Predicting Company Performance Using Anonymous Employee Reviews

Sameer Ahmed
Courant Institute of Mathematical Sciences
New York University
New York, USA
sa6142@nyu.edu

Haseeba Fathiya Fazlur Rahman
Courant Institute of Mathematical Sciences
New York University
New York, USA
hf2313@nyu.edu

Monica Thirunavukkarasu
Courant Institute of Mathematical Sciences
New York University
New York, USA
mt4705@nyu.edu

***Abstract*—Forecasting an organization's growth is critical for investors to make timely choices in order to optimize profits. To that purpose, investors utilize alternative data from other parties to forecast a company's success before it releases its quarterly financial reports. Employee evaluations may be indicative of a firm's performance in the sense that employee happiness may be directly tied to corporate productivity. In this paper, we analyzed the reviews of three companies: Home Depot, General Electric, and Meta. We created various models for each company since each company has unique intrinsic complexity. We used different machine learning algorithms for each organization and examined which model was best suited to that company. We discovered that our models were predictive of corporate performance for some but not all companies. When we used the Decision Tree method, we got the best results for Meta, with an accuracy of 80%. As a result, we concluded that Glassdoor reviews can forecast company success and serve as an alternate data source.**

Keywords—Big Data Science, Data Analysis, Company Performance, Predictive Analytics, Meta, GE, The Home Depot, Machine Learning, CRISP-DM.

I. Introduction

Forecasting an organization's growth is essential for investors to make prompt decisions to maximize their profitability. To this end, investors use alternative data from third parties to predict a company's performance prior to the company releasing their quarterly financial reports. It could be argued that employee reviews can also be indicative of a company's performance, in the sense that employee satisfaction could be directly related to company productivity. We speculate that job postings can be indicators of an organization's future performance, which can be used by institutional investment professionals to make timely decisions.

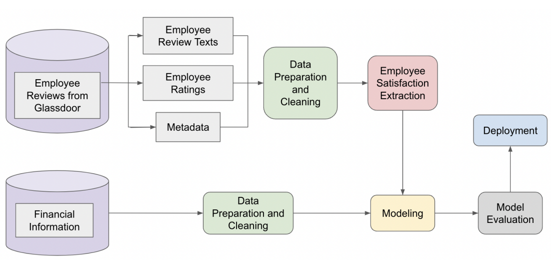


Fig. 1. Architecture Diagram

In this project we have considered 3 companies - Meta, General Electric and Home Depot and got their anonymous employee reviews and revenue data from Jan 2012 till April 2022. We have taken anonymous employee reviews from Glassdoor as our alternative data sources. Glassdoor is a platform where current and former employees anonymously review companies. We utilized web scraping technology to extract the reviews from Glassdoor. To validate our predictions, we used revenue data from SEC Filings which are regulatory documents that companies and issuers of securities must submit to the Securities and Exchange Commission (SEC) on a regular basis.

We tried to visualize the correlation between the employee review and company's revenue quarterly. Based on the analysis report, we prepared the data and cleaned it as per the requirements of our models. We performed various sentiment analysis techniques to extract the employee satisfaction from their review and utilized them as derived features. We experimented with the organization's future performance using various classification models, evaluated their results and got the most efficient classification models specific to each company.

II. Literature Review

A. *Employee Satisfaction and Corporate Performance*

Employees are an asset to a company and it has been found in numerous studies that companies with higher employee satisfaction had better long term performance [1]. A study that used survey data from 35 companies over 8 years found that employee satisfaction has a positive impact on return on assets and earning per share of the companies [2]. A study by Edmans(2011) in which they used "100 Best Companies to Work For in America " (1984 to 2009) survey data as a measurement of employee satisfaction found out that these companies had higher long run stock values in comparison to the overall stock market[5]. O'Reilly et al. (2014) used the survey data from respondents in 32 high-technology companies and showed that CEO personality affects a firm's culture and that culture can affect firm's financial performance. Popadak (2013) conducted a study on employee reviews collected through survey by career intelligence firms and concluded that the corporate culture is positively associated with long-term firm value[3]. A research by Symitsi et al. documented that the collective employee sentiment was a strong predictor of stock market returns with lower future returns following high employee sentiment. This predictive power was more pronounced when the employee sentiment index was constructed using the expectations of employees about the near-term business outlook of their employer [7]. Although, these data points provide significant insights about employee sentiments but they heavily rely on survey and publicly available datasets as a result they suffer from non-trivial selection bias and have limited generality. In addition, they do not delve into the specifics of the cause of employee satisfaction. One of the possible ways to alleviate these limitations is using social media data.

