

ACKNOWLEDGEMENT

We would like to express our deep gratitude to all our friends ,seniors who helped us to successfully complete this project on “Relationship between Employee Sentiment and Stock Prices”. We duly appreciate the efforts made in completing the project.

We specially thank Professor Dr Ashish Kumar Tripathi and TA Tapas Gupta for their mentorship and continuous guidance in the completion of the project. Their commendable knowledge in the field of Machine learning and Finance helped us to achieve the aim of this project .

We are also thankful to our fellow classmates for their efforts ,we benefited from discussion with them regarding methodology and overall flow of the project .

We would like to give special thanks to MNIT Jaipur for providing us essential glassdoor research papers which at last formed the base of our project .We would also like to give thanks to IIT ROORKEE for providing a conducive environment for growth.

Kaustubh Dwivedi (22117066)
Samay Jain (22117124)

Department of Mechanical and Industrial Engineering
B. Tech. 3rd Year
Indian Institute of Technology
Roorkee (Uttarakhand)

Table of Contents

1. Introduction.....	04-05
2. Literature Review.....	06-07
3. Data Collection and Preprocessing.....	08-12
3.1 Data Overview	08
3.2 Data Collection.....	08-10
3.3 Feature Engineering	10-11
3.4 Tools and Libraries Used.....	11-12
4. Proposed Methodology.....	13-18
4.1 Analysis of Daily Data.....	13-14
4.2 Clusters and Insights.....	15-18
5. Analysis of Ordinary Linear Regression.....	19-23
5.1 Variable Description and Dataset.....	19
5.2 Statistical Results of Model.....	19-21
5.3 Analysis and Graphs	22-23
6. Application of Auto Regressive Model.....	24-30
6.1 Introduction.....	24
6.2 Dickey_Fuller Test and Unit Root.....	25
6.3 Auto Correlation Function and its Plot.....	25-27
6.4 Results and Statistical Measures.....	27-30
7. Application of Granger Causality.....	31-33
8. Application of LSTM Model.....	34-37
8.1 Introduction.....	34
8.2 Hyper parameter Tuning.....	34-35
8.3 Result and Related Plots.....	35-36
8.4 Conclusion and Limitations.....	36-37

9. Application of Machine Learning Algorithms.....	38-41
9.1 Support Vector Machines.....	38-39
9.2 Random Forest Method.....	39-40
9.3 Xgboost Method.....	41
10. Conclusions.....	42
10.1 Flowchart.....	43
11. Future Work.....	44-45
11. Future Improvisation	46

END

Introduction

In the dynamic landscape of corporate evaluation, employee sentiment and organizational financial performance have become one of the most intriguing research topics. As traditional metrics for corporate success increasingly give way to more nuanced understandings of organizational health, platforms like Glassdoor have revolutionized how workplace dynamics are perceived and analyzed. This thesis aims to critically investigate the relationship between reviews of employee sentiment and stock market performance, particularly for industries where employee sentiment is a key determinant of an organization's stock prices.

The proliferation of online review platforms has fundamentally transformed how various stakeholders—investors, potential employees, and corporate leadership—gain insight into an organization's culture and performance. Prior research has shown that organizational culture and employee engagement are emerging as indicators of a company's long-term success (Edmans, 2011; Guiso et al., 2015). This study systematically examines the association between employee sentiment and organizations' stock market performance to identify sectors where internal perceptions directly influence external financial outcomes.

The objective of this research is to explore how the effect of employee sentiment varies across industries. The research challenges the assumption of an equalizing effect across sectors and instead employs advanced analytical techniques to map the nuanced ways employee experiences reflect on market performance. This approach will identify which sectors are most sensitive to employee sentiment and how this internal feedback translates into stock market valuations. For example, recent studies suggest that industries with a heavy reliance on human capital, such as technology or healthcare, may exhibit a stronger correlation between employee sentiment and financial outcomes (Li et al., 2020; Wang et al., 2022).

The importance of this study lies in its potential to demonstrate how internal organizational health, reflected through employee sentiment, correlates with or even predicts external financial performance. Traditional financial analysis has focused on quantitative metrics such as revenue, profit margins, and market share, often neglecting human capital dimensions. Emerging research, however, points to employee satisfaction, engagement, and organizational culture as significant leading indicators of corporate success (Huang et al., 2019; Edmans, 2011).

Through a comprehensive analysis of Glassdoor employee reviews and corresponding stock market data, this thesis addresses several critical research questions:

- Do employee reviews consistently correlate with stock performance across industries?
- Are there specific sectors where employee sentiment serves as a more reliable predictor of market performance?
- What underlying mechanisms explain variations in this relationship across industrial contexts?

Using rigorous statistical methodologies and advanced data analysis techniques, this study aims to contribute to the growing body of knowledge intersecting human resources, corporate culture, and financial performance. The findings could have implications for investors, corporate leaders, and researchers seeking to understand the interplay between employee experience and organizational success.

The methodology leverages data from Glassdoor reviews, analyzing key dimensions such as overall satisfaction, leadership effectiveness, pay, career prospects, and work-life balance. This qualitative data will be systematically mapped against quantitative stock market performance to uncover meaningful patterns and cross-industry correlations.

References

1. Edmans, A. (2011). *Does the stock market fully value intangibles? Employee satisfaction and equity prices*. Journal of Financial Economics, 101(3), 621–640.
2. Guiso, L., Sapienza, P., & Zingales, L. (2015). *The value of corporate culture*. Journal of Financial Economics, 117(1), 60–76.
3. Huang, L., Krishnan, V., & Viswanathan, S. (2019). *The role of employee sentiment in predicting organizational performance*. Management Science, 65(8), 3655–3672.
4. Li, H., Lu, L., & Zhang, Q. (2020). *Employee satisfaction and firm value: Evidence from cross-industry analysis*. Review of Corporate Finance Studies, 9(2), 145–172.
5. Wang, Y., Brown, J., & Smith, R. (2022). *Human capital in the digital age: The link between employee reviews and firm performance*. Strategic Management Journal, 43(3), 567–589.

Literature Review

Various theoretical frameworks have investigated the study of employee sentiment impacting organizational performance. The Resource-Based View (RBV) emphasizes human capital as one of the most significant and critical organizational resources and indicates that employees are valued as unique resources for the firm. Stakeholder theory views internal stakeholders, including employees, as drivers of success in organizations. Insights drawn from organizational behavior place the significance of employee sentiment at center stage as a performance indicator since workforce morale and attitudes will determine overall productivity and monetary outcomes.

A growing body of research supports the relationship between employee sentiment and financial performance. For example, Edmans (2011) shows that companies on the "Best Companies to Work For" list systematically outperform the market by 2.3-3.8% per annum. Luthans and Youssef (2007) establish a positive correlation between organizational behavior and financial outcomes, and Huselid (1995) reveals the direct impact of high-performance work practices on firm performance. Another area of study, in which the influence of digital platforms and sentiment analysis is important, is: Datta et al. (2015) showed that online employee reviews substantially impact corporate reputation, and Stieglitz and Dang-Xuan (2013) underlined how emotional content influences information diffusion. Soo et al. (2020) demonstrated that Glassdoor ratings are powerful predictors of organizational performance metrics.

In light of the psychological underpinning of employee sentiment, further elucidation has also been made on its association with performance. According to Judge et al. (2001), job satisfaction would indeed affect workplace productivity. With psychological capital, favorable organizational outcomes are connected by Wright et al. (2005), and Becker and Huselid (2006) emphasize the need for strategic human resource management as a driver of organizational success.

Methodological approaches in this domain have included quantitative analysis of secondary data, longitudinal tracking of organizational performance, multi-dimensional sentiment assessments, and advanced machine learning techniques for sentiment analysis. Despite these advancements, critical gaps remain. There is limited cross-industry comparative analysis and insufficient exploration of the mechanisms linking sentiment to performance. Furthermore, comprehensive predictive models that effectively integrate employee feedback are yet to be fully developed.

This stream is emerging towards the prediction of performance in real time. Real-time performance prediction models and sophisticated machine learning-based advanced sentiment analysis, and it even opens up interdisciplinary studies in collaboration with

management psychology, economics, and related knowledge systems. The Edmans' work (2011), Huselid (1995), Datta et al. (2015), Judge et al. (2001), and Stieglitz and Dang-Xuan (2013) are of utmost value in this burgeoning research, and these pieces offer solid ground for research.

Data Collection and Preprocessing

Data Overview:

This study employs employee reviews gathered from Glassdoor and uses a wide range of variables to add depth to the analysis. The variables employed include financial variables such as P/E ratio, exchange rates, trading volume, and percentage change, geopolitical variables such as Brent crude prices, oil prices, and gold prices, and global event volatility for a complete and balanced perspective.

Data Collection :

Employee review data for the period from 2019 to mid-2023 was sourced from Kaggle. The dataset, titled *Glassdoor Reviews*, is extensive, comprising over 8 billion observations and covering major companies listed on key global stock market indices such as the NYSE, NSE, and LSE. It includes various attributes such as overall ratings provided by employees and textual reviews. Notably, the dataset features two specific columns labeled *Pros* and *Cons*, along with ratings (on a scale of 1 to 5) assessing work-life balance, company culture, and overall satisfaction.


# rating rating (out of 5)	Δ title title of review	Δ status status of reviewer	Δ pros pros	Δ cons cons
	Good 2% Great place to work 1% Other (9593288) 97%	Current Employee 24% Former Employee 16% Other (5955686) 60%	8399149 unique values	8428753 unique values
5.0	Good	Current Employee, more than 10 years	Knowledge gain of complete project	Financial growth and personal growth
4.0	Good	Former Employee, less than 1 year	Good work,good work , flexible, support	Good,work, flexible,good support, good team work
4.0	Supervising the manufacturing the processes, ensuring quality work is done in a safe efficient man...	Current Employee, more than 1 year	This company is a best opportunity for me to learnings core mechanical field.	Monthly Target work,Maintain production scheduled, Evaluate production efficiency.
1.0	terrible	Current Employee, more than 1 year	I wish there were some to list	too many to list here

Figure 1

Figure : Glassdoor review Dataset extracted from Kaggle

Link of the dataset : [Glassdoor Job Reviews 2](#)

Financial variables related to company : Accounting for financial volatility entails that the selected variables exhibit day-to-day changes, which account for the dynamic nature of volatility. The approach is considered robust because all aspects regarding performance on the stock market are adequately portrayed by this set of variables. Such variables as the Price-to-Earnings ratio, exchange rate, stock trading volumes, percentage change in stock prices, and the VIX index are considered relevant enough for this study.

PE-Ratio : The price–earnings ratio, also known as P/E ratio, P/E, or PER, is the ratio of a company's share price to the company's earnings per share. The ratio is used for valuing companies and to find out whether they are overvalued or undervalued .

Exchange rate :An exchange rate is the value of one currency in terms of another. When considering companies listed in a specific country, the exchange rate would indicate how much of the local currency (of the company's country) is equivalent to one US dollar.

Volume : Volume tells the number of stock of a company traded during a day .

Percent -Change : it is the percentage change of a company closing price and opening price of that day .

VIX Index : The VIX Index, also known as the Volatility Index or the Fear Gauge, is a measure of market expectations for near-term volatility .

These data have been obtained from the yahoo finance library .

Link of yahoo finance : [yahoo finance](#)

Economic Variable : To account for market conditions, geopolitical tensions, and global trade dynamics, variables such as oil prices, Brent crude prices, gold prices, and Hedonometer readings were utilized.

