Earnings Conference Call Transcripts - Sentiment Analysis



Brian Higgins

School of Computer Science

University of Galway

*Supervisor* Dr Josephine Griffith

In partial fulfilment of the requirements for the degree of *MSc in Software Engineering and Database Technologies*

August 2024

DECLARATION

I, Brian Higgins, do hereby declare that this thesis entitled ‘Earnings Conference Calls – Sentiment Analysis’ is a bonafide record of research work done by me for the award of MSc in Software Engineering and Database Technologies from the University of Galway. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Acknowledgements*

*I wish to express my eternal gratitude to my wife Teresa and daughter Catherine for their constant encouragement, patience and support.*

*I would like to thank my supervisor Dr Josephine Griffith for all her time, support and guidance in completing this thesis.*

Abstract.

Quarterly and annual Earnings Conference Calls (ECC) are webcasts or conference calls held by publicly traded companies to present and discuss their latest financial results. The participants are company senior executives, institutional investors and financial analysts. Members of the general public may attend in listen mode. ECCs are significant events in the financial year of companies. They offer the first opportunity for companies to discuss the latest financial results with analysts and investors and to answer their questions directly. Transcripts of ECCs are made available on company websites following the calls. Earnings Conference Calls are carefully studied by stock market analysts and investors in attempts to discover new information and which could assist in investment decisions. Information disclosed in Earnings Calls is not confined to technical and fundamental company data, much of which will have been disclosed in the earlier earnings press release. Information is also contained in the sentiment or linguist tone of the management and analysts. Studies have found that the sentiment of earnings calls may have an impact on stock returns. Both positive and negative correlations have been found. This thesis examines if the sentiment expressed in the questions and answers section of Earnings Conference Call transcripts can be related to subsequent stock price changes. Earnings Conference Call transcripts will be web scraped from a single global financial data provider. The sentiment of ECC transcripts will be extracted using FinBERT, a deep learning language model specifically designed for financial text analysis. The research will be implemented in Python 3 version 3.9.8. Selenium WebDriver2 will be used to automatically control the browser. Extensive Python library Pandas3 will be made to manipulate and analyse the data.

**Keywords:** earnings conference calls, sentiment analysis, FinBERT.

CONTENTS

1 Introduction……………………………………………………………………

1.1 Research topic and motivation

1.2 Background.

1.3 Earnings calls.

1.4 Significance of earnings calls.

1.5 Study of earnings calls.

3. Sentiment analysis.

3.1Sentiment analysis methods.

3.2 sentiment analysis models.

3.3 Large Language models.

3.4 FinBERT. 21

3.5 BERT

3.6 Transformers

3.7 FinBERT model used in this research.

1.6 Thesis structure.

2. Literature review.

5. Data

5.1 Data source

5.2 Data gathering

5.3 Data preparation

5.4 Pipeline

6. Analysis of the data.

6.1 FinBERT Model used. 24

6.2 Data input to model.

6.3 Output.

7. Results

7.1 What the results reveal.

8. Limitations.

9. Future work.

Contents

[CHAPTER 1 8](#_Toc172987495)

[1.1 Research topic and motivation. 8](#_Toc172987496)

[1.2 Background 9](#_Toc172987497)

[1.2.1 Earnings Conference Calls 11](#_Toc172987498)

[1.3 Thesis structure 14](#_Toc172987499)

[Chapter 2 14](#_Toc172987500)

[Literature Review. 14](#_Toc172987501)

[Chapter 4 16](#_Toc172987502)

[Data 16](#_Toc172987503)

[4.1 Data description 16](#_Toc172987504)

[4.2 Web scraping the list of available transcripts. 17](#_Toc172987505)

[4.3 Web-scraping the individual transcripts. 22](#_Toc172987506)

[Chapter 5. 23](#_Toc172987507)

[Processing of the transcripts 23](#_Toc172987508)

# CHAPTER 1

Introduction

# 1.1 Research topic and motivation.

Investors and stock market analysts’ study both quantitative and qualitative data when considering investment decisions. Quantitative data includes for example, earnings per share (EPS), profit and loss, share price. Qualitive or unstructured data includes annual and quarterly reports, news reports, social media posts and earnings conference call transcripts. The study of sentiment of earnings conference call transcripts and its relationship to stock market movements is an active area of research. Medya et al show that the semantic characteristics of earnings call transcripts can play a role in the prediction of stock price movements. (Medya *et al.*, 2022). A study by Price et al (Price *et al.*, 2012) found that the linguist tone of conference calls holds incremental information additional to earnings press releases and has use in the prediction of abnormal stock returns and trading volume. The extraction of sentiment of conference call transcripts can be carried out using various methods including lexicon or dictionary-based methods, machine learningmethods and deep learning techniques. Advances in Natural Language Processing (NLP) include the introduction of transformer deep learning models with improved sentiment classification accuracy. This thesis will employ FinBERT a deep learning model, to extract the sentiment from earnings conference call transcripts and correlate it to subsequent stock price movements.

The goal of this research is:

* Extract the sentiment of individual questions and answers from the Q&A section of earnings conference call transcripts using deep learning NLP model.
* Examine if a statistical relationship is evident between the sentiment and later stock price movements.

1.2 Background Understanding the factors affecting stock market movements is of interest to investors and analysts. Methods currently employed for stock market prediction are unreliable. As a result, there is much research in this area.

Two main theoretical hypotheses define market behaviour: the efficient market hypothesis (EMH) (Malkiel & Fama, 1970) and the adaptive market hypothesis (AMH) (Lo, 2005). The notion that markets are random and not predictable is firmly established in the random walk theory Bollen, Mao, and Zeng (2011). ..EMH …(need to describe the EMH…)…. *Behavioural Finance gives another view*… *Behavioural finance investigates stock market movements based on the phycology of* investors ….principles (Picasso, Merello, Ma, Oneto, & Cambria, 2019). T, . *The AMH is based on the belief that investors learn from their mstakes, and* .. people are motivated by self-interest….

*Stock prices are analysed mainly by two methods: Technical and fundamental. Technical analysis consists of modelling of historic stock price data in order to make predictions.*

*Fundamental analysis involves examining a range of factors both internal and external to a company that may affect its share price. Internal factors include the company’s financial ratios…*The fundamental aspects are economic data, financial performance, political and social behaviors, the business environment, and the firm’s financial ratios (Beyaz, Tekiner, Zeng, & Keane, 2018).We point out market capitalization (MC), earnings per share (EPS), the price/sales ratio (P/S), and the debt/equity ratio (D/E) as some of the notable financial ratios.

~~The company will have published its quarter financial results earlier the same day. Financial analysts and investors will be anxious to hear from the senior company management directly and take the opportunity to question the company on the results and future operations. In this way earnings calls provide a direct communication between senior company management, financial analysts and investors.~~

~~They were broken down to three sections. The management discussion and analysis (MD&A), the question-and-answer session (Q&A), and the questions and answers individually. The sentiment of these sections will be extracted separately. The resulting data will be analysed to measure the correlation between the sentiment and subsequent stock price movements.~~

~~The transcripts were web scraped from the Seeking Alpha website~~~~1~~ .

