.556	.565	**.598**	.55	.58	.56	.57	.59	.54	.58	**.56**	.57	**.60**
DEmlp	.547	.552	.55	**.574**	.549	.54	.56	.55	**.58**	.55	.541	.55	.55	**.574**	.55
StockGNN	**.638**	**.606**	**.563**	.56	.55	**.62**	**.609**	**.562**	.56	**.603**	**.544**	**.608**	**.56**	.545	.562

Table 6: Value Based Label (y_v) Results: The accuracy, average precision and recall by different methods in five major sectors. The maximum variances in the results for DEsvm, DElogreg, DEmlp, and StockGNN are .001, .002, .008 and .005 respectively. The best performances are shown in bold. Our main method StockGNN outperforms the baselines in most of the cases.

Labels	VBL (y_v)		SBL (y_s)		IBL ($y_{I,5}$)	
Areas	Train	Test	Train	Test	Train	Test
Fin	15000	2100	1052	174	909	141
Health	9839	1734	3152	673	657	122
Mat	9295	1300	1425	252	603	87
Service	14990	2181	3511	615	946	146
Tech	14182	2397	5000	968	896	161

Table 7: The number of transcripts in specific sectors (areas) with the value (y_v), shock (y_s), and index ($y_{I,5}$) based labels.

(logreg), and a Multilayer Perceptron (MLP). Other classifiers (such as Naive Bayes, k-NN) produce inferior results. We call the entire pipeline of using Doc2Vec and a classifier as DEsvm, DElogreg, and DEmlp when the classifiers are SVM, logreg and MLP respectively. DEmlp is a modified version of the method proposed in [28]. Unlike using prior embeddings of words and aggregating them, we use Doc2Vec [21] to generate the embeddings for the document. Furthermore, we construct features only from transcripts as our aim is to classify with features from transcripts, whereas the method in [28] use the company embeddings as additional features that make the effects of transcript information ambiguous. We find that in most cases StockGNN outperforms these methods.

5.2 Results on Different Labels

We describe the results produced by our models on five major areas/sectors based on the number of transcripts.

5.2.1 Experimental Settings. We train the stock movement prediction methods on the transcripts till 2018. The transcripts in 2019 are used as test data. After removal of missing information, the train/test splits for all labels (y_v, y_s and $y_{I,k}$) are shown in Table 7. We consider $k = 5$ here and get a smaller subset (see Def. 3) of the data. The Doc2Vec method is applied to generate low-dimensional embeddings for the transcripts. We have generated embeddings with dimensions 100, 200, and 300 and report the best results among these three. In most of the cases 300-dimensional vector is the most effective one. The MLP classifier only uses one hidden layer with $32, 64, 128$ nodes for input features with dimensions 100, 200, and 300 respectively and a sigmoid layer to generate the final classification. The StockGNN method uses an initial word vector of 300 dimensions as node (word) features and produces a 96-dimensional vector as final graph-level embedding. We use the pre-trained GloVe[5] embeddings [31] for these initial word (node) vectors. The other settings are described in more details in the Appendix (Section 8.3).

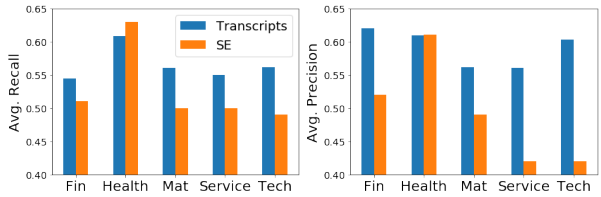
Performance Measures. We measure the quality of our proposed methods by *accuracy*, *average (macro) precision* and *average (macro) recall*. We have also repeated all the experiments at least ten times and reported the mean performances.

5.2.2 Results. Our goal is to show the usefulness of the semantics of the transcripts in stock price movement prediction. We present the results in five major areas/sectors on the index based ($y_{I,5}$), shock based (y_s) and the value based (y_v) labels in Tables 4, 5 and 6 respectively. Our StockGNN method produces the best quality in most of the cases. In particular, StockGNN produces $3\%, 4.1\%$ and 8.9% more average (over all sectors) accuracy than DEmlp, DEsvm,DElogreg respectively for index based label. Competitive

[5]http://nlp.stanford.edu/data/glove.6B.zip

results from the semantic feature based baselines indicate that the transcripts are indeed useful in the stock price movement prediction.