Oil Prices : Dataset has components named as Cushing OK WTI Spot Price FOB \$/bbl

Source : Data source: U.S. Energy Information Administration

Gold Prices :

Source: Bloomberg, Datastream, ICE Benchmark Administration, World Gold Council

Brent oil : It represents crude oil extracted from the North Sea, specifically from fields near Scotland. Brent oil is widely used as a pricing standard for global oil markets, especially in Europe, Africa, and the Middle East.

Source : Data source: U.S. Energy Information Administration

Hedonometer : Hedonometer is a measure of a population's happiness /

Source : [Hedonometer](#)

Feature engineering :

In this section, all datasets related to Glassdoor reviews, financial metrics, and economic variables were merged into a single DataFrame. The following operations were then performed

1.Sentiment Scoring: VADER sentiment analysis was applied on the Pros and Cons columns for calculating sentiment scores. Various mathematical operations were then tested, such as combining overall rating, pros_sentiment and cons_sentiment. Final formula used in the computation was:

Sentiment=(pros_sentiment*overall_rating+cons_sentiment*(5-overall_rating))/5

2. Dimensionality Reduction: Principal Component Analysis (PCA) was applied to the economic variables to consolidate them into a single factor, minimizing multicollinearity and redundancy in the dataset.

3. Generated Lagged Variables: The variable pct_change was lagged because every measure of this variable depends on its past values for every ticker.
4. Data Aggregation: Each ticker had more than one entry for a given date. This was addressed by averaging out all entries for that date.
5. Data Cleansing and Standardization: The data has been checked for null values, then subjected to standardization, the purpose of which is for preparing data for analysis purposes.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	ticker	date	rating	sentiment	lag_pct_ch	Exchange	pct_change	pe_ratio	oil_prices	Volume					
0	AAL	8/26/2020	0.318254	1.349883	-0.03561	-0.01259	-0.13888	-1.37693	3.359493	2.383988					
1	AAL	10/1/2020	0.513072	0.332721	-0.13886	-0.01259	-0.10278	-1.35814	-0.70348	2.685615					
2	AAL	10/2/2020	1.097525	0.837026	-0.10276	-0.01259	-0.06044	-1.39573	0.607741	6.659311					
3	AAL	10/5/2020	0.448132	0.564324	-0.06043	-0.01259	-0.08184	-1.40647	1.561571	2.977049					
4	AAL	10/6/2020	-1.35161	-1.88389	-0.08183	-0.01259	-0.00521	-1.35366	1.254865	4.980679					
5	AAL	10/7/2020	0.707889	0.618888	-0.00521	-0.01259	-0.09549	-1.402	1.394589	3.469256					
6	AAL	10/8/2020	0.756594	1.061694	-0.09548	-0.01259	-0.12714	-1.41005	2.669765	5.207555					
7	AAL	10/9/2020	-0.85065	-0.15516	-0.12712	-0.01259	0.030535	-1.41364	2.348026	3.721818					
8	AAL	10/12/2020	0.9862	0.531439	0.030524	-0.01259	0.015238	-1.38857	2.179665	2.118978					
9	AAL	10/13/2020	0.318254	0.376611	0.01523	-0.01259	-0.17852	-1.32591	1.689149	3.6015					
10	AAL	10/14/2020	0.396181	0.340878	-0.17849	-0.01259	-0.14911	-1.33844	2.087147	1.970692					
11	AAL	10/15/2020	-0.85065	-1.48234	-0.14908	-0.01259	-0.07606	-1.3268	2.598538	1.735963					
12	AAL	10/16/2020	-0.07138	-0.14404	-0.07606	-0.01259	-0.1004	-1.34739	2.287285	1.669223					
13	AAL	10/19/2020	0.318254	-0.11563	-0.10038	-0.01259	-0.33183	-1.35634	2.704478	2.79395					
14	AAL	10/20/2020	-0.85065	-0.41073	-0.33178	-0.01259	-0.20787	-1.37783	4.056529	2.553674					
15	AAL	10/21/2020	0.61048	0.000603	-0.20784	-0.01259	-0.21093	-1.37335	3.991664	1.970919					
16	AAL	10/22/2020	-0.61687	-0.70545	-0.2109	-0.01259	-0.02349	-1.40916	2.263888	5.329663					
17	AAL	10/23/2020	-0.2662	1.067344	-0.02349	-0.01259	-0.19576	-1.35993	1.077891	5.3296					
18	AAL	10/26/2020	0.48524	0.497369	-0.19573	-0.01259	-0.18949	-1.28831	0.539797	4.986806					
19	AAL	10/27/2020	0.61048	-0.24936	-0.18946	-0.01259	-0.16249	-1.23729	1.079788	3.754746					
20	AAL	10/28/2020	0.907707	1.172368	-0.16247	-0.01259	-0.46705	-1.21222	0.944331	3.666273					

Figure 2

Link of the dataset : [Dataset on daily basis](#)

Tools and Libraries Used

For finding a relationship between employee sentiment and companies stock percent change we have used following tools and libraries:

1. Pandas : Pandas is a Python package for data analysis and manipulation. It's an open-source library, very fast and efficient.
2. Matplotlib :Matplotlib is an open-source plotting library for Python. It offers a wide variety of functions to create static, interactive, and animated visualizations in Python.
3. Tensorflow:It provides a comprehensive ecosystem of tools, libraries, and community resources to help researchers and developers build and deploy machine learning applications efficiently.

4. Keras :Keras is an open-source software library that is designed to make it easier to build, train, and deploy deep learning models.
5. Scikit-learn:scikit-learn abbreviated as Sklearn, is Python's open source machine library for learning over large amounts of data or for perception.

Proposed Methodology :

Analysis of Daily Data

The initial dataset was generated on a daily basis, and various algorithms were applied, ranging from Ordinary Least Squares (OLS) to time series models and machine learning models. However, after careful consideration, working with daily data was ultimately avoided due to the following reasons:

Lack of Captured Relationships:

- The daily data analysis did not reveal any significant relationships between the sentiment variable and the percentage change in stock prices. This suggests that the daily fluctuations may be too volatile and subject to random noise, obscuring any underlying patterns that could be identified.

R-squared Values:

- The R-squared values obtained from models using daily data consistently fell below our assumed threshold of 0.7, indicating a weak fit for many companies. An R-squared value below this threshold signifies that a considerable proportion of the variance in percentage change could not be explained by the sentiment variable. This raised concerns about the reliability of our models and their predictive capabilities.
- Only 16 companies are there with $R^2 \geq 0.70$.

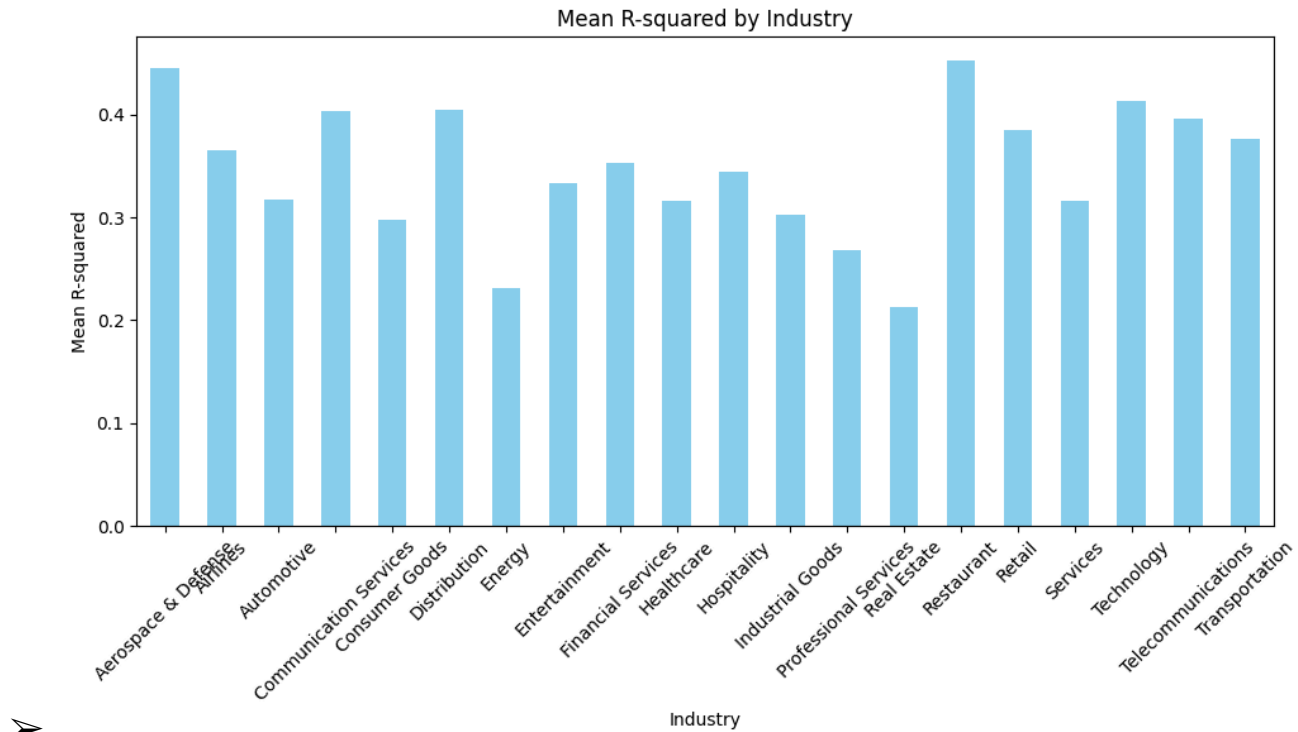


Figure 3

Statistical Measures:

- The log-likelihood was coming to be in the range of negative of several thousands. So it tells a very bad fit.
- Further examination of the statistical measures such as p-values, t-statistics, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) indicated statistical insignificance across many variables. These metrics are critical for evaluating the strength and validity of our model findings. When these values are insignificant, it implies that the relationships being tested do not hold statistical power, leading to doubts about the conclusions we could draw from the analysis.

