

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/349435078>

# Stock Values and Earnings Call Transcripts: a Dataset Suitable for Sentiment Analysis

**Preprint** · February 2021

DOI: 10.20944/preprints202102.0424.v1

CITATIONS

0

READS

584

2 authors:



**Dexter Roozen**

1 PUBLICATION 0 CITATIONS

SEE PROFILE



**Francesco Lelli**

Tilburg University

45 PUBLICATIONS 5,178 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



GridCC [View project](#)

# Stock Values and Earnings Call Transcripts: a Dataset Suitable for Sentiment Analysis

Dexter Roozen  
Tilburg University  
drh.roozen@gmail.com

Francesco Lelli  
Tilburg University  
f.elli@tilburguniversity.edu  
<https://francescolelli.info>

## Abstract:

The dataset reports a collection of earnings call transcripts, the related stock prices, and the related sector index. It contains a total of 188 transcripts, 11970 stock prices, and 1196 sector index values. Furthermore, all of these data originated in the period 2016-2020 and are related to the NASDAQ stock market.

The data have been collected using Yahoo Finance and Thomson Reuters Eikon. Specifically, Yahoo Finance offered daily stock prices and traded volume. At the same time, Thomson Reuters Eikon has been used as source for the earnings call transcripts.

The dataset can be used as a benchmark for the evaluation of several NLP techniques as well as machine learning algorithms for understanding their potential for financial applications. Moreover, it is also possible to expand the dataset by extending the period in which the data originated following a similar procedure.

**Keywords:** dataset, stock, sentiment analysis, nlp, Nasdaq, stock prices

## 1. Introduction:

The dataset contains earning call transcripts and related stock process of 10 popular stocks of the Nasdaq index. These data have been collected for performing a set of “bag of words” analysis in order to evaluate possible correlation between them and stock process.

Several researches try to use yahoo finance Data in order to gather insight in stock prices and volume of trading. This is the case of Rao et al in [2] and authors in [3]. Some of the authors access directly to API. However, in recent times yahoo has made several changes and this practice is becoming more difficult.

Earning call transcript have been collected via Thomson Reuters Eikon. Several authors has use this approach. This is the case of the works presented in [4],[5],[6] and [7] However, the exact procedure for attaining the transcripts is not specified within these papers.

With this preprint we intend to formalize the procedure for collecting the data as well as release an initial version of a dataset that can be extended and reused in order to perform several different analysis.

The table below summarize the data:

<b>Subject</b>	Economics/Linguistics/Computer Science
<b>Specific subject area</b>	<p>The specific subject area of this research is Sentiment Analysis. Sentiment analysis is a natural language processing (NLP) technique to determine the sentiment (positive or negative) behind data.</p> <p>To elaborate, NLP is a field of research that investigates the ability of computers to understand and manipulate natural languages, such as English.</p> <p>A crucial step of textual sentiment analysis is to pre-process the text documents. This pre-processing phase consists of multiple 'pre-processing techniques' of which the effects were studied.</p>
<b>Type of data</b>	<p>Table</p> <p>Text</p>
<b>How data were acquired</b>	<p>The stock values and sector index were acquired through Yahoo Finance (website).</p> <p>The earnings call transcripts were acquired through Thomson Reuters Eikon (software).</p>
<b>Data format</b>	Raw
<b>Parameters for data collection</b>	<p>The related companies of the stock values and earnings call transcripts were chosen based on the condition of being NASDAQ listed.</p> <p>Furthermore, the date range for the stock values and earnings call transcripts is 2016-2020.</p>
<b>Description of data collection</b>	<p>The stock values were acquired by using Yahoo Finance. Yahoo Finance provides news, information, commentary, and reports on the subject of finance. This website lets users search for specific companies with its search bar. When entering a company such as "Apple Inc.", the website will direct the user to a summary of general financial information about the company. Besides this 'Summary' tab, there are also other tabs providing different sorts of information. Selecting the 'Historical Data' tab shows the historical stock values of the searched company. Additionally, it also lets users specify the date period and the frequency with which it shows the stock values. Afterward, the presented stock values can be downloaded as a Microsoft Excel Comma Separated Value (CSV) File.</p> <p>Thomson Reuters Eikon helped acquire the earnings call transcripts. This software provides users with many different sorts of financial</p>

	information. Selecting the 'advanced event search' directs the user to financial information about particular events. Specifying the event type 'Earnings Conference Call' shows information about earnings calls of many different companies. Selecting 'transcript' from the 'Content Type' selector filters out earnings calls without transcripts. Lastly, date preference and company should be specified to find the desired information. With the save batch icon it is possible to download 100 transcripts at a time as text documents.
<b>Data source location</b>	<a href="https://doi.org/10.34894/TJE0D0">https://doi.org/10.34894/TJE0D0</a>
<b>Data accessibility</b>	Repository name: DataverseNL Data identification number: N.A Direct URL to data: <a href="https://doi.org/10.34894/TJE0D0">https://doi.org/10.34894/TJE0D0</a>  Instructions for accessing these data: Data are open access