Discussion: Prediction of stock price movement is a challenging task and often seen as a random walk process [29]. As the investments in trading firms usually happen in large volumes—often in an automated fashion—lead to a significant amount of profit even with a slightly higher chance than random. Compared with a base rate or a random method, our method StockGNN shows significant improvement. Base rate assigns all the data as the label of the largest class whereas the random method assigns labels randomly. Both method have average recall of 0.5. StockGNN achieves up to 23%, 22% and 21% more average recall than a base rate in the index, shock, and value based labels respectively. The performances of other baseline methods that use semantic features also validate that the earnings calls indeed have a correlation with stock price movements even over a period of week.



(a) Avg. Recall **(b) Avg Precision**

Figure 8: Value Based Label (y_v) results: (a) Average Recall and (b) Average Precision in five major areas/sectors. The features are constructed from semantic representations of the earnings call transcripts (Transcripts in blue) and the estimated and actual sales as well as EPS (SE in orange). The maximum standard deviations for Transcripts and SE are .003 and .002 respectively. The semantic features from the transcripts produce more accurate results than SE in at least 80% of the cases and also statistically significant (p-value < 0.01).

5.3 Baseline: Earnings per Share (EPS) & Sales

Earnings per share (EPS) and sales are important quantitative factors that often are used to measure the performance of a company (Section 8.1). More specifically, if a company beats or misses the estimates in the sales and EPS for a given quarter, that usually results in higher and lower stock price movement respectively. Note that EPS (a real value) and earnings calls (a discussion) are different. Next, we show that the features from transcripts are more predictive of the stock price movement than the sales and EPS as features.

5.3.1 Experimental Setting. We have collected the estimated and actual sales and EPS corresponding to the earnings calls and used them as features to predict the value (y_v) as well as the index ($y_{I,5}$) based labels. We perform the classification tasks using a Multilayer Perceptron (MLP). In MLP we only use one hidden layer with 16 nodes and a softmax layer to generate the final classification. The other classification methods have produced results of worse quality.

5.3.2 Results. Figure 8 shows the results in five major areas for the value based label (y_v) prediction (see Fig. 10 for the index based label $y_{I,5}$ in the Appendix). We apply MLP in both scenarios—when the features are created from sales and EPS (denoted as SE) as well as when the features are semantic representations of the transcripts (denoted as Transcripts) produced by the Doc2Vec method (see Section 5.1). In most (80%) of the cases, the semantic features from the transcripts produce more accurate results. In particular, the transcripts produce 15% more AR and 42% more AP in "Tech" and 10% more AR and 33% more AP in "Service" sectors. However, while transcripts produce better results in growing and evolving sectors such as "Tech", the values of the actual/estimated sales and EPS could be important for predicting stock price movement (e.g., "Health"). Furthermore, Fig. 9 presents a visualization of the difference in estimated and actual actual sales as well as EPS and their association with the value based labels (Sec. 8.4 in the Appendix).

6 RELATED WORK & BACKGROUND

Sentiment analysis, tone and readability of the company related documents play a critical role in investments. Sentiment of the financial documents has been useful to predict market action [17]. Even the clarity and readability of the financial documents help the investors in predicting the valuation whereas a complex and long text results in confusion and subsequent volatility [27]. Lee et al. [22] studied the reaction of the market in response to the reported 8-K documents. They found that the earning surprise is the most important feature in predicting price movements. Larcker and Zakolyukina [20] classified the earnings calls as "truthful" or "deceptive" based on subsequent statements from the company and shown that the language of deceptive executives exhibits more references to general knowledge, fewer non extreme positive emotions, and fewer references to shareholder value.

In financial documents, the tone has also played an important role for stock market prediction [5, 7, 13, 26, 32]. The tone in earnings press releases can influence investors [13]. Price et al. [32] found that the conference call linguistic tone is a significant predictor of abnormal returns and trading volumes. With analyses on the 10-Q and 10-K form, Feldman et al. [11] showed that market reactions in a short window around the SEC filing are significantly affected by the tone change, even after controlling for accruals and earnings surprises. Furthermore, the tone change can also help to predict the subsequent quarter's earnings surprise. Moreover, the audio of the earnings call can also determine managerial emotional states [30]. Recently, Qin et al. [33] used the audio and text of the earnings calls to predict stock volatility.