1*https://seekingalpha.com/*

2 https://www.selenium.dev/documentation/webdriver/

3 https://pandas.pydata.org/

4 https://huggingface.co/ProsusAI/finbert

## [1.2.1 Earnings Conference Calls](#_1._Earnings_Conference)

Earnings Conference Calls are quarterly conference calls/webcasts hosted by publicly traded companies to present and discuss the latest quarterly or annual earnings reports and outlook with financial analysts and investors. Companies usually issue quarterly earnings press releases within four to six weeks following the end of the previous quarter. Press releases which disclose details of the company performance are followed by earnings conference calls which usually take place on the same day or the next day. The participants are typically the company senior management, normally the CEO and CFO, who present the latest financial reports and who will later take questions and financial analysts, institutional investors and financial journalists who may question the management in regard to the company performance and plans. The financial analysts will normally be those who cover the particular company or the industrial sector in which the company operates. Members of the general public can attend, normally in listen mode only. Earnings Conference Calls can be freely accessed via the companies’ websites, usually on the ‘Investor Relations’ page. They typically last between 30-60 minutes. Notification of the date and time is posted a few days in advance of the earnings press release, Transcripts and audio recordings are normally made available on company websites following the conference calls. As an example can be found on the Microsoft Inc website (microsoft.com).

Reporting requirements.

All companies listed on a US stock exchange are legally required to publish detailed financial information regarding the current performance of the company on a quarterly basis (Securities Act of 1933). At the end of each of the three financial quarters Earnings Reports must be filed with the U.S. Securities and Exchange Commission (SEC). The three quarterly Earnings Reports are known a 10Q reports and contain unaudited financial statements and review of the quarter’s operations. The fourth quarter report is the annual and more detailed Earnings report which is filed following the end of the fiscal year. The annual Earnings report is known as the 10K report and contains audited financial statements, annual review of markets, operations and company organisation and history, its products and services.

10Q Reports must be filed with the SEC within 45 days of the quarter end. 10K reports must be filed within the SEC 60 days of the end of the fiscal year.

They provide detailed financial information regarding the company performance for the most recent quarter including balance sheet, sales, cash flow, earnings per share and include a comparison with the same quarter the previous year.

Quarterly company reports to the SEC in the USA are mandatory, but conference calls are not, however most of the larger listed companies conduct earnings conference calls.

Earnings Seasons Many companies in the US align their financial year with the calendar year. This leads to a tendency for earnings reports to be released in clusters within specific date ranges. These date ranges are called ‘Earnings Seasons’. Mid-April to Mid-May 1’st QTR, Mid July to Mid Aug -2’nd QTR, Mid Oct – Mid Nov – 3’rd Qtr, Mid Jan – Mid Feb -4’th Qtr. It is often the case that hundreds of companies could be releasing earnings reports and hosting Earnings Calls on the same day(forbes.com, 2023).

Earnings Call Schedule. Companies announce in advance the date when earnings will be released. At the same time the date and time of the related earnings call/webcast is announced with details of how to attend, Access to the call is usually via the company website.

Earnings Press releases are issued a few weeks after quarter or year-end in advance of the SEC filings. The Earnings Press release contains a summary of the financial performance for the quarter or year and includes ley metrics such as Earnings Per Share (EPS), Revenue, Net Income and Sales.

Earnings Conference Calls are usually held on the same day or next day following the press releases. There is no legal requirement to hold earnings calls conferences, but most companies hold them in the interests of transparency and investors relations. In the period considered by this thesis there are records of conference calls by 499 companies that make up the S&P 500 index. The exception being Berkshire Hathaway Inc (BRK), the holding company run by Warren Buffet, which does not hold earnings calls.

At the time of the earnings press releases the company has some idea of how the current quarter is progressing and often will incorporate projections for the coming quarter known as forward guidance.

## 1.3 Thesis structure

This thesis consists of seven chatpers.

The introduction and background is given in chapter 1. Chapter 2 reviews the current research activity in the study of earnings conference call transcripts. The data used in this research is described in Chapter 4

# Chapter 2

## Literature Review.

Price et al (Price *et al.*, 2012) found that the tone of language used in earnings calls can be correlated to the subsequent abnormal stock returns by comparing the actual returns to the expected returns. They examined earnings call transcripts and quantified the textual tone of the MD&A section. The relationship between the tone and subsequent stock returns was investigated. Tone was quantified using the Loughran-McDonald financial sentiment dictionary and the Harvard IV-4 psychosocial dictionary. A positive tone was found to be correlated to positive abnormal stock returns. A negative tone was found to be correlated to negative abnormal stock returns. The conference paper ‘Forecasting Earnings Surprises from Conference Call Transcripts’ by Koval et al describes the development of a deep learning model to predict earnings surprises. Earnings surprises are a measure of the difference between the actual Earnings Per Share (EPS) and the analysts’ consensus estimates. The sentiment of Earnings Call transcripts was extracted using a number of different traditional models and hierarchical transformer models. They link each transcript to the subsequent earnings report to determine if an earnings surprise occurred. Their results show that it is possible to predict earnings surprises with what they call ‘reasonable accuracy’ using only the text of earnings conference call transcripts. The authors acknowledge certain limitations of the work. Nagendra et al (Bv *et al.*, 2023) in their conference paper ‘Deploying NLP Techniques for Earnings Call Transcripts for Financial Analysis: A Reverse Phenomenon Paradigm’ observe what they term the ’Inverse Effect’ where stock price movements are in the opposite direction to the sentiment of the corresponding Earnings Call transcripts. In order to extract the sentiment, they used the VADER (Valenc Aware Dictionary and Sentiment Reader) which is a lexicon-based analysis model which classifies text as positive, negative or neutral sentiment. The results show that there are significant instances of positive sentiment followed by negative stock price movement. They term this as the ‘Inverse Effect’.

.( In sharp contrast to the optimism revealed by MANAGER TONE, ANALYST TONE is either negative (mean = –0.01) or neutral (median = 0.00)( Differences in Conference Call Tones: Managers vs. Analysts Paul Brockman Pg 9))

# Chapter 4

## Data

### 4.1 Data description

The data consists of earnings conference call transcripts and historical stock price data. Transcripts were web scraped from the Seeking Alpha website2.A total of 5126 transcripts covering the period November 2021 to February 2024 were obtained. This covers financial quarters 2021\_03 to 2023\_04. Of these 3956 are earnings conference call transcripts. The remaining 1,170 are not earnings call transcripts but are transcripts from other events. For example, transcripts of company presentations at financial conferences. These were discarded from the analysis.

Seeking Alpha is a privately owned crowd-sourced financial data content provider. It publishes news, analysis and research on financial markets. Published Earnings Conference Call Transcripts are gathered globally by Seeking Alpha from company websites and made available on their website. Access is either by free access with viewing limited of one transcript per day or via a number paid plans with varying degrees of access. For this research the data was web scraped directly from the website. The process will be described below. This research focusses on S&P 500 companies. The reason for this was twofold. Firstly, these are the stocks that make up the S&P 500 index. The S&P 500 index is considered to be the numerical indicator or barometer of the U.S. economy. It is made up of 500 of the largest companies listed on the NYSE, Nasdaq, or CBOE. (Investopedia, 2023) and covers the. Secondly it was necessary to restrict the amount of data in preparation for the processing of long documents and the likely lengthy processing time. Historical stock price data for each of the 500 companies that make up the S&P 500 was obtained from YahooFinance.com3 .

### 4.2 Web scraping the list of available transcripts.

Earnings Call Transcripts were web scraped from the Seeking Alpha website3. Details of the entire set transcripts available on the website can be found by directly accessing a Seeking Alpha API endpoint (Application Programming Interface). API endpoints are the final locations from where information is sent and received by the server. A website can have one or multiple endpoints. This endpoint here is the Seeking Alpha API endpoint4 URL holding lists of transcript details. The data is presented as a list of 50 JSON objects per page (JavaScript Object Notation). Each JSON object contains detailed individual transcript information. The JSON data is structured to provide a comprehensive set of details about each transcript, including metadata (such as publish date and title), related entities (such as author and tickers), and links to the full content.

