An Exploratory Study of Stock Price Movements from Earnings Calls

Sourav Medya
Kellogg School of Management
Northwestern University, USA
sourav.medya@kellogg.northwestern.edu

Mohammad Rasoolinejad
Signature Consultants
USA
Rasoolinejad@u.northwestern.edu

Yang Yang
School of Information Studies
Syracuse University, USA
yyang87@syr.edu

Brian Uzzi
Kellogg School of Management
Northwestern University, USA
uzzi@kellogg.northwestern.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.

CCS CONCEPTS
• Applied computing → Electronic commerce.

KEYWORDS
Earnings call, stock price movement, natural language processing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation \( \rightarrow \) on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

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
We formulate the stock price movements as binary labels. Let \( T \) be a set of transcripts denoted by \( \{T_1, T_2, \ldots, T_m\} \) where \( T_d \) represents the earnings call transcript on day \( d \). The value of \( t \) can vary depending on the history of earnings calls.

2. **STOCK PRICE MOVEMENTS**

We formulate the stock price movements as binary labels. Let \( C = \{C_1, C_2, \ldots, C_m\} \) be the set of companies. We denote the closing stock price of the company \( c \) on the day \( d \) as \( S_d \). 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 \( T_c = \{T_{d_1}, T_{d_2}, \ldots, T_{d_t}\} \) where \( T_{d_i} \) 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 \( y_v(T_d^c) \in \{0, 1\} \) for a transcript \( T_d^c \) of a company \( c \) on the day \( d \) as follows:

\[
y_v(T_d^c) = \begin{cases} 
1, & \text{if } S_{d+1}^c - S_d^c > S_{d-1}^c - S_d^c \\
0, & \text{otherwise}
\end{cases}
\]

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 \( y_s(T_d^c) \in \{0, 1\} \) for a transcript \( T_d^c \) of a company \( c \) on the day \( d \) as follows:

\[
y_s(T_d^c) = \begin{cases} 
1, & \text{if } \frac{S_{d+1}^c - S_d^c}{S_d^c} \geq \tau \\
0, & \text{if } \frac{S_{d-1}^c - S_d^c}{S_d^c} \geq \tau
\end{cases}
\]

where \( d + 1 \) and \( d - 1 \) denote the next and previous business days respectively of the day \( d \). \( \tau \) is a threshold (we set it as 0.05 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 earnings 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
IYC
Ref. Index
12.6
18.3
XLV
SP500
0.78
11.3
XLI
XLK
Percentage
18.5
18.1
8.8
XLY
9.2
XLB

Figure 2: Value Based Label ($y_{c,k}$): 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.

Figure 3: Shock Based Label ($y_{c,k}$): 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.

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

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 ($y_{c,k}$) 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 $L$. While the daily (value and shock) labels are relevant, they are myopic 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 myopic). 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

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

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

We use the Linguistic Inquiry and Word Count (LIWC) dictionary to identify the sentiment traits in the transcripts. We include positive and negative emotions, anxiety, and sadness as affect traits

1https://seekingalpha.com/
2https://www.zacks.com/
3https://wrds-www.wharton.upenn.edu/pages/about/data-vendors/zacks/
4https://www.kovcomp.co.uk/wordstat/LIWC.html
We show descriptive analysis of the data on the distribution of the
WWW ’22 Companion, April 25–29, 2022, Virtual Event, Lyon, France
the index label ($y_5$). Among others. Figs. 2 and 3 show the year-wise
among different companies. Figs. 4 and 5 represent the number and fraction of the.
3.2 Descriptive Analysis on Label Distributions
among others. Figs. 2 and 3 show the year-wise sentiments in
Unsurprisingly, the number of transcripts increase (Figs. 4a, 4c, and 5a) over the years for all label functions as the number of company
variables. Significance levels marked with stars: *** p<0.001,
4 PILOT REGRESSION ANALYSIS
among others. Figs. 2 and 3 show the year-wise sentiments in
among others. Figs. 2 and 3 show the year-wise sentiments in
Unsurprisingly, the number of transcripts increase (Figs. 4a, 4c, and 5a) over the years for all label functions as the number of company
variables. Significance levels marked with stars: *** p<0.001,
4 PILOT REGRESSION ANALYSIS
among others. Figs. 2 and 3 show the year-wise sentiments in
Unsurprisingly, the number of transcripts increase (Figs. 4a, 4c, and 5a) over the years for all label functions as the number of company
variables. Significance levels marked with stars: *** p<0.001,
4 PILOT REGRESSION ANALYSIS
among others. Figs. 2 and 3 show the year-wise sentiments in
Unsurprisingly, the number of transcripts increase (Figs. 4a, 4c, and 5a) over the years for all label functions as the number of company
variables. Significance levels marked with stars: *** p<0.001,
using LIWC dictionary. There are two categories: 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 the frequency of certainty, insight and cognitive process words.