New advancements in natural language processing techniques allow new analysis of earning calls. Our paper is inspired by the work of Ma et al. [28] that addresses the problem of identifying the stock price movement via earnings calls. However, our classification task that measures weekly impact (Def. 3) of an earnings call is novel and more practical. Our proposed graph neural network based method also outperforms the method in [28] (more details are in Sec. 5.1). Furthermore, our empirical analyses on a larger dataset are more elaborated, showing comparisons with external hard data such as sales and earnings per share (EPS) as well as the recommendations from the analysts that are missing in [28].

Additional Details: More details about Earnings per Share (EPS), sales, and analysts' ratings are provided in Sec. 8.1 (Appendix).

7 CONCLUSION

In this paper, we have studied the association between public companies' earnings calls and stock price movements with a dataset of

97,478 transcripts from 6,300 different companies. We have demonstrated three novel findings using our method that exploits the semantic features of the earnings call transcripts. The semantic characteristics of transcripts are relatively more accurate to predict stock price movements. We have also established that the transcripts have more predictive power than traditional hard data such as actual and estimated values of sales and earnings per share. Interestingly, the recommendations from the analysts made prior to the earnings are unrelated to stock price movements after the earnings call whereas the sentiments in transcripts are valuable.

REFERENCES

[1] Paul Asquith, Michael B Mikhail, and Andrea S Au. 2005. Information content of equity analyst reports. *Journal of financial economics* 75, 2 (2005), 245–282.

[2] Mark Bagnoli, Sanjay Kallapur, and Susan G Watts. 2001. Top line and bottom line forecasts: a comparison of internet firms during and after the bubble. *Available at SSRN 274178* (2001).

[3] Brad Barber, Reuven Lehavy, Maureen McNichols, and Brett Trueman. 2001. Can investors profit from the prophets? Security analyst recommendations and stock returns. *The Journal of Finance* 56, 2 (2001), 531–563.

[4] Brad M Barber, Reuven Lehavy, and Brett Trueman. 2007. Comparing the stock recommendation performance of investment banks and independent research firms. *Journal of financial economics* 85, 2 (2007), 490–517.

[5] Benjamin M Blau, Jared R DeLisle, and S McKay Price. 2015. Do sophisticated investors interpret earnings conference call tone differently than investors at large? Evidence from short sales. *Journal of Corporate Finance* 31 (2015), 203–219.

[6] Gerda Claeskens and Nils Lid Hjort. 2008. *Model Selection and Model Averaging*. Cambridge University Press.

[7] Angela K Davis, Weili Ge, Dawn Matsumoto, and Jenny Li Zhang. 2015. The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies* 20, 2 (2015), 639–673.

[8] Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2019. Learning dynamic context graphs for predicting social events. In *KDD*.

[9] WM Donders, Monique, Roy Kouwenberg, and CF Vorst, Ton. 2000. Options and earnings announcements: an empirical study of volatility, trading volume, open interest and liquidity. *European Financial Management* 6, 2 (2000), 149–171.

[10] Yonca Ertimur, Joshua Livnat, and Minna Martikainen. 2003. Differential market reactions to revenue and expense surprises. *Review of Accounting Studies* 8, 2-3 (2003), 185–211.

[11] Ronen Feldman, Suresh Govindaraj, Joshua Livnat, and Benjamin Segal. 2010. Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies* 15, 4 (2010), 915–953.

[12] Jennifer Francis and Leonard Soffer. 1997. The relative informativeness of analysts' stock recommendations and earnings forecast revisions. *Journal of Accounting Research* 35, 2 (1997), 193–211.

[13] Elaine Henry. 2008. Are investors influenced by how earnings press releases are written? *The Journal of Business Communication (1973)* 45, 4 (2008), 363–407.

[14] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *WSDM*. 261–269.

[15] Lianzhe Huang, Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2019. Text level graph neural network for text classification. In *ACL*.

[16] Zoran Ivković and Narasimhan Jegadeesh. 2004. The timing and value of forecast and recommendation revisions. *Journal of Financial Economics* 73, 3 (2004), 433–463.

[17] Colm Kearney and Sha Liu. 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis* 33 (2014), 171–185.

[18] Thomas N Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907* (2016).

[19] Mert Kosan, Arlei Silva, Sourav Medya, Brian Uzzi, and Ambuj Singh. 2021. Event Detection on Dynamic Graphs. *arXiv preprint arXiv:2110.12148* (2021).

[20] David F Larcker and Anastasia A Zakolyukina. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research* 50, 2 (2012), 495–540.

[21] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*. 1188–1196.

[22] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. 2014. On the Importance of Text Analysis for Stock Price Prediction.. In *LREC*, Vol. 2014. 1170–1175.