Attributes include a unique seven-digit transcript id, company name, ticker symbol, date, type of transcript and fiscal period and further metadata relevant to each transcript. The JSON data structure simplifies access, analysis and extraction of various pieces of the data.

*2 https://seekingalpha.com/*

*3 https://finance.yahoo.com/quote/AAPL/*

*4 https://seekingalpha.com/api/v3/articles?filter[category]=earnings%3A%3Aearnings-call-transcripts&filter[since]=0&filter[until]=0&include=author%2CprimaryTickers%2CsecondaryTickers&isMounting=true&page[size]=50&page[number]=1*

Below is an example of a JSON object from the Seeking Alpha API endpoint URL. It relates to transcript id:4619723. This is the Texas Instruments Incorporated (TXN) Q2 2023 Earnings Call Transcript published on 2023-07-25 at 18:47:06 local time with an offset from GMT of minus 4 Hrs.

{"id":"4619723","type":"transcript","attributes":{"publishOn":"2023-07-25T18:47:06-04:00","isLockedPro":false,"commentCount":0,"gettyImageUrl":null,"videoPreviewUrl":null,"videoDuration":null,"themes":{},"title":"Texas Instruments Incorporated (TXN) Q2 2023 Earnings Call Transcript","isPaywalled":false},"relationships":{"author":{"data":{"id":"44211","type":"author"}},"sentiments":{"data":[]},"primaryTickers":{"data":[{"id":"1051","type":"tag"}]},"secondaryTickers":{"data":[]},"otherTags":{"data":[{"id":"96991","type":"tag"},{"id":"49","type":"tag"},{"id":"586376","type":"tag"},{"id":"326","type":"tag"}]}},"links":{"self":"/article/4619723-texas-instruments-incorporated-txn-q2-2023-earnings-call-transcript"}}.

In a more readable form, it looks like:

*json\_data = '''{*

*"id": "4619723",*

*"type": "transcript",*

*"attributes": {*

*"publishOn": "2023-07-25T18:47:06-04:00",*

*"isLockedPro": false,*

*"commentCount": 0,*

*"gettyImageUrl": null,*

*"videoPreviewUrl": null,*

*"videoDuration": null,*

*"themes": {},*

*"title": "Texas Instruments Incorporated (TXN) Q2 2023 Earnings Call Transcript",*

*"isPaywalled": false*

*},*

*"relationships": {*

*"author": {*

*"data": {*

*"id": "44211",*

*"type": "author"*

*}*

*},*

*"sentiments": {*

*"data": []*

*},*

*"primaryTickers": {*

*"data": [*

*{*

*"id": "1051",*

*"type": "tag"*

*}*

*]*

*},*

*"secondaryTickers": {*

*"data": []*

*},*

*"otherTags": {*

*"data": [*

*{*

*"id": "96991",*

*"type": "tag"*

*},*

*{*

*"id": "49",*

*"type": "tag"*

*},*

*{*

*"id": "586376",*

*"type": "tag"*

*},*

*{*

*"id": "326",*

*"type": "tag"*

*}*

*]*

*}*

*},*

*"links": {*

*"self": "/article/4619723-texas-instruments-incorporated-txn-q2-2023-earnings-call-transcript"*

*}*

*}'''*

In summary it contains the following information:

Transcript ID: 4619723

Type: transcript

Publish Date: 2023-07-25T18:47:06-04:00

Is Locked Pro: False

Comment Count: 0

Title: Texas Instruments Incorporated (TXN) Q2 2023 Earnings Call Transcript

Is Paywalled: False i.e. Not behind a paywall – content is accessable without payment.

Author ID: 44211

Primary Tickers: ['1051'] - ID used in the ‘primary’ stock market area.

Other Tags: ['96991', '49', '586376', '326']

Self Link: /article/4619723-texas-instruments-incorporated-txn-q2-2023-earnings-call-transcript

(link to the transcript.)

The endpoint URL is paged. This allows the transcript data to be retrieved in a managable manner. By iterating through the page numbers and extracting data that is required it is possible to create a list of all of the available transcripts on the website. A list containing the unique transcript id, company\_name, ticker symbol, published date and fiscal period of all the available transcripts from every JSON object from every page. Pages 1 up to page 1000 are accessible. The pages are dynamic, with the oldest data, on page 1000 being dropped each day and the latest data being placed on page 1.

A list of all transcripts available on the Seeking Alpha website was constructed. It contained the unique id, ticker symbol, company name, date and time of earnings call, and fiscal period for each transcript.

Preparation of the transcript list was carried out as follows.

A web scraping tool was constructed in Python making use of the Selenium browser control package and the requests package. The data-scraper automatically iterated over pages 1 to 1000 of the Seeking Alpha endpoint extracting the ‘id’, ‘publishOn’and ‘title’ in regard to each of the 50 transcripts on each page. This data was written to the file ‘updated\_transcript\_data.csv’. In order to avoid blocking by the website, the program was designed with random time break between each group of fifteen transcripts and additionally a random time disconnection and reconnection to the website after each 350 batch of transcripts.

This process yielded a list of 50k transcript ids and the associated data as above.

Identifying transcripts relating to S&P 500 companies. The transcript list contains transcript data in regard to all companies listed on US stock exchanges who hold earnings conference calls. This thesis focusses on S&P 500 companies. In order to identify this sub-set, a list of S&P 500 ticker symbols was obtained from a stock research website stockanalysis.com5. This list was used to update a new field ‘S&P500\_Company’ in the csv file. This field was updated to ‘yes’ in the case of a ‘S&P500\_Company’ otherwise ‘no’.

### 4.3 Web-scraping the individual transcripts.

To obtain the individual transcripts another API endpoint on the Seeking Alpha website was accessed6. Each page on this endpoint holds a single transcript with its associated meta data. The URL for each individual transcript can be identified by the unique transcript id referred to above. To fetch and save each of the individual 5808 transcripts a second web data scraper was designed. This data scraper accessed the URL of each transcript in turn by iterating over each id which had been labelled as ‘S&P500\_Company’ in the csv file created in stage one. This ensured that only S&P500 company transcripts were fetched. The transcripts were written to a csv file xxxx.csv. The full set of transcripts was written to a series of csv files for simple access and easy checking of the data.

After completing this process **(*provide code/pipeline*)** two blocks of data were available to hand.

(1) A complete list of available earnings calls transcripts from the Seeking Alpha website. (2) A complete set of Earnings Calls transcripts of S&P 500 companies between Qtr3 2021 and Qtr1 2024.

5 https://stockanalysis.com/list/sp-500-stocks/ 6https://seekingalpha.com/api/v3/articles/4635802?include=author%2CprimaryTickers%2CsecondaryTickers%2CotherTags%2Cpresentations%2Cpresentations.slides%2Cauthor.authorResearch%2Cauthor.userBioTags%2Cco\_authors%2CpromotedService%2Csentiments

# Chapter 5.

## Processing of the transcripts

The transcripts were in HTML format when web scraped. They were automatically written to a series of csv files by the data-scraper.

Earnings call transcripts are long documents with word counts ranging between 6000 and 12000 words. They were written to the csv files in chunks of 8,000 characters to a row. This method resulted in each transcript being spread over a number of rows of the csv file. Typically, a transcript of 11,000 words would have approx 80k characters incl spaces. Such a transcript in its raw scraped html state would be spread over ten rows of the csv file when HTML Tags are included. Earnings calls are normally conducted in two distinct sections. Section one, the management discussion and analysis section (MD&A), is a prepared and scripted presentation by management of the financial results, plans and outlook for the company. The second section is the Q&A section. In the QA section analysts question management on the latest quarter results and future prospects and plans. This section is unscripted as analysts present their own questions . Brockman et al found the Q&A section to be more informative than the MD&A section in prediction of future returns (Brockman, Li and Price, 2015). The length of the transcripts made them unsuitable to attempt to place a complete transcript on a single csv file row. On average each transcript occupied eight csv file rows. This series of csv files contained the raw HTML formatted transcripts.