B. *Use of Glassdoor.com data*

Since the past decade, social media has been impactful in several areas, consumer opinions and sentiment extracted through online reviews, blogs and tweets are immensely used to predict sales and performance of companies. Twitter has been frequently used to gauge the sentiments of masses and predict election results[4]. Glassdoor is a similar social media platform founded by Robert Hohman, Rich Barton and Tim Besse in 2007 where current and former employees review the company and management anonymously. The attributes in data on glassdoor includes Star_Rating, CEO_Approval, Pros, Cons, Advice_to_Management, etc . Some of these attributes are numerical and some are textual. We have tried predicting company performance with numerical values like Star_Rating and later we extracted sentiment out of textual fields like pros and cons, Advice to Management and incorporated these into our model which resulted in better results.

III. Data Collection And Understanding

A. *Datasets*

Glassdoor is a platform where current and former employees anonymously review companies; hence this was our main source of alternative data. The anonymous employee reviews from Glassdoor were collected using web scraping technology. Employee reviews were collected for three companies: The Home Depot, Meta, and General Electric (GE). The dataset contained 34,721 reviews for The Home Depot, 11,873 reviews for GE and 4,855 reviews for Meta. The features and datatypes of the collected data from Glassdoor is shown in Fig. 2.

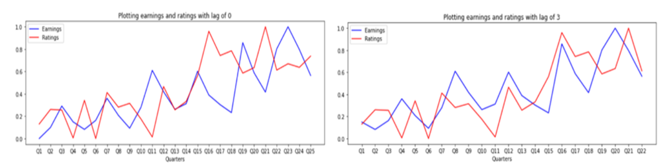
Company	object
Star_Rating	float64
Current_Employee	int64
Work_Duration	object
Review_Title	object
Date_Posted	datetime64[ns]
Job_Title	object
Job_Location	object
Recommend	int64
CEO_Approval	int64
Business_Outlook	int64
Pros	object
Cons	object
Advice_to_Management	object

Fig. 2. Features and datatypes of Glassdoor data

To validate our predictions, revenue data from SEC Filings were used which are regulatory documents that companies and issuers of securities must submit to the Securities and Exchange Commission (SEC) on a regular basis. Quarterly revenues and earnings of the three companies from 2012 to 2022 were collected. Ultimately, the revenues and earnings dataset consisted of 40 rows, one for each quarter through the ten years of data.

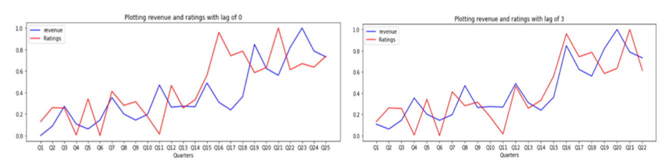
B. *Feature Selection and Visualization*

Two features: Job_Title and Job_Location were not considered in this analysis since only reviews were going to be considered, hence these features were eliminated from the analysis. To check if it is feasible to predict company performance by using reviews, we mapped the quarterly average Star_Ratings from Glassdoor against the quarterly earnings and quarterly revenues of the companies. The results obtained are shown from Fig. 3 to Fig. 7.