[23] Baruch Lev. 1989. On the usefulness of earnings and earnings research: Lessons and directions from two decades of empirical research. *Journal of accounting research* 27 (1989), 153–192.

[24] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493* (2015).

[25] Roger K Loh and René M Stulz. 2011. When are analyst recommendation changes influential? *The review of financial studies* 24, 2 (2011), 593–627.

[26] 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 (2011), 35–65.

[27] Tim Loughran and Bill McDonald. 2014. Measuring readability in financial disclosures. *The Journal of Finance* 69, 4 (2014), 1643–1671.

[28] Zhiqiang Ma, Grace Bang, Chong Wang, and Xiaomo Liu. 2020. Towards Earnings Call and Stock Price Movement. In *SIGKDD MLF Workshop*.

[29] Burton G Malkiel. 2003. The efficient market hypothesis and its critics. *Journal of economic perspectives* 17, 1 (2003), 59–82.

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

[31] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[32] S McKay Price, James S Doran, David R Peterson, and Barbara A Bliss. 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance* 36, 4 (2012), 992–1011.

[33] Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *ACL*. 390–401.

[34] Scott E Stickel. 1995. The anatomy of the performance of buy and sell recommendations. *Financial Analysts Journal* 51, 5 (1995), 25–39.

[35] Ronald L Wasserstein and Nicole A Lazar. 2016. The ASA statement on p-values: context, process, and purpose.

[36] Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *ACL*. 1970–1979.

[37] Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. 2020. Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks. In *ACL*.

8 APPENDIX

8.1 Sales, Earnings & Analysts' Ratings

Earnings per Share (EPS) and Sales: EPS and sales are two important quantitative factors that are often used to evaluate the performance of a company. The EPS indicates whether the company is currently profitable and can be self-sustained. The investors are often rewarded based on the EPS of a company. Moreover, the stock price is usually a multiplier (i.e., price earning ratio) of the EPS. The EPS often plays an important role in the stock movement especially for the established companies with small growth [2]. Note that the EPS (a real value) and earnings calls (a discussion) are different. On the other hand, sales is indicative of the market share of a company and an important measure especially for growth companies [2]. Sales might influence investors to overlook or ignore the earnings while having the promise of future growth [23].

Sales and EPS are important quantitative factors that often affect the perception of the investors about corresponding companies. Usually before the earnings call happens, the market forms a general consensus (estimates) on the EPS and sales. If a company beats (or misses) the estimates in the sales and EPS for a given quarter, that usually results in higher (or lower) stock price movement. Past research has shown that earning surprise due to surprise in sales has more effect on the price than due to cutting costs [10]. In this study, we demonstrate that the transcripts have more predictive power than actual and estimated values of sales and EPS.

Analysts' Ratings: Analysts are the financial experts who often post their opinions about companies and play a key role in impacting or modifying the perceptions of the investors about the future prospects of a business [1, 3, 12]. Past research has demonstrated that analysts from popular banking and investment firms can generate significant stock price movements [4, 25, 34]. Analysts' recommendations become even more powerful when they are accompanied by forecasts of the earnings of the company [25]. Moreover, the timing of these recommendations with respect to the earnings announcements affects their impact [16]. In this work, we analyze the impact of the recommendations by the analysts in stock price movements locally after the occurrence of the earnings call.

8.2 Additional Regression Results

In Table 8 we present the same analyses as in Table 2 for the shock based label y_s as a dependent variable. The analyses produce similar results as in the case of value based label y_v .

8.2.1 Analysts' Recommendations. With the same set of independent variables as in Table 3, we present the regression results for the for $y_{I,k}$ (Index Based Label Function) when $k = 3$ in Tables 9 ($k = 4, 5$ produce similar results). Different from results in Table 3 for y_v, we find that only the positive emotion in the transcripts remains significantly (low p-value) indicative of the label $y_{I,3}$ when other control variables are added into the model incrementally. However, some results are consistent with the case of y_v. Adding more control variables does not improve the fitness of the model. BIC in Model 2 is larger (worse) than that of Model 1, which suggests that prior analysts' recommendations do not provide additional information to the model.