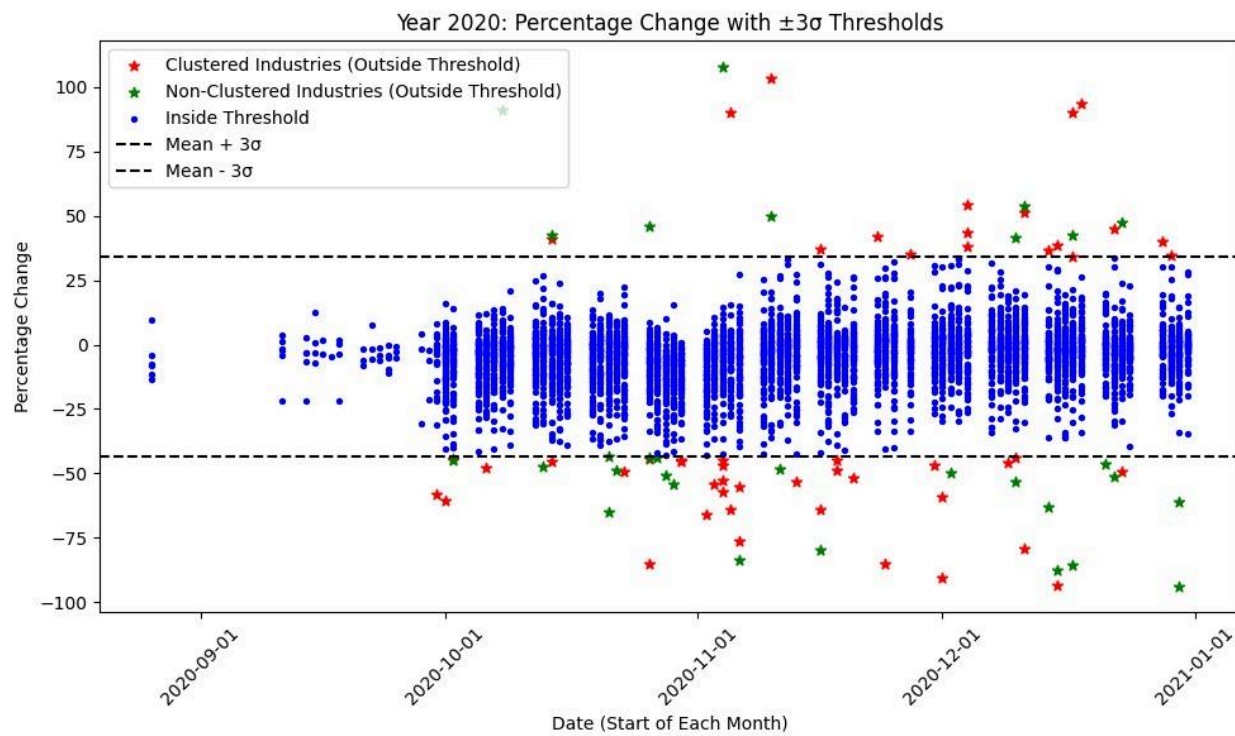
Clusters and Insights:

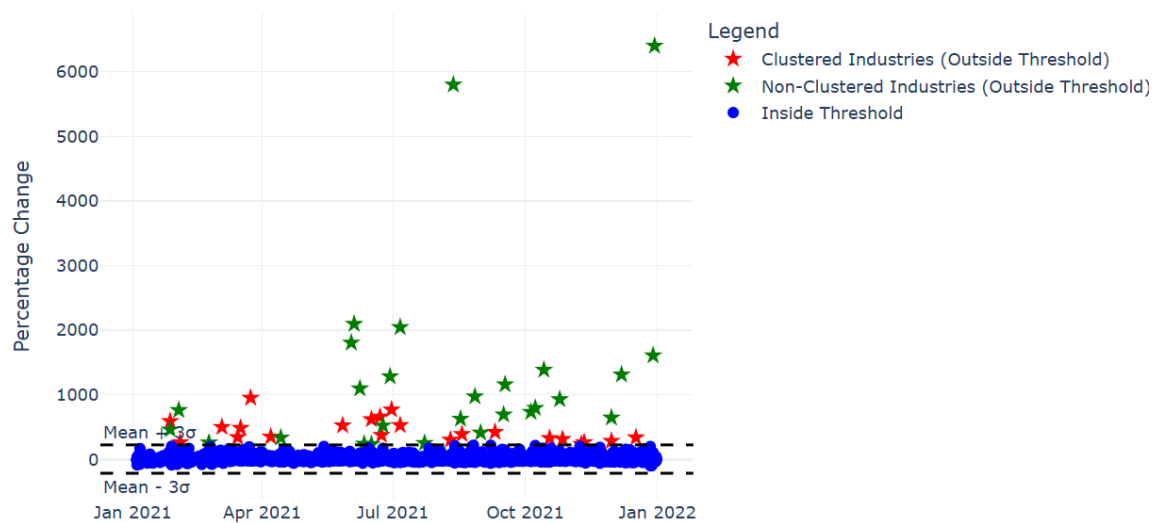
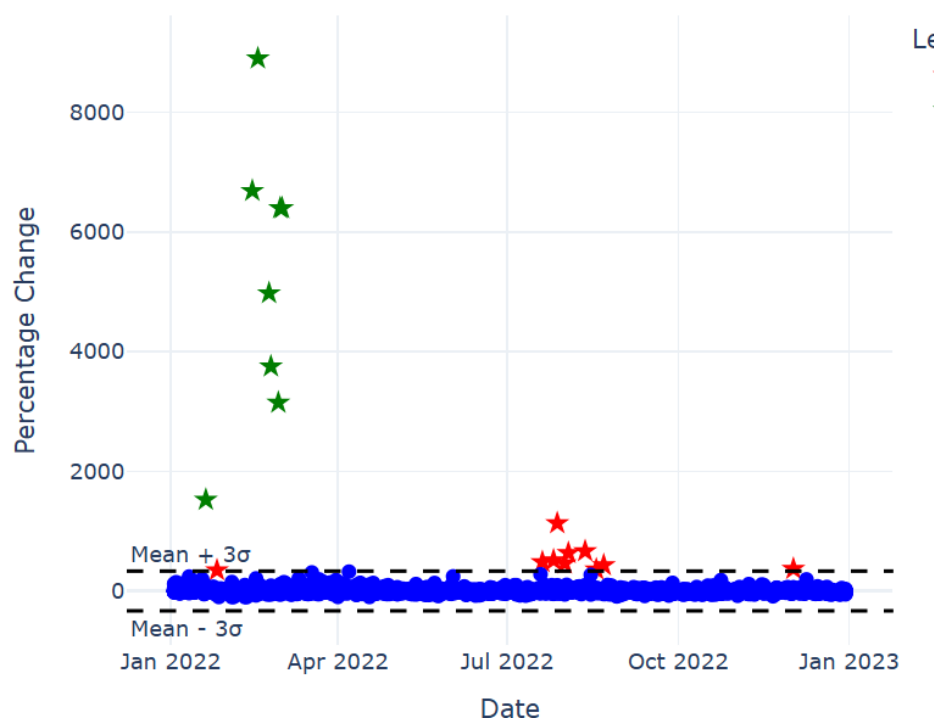
- The daily analysis also failed to produce meaningful clusters of companies, which are essential for targeted analysis and decision-making. Without identifiable groups, we could not derive actionable insights or stratify our analysis based on company performance or behavior.

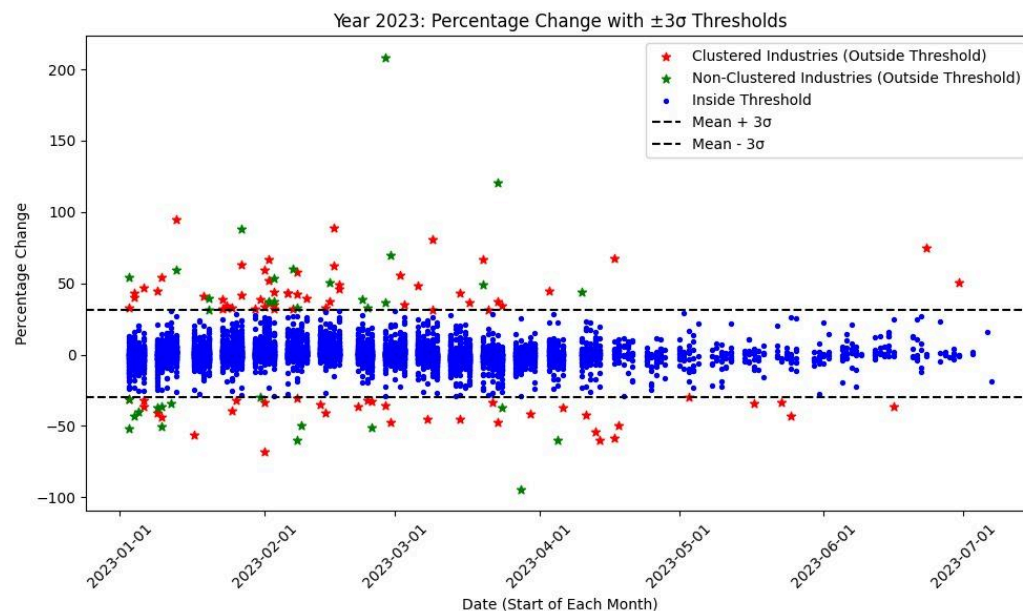
Transition to monthly analysis : In this the approach used was to club data onto a monthly basis by taking average and then performing various algorithms to analyze results .

Analysis of daily data : Here a methodology is applied where top positive and negative values of sentiment were shortlisted and then corresponding stock pct_change was plotted against the date in which sentiment was analyzed .

After that a threshold line is drawn with Lower bound as $\mu - 3\sigma$ and upper bound as $\mu + 3\sigma$ and then identifying those points lying outside these thresholds .



Year 2021: Percentage Change with $\pm 3\sigma$ ThresholdsYear 2022: Percentage Change with $\pm 3\sigma$ Thresholds



summary_df

	Year	Clustered Industries (Outside Threshold)	Non-Clustered Industries (Outside Threshold)
0	2020	51	30
1	2021	22	27
2	2022	10	8
3	2023	86	34

Figure 4 Figure 5 Figure 6 Figure 7 Figure 8

This analysis emphasises that employee sentiment gives better results on result clustered industries rather than all industries .

Analysis of Ordinary Linear Regression:

Equation for OLS Algorithm :

$$Y = \beta_0 + \beta_1(\text{sentiment_sum_1}) + \beta_2(\text{rating}) + \beta_3(\text{Volume}) + \beta_4(\text{oil_prices}) + \beta_5(\text{pe_ratio}) + \beta_6(\text{lag_pct_change}) + \beta_7(\text{ExchangeRate}) + \epsilon$$

```
dependent_var = 'pct_change'
independent_vars = ['sentiment_sum_1', 'rating', 'Volume', 'oil_prices',
                    'pe_ratio', 'lag_pct_change', 'ExchangeRate']
```

https://drive.google.com/file/d/1mkZ9NYwDTSKm0Y9ovfvbuZu-BQCrT8xi/view?usp=sharing-detailed_results_csv

Statistical Results of Model

i) R squared: This statistic indicates that approximately what % of the variance in the dependent variable (in this case, **pct_change**) can be explained by the independent variables in the model. A higher R-squared value indicates a better fit of the model.

Industry	Count
Technology+Financial Services+Retail	36
Rest industry	18

ii) F-statistic: The F-statistic tests the overall significance of the model. It compares the model with no predictors to the specified model. A higher F-statistic indicates that at least one of the predictors is significantly related to the dependent variable.

->we will calculate the F-statistic overall as the average of all companies as well as for those for which sentiment affects the most .

->calculate the F-statistic average of all companies, average of top 40 % companies and also for companies who belong some special industries like Technology Financial services ,Retail - as sentiment affects them the most.

Average F-statistic of all companies: 7.066263382874459

Average F-statistic of the top 40% companies: 11.515498775518127

Average F-statistic of special industries: 7.887039766312191

(Here we have done analysis separately for all companies , top 40% companies and special industries (which will form our cluster lately) in order to conclude that sentiment does not affect all sectors of companies significantly, it only affects some identified sectors.)

The F statistic for our cluster is less than the top 40 % of the companies which shows it is more significant to choose our cluster over other industries as far as sentiment variable is considered.

iii)Log-likelihood: This value is used in the context of maximum likelihood estimation. It indicates how likely it is to obtain the observed data given the model parameters. The closer this value is to zero, the better the model fits the data.