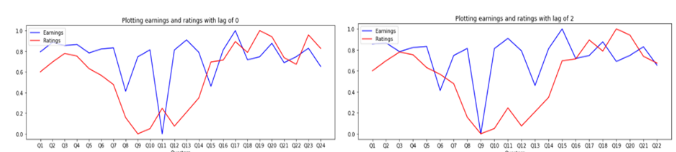
(a) (b)

Fig. 3. The Home Depot – Earnings Vs. Ratings (a) Correlation when no lag = 0.47 (b) Correlation with lag of 3 quarters = 0.70



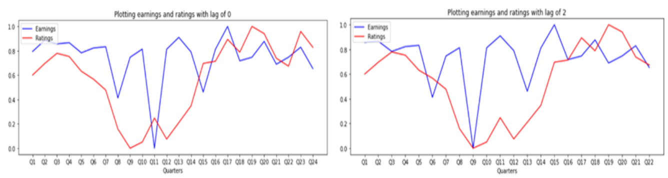
(a) (b)

Fig. 4. The Home Depot – Revenues Vs. Ratings (a) Correlation when no lag = 0.58 (b) Correlation with lag of 3 quarters = 0.77



(a) (b)

Fig. 5. GE – Earnings Vs. Ratings (a) Correlation when no lag = 0.275 (b) Correlation with lag of 3 quarters = 0.33



(a) (b)

Fig. 6. GE – Revenues Vs. Ratings (a) Correlation when no lag = 0.239 (b) Correlation with lag of 3 quarters = 0.156

Fig. 7. Meta – Earnings Vs. Ratings: Correlation with lag of 1 quarter = -0.10

From Fig.3 and Fig.4, it is clear that for Home Depot, there is a strong positive correlation between employee Star_Rating and company revenues and earnings, whereas the correlation between the three variables for the remaining companies is much weaker. Therefore, Star_Ratings alone cannot be used as a good indicator for company performance; features need to be derived from other textual features to get decent results.

IV. Data Preprocessing And Preparation

A. *Data Preprocessing*

The process of cleaning and transforming raw data before processing and analysis is known as data preparation. It is a critical stage before processing that involves reformatting data, making data corrections, and integrating data sets to enrich data. It entails normalization, replacing missing values, resampling, etc.

For this analysis, some columns had to be converted to numerical or discrete variables so that they can be used as input to the machine learning models. To extract the numerical information from the Work_Duration column, we replaced the text with the number of whole years worked. For example, if the value was “more than 1 year”, the value was reduced to 1. Similarly, If the value was “less than 3 years”, the value was reduced to 2. After this, for every column, the missing values were replaced with the averages of the respective column.

While aggregating the polynomial columns Recommend, Ceo_Approval and Business_Outlook for each quarter, the mode was taken to represent that value from that quarter. For example, if there are 100 reviews in a particular quarter out of which 55% were positive, then the value for that quarter is considered positive (i.e. 2).

For Meta and GE, both of which were companies for which the Star_Rating did not correlate with company earnings, more features were derived from the nominal columns. The percentages of positive, negative and “neutral” opinions were computed for each quarter. This was done so that we do not lose the opinions of those who were not in the majority group.

This gave us the following new additional features which all are real values: Recommend_minus, Recommend_same, Recommend_plus, Ceo_minus, Ceo_same, Ceo_plus, Business_minus, 'Business_same, Business_plus

With these new derived features, the correlations with earnings were better than the features we originally had. All the features for all companies were normalized since they were of different scales.

B. *Sentiment Analysis*

We have performed sentiment analysis on employee reviews using 3 different sentiment analyzers on ‘Review Titles’, ‘Pros’ and ‘Cons’ features in the original dataset. We calculated sentiment scores for them. As preprocessing, we converted the input string to lowercase. Then, we extracted only the letters from this string. This task is performed using Regular Expressions. Next, we tokenized the string. After tokenizing, we filtered the text of stop words. At last, we lemmatized the remaining words and returned these words. Then the ‘Pros’ and ‘Cons’ features columns were combined together. i.e we concatenated these pros and cons text fields to find their sentiment scores.