Variables	Model 1	Model 2	Model 3	Model 4	Model 5
Positive Emotion		0.32***	0.31***	0.33***	0.42***
		(0.027)	(0.027)	(0.028)	(0.036)
Negative Emotion		-0.80***	-0.29*	-0.31**	-0.45***
		(0.069)	(0.11)	(0.11)	(0.15)
Anxiety Score			-0.43*	-0.38	-0.12
			(0.22)	(0.22)	(0.29)
Anger Score			0.042	0.080	0.17
			(0.25)	(0.25)	(0.31)
Sad Score			-1.24***	-1.26***	-1.78***
			(0.19)	(0.19)	(0.26)
Certainty Score			0.041	0.083	0.18
			(0.068)	(0.074)	(0.098)
Cognitive Score				-0.032*	-0.058*
				(0.015)	(0.021)
Insight Score				-0.060	-0.11
				(0.043)	(0.059)
Causation Score				-0.11*	-0.00077
				(0.042)	(0.053)
Discrepancy Score				0.19*	0.19*
				(0.063)	(0.082)
Actual Sales					0.00052
					(0.00028)
Estimated Sales					-0.00053
					(0.00028)
Actual EPS					0.042
					(0.036)
Estimated EPS					-0.033
					(0.029)
Sector	YES	YES	YES	YES	YES
Year	YES	YES	YES	YES	YES
Observations	20,020	19,965	19,965	19,965	12,946
BIC	27,887	27,463	27,437	27,445	17,695

Table 8: Regression of Shock Based Label Function (SBL) $y_s(T_d^c)$ on Text Features.

Variables	Model 1	Model 2	Model 3	Model 4	Model 5
positive	0.10***	0.10***	0.10***	0.10***	0.12**
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
negative	-0.10	-0.10	-0.14	-0.14	-0.07
	(0.12)	(0.12)	(0.13)	(0.13)	(0.18)
anxiety	0.28	0.28	0.21	0.19	0.09
	(0.22)	(0.22)	(0.24)	(0.24)	(0.34)
anger	0.26	0.26	0.42	0.41	0.54
	(0.24)	(0.24)	(0.24)	(0.24)	(0.36)
sad	-0.02	-0.03	0.03	0.03	-0.06
	(0.18)	(0.18)	(0.19)	(0.19)	(0.29)
certain	0.12	0.12	0.11	0.11	0.08
	(0.07)	(0.07)	(0.07)	(0.07)	(0.10)
MAR_{1m}		YES	YES	YES	YES
Sector			YES	YES	YES
Year				YES	YES
MAR_{5d}					YES
Observations	18,545	18,545	18,545	18,545	9,097
BIC	25733	25771	25836	25893	12823

Table 9: Index Based Label Function ($y_{I,3}$) results on Analysts' Recommendations within 1 month (MAR_{1m}) prior and 5 days (MAR_{5d}) after the earnings call date and sentiment of the transcripts.

8.3 Experimental Settings of Section 5.2

The Doc2Vec method generates low-dimensional embeddings for the transcripts and uses standard settings as follows: "distributed memory" (PV-DM), the initial learning rate as .001, and the negative sample as 5. We ignore all words with total frequency lower than 2. In the StockGNN method we use the window size of 3 to build the graph from a document. The number of iterations (steps, k) that the gated graph neural network (GNN) uses inside the method is 2. We use the batch size of 128 in all the experiments.

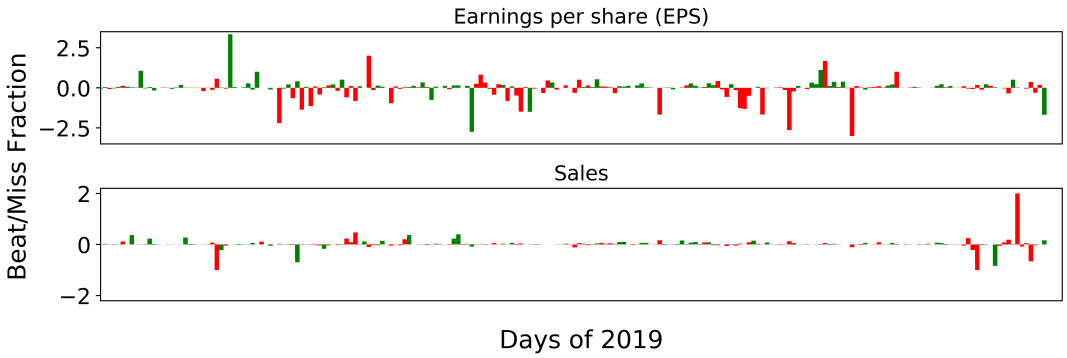
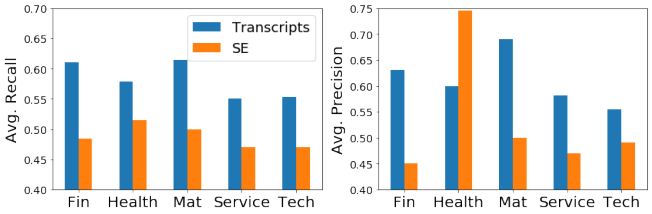
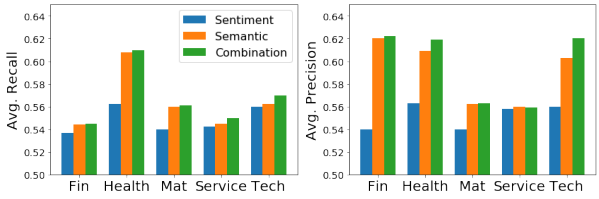


