Predicting Stock Prices using Customer and Employee Ratings

Sanat Batra
Department of Computer Science
Courant Institute, New York University
New York, USA
sab1086@nyu.edu

Abstract—Stock price prediction is a widely researched topic with many different methods to go about it but is always challenging due to the randomness of the stock market. In this paper, we investigate the relationship between consumer and employee ratings and stock market trends. We use Yelp as our primary source for customer ratings and similarly Glassdoor for employee ratings. There are articles and research which show that a positive increase in a company's Glassdoor rating reflects on the company's annual earnings. We capitalise on this and also discover a positive correlation between Yelp and Glassdoor ratings and put forward a method and an LSTM model for combining both customer and employee ratings and predicting future stock prices.

Index Terms—stock market, ratings, prediction, trend, employee, customer

I. Introduction

Researchers have empirically used historical stock prices, textual data from a company's financial statements or news articles to predict future stock prices. But one may also follow a different school of thought and use alternative data sources to predict future stock prices.

Alternative data are innovative sources of data usually published by sources external to the company. Current popular alternative data sources in finance are geolocation (foot traffic), credit card transactions, product reviews and online browsing activity. We have chosen the business directory service and crowd-sourced review forum, Yelp as one alternative data source and the job search and employee review website, Glassdoor as another source.

Yelp ratings are indicative of the customer's perception of the company and highly influence its earnings especially in the restaurant business. But people also tend to use Yelp to rate shops and thus it also has thousands of ratings for department stores and retail outlets. These ratings can be assumed to be the public sentiment about that company.

Similarly, Glassdoor ratings are indicative of the employees perception or overall view of the company and its working and culture. Thus, it is a quantitative score for employee satisfaction and confidence in the company. In various studies, it has been shown that there is a positive correlation between the earnings of a company and employee satisfaction.

By using both these ratings we get an inside(Glassdoor) and an outside(Yelp) view which tells us the entire story of the company.

We show that both these views can be used in tandem to predict future stock prices.

II. Related Work

Many hedge funds develop models that use alternative data to support investment decisions but don't make any specific methods or processes public. There is a scarcity of information on how to use customer and employee ratings to forecast stock prices however there are articles which state the influence of these ratings on other factors.

[1]An article by Dr. Andrew Chamberlain gives a concise aggregation of different studies on the effect of Glassdoor ratings on different factors. He states that a certain study found that having a 1-star higher rating on Glassdoor predicts about a 1 percent higher annual return on company assets which is quite significant considering the size of the company. Many articles suggest the use of ratings instead of reviews since the rating itself is indicative of the review sentiment.

[2]This article states that around 54% of people feel that numeric ratings are the key consideration when reading a review. A Cornell University study states that in the eyes of consumers, the numerical ratings are seen as an objective measure of a business' quality. Oftentimes, the online rating of a business is taken as a proxy for its objective quality rating.

Symitsi et al. (2018) performed a portfolio analysis but used online reviews from Glassdoor to decide which US stocks to include in the portfolio. They found a positive and significant portfolio alpha. The researcher further found that there's a positive relation between employee satisfaction and company profitability using regression analysis. They found an association between employee ratings and firm performance in terms of profitability.

The chief Data scientist at Neuberger Berman also stated that Glassdoor ratings help in tracking poor management before the warning signs emerge and help predict the increase or decrease in stock prices.

III. Data Extraction

We built our own web scrapers for Yelp and Glassdoor using Python and ran those for the following list of companies:
Nike, Apple, Home Depot, McDonald's, Macy's, Gap
These companies were selected on the basis that they have a high number of ratings on both Yelp and Glassdoor. This also indicates that they are popular and have good social media recognition.