An example of some rows of one of the csv files is shown below:



The raw transcripts were then processed to remove the HTML tags leaving plain text. Further pre-processing split the transcripts into company statement (MD&A) and Q&A sections. The two sections were written to the csv file in row pairs with the ‘Q&A’ written over a number of rows and identified as ‘Q&A Session’ in the column ‘Call\_Section’. The MD&A was written to the rows immediately below in the same manner and identified as ‘company\_statement’. The one ID in the ‘ID’ column identified all the rows which formed part of the same transcript. Kimbrough(Kimbrough, 2005) in his analysis of earnings calls contends that The MD&A is a prepared and scripted reiteration of the earlier Earnings Press Release. He splits the analysis of Earnings Calls into two sections, the MD&A and the Q&A. Similarly Price (Price *et al.*, 2012) in their analysis split transcripts into what they call INTRO and Q&A sections.

~~Stop word removal, stemming or other such pre-processing was not carried to avoid possible loss of context.~~

A sample is shown below:



A number of the transcripts were found to relate to events other than quarterly earnings calls and were excluded from the analysis. For example, transcripts of company presentations at financial conferences. The excluded number amounted to 1552. This left 4256 transcripts for analysis.

Input of the transcripts to FinBERT.

FinBERT

FinBERT is an open-source Natural Language Processing model (NLP) specifically trained for financial sentiment analysis. It is based on the BERT language model and is constructed by further training BERT on the unlabelled 1.1 million word Reuters news corpus. This was followed by fine tuning on a human annotated 5,000 sentence labelled financial text dataset.

BERT.

BERT (Bidirectional Encoder Representations from Transformers) is an open-source transformer-based machine learning architecture for natural language processing (NLP) introduced by Google in 2018. It is designed to help computers understand naturally spoken human language. In order to understand the meaning of words BERT uses surrounding text to determine the context and deal with ambiguity.

BERT splits or tokenizes the input text into sub word units (tokens) in a process called WordPiece tokenization. Tokens are converted to high dimensional (768 dimensions) numerical vector representations or embeddings. The initial token embeddings are augmented with positional embeddings which capture the token position in a sentence or sequence, and segment embeddings which distinguish between sentences. The token embeddings are processed by transformer encoder layers which apply a process called self-attention which takes into account the context in which a token appears. The result is a final embedding which is dynamically context aware.

Transformers.

BERT makes use of a transformer architecture.

Transformers are a type of neural network designed to process sequential data.

BERT is one of a class of large language models or LLMs. It employs two key innovations in language modelling. First the transformer architecture (T) is used for modelling long term dependencies and second it uses Masked Language Modelling (MLM) in which a random portion of the tokens or words are masked and the model predicts them, leading to bi-directionality (B).

FinBERT model used in this research.

FinBERT is an NLP model based on BERT and designed specifically for financial text analysis. The language used in the financial domain has its own characteristics particular to that domain and has differences from general purpose domain. Models developed for the general domain often struggle when applied to specialized language used in the finance domain. For example, “ACME Inc beats earnings forecast” might be interpreted as a negative sentiment in normal language but in the finance domain it is positive.

The FinBERT model used in this research is produced by Pross AI.

It is available on the Hugging Face model hubxx. The model was developed by Araci (Araci, 2019) to address the problem of financial sentiment analysis. Financial text contains specialized financial language not usually found is normal everyday language. General purpose sentiment analysis models do not perform well when attempting to classsify financial text. Araci used the pretrained BERT language model and provided it with further training on a large financial corpus

FinBERT is an NLP model based on BERT and designed specifically for financial text analysis. The language used in the financial domain has its own characteristics particular to that domain and has differences from general purpose domain. Models developed for the general domain often struggle when applied to specialized language used in the finance domain. For example, “ACME Inc beats earnings forecast” might be interpreted as a negative sentiment in normal language but in the finance domain it is positive.

Verifying the accuracy of the model.

The operation and accuracy of the model was checked using the Financial Phrase Bank as input and comparing the model predictions of sentiment to the human annotated values on which the model had been fine tuned. Tbl xxx shows the test results. The accuracy of xxx indicates that the model is performing well allowing for model limitations, generalizations, human annotation errors and data variability.

xx https://huggingface.co/ProsusAI/finbert

Operation of the model.

ECC transcripts are long documents and cannot be input to FinBERT as a complete unit. The maximum token length that can be input to a BERT model is 512 tokens. This includes 510 text tokens plus special start and end tokens inserted by the model. In order to input long documents to FinBERT two approaches are often used. Summarise the text or split the text into 512 length tokens using a sliding window method. The second method is used here.

Pipeline:

The data acquisition and analysis pipeline is shown in Fig xx.

Fig 51

Description of the pipeline modules.

Module\_1.

Web scraping the transcripts list.

Web scraping the list of available transcripts from the Seeking Alpha website was achieved by accessing an API endpoint holding transcript data. This is the URL where a list of JSON objects containing transcripts data is located. The endpoint is paged from 1 to 1000 with nominally 50 JSON objects per page. A web scraping tool was developed to access the endpoint and iterate through pages 1 to 1000 extracting relevant transcript data from each JSON object in turn. The data extracted consisted of the transcript ID, Stock market ticker symbol, financial quarter, description of transcript and the date of the conference call. This data was written automatically to a csv file ‘transcripts\_list\_data.csv’. This file was then updated by adding a column ‘S&P 500’. This column was updated from file ‘S&P 500 Index Stocks List.csv’ to mark the S&P 500 companies based on the ticker. Selenium WebDriver was used automate the browser and navigate to the API endpoint. The browser was configured to iterate through 150 pages per run. A random time delay of between 4 and 9 seconds was applied to ensure the page loaded and to prevent detection as an automated tool. A further random time delay of between 4 and 9 seconds was applied between page pulls to further reduce the chance of blocking by the website. BeautifulSoup and JSON were used to parse the HTML content and extract the data.

Module\_2.

Web scraping the individual transcripts. In order to obtain the individual transcripts a different Seeking Alpha API endpoint was accessed. This endpoint is the URL of each individual transcript location and is identified by the unique transcript ID. Using the transcript list from module\_1 it was possible to access each S&P 500 transcript by its unique ID and web scrape it to a csv file. Transcripts are long documents, typically between 8000 and 12000 words. Attempting to place them on one csv file row might cause problems. To overcome this the web scraping tool was designed to write the transcripts to the csv files in chunks of 8000 characters per row. In this way each transcript chunk fitted comfortably on a row with each entire transcript then occupying a number of rows of the csv file. Each transcript was identified by its unique ID placed in the ID column as well as the ticker, date, quarter and title on each line the transcript occupied. The transcripts were in HTM format. Selenium was used to automate the browser and navigate to the relevant URLs. The process was carried out in batches of 150 with Selenium configured to close the WebDriver and disconnect from the website before reopening it after a random period of between seven and nine minutes between batches to avoid being blocked. This web scraping process was set to run automatically.

Module \_3

Transferring the transcripts to plain text and splitting into sections.

The extracted HTML was parsed using BeautifulSoup. The text lying between the <body> tags was extracted in HTML format. The transcripts were later extracted in plain text and split into two sections. The MD&A section and the Q&A section. This was achieved by parsing the HTML using the python Beautiful Soup library. The MD&A section can be identified as lying between xxx and ‘Question and Answer Session’. The Q&A section can be identified as that section lying between ‘Question and Answer Session’ and ‘Twit Content’.