Average Log-likelihood of all companies: -72.24130052937495

Average Log-likelihood of the top 40% companies: -66.46853964075643

Average Log-likelihood of special industries: -68.34558562875252

It shows that the chosen cluster is performing better than other industries as far as the statistical value of Log-likelihood of the model is considered.

iv)AIC (Akaike Information Criterion) : AIC is used for model selection. It penalizes the number of parameters in the model to discourage overfitting. Lower AIC values suggest a better-fitting model when comparing multiple models.

Average AIC of all companies: 158.64520268476616

Average AIC of the top 40% companies: 147.1003445876353

Average AIC of special industries: 150.91339347972723

The Akaike Transformation criteria is lesser for the identified cluster so it can be evidently said that the identified cluster performs better than other industries as far as sentiment is concerned.

v)BIC (Bayesian Information Criterion): Similar to AIC, BIC also penalizes for the number of parameters but does so more heavily than AIC. It is another criterion for model selection. Lower BIC values indicate a better-fitting model.

Average BIC of all companies: 168.3333227732003

Average BIC of the top 40% companies: 156.5434253947891

Average BIC of special industries: 160.58195938323672

The Bayesian Transformation criteria is lesser for the identified cluster so it could be evidently said that the identified cluster performs better than other industries as far as sentiment is concerned.

vi) T-Stat: In statistics, a t-statistic is used to assess the statistical significance of a variable, typically in a hypothesis testing framework. The criteria for determining statistical significance depend on the chosen **confidence level** or **p-value threshold**. Commonly used thresholds are:

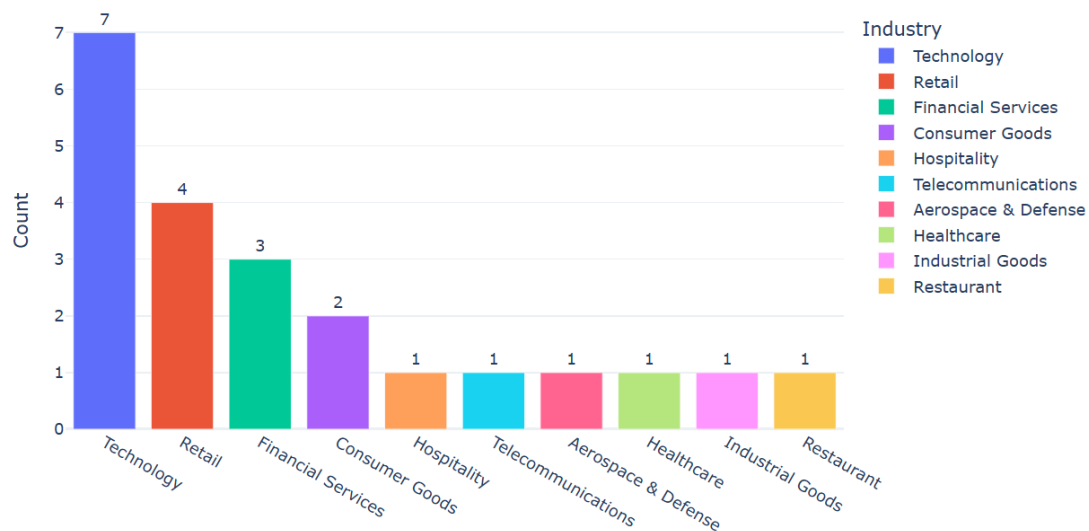
General Criteria for Significance

1. Two-Tailed Test:

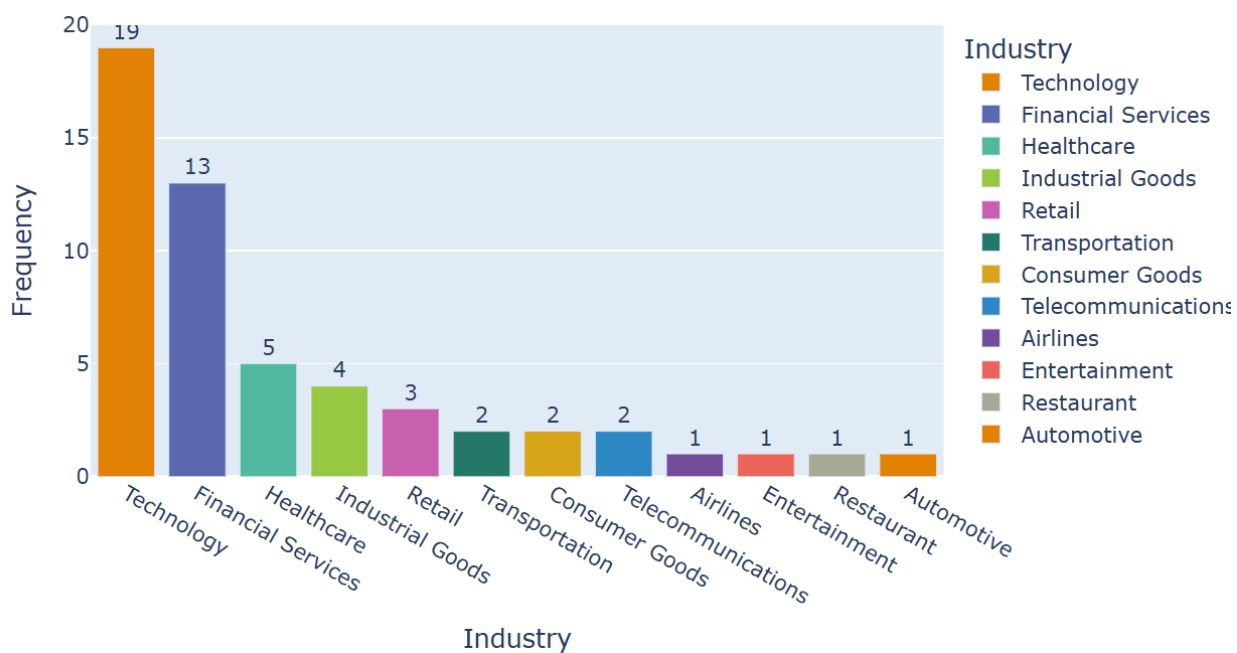
- At a **5% significance level ($\alpha = 0.05$)**:
 - The critical values are approximately **± 1.96** .
 - If the absolute value of the t-statistic (**$|t\text{-stat}|$**) is **greater than 1.96**, the result is statistically significant.

Graphs for Statistical Measures:

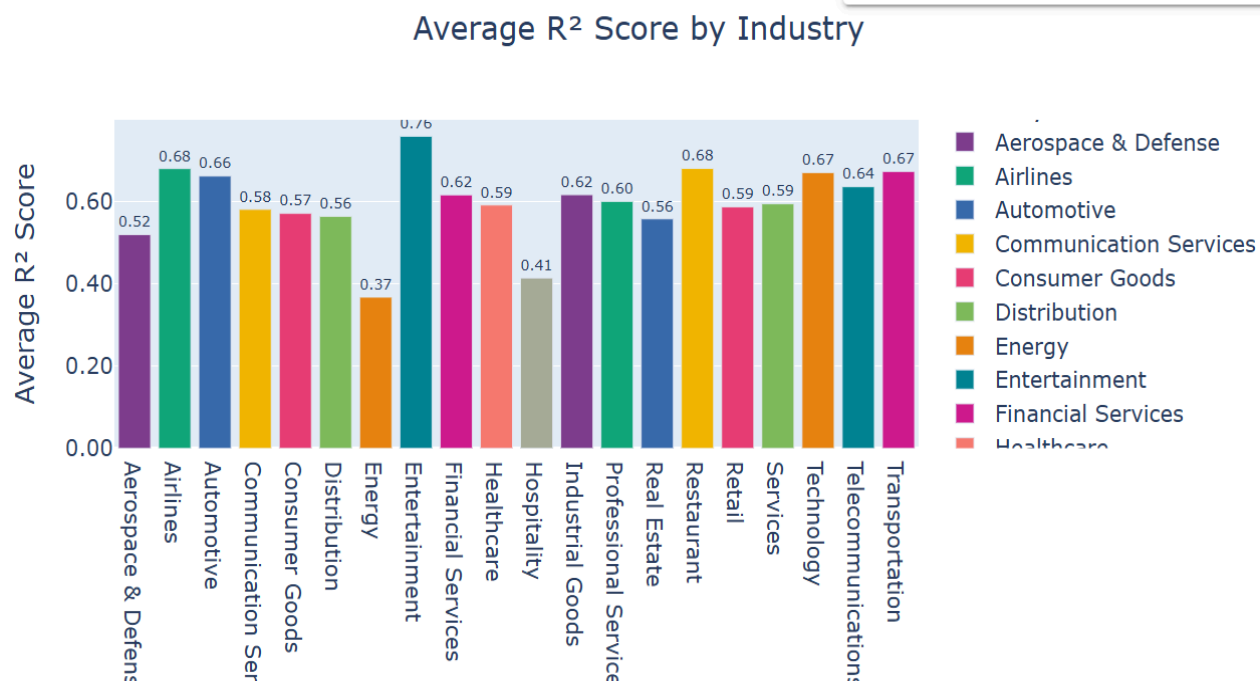
Count of companies with significant t-stat for sentiment: 22 out of which 15 belongs to our specified cluster .



a) Significance test



b) R^2 -square > 0.70



(c) average R²-square of Industries

Figure 9 (a),(b),(c)

Application of Auto Regressive Model

Introduction

The dataset is surely a time series dataset. It would be suggested to utilize the method of Auto-Regressive Distributed Lag (ARDL) and model the connections between lagged versions of `pct_change` variable, amongst others, which include the `sentiment` variable, to `pct_change`. This makes it suitable in order to capture both dynamics at shorter and longer levels of relation between variables concerned.

The ARDL model can be expressed in the following form:

$$Y_t = \alpha + \sum_{i=1}^p \beta_i Y_{t-i} + \sum_{j=0}^q \gamma_j X_{t-j} + \epsilon_t$$

Where:

- Y_t is the dependent variable at time t .
- X_t represents one or more independent variables.
- α is the constant term (intercept).
- β_i are the coefficients for the lagged values of the dependent variable Y .
- γ_j are the coefficients for the current and lagged values of the independent variable(s) X .
- p is the number of lagged terms for the dependent variable.
- q is the number of lagged terms for the independent variable(s).
- ϵ_t is the error term (white noise).



Figure 10

Assumptions assumed before applying ARDL :

- 1.Data set is assumed to be having properties of linearity,Independence ,No Multicollinearity, Homoscedasticity .
- 2.No integration is present between the variables .
- 3.Independent variables are chosen such that co-integration becomes negligible.

Dickey fuller Test->

The Dickey-Fuller test is a statistical test used to determine whether a time series is stationary or has a unit root, which indicates non-stationarity. Stationarity is an important property in time series analysis, as many statistical methods assume that the underlying data is stationary.

Unit Root:

A time series has a unit root if it shows a systematic pattern that is not predictable over time, often indicated by a stochastic trend.

Null Hypothesis (H0): —>The time series has a unit root (i.e., it is non-stationary).

Alternative Hypothesis (H1): —> The time series does not have a unit root (i.e., it is stationary)

Results of Dickey Fuller Test

```
Dickey-Fuller Statistic:      3.610877309415868
p-value: 0.005556449721806796
Critical Values: 1%:        3.6889256286443146
5%:    2.9719894897959187
10%:   2.6252957653061224
```

According to our data we are good to go on this data as it is stationary.