Figure 9: Value Based Label (y_v) results: It presents a visualization of the difference in estimated and actual sales as well as earnings per share (EPS) and their association with the labels. The X-axis shows a total of 212 days which has at least one occurrence of earnings call from any company in 2019. The Y-axis shows the factor by which the actual value beats (positive and upward) or misses (negative and downward) the estimated value. Some of the high positive (or negative) values of these factors predict the class 1 (or 0) labels, denoted by upward green (or downward red), accurately. However, there are several mis-classifications in the form of upward red and downward green in both cases (Sales and EPS).



(a) Avg. Recall **(b) Avg Precision**

Figure 10: IBL ($y_{I,5}$) results: (a) Average Recall and (b) Average Precision in five major areas/sectors. The features are constructed from the earnings calls transcripts (Transcripts, denoted by blue) and the estimated and actual sales as well as EPS (SE, denoted by orange). The semantic features from the transcripts produce more accurate results than SE in at least 80% of the cases.



(a) Avg. Recall **(b) Avg. Precision**

Figure 11: Value Based Label (y_v) results: (a) Average Recall and (b) Average Precision of predicting the labels in five major sectors. The features are constructed from sentiments (blue), semantic representations of the transcripts (orange), and combination of them (green). The results show that though the semantic features are relatively the most predictive features, sentiments often play a critical role.

8.4 Results: Earnings per Share and Sales

Visualizations: In Figure 8 (Section 5.3), we have already shown that the features with estimated/actual sales and EPS as features are inferior than semantics of transcripts in predicting stock price movements for the value based labels (y_v). Figure 9 presents a visualization of the differences in estimated and actual sales as well as earnings per share (EPS) and their association with the value based labels (y_v). The X-axis shows a total of 212 days on which there is at least one occurrence of earnings call in 2019. If there are multiple earnings calls on the same date, we randomly choose one. The Y-axis shows the factor by which the actual value beats (positive) or misses (negative) the estimated value. For instance, 0.5 would mean the actual value beat the estimated one by 50%. The figure demonstrates that some of the high positive (or negative) values of these factors predict the class 1 (or 0) labels, denoted by upward green (or downward red), accurately. However, there are several misclassifications in the form of upward red and downward green in both cases (sales and EPS). These imply that beating (missing) the estimated value is not always associated with upward (downward) stock movement for both sales and EPS.

Results on index based label $y_{I,5}$: We present additional results for the index based label $y_{I,5}$ in five major sectors in Figure 10.

The classification results are produced by a Multilayer Perceptron (MLP) in two scenarios—when the features are created from sales and EPS (denoted as SE) as well as when the features are semantic representations of the transcripts (denoted as Transcripts) produced by the Doc2Vec method (see Section 5.1). In 80% of the cases, the semantic features from the transcripts produce more accurate results. In particular, the transcripts produce up to 21% more AR and 40% more AP in finance ("Fin") sector.

8.5 Effect of Sentiments and Semantics

The previous results indicate that the sentiment variables from transcripts might be useful for stock price movement prediction. We further use those six sentiment variables as features and compare the predictions with the ones that use the semantic features produced by our StockGNN method (see Section 5.1). To investigate whether these two features capture different information, we also present the results of the combination of both as features. Figure 11 presents the classification results produced by a Multilayer Perceptron (MLP). As performance measures, we show average precision and average recall. Though the semantic features are the most useful ones in all of the five sectors, sentiment in transcripts is also useful for prediction. For instance, in the "Service" sector, the sentiment features produce competitive results compared to the semantic ones.