Prior to input to FinBERT the text was cleaned to remove a number of words and phrases such as ‘Good morning’, ‘Great question’ which do not reflect the sentiment of the question or answer. Tbl xxx shows the list of removals.

Tbl xxx. Phrases and words cleaned from text

|  |  |  |  |
| --- | --- | --- | --- |
| Phrases and words removed from transcripts prior to FinBERT processing | | | |
| *'good morning'* |  |  |  |
| *'good evening'* |  |  |  |
| *'good afternoon'* |  |  |  |
| *'good question'* |  |  |  |
| *'great question'* |  |  |  |
| *'Thanks for taking the question'* | |  |  |
| *'Thanks for the question'* |  |  |  |
| *'Thank you'* |  |  |  |
| *'Thanks'* |  |  |  |
| *'Great'* |  |  |  |

Module\_4

Input to FinBERT

The cleaned questions and answers extracted from the Q&A sections of the ECC transcripts were contained in a series of csv files. Each question and answer appeared on a separate row. The files were input to the model which classified the sentiment each question and answer as either positive, negative or neutral. This pipeline returned the input file with three new columns, count of +ve, count of negative and count of neutral. The counts referred to how many 510 token length sections of text in each question or answer were classified as +ve, -Ve or neutral.

Module\_6

Combining the output files from FinBERT

The series of output files from the model were aggregated and grouped on the ID and ticker fields. The result was a combined file with the complete set of transcript IDs, in each case split into questions (Q) and answers (A) on two separate rows. A sample of the first few rows is shown in Fig xxx.

The three columns ‘sum of +ve’, ‘sum of -ve’ and ‘sum of neutral’ show the number of questions and answers that were classified into each of the sentiment categories.



Fig XXX

The following columns were then added to the file.

Module\_7

Updating stock price data and tone in respect of each transcript.

Tone (positivity) was defined as:

(sum of +ve – sum of -ve) / (sum of +ve + sum of -ve).

Stock price data was contained in ‘stock\_data.csv’ files for each of the 500 S&P500 companies. As an example, the stock price data for Apple Inc was contained in file AAPL.csv and so on.

Module\_8

Correlation calculations and charts.

The correlation of sentiment and stock price movements was carried out as follows.

https://stockanalysis.com/list/sp-500-stocks/

Results.

The results data is contained in the results file ‘updated\_combined\_file\_with\_compatible\_date .csv ‘ It contains the following fields:



The fields ‘sum of +ve’, ‘sum\_of\_-ve and sum of neutral contain the sum of the classifications of questions and answers by the model in the case of each transcript. The results file was constructed as follows. The transcripts dataset was input to FinBERT in a series of csv files with questions and answers accommodated separately on each row in sequence. The model classified each row as positive, negative or neutral, and output the same file with four additional fields: Sentiment, count of positive, count of negative and count of neutral. As the model used was on the Kaggle website the size of input file was restricted. In order to get around this the input dataset was split into a manageable size files and input to the model in stages. All of the output files were then combined and summed for each unique ID and its associated question and answer.

Example of output file

Processing on Kaggle.

The files were input to the model using an account created on Kaggle xxxxx. Kaggle, a subsiriary of Google, is an online community of data scientists and machine learning engineers. It provides users with access to and use of use of AI models such as FinBERT leveraging the power of its processers. A account was created with Kaggel and the FinBERT model was accessed. Processing of the transcripts through the model requires the use of a GPU. Various GPU options are available on Kaggle. The Nvidia Tesla P100 GPU was chosen. Processing time for the entire Q&A transcript data set using this GPU was approximately eleven hours.

Verification of the model.

Operation of the model was tested by inputting the Financial PhraseBank datset and comparing the results. This is one of the datasets used by the developers of the model to fine tune it for use on financial text. The test produced an accuracy of xxxx. This test confirmed the correct working of the model. Details of the testing are shown in appendix xxx.

Tone

Following the method of Brockman et al (Brockman, Li and Price, 2015) a measure for linguistic tone (Tone) was defined. Linguist Tone (Tone) is defined as the difference between the sum positive and negative counts divided by the sum of the two.

Tone = (sum of Positive – sum of Negative) / (sum of Positive + sum of Negative). This definition provides a measure of the relative positivity and is bounded between –1 and +1.

The results file was updated with a new field ‘tone’ whose value was calculated as above for each row. This update provides the tone of the questions set and the answers set for each transcript.

Values of tone found in the results.

Tbl xxx. Tone of transcripts sections.



*Tbl xxx* shows the values of tone obtained for the various sections of the Q&A section of transcripts. The overall tone of answers was found to be positive with a value of +0.534. Questions had a negative tone of -0.197. Overall the Q&A section had a positive tone value of +0.288.

This is in line with the results of Brockman et al (Brockman, Li and Price, 2015) who measured the sentiment of conference call transcripts based on a lexicon method using the Loughran and McDonald (2011) dictionary. Brockman et al found that the most optimistic tone during an entire conference call was found to be the MD&A. During the Q&A session, the tone of answers were found to be significantly more positive than the tone of questions.

Correlation of tone with stock price change.

The results file contains the fields ‘price\_chng\_1day’, ‘price\_chng\_2day’ and ‘price\_chng\_5day’. The values are the difference in the mean stock market close price of the stock over the five/two/one days immediately prior to the ECC (Earnings Conference Call) date and the five/two/one days immediately following the ECC date. The Pearson correlation of tone and price change was calculated for the using the Pearson correlation package in Python.

Tbl xxx shows the findings.

Tbl XXX. Correlation of Tone and price change

The correlation of tone and price change was measured using the Pearson corelation coefficient. The pearson correlation coefficient is a measure of the linear correlation of two sets of data(Wiki – change this wording!)



Tb1xxx shows the correlation matrix for tone and price change. The highest Pearson correlation is for questions and the one-day price change. The value of 0.0749 indicates a weak positive relationship between the tone of questions and the following day stock price change. The associated scatter plots show a weak positive correlation. This is in line with the results of Blockman et al.