Auto correlations Function

The **Partial Autocorrelation Function (PACF)** is the measure of the correlation at different lags between the observations in the time series after accounting for the contributions from shorter lags. The task now is to estimate the PACF and plot this against the lags and identify at which lags are significant for the subsequent decisions on what lags to include. Another observed is the **Autocorrelation Function (ACF)**, which is

decreasing for most tickers. The ACF captures the relationship between lagged versions of the series and their own past values.

The ARDL model can be expressed in the following form:

$$Y_t = \alpha + \sum_{i=1}^p \beta_i Y_{t-i} + \sum_{j=0}^q \gamma_j X_{t-j} + \epsilon_t$$

Where:

- Y_t is the dependent variable at time t .
- X_t represents one or more independent variables.
- α is the constant term (intercept).
- β_i are the coefficients for the lagged values of the dependent variable Y .
- γ_j are the coefficients for the current and lagged values of the independent variable(s) X .
- p is the number of lagged terms for the dependent variable.
- q is the number of lagged terms for the independent variable(s).
- ϵ_t is the error term (white noise).

Figure 11

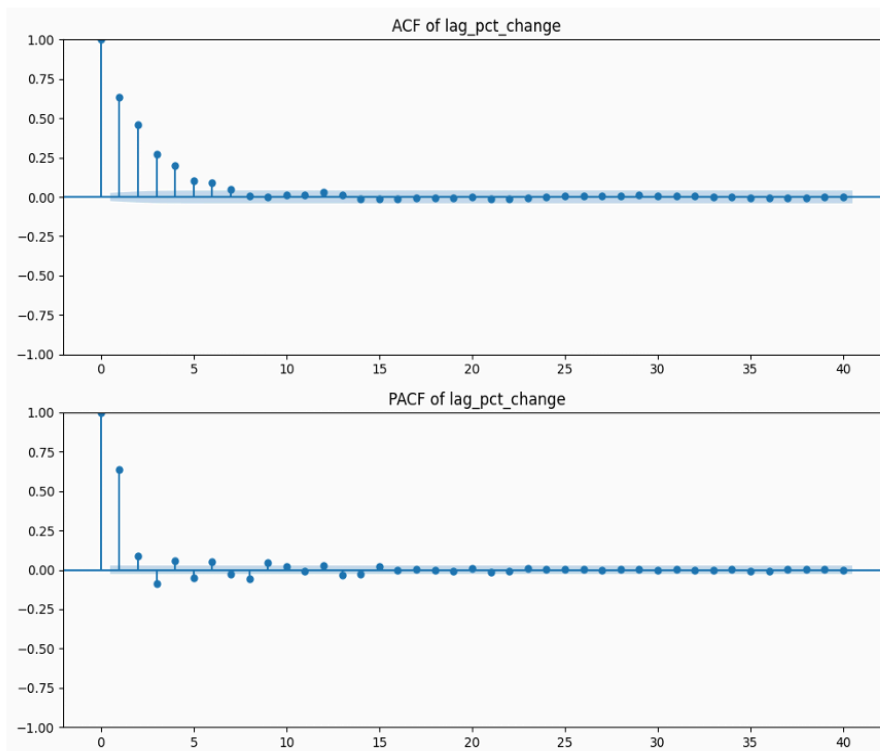


Figure 12

This graph is shown here for one particular ticker. The ACF plot evidently shows that the ACF is Decaying in nature which proves that it is good to use auto Regressive lagged models that is ARDL.

The **PACF plot** of each ticker determines the correct number of lags for the ARDL model. Significant lags are those that lie outside the blue zone of insignificance. Those lags are selected and used as parameters in training the ARDL model, such that only the most relevant lagged relationships are included.

Results and Statistical Measures

Number of companies with R-squared > 0.85 : 43

Number of companies with R-squared > 0.8 : 113

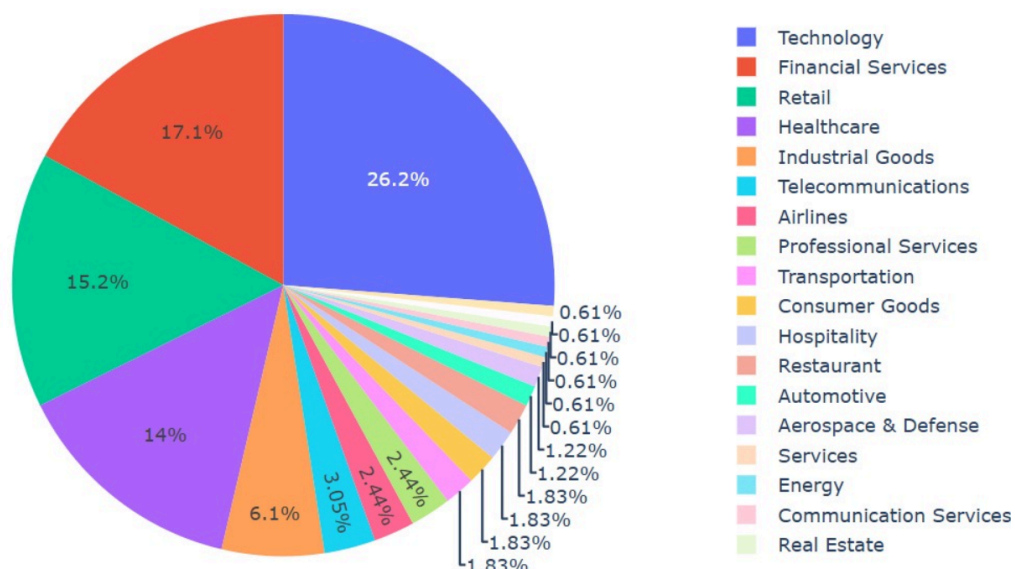
Number of companies with R-squared > 0.7 : 164

Below is the industry-wise plot where the significant sentiment effect on each industry is analyzed using the p-value hypothesis. The results are that the top-performing industries, based on sentiment impact, are **Financial Services, Retail, Healthcare, and Technology**. These industries have the most significant relationship between sentiment and their performance, which indicates the possible influence of sentiment on the stock performance within these sectors.

The PIE chart shows the percentage of companies from each industry having $R^2 \text{ square} \geq 0.70$.

Based on the pie chart below it is shown that the percentage of companies from the above mentioned industries constitute most of the thing.

Industry Distribution (ARDL R-Squared > 0.7)

**Figure 13**

The next objective is to determine the number of companies in each sector where the **Sentiment** variable has been important. This can be found by taking the p-value of the `sentiment` variable for each ticker within the model summary. The p-value being below 0.05 means that, for that particular company's performance in its sector, the `sentiment` variable is significant to the model..

H0_Null hypothesis ->The sentiment variable is not significant and in turn contains no relationship with the pct_change variable .

H1_Alternative Hypothesis ->The sentiment variable must not be neglected .It has a direct relationship with the pct_change variable.

If the p-value is more than 0.05, the null hypothesis is rejected as the sentiment variable has no effect. However, if the p-value is equal to or lesser than 0.05, then the null hypothesis is not rejected, meaning that the sentiment variable has an effect.

Based on this analysis, the model summary was inspected and the number of companies within each industry having a large sentiment variable plotted. The plot displays the distribution of companies in various industries for which the sentiment variable is major, allowing for the determination of those industries most affected by sentiment.

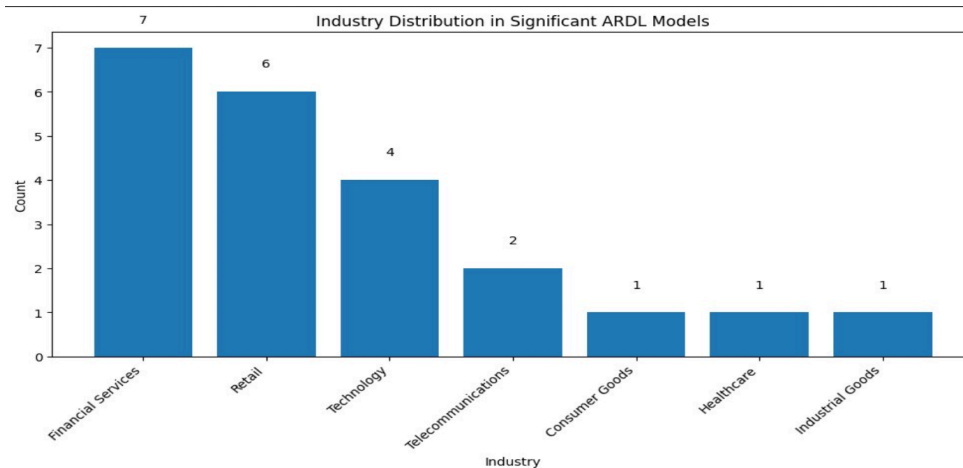


Figure 14

So it proves the fact that the identified cluster (Technology ,Financial Services ,Retail) is statistically performing better than other industries as far as sentiment variable is considered.

i) The F-statistic is often derived from the ratio of the model's explained variance to its unexplained variance.

Average F-statistic for all companies:5.6702687398

Average F-statistic for top 50% companies:8.310834

Average F-statistic for special industries

(Technology, Financial Services, Healthcare):

6.06390832274307

Above mentioned statistic proves evidently that our cluster is performing better than other industries .

ii) Log-likelihood :This value is used in the context of maximum likelihood estimation. It indicates how likely it is to obtain the observed data given the

model parameters. The closer this value is to zero, the better the model fits the data.

Average Log Likelihood for all companies: -71.6982454

Average Log Likelihood for top 50%

companies:-57.997996

Average Log Likelihood for special industries (

Technology , Financial Services, Healthcare

):-69.87830

It shows that the identified cluster is performing better than other industries as far as the statistical value of Log-likelihood of the model is considered.

iii))BIC (Bayesian Information Criterion) :Similar to AIC, BIC also penalizes for the number of parameters but does so more heavily than AIC. It is another criterion for model selection. Lower BIC values indicate a better-fitting model.

Average BIC for all companies: 176.89986294909207

Average BIC for top 50% companies: 149.3818999055978

Average BIC for special industries (Technology,
Financial Services, Healthcare): 173.2622118962776

The Bayesian Transformation criteria is lesser for the identified cluster so we can evidently say that choosing our cluster over other industries would be good if sentiment variable is considered.

Application of Granger causality and its analysis

Granger causality is a statistical hypothesis test used to determine whether one time series can predict another. The test is based on the idea that if variable X "Granger-causes" variable Y, then past values of X should contain information that helps predict future values of Y, over and above the information already contained in the past values of Y.

Key points about Granger causality:

1. It does not imply a true cause-and-effect relationship but rather a predictive relationship.
2. The test typically involves lagged values of the time series to see if they help in forecasting the other series.
3. If past values of X significantly improve the prediction of Y, then X is said to "Granger-cause" Y.

The process involves:

1. Regressing the dependent variable (Y) on its past values and the past values of the independent variable (X).
2. Testing whether the coefficients of the lagged values of X are statistically significant.

The null hypothesis for the test is that X does not Granger-cause Y. If the null hypothesis is rejected, it suggests that X contains information that helps predict Y.

Causality vs Correlation

There is a difference between Causality and correlation which could be understood by an example described below .

If the oil Prices increase it might cause stock prices to decrease because of the increased cost of companies reliant on oil. But Correlation simply indicates the relationship between one variable and another without describing whether one is dependent on the other .

After applying **Granger causality** on the dataset, it checks for the p-value to find out how many companies are showing a significant relationship with the **sentiment** variable. A company is said to be significant if the p-value is less than 0.1.

