### Datasheet for Input Data: Earnings Transcripts and Stock Prices

This datasheet provides a comprehensive overview of the input data used for classifying the sentiment of earnings transcripts and their impact on stock prices.

#### Motivation

* **Purpose of Dataset**: To classify the sentiment of earnings transcripts and their impact on stock prices. The dataset aims to fill the gap in predicting stock movements based on earnings call sentiment.
  + **Specific Task**: Sentiment classification of earnings calls to generate trading signals.
* **Creator and Funding**: Created by an independent contributor using publicly available resources.

#### Composition

* **Instances**: The dataset consists of earnings transcripts (text documents) and daily stock prices (numerical values).
* **Total Instances**: 3,198 data points.
* **Representativeness**: Sample from S&P 500 companies, covering a diverse range of industries and sectors. The dataset is representative of the larger set of companies in the S&P 500.
* **Instance Details**:
  + **Transcripts**: Raw text from earnings calls.
  + **Stock Prices**: Historical daily closing prices.
* **Labels**: Each instance is labeled based on stock price movement:
  + **Positive (2)**: Price change ≥ +2%
  + **Neutral (1)**: Price change between -2% and +2%
  + **Negative (0)**: Price change ≤ -2%
* **Missing Information**: No missing data; ensured completeness.
* **Relationships**: Explicit relationships between transcripts and corresponding stock prices.
* **Data Splits**:
  + **Training**: April 2019 to December 2021.
  + **Test**: January 2022 to February 2023.
* **Self-Contained**: Dataset includes all necessary data without relying on external sources.
* **Identifiability**: Companies are identifiable by their ticker symbols, but no individual personal data is included.
* **Sensitive Data**: No sensitive information present.
* **Other Comments**: Comprehensive dataset aimed at financial sentiment analysis.

#### Collection Process

* **Data Acquisition**:
  + **Transcripts**: Scraped from [Kaggle - Motley Fool Scraped Earnings Call Transcripts](https://www.kaggle.com/datasets/tpotterer/motley-fool-scraped-earnings-call-transcripts).
  + **Stock Prices**: Downloaded from Yahoo Finance.
  + **Tickers**: Collected from Wikipedia's list of S&P 500 companies.
* **Sampling Strategy**: Comprehensive coverage of all available transcripts and stock prices for the given period.
* **Time Frame**: Data collected from April 2019 to February 2023.
* **Ethical Review**: Not applicable as data is publicly available.
* **Consent**: Not applicable for publicly available data.
* **Data Protection Impact Analysis**: Not conducted as data is publicly available.
* **Other Comments**: Data collection ensured accuracy and completeness.

#### Preprocessing/Cleaning/Labeling

* **Preprocessing**:
  + **Text**: Tokenization, duplicates removal, and conversion to embeddings using RoBERTa.
  + **Stock Prices**: Daily closing prices normalized and synchronized with transcripts.
* **Labels**: Based on stock price changes as described in the composition section.
* **Raw Data**: Saved alongside preprocessed data for future use.
* **Other Comments**: Ensured data consistency and quality.

#### Uses

* **Other Tasks**: Sentiment analysis, financial forecasting, natural language processing tasks.
* **Impact on Future Uses**: Preprocessing steps are crucial for replicability; users should be aware of potential biases in labeling.
* **Restrictions**: Not suitable for personal financial advice or prediction without further validation.

#### Distribution

* **Third-Party Distribution**: Potential for sharing with academic institutions or researchers upon request.
* **Distribution Method**: Via secure download links or data sharing platforms.
* **Distribution Timeline**: Available upon request.

#### Maintenance

* **Maintainer**: Independent data scientist.
* **Other Comments**: Plans to extend the data history to improve accuracy.