Table 1. Correlation of tone (questions) and avg price change



Fig 10 Correlation matrix

Tone (questions) and Price change

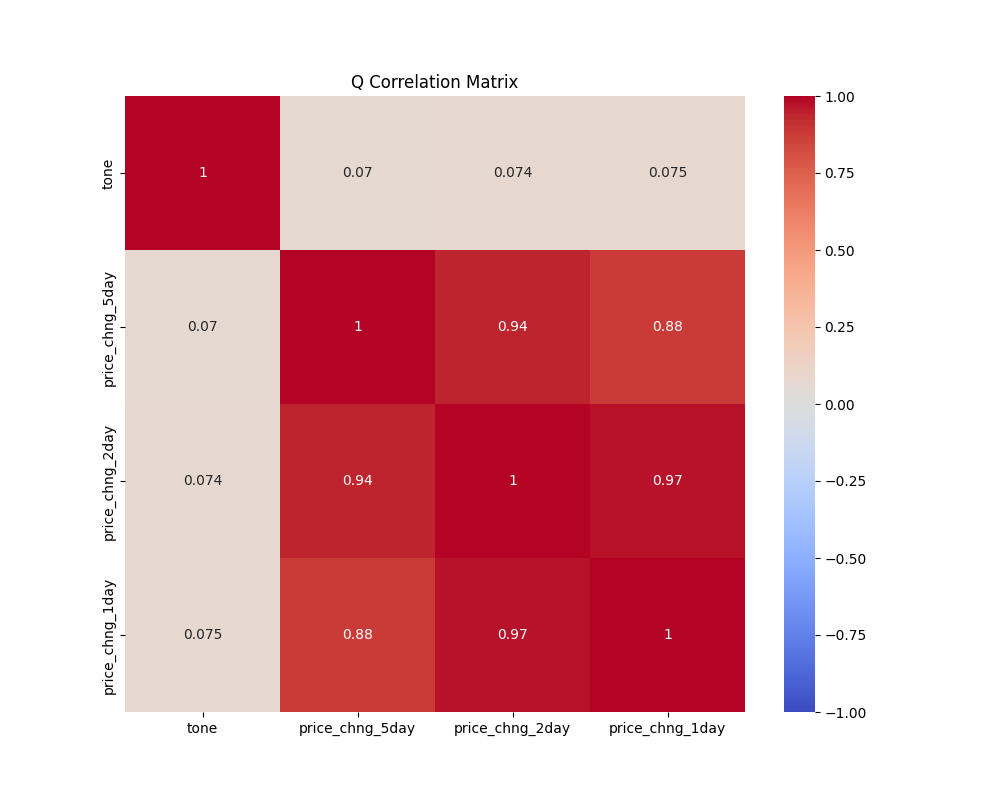


Fig xxx Questions

Scatter plot of Tone Vs Price Change 1day

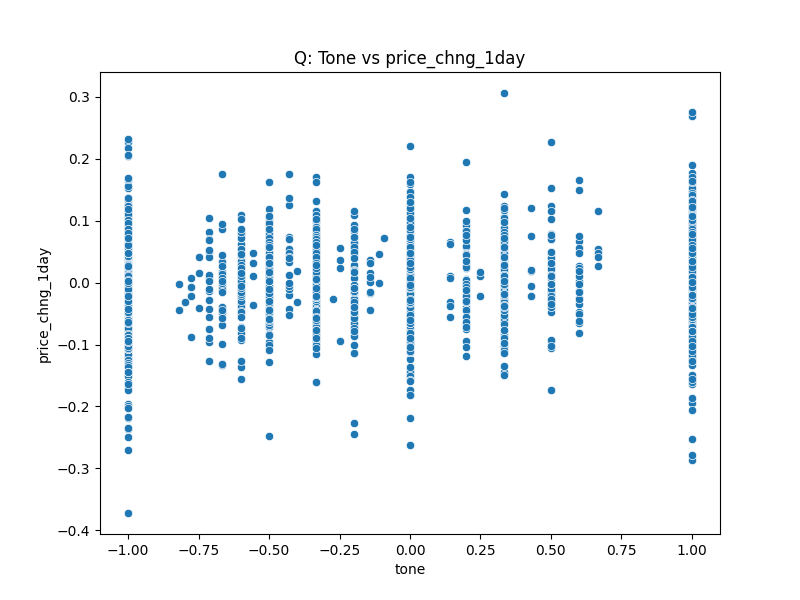


Fig xxx. Questions

Scatter plot of Tone Vs Price Change 2day

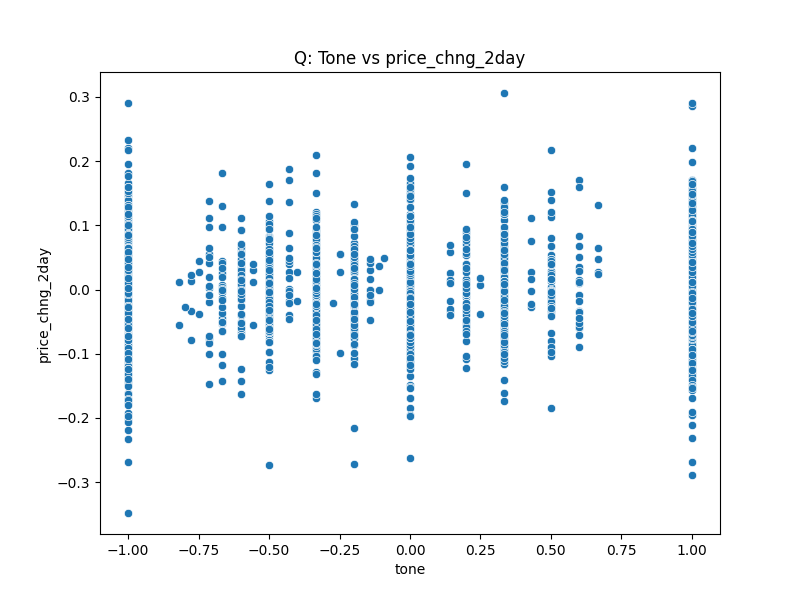
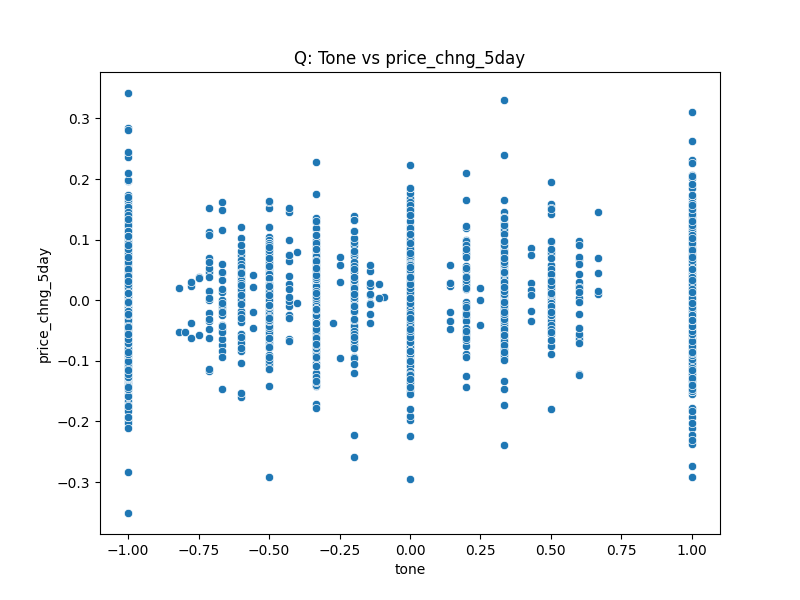
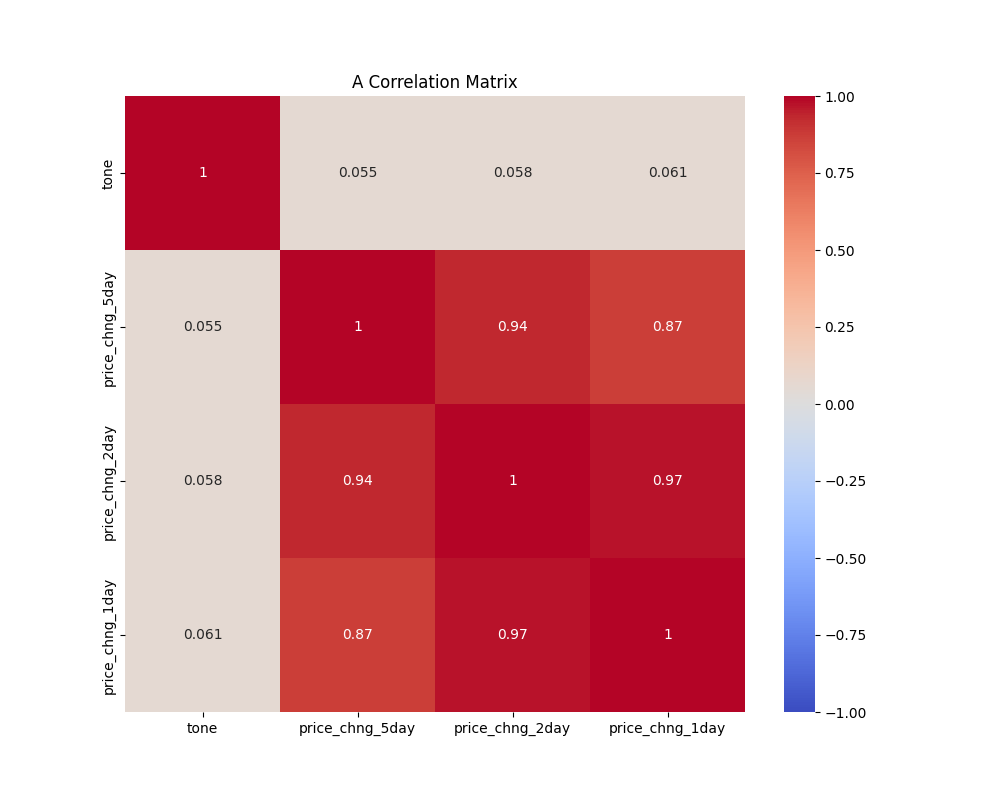