## 2. Methods for data acquisition

All of the stock values and the sector index were acquired by utilizing the Yahoo Finance search bar. Searching for a company such as Apple Inc. results in a summary of financial information about this company. However, the stock values and sector index within the dataset are presented in the "Historical Data" tab. Selecting this tab and specifying the time period January 1<sup>st</sup>, 2016 – October 1<sup>st</sup>, 2020 and selecting "Apply" will show the data presented in this dataset. Lastly, selecting "Download" provides a CSV file containing all of this data.

The earnings call transcripts were acquired through Thomson Reuters Eikon. Selecting the "advanced event search" option shows unfiltered financial information about many different sorts of events. Specifying the event type by selecting "Earnings Conference Call" will filter this information by only showing information about earnings calls. Additionally, selecting "Transcript" from the "Content Type" selector will show only earnings calls that can be provided together with a transcript of the earnings call. Lastly, specifying the company and time period will show a list with the earnings call transcripts contained in this dataset. For efficiency purposes, the save batch icon makes it possible to download this whole list of transcripts.

## 2.1 List of resources used for the collection of the data

The following sources were used in retrieving the historical stock values of the NASDAQ listed companies and the sector index:

- NASDAQ. (2020). Apple Inc. (AAPL). [Historical stock values, 2016-2020]. Retrieved from <https://finance.yahoo.com/quote/AAPL/history?p=AAPL>
- NASDAQ. (2020). Advanced Micro Devices, Inc. (AMD). [Historical stock values, 2016-2020]. Retrieved from <https://finance.yahoo.com/quote/AMD/history?p=AMD>
- NASDAQ. (2020). Amazon.com, Inc. (AMZN). [Historical stock values, 2016-2020]. Retrieved from <https://finance.yahoo.com/quote/AMZN/history?p=AMZN>
- NASDAQ. (2020). ASML Holding N.V. (ASML). [Historical stock values, 2016-2020]. Retrieved from <https://finance.yahoo.com/quote/ASML/history?p=ASML>
- NASDAQ. (2020). Cisco Systems, Inc. (CSCO). [Historical stock values, 2016-2020]. Retrieved from <https://finance.yahoo.com/quote/CSCO/history?p=CSCO>
- NASDAQ. (2020). Alphabet Inc. (GOOGL). [Historical stock values, 2016-2020]. Retrieved from <https://finance.yahoo.com/quote/GOOGL/history?p=GOOGL>
- NASDAQ. (2020). Intel Corporation (INTC). [Historical stock values, 2016-2020]. Retrieved from <https://finance.yahoo.com/quote/INTC/history?p=INTC>
- NASDAQ. (2020). Microsoft Corporation (MSFT). [Historical stock values, 2016-2020]. Retrieved from <https://finance.yahoo.com/quote/MSFT/history?p=MSFT>
- NASDAQ. (2020). Micron Technology, Inc. (MU). [Historical stock values, 2016-2020]. Retrieved from <https://finance.yahoo.com/quote/MU/history?p=MU>
- NASDAQ. (2020). NVIDIA Corporation (NVDA). [Historical stock values, 2016-2020]. Retrieved from <https://finance.yahoo.com/quote/NVDA/history?p=NVDA>
- NASDAQ. (2020). NASDAQ Composite (^IXIC). [Historical stock values, 2016-2020]. Retrieved from <https://finance.yahoo.com/quote/%5EIXIC/history?p=%5EIXIC>

The following database provided all of the earnings call transcripts from the selected NASDAQ companies:

- Thomson Reuters Eikon. (2020). [Earnings call transcripts, 2016-2020]. Available at: Thomson Reuters (Accessed: November 19 2020).

### 3. Results: Data Description

The folder named “Stock Values and Sector Index” in the dataset contains all of the CSV files that were acquired through the before mentioned method. These files consist of individual tables for each NASDAQ Company and the NASDAQ sector index. The folder is structured as portrayed in table 1.