Performed sentiment analysis using:

- Using positive and negative word counts with normalization for calculating sentiment scores
- Using Vader Sentiment Analyzer
- Flair Sentiment Analyzer, which is a pre trained machine learning model

Method 1: Using positive and negative word counts with normalization for calculating sentiment scores

It is a rule based sentiment analysis technique. In this method, we will calculate the Sentiment Scores by classifying and counting the Negative and Positive words from the given text and taking the ratio of the difference of Positive and Negative Word Counts and Total Word Count. We defined the generic Negative Words and Positive Words list for which we have taken the positive-text.txt and negative-text.txt files or generally known as Opinion Lexicon. Then got a count of positive and negative words for the text and calculated the sentiment scores.

Method 2: Using Vader Sentiment Analyzer

It is also a rule-based sentiment analysis method. VADER means Valence Aware Dictionary and sEntiment Reasoner. It is more robust and precise in terms of Internet slang, which is widely used today. It uses a list of lexical features (e.g. word) which are labeled as positive or negative according to their semantic orientation to calculate the text sentiment. Vader sentiment returns the probability of a given input sentence to be Positive, negative, and neutral and compound prob of these 3. Here we have used the compound probability, it is the sum of positive, negative & neutral scores which is then normalized between -1 (most extreme negative) and +1 (most extreme positive)

Method3: Flair Sentiment Analyzer, which is a pre trained machine learning model

It is a pre-trained machine learning model. Flair allows you to apply state-of-the-art natural language processing (NLP) models to sections of text. It works quite differently to the previously mentioned models. Flair utilizes a pre-trained model to detect positive or negative comments and print a number in brackets behind the label which is a prediction confidence. The previous methods outputs a sentiment score between -1 and 1 but flair sentiment outputs the predicted label with a confidence score. The confidence score ranges from 0 to 1, with 1 being very confident and 0 being very unconfident.

On analysis we found that the Vader Sentiment Analyzer performed better. The reason is that it can very well understand the sentiment of a text containing emoticons, slangs, conjunctions, capital words, punctuations and also it works excellent on social media text. The drawback of using a flair pre-trained model for sentiment analysis is that it is trained on IMDB data and this model might not generalize well on data from other domains like twitter. But the main drawback with the rule-based approach for sentiment analysis is that the method only cares about individual words and completely ignores the context in which it is used.

C. Feature Engineering - Text Analysis

For the case of The Home Depot, it is not necessary for further features to be extracted since there is already a set of features from the raw dataset that showed good correlation with company performance. However, this was not the case for Meta and GE. Consequently, features were engineered from the review text as well. Plotting word clouds on the Pros_And_Cons for GE and Meta showed that common keywords persisted throughout all employee reviews. The word cloud for GE is shown in Fig. 8.

Fig. 8. Word Cloud for GE Pros and Cons.

The review text was first preprocessed by removing stop words and performing lemmatization. To extract popular topics, n-gram analysis was conducted using bigrams and trigrams. Since the bigrams and trigrams were similar, trigrams were not used to make the final features that would be input into the model. Hence, using bigrams, the 20 most common topics from all the reviews in the remaining two companies (Meta, GE) were extracted. Next, each sentence of a review was classified into one of these twenty topics by keyword matching, and then the sentiment of that review for that topic was calculated. In this manner, we get twenty additional columns for Meta and GE.

D. Predictive Column

The aim of this analysis is to predict the company performance using anonymous employee reviews. To be precise, we would like to predict if there will be an increase or decrease in the company's performance.

In the data visualization phase, correlations between Star_Rating, “earnings” and “revenues” were plotted to see if Star_Rating can be predictive of “earnings”. Since “earnings” can be considered as a derived column of “revenues”, “earnings” was selected as the final measure of company performance.

The predictive column of the dataset was made by comparing the earnings of the previous quarter and the current quarter. If the earnings of the current quarter were higher than that of the previous quarter, then the predictive value is ‘1’, which signifies “increase”. Similarly, if the earnings of the current quarter were lower than that of the previous quarter, then the predictive value is ‘0’, signifying “decrease”.

V. Modeling And Evaluation

We have created different models specific to each company. We have used ‘RapidMiner Studio’ to build our models. We took the training data from 2012 Quarter 1 to 2021 Quarter 4 and the test data.