A. Glassdoor Scraping

We used BeautifulSoup to extract data from Glassdoor. BeautifulSoup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML. Using this package we scraped the Glassdoor ratings on every page for the above list of companies.
The features extracted were date, overall rating, worklife balance rating, culture and values rating, career opportunities rating, compensation benefits ratings and senior management rating. Along with these location, employee type and helpful votes were extracted but not used. The overall rating was the most important feature since it was the overall star rating and was given by every user. The rest of the ratings are subratings and have missing values at times. These values were replaced by "-1" while extracting the information. The ratings were stored in a csv file which undergoes further processing. Approximately 5000-17000 ratings were extracted for each company.

date	overallStar	workLifeStar	cultureStar	careerOppStar	comBenefitsStar	srManagementStar
Aug 28, 2014	4	1	3	3	3	3
Nov 3, 2019	4	3	5	4	4	4
Nov 1, 2019	5	5	5	5	5	5
Nov 1, 2019	5	4	5	4	4	5
Nov 1, 2019	4	4	4	3	5	4
Nov 1, 2019	4	2	4	4	4	4
Oct 31, 2019	5	3	2	4	2	2
Oct 31, 2019	5	5	5	5	5	5
Oct 29, 2019	4	3	5	4	4	3
Oct 29, 2019	4	-1	-1	-1	-1	-1
Oct 29, 2019	4	3	5	4	5	4
Oct 29, 2019	4	4	5	4	4	5
Oct 28, 2019	5	4	5	4	5	4
Oct 29, 2019	5	5	5	5	5	5

Fig. 1. Glassdoor data scraped for Apple.

B. Yelp Scraping

A similar process was followed for extracting Yelp ratings. BeautifulSoup was used to parse the HTML and extract ratings for the above list of companies. There are a lot of stores throughout the country and a lot of reviews per store. Thus to limit the amount of data extracted we selected only the stores within major cities namely New York, Los Angeles, Chicago, Boston, Houston, San Francisco, Phoenix, Dallas, San Jose, Austin and San Diego. The date and rating out of 5 was extracted. The number of friends, reviews and photos were also extracted but not used. All the star ratings were populated and we got approximately 2000-7000 ratings per company. These ratings were stored in a csv file which undergoes further processing.

date	rating	link
2019-09-10	5	/biz/adidas-los-angeles-3?osq=adidas
2018-11-26	2	/biz/adidas-los-angeles-3?osq=adidas
2018-06-17	3	/biz/adidas-los-angeles-3?osq=adidas
2019-07-27	1	/biz/adidas-los-angeles-3?osq=adidas
2018-06-21	1	/biz/adidas-los-angeles-3?osq=adidas
2019-05-06	1	/biz/adidas-los-angeles-3?osq=adidas
2018-12-05	4	/biz/adidas-los-angeles-3?osq=adidas

Fig. 2. Yelp data scraped for Adidas.

C. Stock Data

The end of day prices for each company were downloaded from Yahoo Finance.

IV. Data Preprocessing

The raw data extracted during the data extraction step is then put through certain processes before it can be directly used for modelling.
We aggregated the data into chunks of three months and assigned the corresponding end of day stock price to the final day within that chunk. The steps required for aggregating the data for Glassdoor and Yelp are given below:

A. Glassdoor Data Preprocessing

For each company in our list of company we fetched the corresponding list of Glassdoor ratings and stock prices. For each Glassdoor rating, the dates are split into intervals ie since we are working in chunks of three months, each date would be assigned to a chunk. For example any date within the months of January, February and March would fall within the same chunk.Similarly for April, May, June and so on. This chunk number along with the year is concatenated and hereafter used as the key for referring to any specific chunk.
We initialise 7 dictionaries namely avg _ratings, factor _sum, one_rating,two_rating,three_rating,four_rating and five_rating. avg_ ratings contains the sum of all the combined ratings within a chunk. It is later replaced by the average combined rating of that chunk. factor_sum contains the total number of ratings within that chunk. one_rating,two_ rating,three_rating,four_ rating and five_rating contain the number of combined ratings which lie within a particular range. These ranges will be explained ahead.
Now, Glassdoor contains an overall rating and 5 subratings namely work life, culture, career opportunities, compensation benefits and senior management. The combined rating for each date in a chunk is calculated as the sum of each rating type multiplied by a weighting factor. The overall rating is given a

weighting factor of 7 since it is the most important value and doesn't contain any null values. The subratings are given a weighting factor of 1.5 each and if any subrating is null ie -1 then it is ignored ie set to 0. This combined rating for each date is added up for all dates in a chunk to get the total combined rating for that chunk. This value is divided by factor_sum ie the total number of ratings in a chunk to get the average combined rating of a chunk.