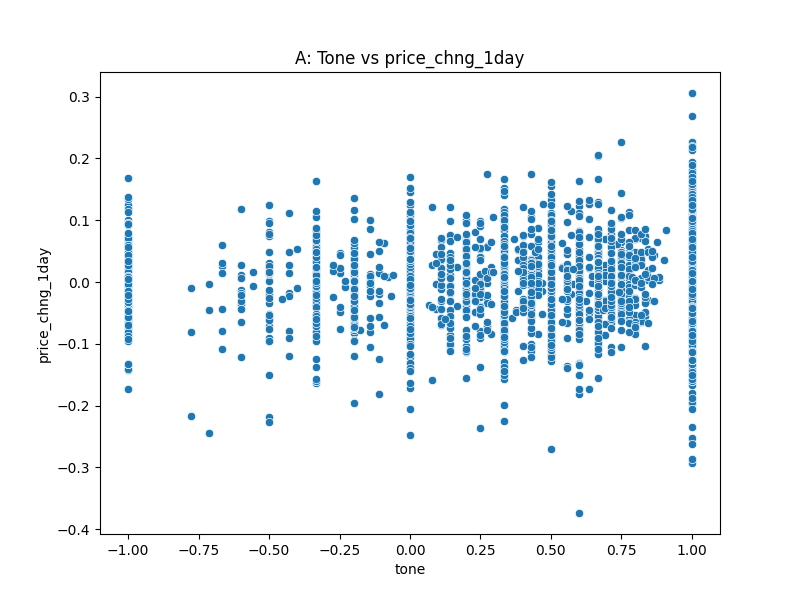
Fig xxx Questions

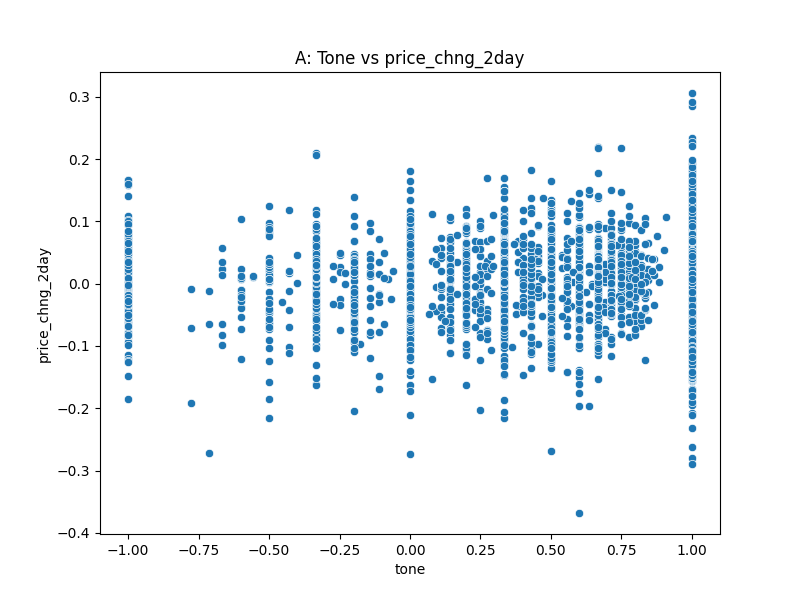
Scatter plot of Tone Vs Price Change 5day