For rejecting the null hypothesis (the condition here is that **sentiment** doesn't influence **pct_change**) p-value<0.1 is demanded, suggesting a causal effect between companies with respect to this phenomenon **sentiment** , stock price percent change is shown to be impacted on **pct_change** in this respect for those companies; however, the number of firms possessing such a significant value is counted that helps to quantify how significant is that association.

the number of companies with granger causality-p value <0.1 is 16

Plot

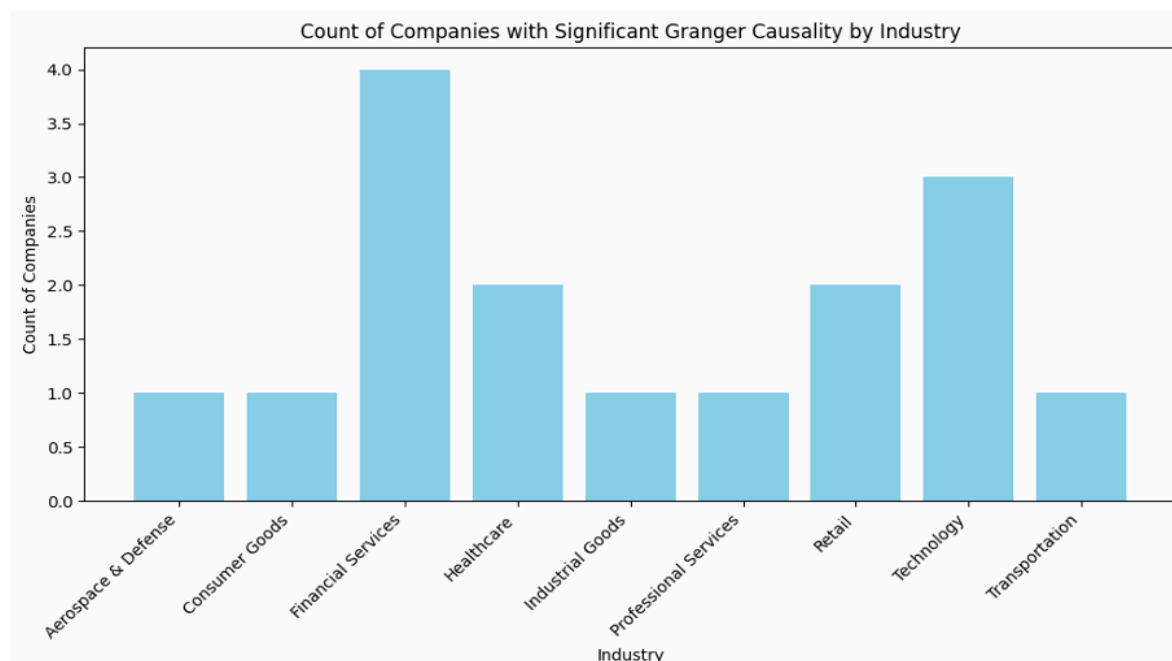


Figure 15

So we can see that the results correspond to maximum dependence in the special industries-(Technology , Financial services ,Retail).

Application of LSTM Model on data

Introduction :

LSTMs (Long Short-Term Memory networks) are a type of recurrent neural network (RNN) specifically designed to model and capture long-term dependencies in sequential data.

LSTM captures the non linearity in data which is now our aim and also it is very effective in modelling the time series data.

Unlike traditional RNNs, LSTMs use memory cells and gating mechanisms to retain important information over long sequences while mitigating the vanishing gradient problem, making them ideal for learning patterns in complex, temporal datasets.

We want to capture intricate relationships between the dependent_variable and independent variable and thus the LSTM model becomes a good option for modelling the sequential data .

We analyse the training error occurred while applying the LSTM model on monthly grouped data

.our purpose is to identify the change in training error occurred when taking sentiment into account and when dropping sentiment variable and we plot the stats regarding it.

Hyper-parameter tuning

Key Hyperparameters for LSTMs

1. Network Architecture

- NumberofLayers: Increasing layers can help capture higher-level patterns but may result in vanishing gradients.

50 layers

2. Training Parameters:

- Learning Rate: Controls the step size during optimization. Too high leads to divergence; too low slows convergence.

We got a 0.001 learning rate as optimum learning rate for most of the companies

3. Sequence and Input Settings:

- Sequence Length: ->

Number of time steps fed into the LSTM. Longer sequences capture more context but increase computation. Different for different companies so it depends on dataset availability.

Results and related plots :

When we take sentiment into account

The average error is 6.3215879

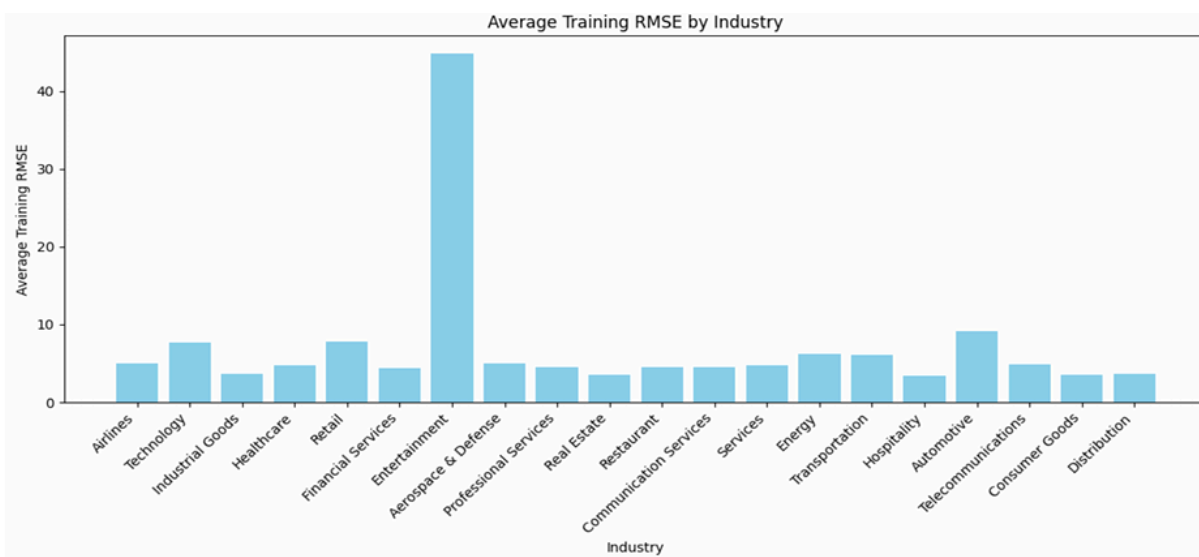


Figure 16

And when we drop this variable

The average error is 8.64789545

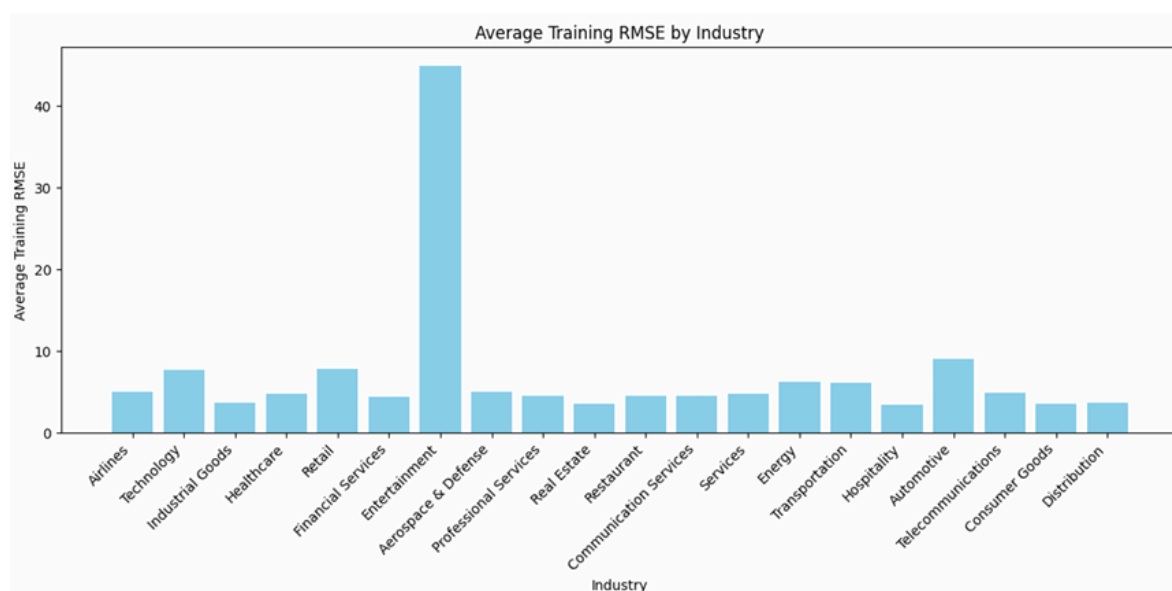


Figure 17

The difference between the rmse->

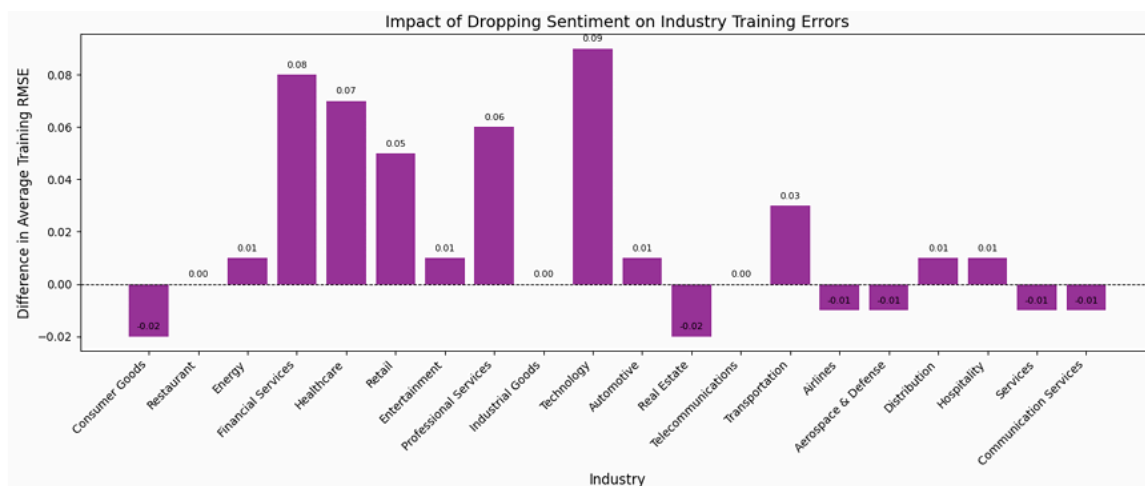


Figure 18

Conclusion and Limitations :

After doing the LSTM analysis, it was found that the cluster of Technology, Financial Services, and Retail industries has a better performance in terms of sentiment variable. This is because the Root Mean Squared Error (RMSE) becomes significantly low when the model includes sentiment. This indicates that including sentiment data enhances the model's predictive ability for these industries.

The LSTM model is no longer valid because the dataset size is not large enough to catch the variance of the dependent variable in relation to the sentiment variable. Therefore, the R-squared value for the LSTM model turns out to be negative, implying that the model is not delivering a meaningful fit and does not apply to this study. Therefore, alternate approaches need to be considered to capture the relationship between the sentiment variable and the dependent variable.