Company	File Type	Size	Source
Apple Inc. (AAPL)	CSV	1197 rows x 1 column	NASDAQ (2020)
Advanced Micro Devices, Inc. (AMD)	CSV	1197 rows x 1 column	NASDAQ (2020)
Amazon.com, Inc. (AMZN)	CSV	1197 rows x 1 column	NASDAQ (2020)
ASML Holding N.V. (ASML)	CSV	1197 rows x 1 column	NASDAQ (2020)
Cisco Systems, Inc. (CSCO)	CSV	1197 rows x 1 column	NASDAQ (2020)
Alphabet Inc. (GOOGL)	CSV	1197 rows x 1 column	NASDAQ (2020)
Intel Corporation (INTC)	CSV	1197 rows x 1 column	NASDAQ (2020)
Microsoft Corporation (MSFT)	CSV	1197 rows x 1 column	NASDAQ (2020)
Micron Technology, Inc. (MU)	CSV	1197 rows x 1 column	NASDAQ (2020)
NVIDIA Corporation (NVDA)	CSV	1197 rows x 1 column	NASDAQ (2020)
NASDAQ Composite (^IXIC)	CSV	1196 rows x 1 column	NASDAQ (2020)

**Table 1:** Folder structure stock values and sector index

Accessing these files can be done through Microsoft Excel. A snippet of what these files look like when opened in Excel is portrayed in figure 1.

	A	B	C	D	E	F	G	H
1	Date,Open,High,Low,Close,Adj Close,Volume							
2	2016-01-04	25.652500	26.342501	25.500000	26.337500	24.443037	270597600	
3	2016-01-05	26.437500	26.462500	25.602501	25.677500	23.830513	223164000	
4	2016-01-06	25.139999	25.592501	24.967501	25.174999	23.364161	273829600	
5	2016-01-07	24.670000	25.032499	24.107500	24.112499	22.378082	324377600	
6	2016-01-08	24.637501	24.777500	24.190001	24.240000	22.496408	283192000	
7	2016-01-11	24.742500	24.764999	24.334999	24.632500	22.860676	198957600	
8	2016-01-12	25.137501	25.172501	24.709999	24.990000	23.192465	196616800	
9	2016-01-13	25.080000	25.297501	24.325001	24.347500	22.596178	249758400	
10	2016-01-14	24.490000	25.120001	23.934999	24.879999	23.090370	252680400	
11	2016-01-15	24.049999	24.427500	23.840000	24.282499	22.535858	319335600	
12	2016-01-19	24.602501	24.662500	23.875000	24.165001	22.426809	212350800	
13	2016-01-20	23.775000	24.547501	23.355000	24.197500	22.456970	289337600	
14	2016-01-21	24.264999	24.469999	23.735001	24.075001	22.343281	208646000	
15	2016-01-22	24.657499	25.365000	24.592501	25.355000	23.531210	263202000	
16	2016-01-25	25.379999	25.382500	24.802500	24.860001	23.071817	207178000	
17	2016-01-26	24.982500	25.219999	24.517500	24.997499	23.199427	300308000	
18	2016-01-27	24.010000	24.157499	23.334999	23.355000	21.675074	533478800	
19	2016-01-28	23.447500	23.629999	23.097500	23.522499	21.830519	222715200	
20	2016-01-29	23.697500	24.334999	23.587500	24.334999	22.584578	257666000	
21	2016-02-01	24.117500	24.177500	23.850000	24.107500	22.373442	163774000	
22	2016-02-02	23.855000	24.010000	23.570000	23.620001	21.921011	149428800	
23	2016-02-03	23.750000	24.209999	23.520000	24.087500	22.354887	183857200	
24	2016-02-04	23.965000	24.332500	23.797501	24.150000	22.534504	185886800	
25	2016-02-05	24.129999	24.230000	23.422501	23.504999	21.932650	185672400	
26	2016-02-08	23.282499	23.924999	23.260000	23.752501	22.163597	216085600	
27	2016-02-09	23.572500	23.985001	23.482500	23.747499	22.158932	177324800	
28	2016-02-10	23.980000	24.087500	23.525000	23.567499	21.990967	169374400	
29	2016-02-11	23.447500	23.680000	23.147499	23.424999	21.858000	200298800	

Figure 1: Stock values in Excel

Furthermore, the folder named “Transcripts” in the dataset contains multiple folders named after each company of which the earnings call transcripts were acquired. The structure of these folders is portrayed in table 2.

Company Folder	Contents	Source
Apple Inc. (AAPL)	19 Text Documents	Thomson Reuters Eikon (2020)
Advanced Micro Devices, Inc. (AMD)	19 Text Documents	Thomson Reuters Eikon (2020)
Amazon.com, Inc. (AMZN)	19 Text Documents	Thomson Reuters Eikon (2020)
ASML Holding N.V. (ASML)	19 Text Documents	Thomson Reuters Eikon (2020)
Cisco Systems, Inc. (CSCO)	19 Text Documents	Thomson Reuters Eikon (2020)
Alphabet Inc. (GOOGL)	19 Text Documents	Thomson Reuters Eikon (2020)
Intel Corporation (INTC)	19 Text Documents	Thomson Reuters Eikon (2020)
Microsoft Corporation (MSFT)	19 Text Documents	Thomson Reuters Eikon (2020)
Micron Technology, Inc. (MU)	17 Text Documents	Thomson Reuters Eikon (2020)
NVIDIA Corporation (NVDA)	19 Text Documents	Thomson Reuters Eikon (2020)