A. The Home Depot

TABLE I. Home Depot - Selected Features

Feature Name	***Type***	***Description***
ratings	real	star rating given by employees
title_senti	real	VADER result of title_review
pnc_senti	real	VADER result of pros & cons
work_duration	real	work duration of employees
current_employee	binomial	is the employee currently working 1 - current employee 0 - not a current employee
recommend	polynomial	would employee recommend this company -1 - No 0 - Answer unknown 1 - No opinion/neutral 2- Yes
ceo_approval	polynomial	do employee approve their ceo -1 - No 0 - Answer unknown 1 - No opinion/neutral 2- Yes
business_outlook	polynomial	employee’s prediction of the company’s performance within the next six months -1 - Negative 0 - Answer unknown 1 - No opinion/neutral 2- Positive
earnings	real (not used)	Earnings of the company in that quarter

Feature Name	*Type*	*Description*
pred_column	binomial (label)	0 - company's performance decreases 1 - company's performance increases

a.

We used all these features except earnings and used cross validation with 10 numbers of folds. Used different modeling techniques

- Tree - Random Forest, Random Tree, Decision Tree
- Bayesian - Naive bayes, Naive bayes (Kernel)
- Regression - Logistic Regression, Logistic Regression (Evolution)
- Neural Nets - Deep Learning

We recorded the performance of these models and got accuracy, precision, recall and f-measure for all of them. Also we tried by lagging the revenues by 2 and 3 quarters. For example, if we take 2 quarters lag, we linked the 2015 Quarter 1 reviews with 2014 Quarter 3. Among these we found Random Forest, Random Tree and Naive Bayes had better accuracy. The lag with 2 and 3 quarters did give less accuracy when compared to the models with no lags.

TABLE II. HOME DEPOT - ACCURACIES

Model Name	*No Lag (%)*	*Lag 2(%)*	*Lag 3(%)*
Random Forest	58.33	50.8	46.67
Logistic Regression	56.67	40.83	51.67
Random Tree	55.83	43.3	38.33
Deep Learning	51.67	35	54.17
Naive Bayes	54.17	46.67	50
Logistic Regression Evaluation	49.17	52.5	49.17
Decision Tree	40.83	52.5	46.67

TABLE III. HOME DEPOT - PRECISION, RECALL, F1-SCORE

Model Name	*Precision (%)*	*Recall(%)*	*F1-Score(%)*
Random Forest	56.67	60	60
Logistic Regression	56.67	65	58.54
Random Tree	60	15	25

With these trained models we checked the 2022 quarter 1 prediction and most of the models predicted that the Home Depot's performance would decrease except logistic regression. Also according to the quarter 1- 2022 revenue report we noted that the company's revenue decreased.

B. Meta

In the raw dataset, we had 8 features, which were used for The Home Depot. In addition to these features, sentiments of the Review_Title and Pros_And_Cons were added as two more features. Features were extracted from the polynomial features Recommend, CEO_Approval and Business_Outlook. The percentages of positive, negative and "neutral" opinions from these categorical variables were computed for each quarter and added to the final dataset. Finally, twenty additional columns were added which are the sentiments calculated for the twenty most common topics. The 4 nominal columns were removed since features were derived from them. Therefore, for the final dataset, we have 33 final features which are all numerical, and a binomial predictive column. We used all these features except earnings and used cross validation with 10 numbers of folds. Modeling techniques used:

- Tree - Random Forest, Random Tree, Decision Tree
- Bayesian - Naive bayes, Naive bayes (Kernel)
- Regression - Logistic Regression, Logistic Regression (Evolution), Logistic Regression (SVM)
- Neural Nets - Deep Learning, Neural Net, Perceptron
- Discriminant Analysis - Linear Discriminant Analysis

For all these models, the accuracy, precision, recall and f-measure were recorded, which are depicted in the table below. The models which had the highest accuracies were: Decision Tree, Random Forest.