In Table 2, we present the estimates (coefficients, standard errors, significance) of the variables in five regression models where $y_d$ 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 $y_d$ (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 $y_t$ 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.

#### 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 ($y_d$) label (the results on index ($y_{f,k}$) based label are in Sec. 9.2.1 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 corresponding 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 $MAR_{sd}$.

#### 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 $MAR_{sd}$ 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 $y_d$. 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 ($MAR_{sd}$) 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 following: the sample size (N) is different; and Model 5 includes posterior analysts’ recommendations $MAR_{sd}$, which intuitively should have better explanation power of the dependent variable $y_d$.

**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.

### Table 3: Regression Results: Value Based Label Function ($y_d$) on Analysts’ Recommendations within 1 month ($MAR_{1m}$) prior and 5 days ($MAR_{sd}$) 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.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>0.23*** (0.03)</td>
<td>0.22*** (0.03)</td>
<td>0.20*** (0.03)</td>
<td>0.22*** (0.03)</td>
<td>0.31*** (0.04)</td>
</tr>
<tr>
<td>negative</td>
<td>-0.43*** (0.12)</td>
<td>-0.43*** (0.12)</td>
<td>-0.42*** (0.13)</td>
<td>-0.42*** (0.13)</td>
<td>-0.63*** (0.19)</td>
</tr>
<tr>
<td>anxiety</td>
<td>0.35 (0.23)</td>
<td>0.35 (0.23)</td>
<td>0.28 (0.24)</td>
<td>0.28 (0.24)</td>
<td>0.07 (0.35)</td>
</tr>
<tr>
<td>anger</td>
<td>0.56* (0.24)</td>
<td>0.54* (0.24)</td>
<td>0.57* (0.25)</td>
<td>0.56* (0.25)</td>
<td>0.62 (0.36)</td>
</tr>
<tr>
<td>sad</td>
<td>-0.66*** (0.19)</td>
<td>-0.58*** (0.19)</td>
<td>-0.63*** (0.20)</td>
<td>-0.63*** (0.20)</td>
<td>-0.57*** (0.29)</td>
</tr>
<tr>
<td>certain</td>
<td>-0.01 (0.07)</td>
<td>-0.02 (0.07)</td>
<td>-0.01 (0.07)</td>
<td>-0.06 (0.07)</td>
<td>0.12 (0.11)</td>
</tr>
<tr>
<td>MAR$_{1m}$</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Sector</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Year</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>$MAR_{sd}$</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Observations</td>
<td>18,289</td>
<td>18,289</td>
<td>18,289</td>
<td>18,289</td>
<td>8,994</td>
</tr>
<tr>
<td>BIC</td>
<td>25185</td>
<td>25201</td>
<td>25271</td>
<td>25346</td>
<td>12474</td>
</tr>
</tbody>
</table>
Based on GNN

To test the predictive power of the content in the earnings call, we first build a graph. The gated GNN produces an embedding of or constructed features embeddings to learn the node (word) embeddings. We start with initial (given or constructed) embeddings to achieve higher order (k-hop neighbours) interactions, and the words appear in its neighbourhood (or context) become its particular tools in several natural language processing and related tasks.

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

(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) labels. We begin with descriptions of our approaches to solve these classification problems.