Table 2: Folder structure transcripts

The earnings call transcripts are text documents that are all structured in the same manner. Date, time, participants, and the words that are spoken during the earnings call are all registered within these documents. Accessing these data can be done through standard software such as Notepad on Windows devices. A snippet of what these documents look like in Notepad is portrayed in figure 2.

```
=====
Conference Call Participants
=====

* Katy Huberty
  Morgan Stanley - Analyst
* Gene Munster
  Piper Jaffray & Co. - Analyst
* Rod Hall
  JPMorgan - Analyst
* Shannon Cross
  Cross Research - Analyst
* Toni Sacconaghi
  Bernstein - Analyst
* Simona Jankowski
  Goldman Sachs - Analyst
* Steve Milunovich
  UBS - Analyst

=====
Presentation
-----
Operator      [1]
-----

      Good day, everyone, and welcome to the Apple, Incorporated second-quarter FY16 earnings r

-----
Nancy Paxton, Apple Inc. - Senior Director of IR      [2]
-----

      Thank you. Good afternoon, and thanks to everyone for joining us today. Speaking first is
Please note that some of the information you'll hear during our discussion will consist of forward-
For more information, please refer to the risk factors discussed in Apples Form 10-K for 2015, the
In addition, today's comments will refer to a metric we describe as the purchase value of services
```

**Figure 2:** Snippet earnings call transcript

### 3.1 Determining positive and negative transcripts

Formulas are used to determine whether a transcript is positive or negative. Firstly, the stock ratio formula, which has the following form:

$$\text{stock ratio} = \text{stock value one day after earnings call} / \text{stock value } n \text{ days before earnings call}$$

The stock ratio shows the percentage increase or decrease of the stock value. However, a percentage increase in stock value does not immediately imply that the earnings call is positive as there are other variables to consider. To factor in an additional variable called investor mood, a second formula is defined:

$$\text{sector ratio} = \text{sector value one day after earnings call} / \text{sector value } n \text{ days before earnings call}$$

Sector refers to the NASDAQ composite. The sector ratio is taken into account to consider the mood of the sector index. If the increase in stock ratio turns out to be higher than the increase of the sector ratio, the earnings call can be determined positive. If not, the transcript is deemed negative.



## 4. Discussion

The collected dataset provides the following value:

- These data can prove useful as they may help to further uncover dynamics related to correlational relationships between stock values and earnings call transcripts.
- Furthermore, the data can easily be expanded by i.e. extending the date range. Additionally, the data is easy to use and readable by multiple programming languages.
- Both practitioners at companies as well as scholars can benefit from the use of these data. Every company and scholar uses homemade datasets with consequential discrepancies. The adoption of a shared dataset for benchmarking analysis will promote a homogeneous evaluation of the results.
- The data was used primarily for the application of a limited amount of NLP techniques and machine learning algorithms. Consequently, this dataset offers the possibility to explore different approaches.

## 5. Conclusions

In this preprint we presented a dataset that has been designed for performing sentiment analysis in the stock market. Information regarding daily price and volume has been collected using yahoo finance. At the same, Thomson Reuters has been used for collecting earning transcripts. Details about the procedure has been described and presented in the previous sections. The dataset contains 11970 stock prices, and 1196 sector index values. Furthermore, all of these data originated in the period 2016-2020 and are related to the NASDAQ stock market.

The dataset can be used for developing and benchmarking NLP techniques and machine learning algorithms.

### Competing Interests

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

## 6. References

- [1] Roozen, D. R. H. (2021). Correlations between Earnings Call Transcripts and Stock Price: a Sentiment Analysis Approach. Master thesis at Tilburg University
- [2] Rao, T., & Srivastava, S. (2012). Analyzing stock market movements using twitter sentiment analysis.

- [3] Kordonis, J., Symeonidis, S., & Arampatzis, A. (2016). Stock price forecasting via sentiment analysis on twitter. In *Proceedings of the 20th Pan-Hellenic Conference on Informatics* (pp. 1-6).
- [4] Jha, V., Blaine, J., & Montague, W. (2015). Finding Value in Earnings Transcripts Data with AlphaSense. Available at [extractalpha.com](http://extractalpha.com)
- [5] Heinrichs, A., Park, J., & Soltes, E. F. (2019). Who consumes firm disclosures? Evidence from earnings conference calls. *The Accounting Review*, 94(3), 205-231.
- [6] Theil, C. K., Broscheit, S., & Stuckenschmidt, H. (2019). PRoFET: Predicting the Risk of Firms from Event Transcripts. In *IJCAI* (pp. 5211-5217).
- [7] Jenkins, P. (2020, April). Structured Paragraph Embeddings of Financial Earnings Calls. In *Companion Proceedings of the Web Conference 2020* (pp. 264-268).