TABLE IV. META - ACCURACY, PRECISION, RECALL, F1-SCORE

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Decision Tree	80	80.83	96.67	87.52
Logistic Regression	76.63	94.17	76.67	80.57
Random Forest	70.83	75.83	88.33	80.71
Linear Discriminant Analysis	74.17	83.33	83.33	80.14
Naive Bayes	70.83	76.67	88.33	81.10
Naive Bayes (Kernel)	62.5	72.5	75	72.71
Random Tree	60.83	66.67	85	73.76
Logistic Regression - Evaluation	58.33	70	70	69.14
Deep Learning	56.67	68.33	63.33	64.15

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Neural Network	56.67	65.83	73.33	70.18
Logistic Regression - SVM	58.33	64.17	66.67	63.24
Perceptron	36.67	58.82	38.33	44.44

According to the revenue report submitted by Meta, the earnings of the company decreased for the first quarter of 2022. We can conclude that the best performing model was Decision Tree for Meta, which correctly predicted that for the first quarter of 2022, the company performance would decrease.

C. General Electric (GE)

The features for GE were selected in a similar manner as for Meta, hence we have 33 final features which are all numerical, and a binomial predictive column in the final dataset of GE. Just like in Meta, we used all these features except earnings and used cross validation with 10 numbers of folds. We experimented with the same models used in Meta, for which the accuracy, precision, recall and f-measure were recorded. The models which had the highest accuracies were: Random Forest, Random Tree, and Naive Bayes.

TABLE V. GE- Accuracy, Precision, Recall, F1-Score

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Decision Tree	56.67	37.5	15	26.09
Logistic Regression	41.67	32.5	60	41.03
Random Forest	65	55.56	30	41.67
Linear Discriminant Analysis	44.17	31.67	55	38.89
Naive Bayes	56.67	42.86	40	41.38
Naive Bayes (Kernel)	49.17	33.33	30	33.33
Random Tree	61.67	50	10	21.05
Logistic Regression - Evaluation	50.83	40	80	55.81
Deep Learning	39.17	15.38	10	14.29
Neural Network	41.67	27.78	30	30.3
Logistic Regression - SVM	40.83	21.43	15	20.69
Perceptron	49.17	27.27	20	23.08

It is evident that this method of data preparation did not work well for GE. According to the company's revenue report for the first quarter of 2022, revenue grew. Random Forest, Random Tree, Naive Bayes were some of the models which predicted this result correctly.

VI. Conclusion

In our work, we made different models for Home Depot, General Electric and Meta as different companies have different inherent intricacies specific to that company. We utilized multiple machine learning algorithms for each of the companies and analyzed which model is suitable to that specific company. We found that our models were predictive of the company performance for some companies, but not for all. The best result we obtained was for Meta, with an accuracy of 80% when the algorithm was Decision Tree. Therefore, we can conclude that Glassdoor reviews can be predictive of company performance and can be used as an alternative data source.

References

[1] Luo, N., Zhou, Y. and Shon, J., 2016. Employee satisfaction and corporate performance: Mining employee reviews on glassdoor.com.

[2] Popadak, J., 2013. A corporate culture channel: How increased shareholder governance reduces firm value. Available at SSRN, 2345384.

[3] Jahanbakhsh, K. and Moon, Y., 2014. “The predictive power of social media: On the predictability of US Presidential Elections using Twitter.” arXiv preprint arXiv:1407.0622.

[4] Edmans, A., 2011. Does the stock market fully value intangibles? Employee satisfaction and equity prices. Journal of Financial economics, 101(3), pp.621-640.

[5] Chinazzo, G., 2021. Investigating the indoor environmental quality of different workplaces through web-scraping and text mining of Glassdoor reviews.

[6] Symitsi, E. and Stamolampros, P., 2021. Employee sentiment index: Predicting stock returns with online employee data. Expert Systems with Applications, 182, p.115294.

[7] Beauregard, T.A. and Henry, L.C. 2009. “Making the link between work-life balance practices and organizational performance.” Human resource management review, (19:1), pp.9-22.

[8] Glassdoor. 2016. "About Us", Glassdoor, available at https://www.glassdoor.com/about/index_input.htm; retrieved May 9, 2016.

[9] Harter, J.K., Schmidt, F.L. and Hayes, T.L. 2002. “Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: a meta-analysis.” Journal of applied psychology, (87:2), p.268.