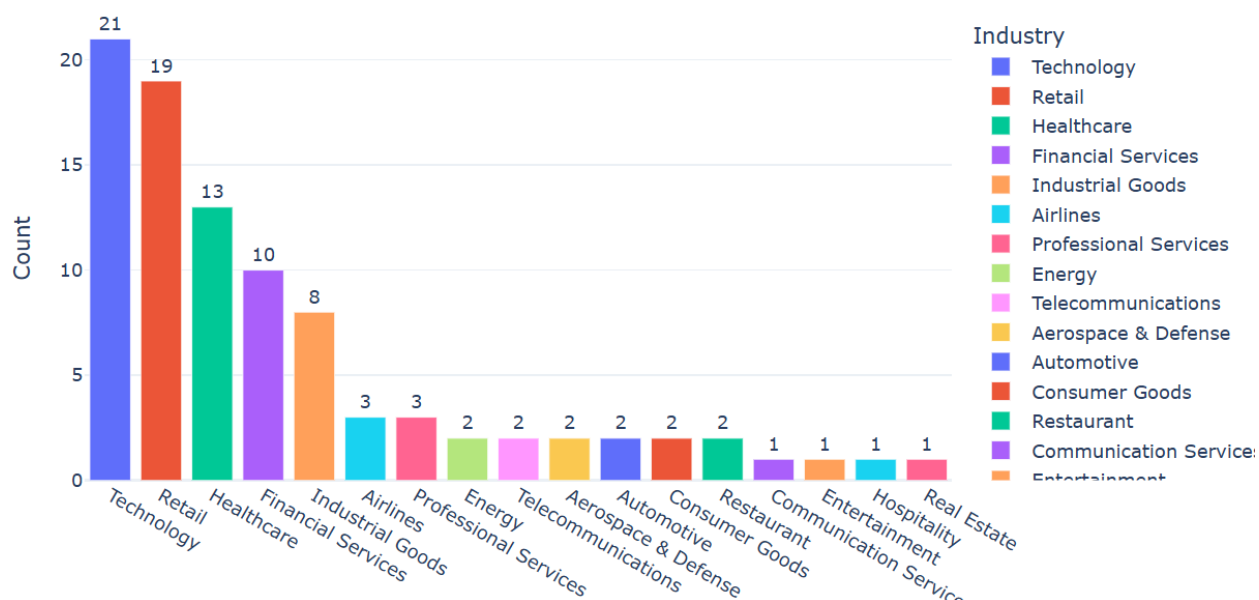
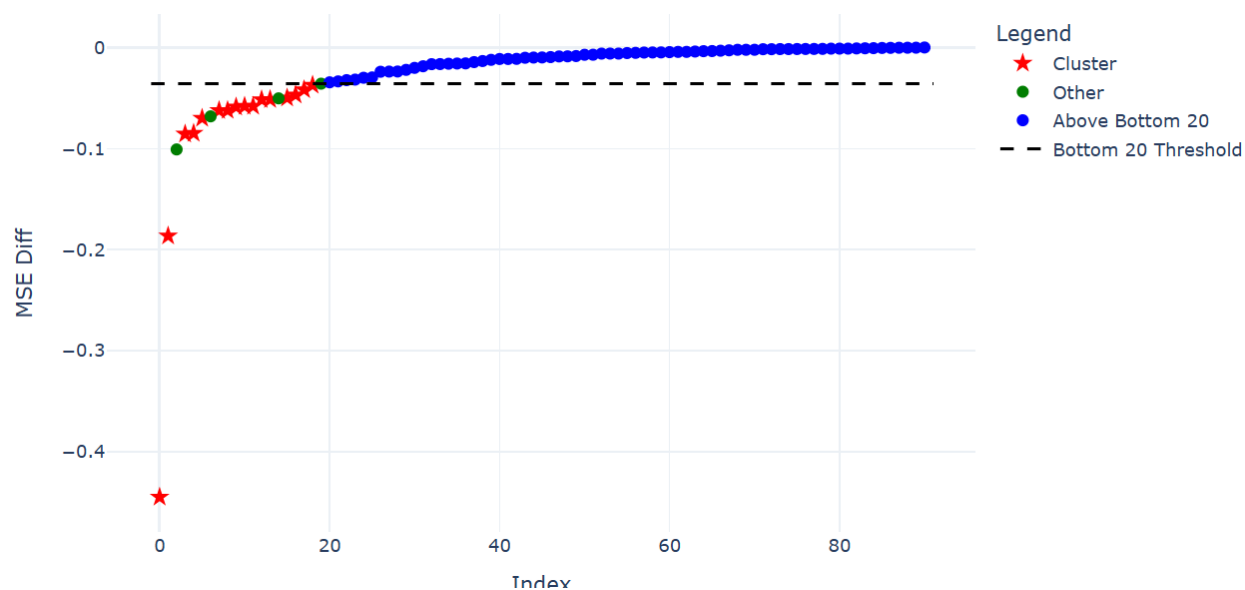
Application of Machine Learning Algorithms :

Because of the lack of data, it was not possible to evaluate the R-squared score. Instead, Mean Squared Error (MSE) was used as a measure for analysis and hypothesis validation. To evaluate the error, XGBoost, Support Vector Machine (SVM) and Random Forest were used.

Support Vector Machine : Hyperparameters used are {'regressor_C': 1.1789142699330446, 'regression_epsilon': 0.04142918568673425, 'regression_gamma': 'scale', 'regressor_kernel': 'linear'}

Analysis of mse error when it is less when sentiment is included as compared to when it is absent .

Industry Count Plot

**Figure 19****Figure 20**

PLot showing mean square diff when sentiment is included and then neglected.

Random Forest :

a) Calculating sentiment Importance as a feature for modeling stock prices percent changes. Results are Technology, Financial Services and retail have significant companies as compared to other industries.

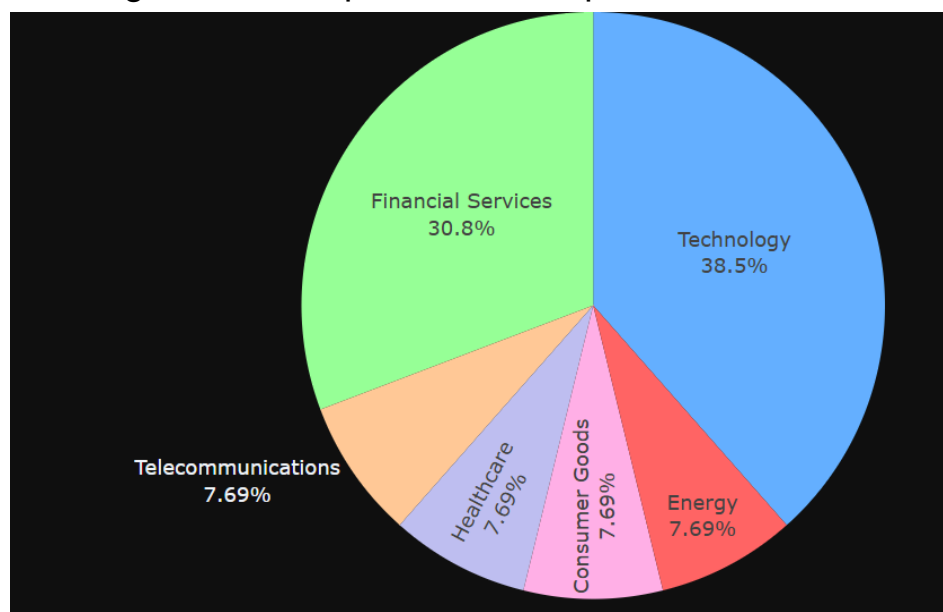


Figure 21

b) Measuring avg r^2 score diff for each industry taking sentiment into account and then not taking it and then analyzing top 5 industries.

Top 5 Industries with the Largest Positive R^2 Difference (Sentiment Present > Absent)

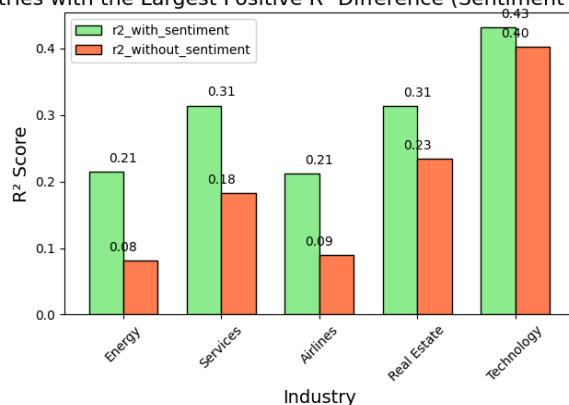


Figure 22

c) Error difference analysis has shown that MSE was higher in the absence of sentiment and thus significance is highlighted. The top 5 industries which have the highest error due to the lack of sentiment are plotted to highlight

the role of sentiment in improving the accuracy of the model.

Top 5 Industries with the Largest Error Difference (Sentiment Absent vs Present)

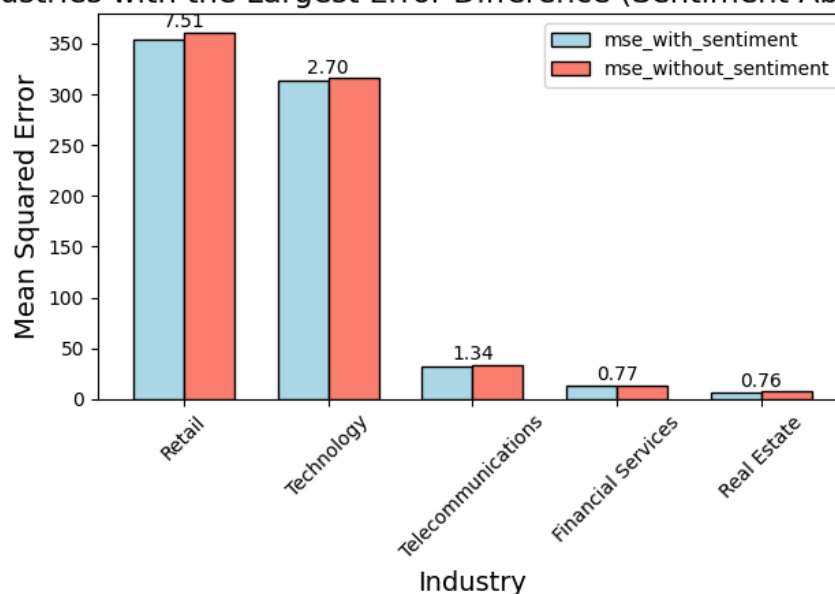


Figure 23

Xgboost :

XGBoost, short for *Extreme Gradient Boosting*, is a machine learning algorithm designed to optimize both computational efficiency and model performance. It belongs to the family of gradient boosting algorithms and is widely used in both regression and classification tasks due to its ability to handle a variety of data structures, including sparse datasets.