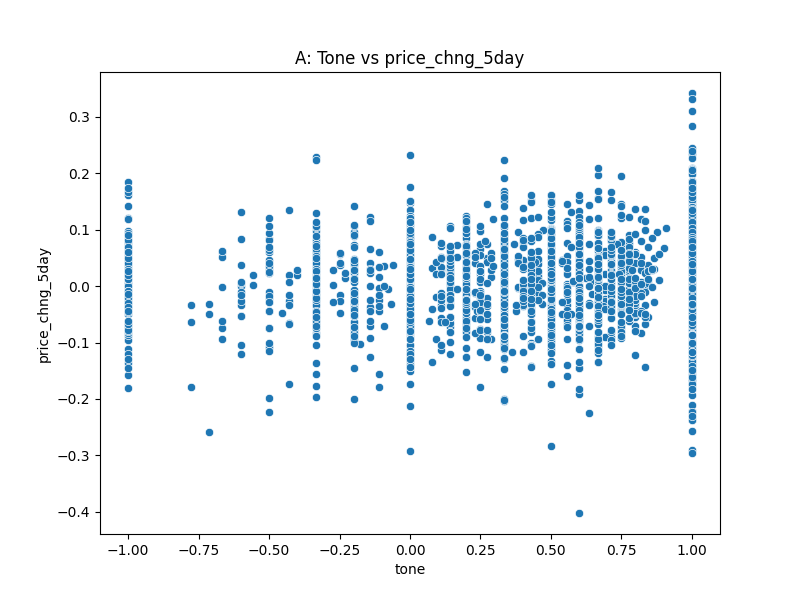




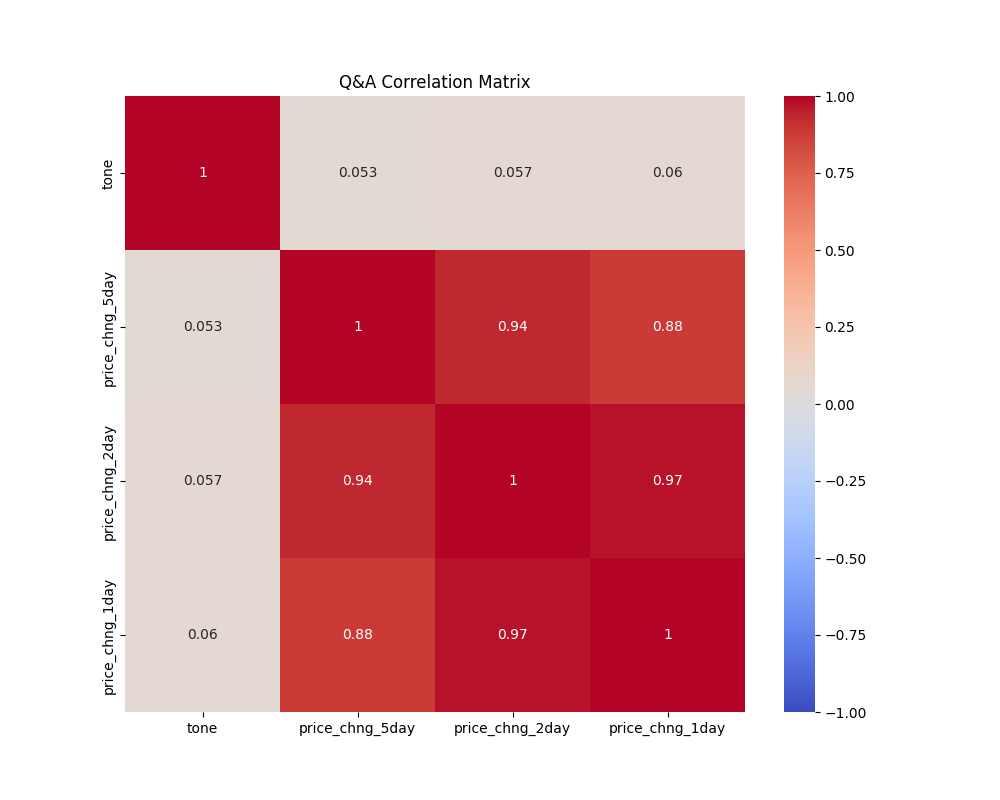


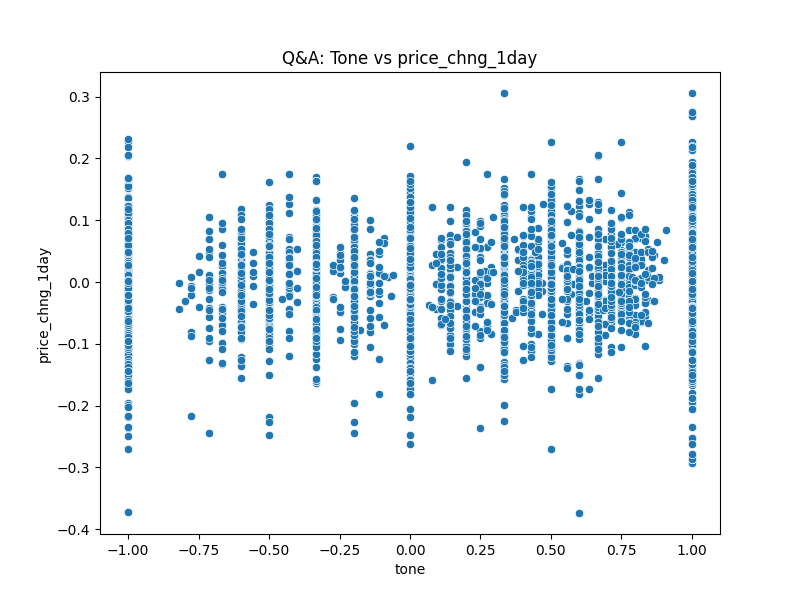


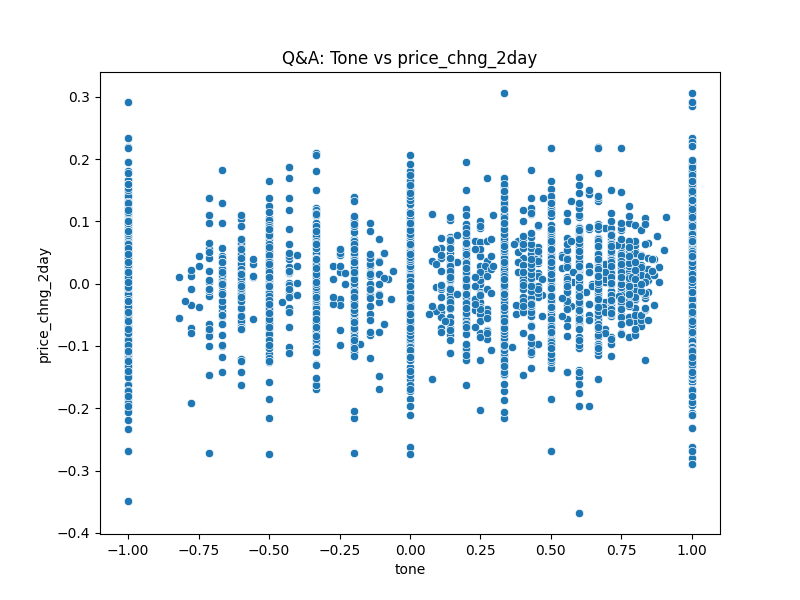


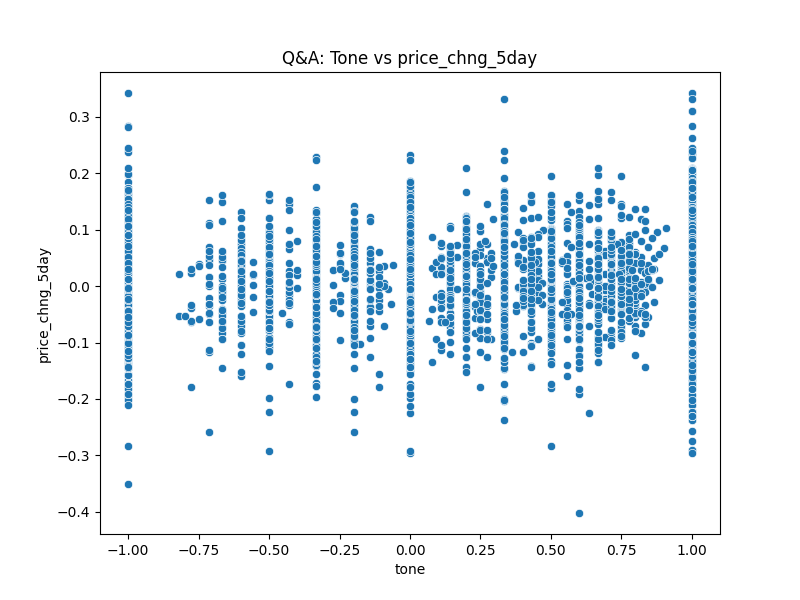












References.

Araci, D. (2019) ‘FinBERT: Financial Sentiment Analysis with Pre-trained Language Models’. Available at: https://doi.org/10.48550/ARXIV.1908.10063.

Brockman, P., Li, X. and Price, S.M. (2015) ‘Differences in Conference Call Tones: Managers vs. Analysts’, *Financial Analysts Journal*, 71(4), pp. 24–42. Available at: https://doi.org/10.2469/faj.v71.n4.1.

Bv, N. *et al.* (2023) ‘Deploying NLP Techniques for Earnings Call Transcripts for Financial Analysis: A Reverse Phenomenon Paradigm’, in *2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. *2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Kirtipur, Nepal: IEEE, pp. 368–375. Available at: https://doi.org/10.1109/I-SMAC58438.2023.10290494.

invetsopedia, 2023. S&P 500: The Index You Need to Know https://www.investopedia.com/articles/investing/090414/sp-500-index-you-need-know.asp#:~:text=The%20index%20includes%20500%20of,are%20selected%20by%20a%20committee.

Kimbrough, M.D. (2005) ‘The Effect of Conference Calls on Analyst and Market Underreaction to Earnings Announcements’, *The Accounting Review*, 80(1), pp. 189–219. Available at: https://doi.org/10.2308/accr.2005.80.1.189.

Medya, S. *et al.* (2022) ‘An Exploratory Study of Stock Price Movements from Earnings Calls’, in *Companion Proceedings of the Web Conference 2022*. *WWW ’22: The ACM Web Conference 2022*, Virtual Event, Lyon France: ACM, pp. 20–31. Available at: https://doi.org/10.1145/3487553.3524205.

Price, S.M. *et al.* (2012) ‘Earnings conference calls and stock returns: The incremental informativeness of textual tone’, *Journal of Banking & Finance*, 36(4), pp. 992–1011. Available at: https://doi.org/10.1016/j.jbankfin.2011.10.013.