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 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 neighbours. 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 \( \mathbf{z} \in \mathbb{R}^{|V| \times d} \), a vector of size \( d \) of the nodes or words. The gated GNN produces an embedding of node \( v \) 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:

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

Here, \( \odot \) is element-wise multiplication, \( \sigma \) 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 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 \( \mathbf{z}_v \) for the entire graph or document. The aggregation function is as follows:

\[
z_G = \frac{1}{|V|} \sum_{v \in V} z_v \quad \text{where} \quad z_v = \sigma(\text{MLP}(z_v^k)) \odot \tanh(\text{MLP}(z_v^k))
\]

where \( z_v \) is embedding for the node \( v \) and MLP is Multi-layer 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 the 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 from the transcripts and to measure the accuracy of the proposed StockGNN method compared to baselines. To achieve that we propose several baselines with two phases. First, we generate unsupervised low-dimensional embeddings via the well-known Doc2Vec method [21]. Intuitively, similar documents will generate unsupervised low-dimensional embeddings via the well-known Doc2Vec method [21].
An Exploratory Study of Stock Price Movements from Earnings Calls

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.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Accuracy</th>
<th>Avg. Precision</th>
<th>Avg. Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fin</td>
<td>Health</td>
<td>Mat</td>
</tr>
<tr>
<td>DESVM</td>
<td>.624</td>
<td>.549</td>
<td>.74</td>
</tr>
<tr>
<td>DELOGREG</td>
<td>.595</td>
<td>.524</td>
<td>.632</td>
</tr>
<tr>
<td>DEMLP</td>
<td>.666</td>
<td>.540</td>
<td>.724</td>
</tr>
<tr>
<td>STOCKGNN</td>
<td>.582</td>
<td>.60</td>
<td>.761</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Accuracy</th>
<th>Avg. Precision</th>
<th>Avg. Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fin</td>
<td>Health</td>
<td>Mat</td>
</tr>
<tr>
<td>DESVM</td>
<td>.563</td>
<td>.603</td>
<td>.52</td>
</tr>
<tr>
<td>DELOGREG</td>
<td>.591</td>
<td>.59</td>
<td>.52</td>
</tr>
<tr>
<td>DEMLP</td>
<td>.592</td>
<td>.57</td>
<td>.53</td>
</tr>
<tr>
<td>STOCKGNN</td>
<td>.60</td>
<td>.61</td>
<td>.62</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Measures</th>
<th>Accuracy</th>
<th>Avg. Precision</th>
<th>Avg. Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fin</td>
<td>Health</td>
<td>Mat</td>
</tr>
<tr>
<td>DESVM</td>
<td>.544</td>
<td>.582</td>
<td>.534</td>
</tr>
<tr>
<td>DELOGREG</td>
<td>.55</td>
<td>.584</td>
<td>.556</td>
</tr>
<tr>
<td>DEMLP</td>
<td>.547</td>
<td>.552</td>
<td>.55</td>
</tr>
<tr>
<td>STOCKGNN</td>
<td>.638</td>
<td>.606</td>
<td>.563</td>
</tr>
</tbody>
</table>

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. 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.

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 models 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 \((y_v, y_s, y_{I,k})\) 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. 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 embeddings [31] for these initial word (node) vectors. The other settings are described in more details in the Appendix (Section 9.3).

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 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

3http://nlp.stanford.edu/data/glove.6B.zip
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 methods 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.

Figure 8: Value Based Label (\(y_v\)) results: (a) Average Recall and (b) Average Precision. The features are from semantic representations of the transcripts (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 9.1). 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.

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 Multi-layer 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 sales as well as EPS and their association with the value based labels (Sec. 9.4 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, 16, 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. 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 earnings 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].

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.

8 ACKNOWLEDGMENTS

The authors would like to thank Arlei Silva for the helpful comments on this paper and the Kellogg Graduate School of Management as well as the Northwestern University Institute on Complex Systems (NICO) for their financial support of the research.
An Exploratory Study of Stock Price Movements from Earnings Calls

REFERENCES


9 APPENDIX

9.1 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.

9.2 Additional Regression Results

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

9.2.1 Analysts’ Recommendations. With the same set of independent variables as in Table 3, we present the regression results for the for $y_{t,k}$ (Index Based Label Function) when $k = 3$ in Tables 8 ($k = 4, 5$ produce similar results). Different from results in Table 3 for $y_t$, we find that only the positive emotion in the transcripts remains significantly (low $p$-value) indicative of the label $y_{t,3}$ 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.

9.3 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.
9.4 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).

9.5 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.