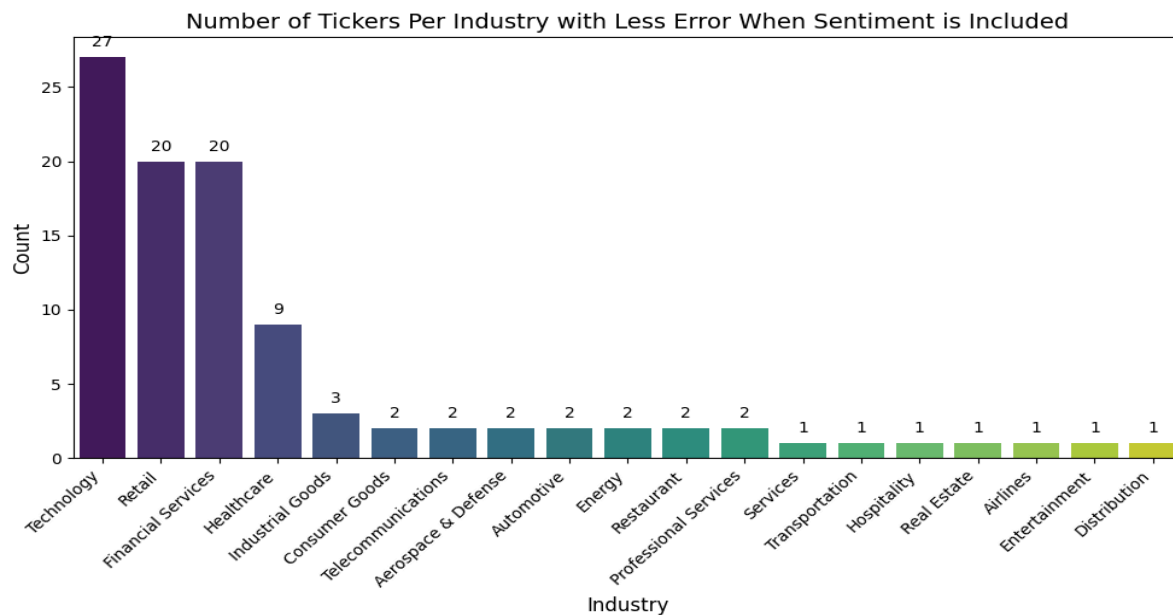


Figure 24

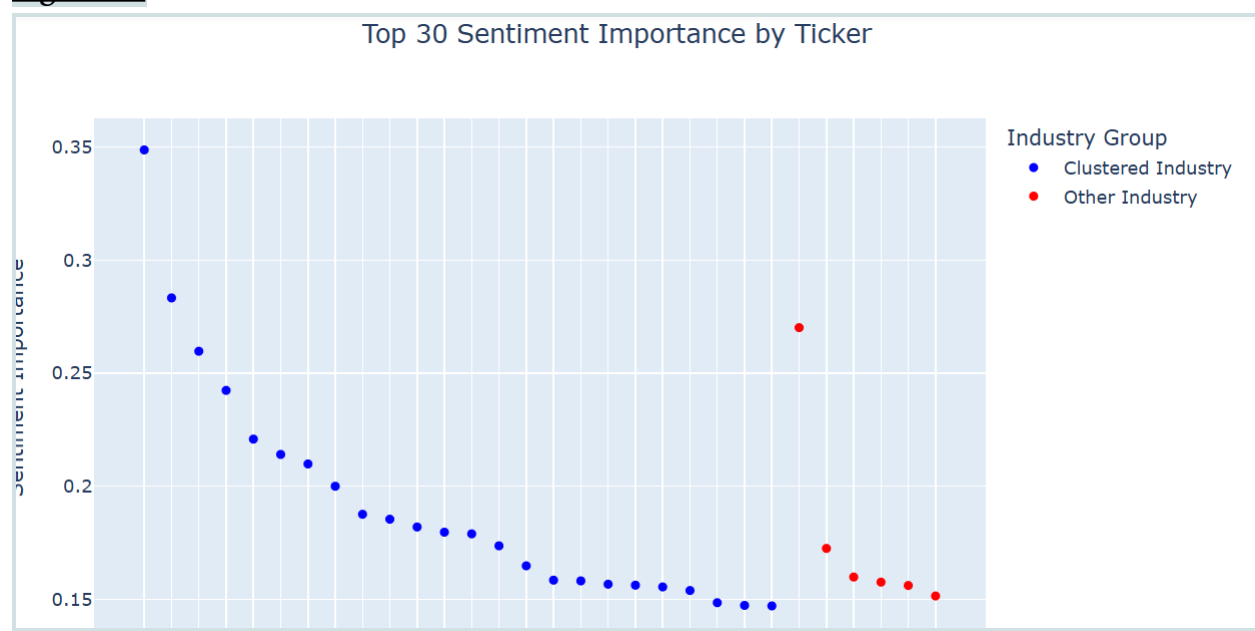


Figure 25

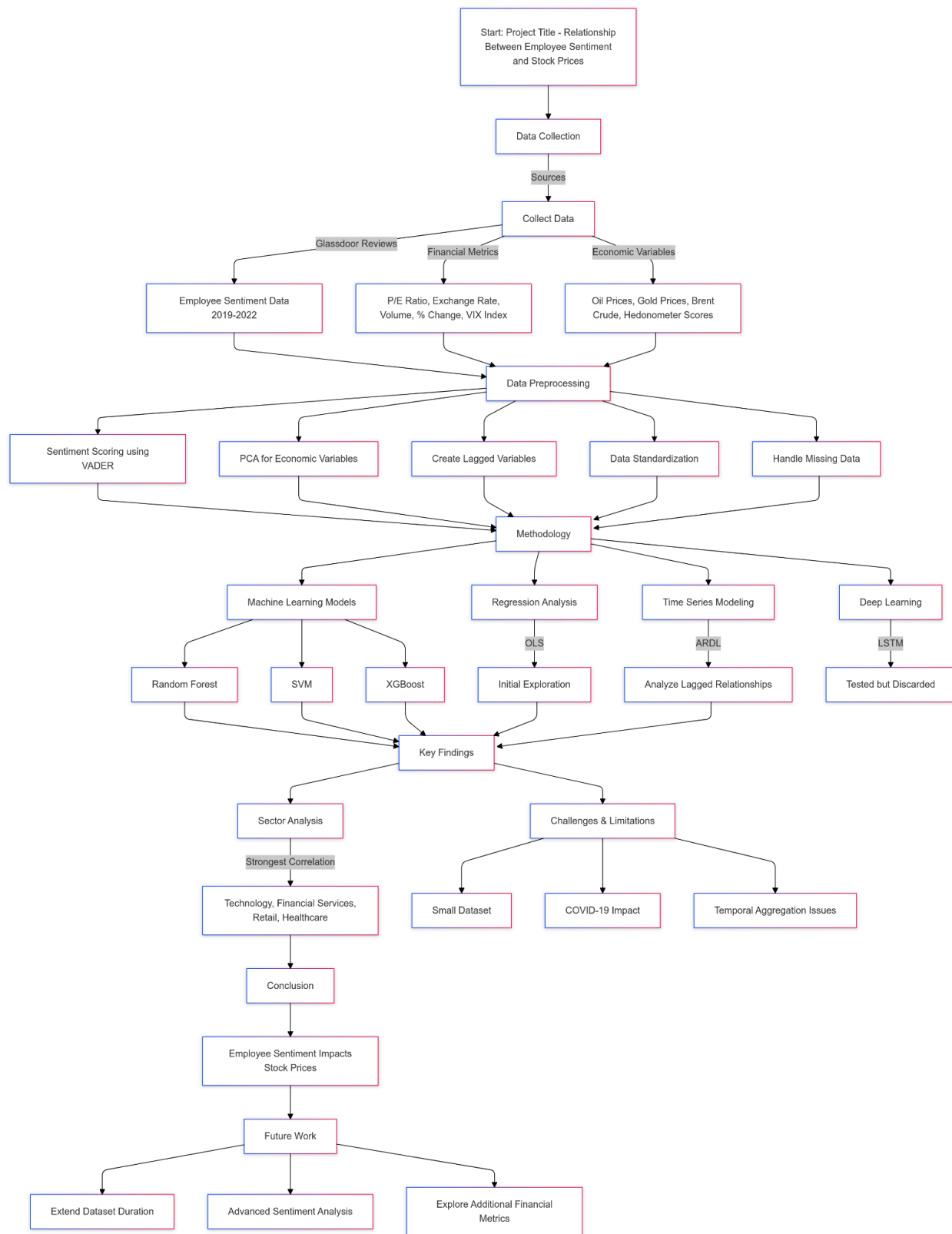
Conclusion

It exhibits how employee sentiment controls its large role in stock prices, rather measuring them for the months, not on daily periodicals, instead. In order to detect those specific sectors, analysis was rather intense to prove

the influence of employees' sentiment on stock price. Among all of them Technology, Financial Services, Retail, and Healthcare were found to be more highly correlated between stock prices and operational employee sentiment.

It shows the fact that for the companies surveyed, the effect of sentiment analysis on stock prices depreciates at an average level across all the companies yet becomes more notable at its identifiable clusters. This backs up the argument to take a sectoral approach to the application of sentiment analysis into stock price models prediction.

Finally, based on the above research, one can validate the alternative hypothesis instead of the null hypothesis; however, it is only applicable to the particular clusters. Broader findings would require further studies across other industries.



Future work :

Limitations of the Current Methodology along with limitations of glassdoor dataset :

1. Overfitting risks : Since application is on date clubbed on monthly average there is a high chance of overfitting .
2. Variance reading : Due to the final dataset being averaged monthly it is very much possible that we have reduced variability in the dataset thus improving results which may be questionable.
3. Data Availability and Coverage
Employee sentiment data, from 2019 up to mid-2023, was difficult to analyze because of its short historical coverage. This meant that long-term trends or cycles between employee sentiment and stock prices could not be easily identified.
4. Effect of Externally Triggered Events
The analysis period had prominent global events, namely the COVID-19 pandemic between 2020 and 2021 and its subsequent events. These events must have brought anomalies in the patterns of employee sentiment and stock price movement, making it even challenging to separate the direct relation between the two variables.
5. Data Gaps and Asynchronous Updates
Within the period considered, employee sentiment data would not have been uniformly available throughout all months or companies. Adding this to the inclusion of monthly stock price averages made it necessary to deal with missing or irregular data, which could create biases or reduce the robustness of results.
6. Temporal Aggregation
The reduction of the dataset to monthly averages, though necessary for uniformity, may have masked short-term interactions or causality between employee sentiment and stock prices. This level of aggregation may have reduced the effectiveness of capturing immediate market reactions to sentiment changes.
7. Lack of Data Sensitivity in the Model

Advanced models like LSTM require big datasets to work well. In this case, the dataset consists of only 54 months (from January 2019 through June 2023), and even that gets diminished further by aggregation. So perhaps, it was not the kind of dataset to leverage deep learning techniques properly.

8. Shifting Market Dynamics

The time span covered phases of significant market transitions, including economic stimulus measures, inflationary pressures, and rising interest rates. These may have affected the stock prices in ways that are independent of employee sentiment, thus contributing to noise in the analysis.

Future Improvisation:

- Working on a dataset having more number of years so that a true reflection of the relationship between sentiment and stock prices could be established .
- Use of advanced Deep Learning Algorithms like Autoencoders , Gated Recurrent Unit, stacked Lstm and LLM models like GPT, Fin Bert could be applied .
- Other Financial metrics of a company like Revenue Growth, Profit Margins, Return on Assets (ROA) and Return on Equity (ROE), Earnings Per Share (EPS) which are unexplored could be explored and a research around them could be done .
- LinkedIn can also be used for employee reviews and their sentiment.
- Investigating the relationship between employee sentiment and innovation outcomes, such as patent filings, R&D performance, and product launches, could reveal its role in fostering creativity and competitive advantage.