Now, while computing the combined rating for each date we check whether it lies within a specific range. If it lies between 4.5 and 5 we consider it to be a 5 rating, 3.5 and 4.5 we consider it to be a 4 rating and so on. Basically we are dividing each combined rating further into types and generalising the combined ratings for each dates into 5 discrete categories (one_rating, two_rating, three_rating, four_rating, five_rating). When it is categorised, the count in that dictionary for that chunk increases by 1. This means that the number of high rated or low rated ratings has increased by 1. So basically each of these dictionaries represents the number of high, medium or low rated ratings within a particular chunk ie five_rating represents the number of very high ratings, four_rating represents the number of high but not very high ratings and one_rating represents the number of very low ratings.

Now after all this processing, avg_rating contains the average combined rating for that chunk and one _rating to five_rating contain the values discussed above. All these values for each chunk are stored in a single dictionary known as ratings with the keys as each chunk identifier ie the month number % number of chunks and year.

Each row in ratings corresponds to a chunk. So each column of ratings is normalised to bring the values within a common scale. Min max normalisation is carried out. The equation is as follows

$$X = \frac{X - X_{min}}{X_{max} - X_{min}}$$

This processing is carried out for the data of each company and stored in separate csv files.

B. Yelp Data Preprocessing

The steps for preprocessing Yelp data are exactly the same as above except for a few minor changes. Yelp data only contains the date and star rating. So the average value for each chunk is computed by directly summing up the ratings for each date within that chunk and dividing by the number of ratings in that chunk. Since Yelp does not have star ratings, the combined rating is an integer and thus the one_rating to five_rating dictionaries directly correspond to the number of one star ratings or number of five star ratings in that chunk. Min max normalisation is carried out on these values. Similarly, this processing is carried out for each company's data and stored in separate csv files.

C. Combining Yelp and Glassdoor Data

The company-wise output files for Yelp and Glassdoor are combined by merging both files on the basis of company name and chunk identfier. Having the same chunk identifier means the group of ratings is for the same time period or group of months. So the information for Glassdoor and Yelp is combined into a single file by matching the above two criteria. Further analysis of the data is done using the five_rating attribute ie glassdoor_5 and yelp_5.

period	company	normalized_yelp_avg_rating	yelp_no_of_reviews	yelp_5	percentage_yelp_5	yelp_4
2018_2	Nike	0.5797101449275363	0.8947368421052632	0.6944444444444444	0.6227355072463768	0.28571428571428
2018_3	Nike	0.5191146881287726	0.9210526315789473	0.6111111111111112	0.5325704225352113	0.28571428571428
2011_1	Nike	0.6666666666666667	0.1447368421052631	0	0	0.33333333333333

yelp_3	yelp_2	yelp_1	normalized_glassdoor_avg_rating	glassdoor_no_of_reviews	glassdoor_5
0.6	0.9090909090909091	0.6285714285714286	0.4910043558279394	0.8595505617977528	0.859375
0.9	0.4545454545454545	0.8285714285714286	0.2800579883170606	0.8370786516853933	0.5
0.3	0.0909090909090909	0.0285714285714285	0.5273126039312666	0.0337078651685393	0.015625

glassdoor_4	glassdoor_3	glassdoor_2	glassdoor_1	months_after_stock_price
0.6842105263157895	0.5925925925925926	0.9333333333333333	0.5714285714285714	76.910004
0.7236842105263158	1	0.4666666666666667	0.7142857142857143	75.040001
0.0789473684210526	0.0555555555555555	0	0	20.58

Fig. 3. Features extracted by the pre-processing process.

V. Modelling

A. Causality

Before modelling our data, we tested to see if the data we have extracted and pre-processed into different features has predictive power towards forecasting the stock price of the respective company. For this we used the Granger Causality test which checks to see if the past values of a certain variable Y combined with the past values of a variable X are more predictive of future values of X than past values of X alone. We tested all the features extracted separately to see which ones can be deemed to be most predictive.

```
Granger Causality
('number of lags (no zero)', 8)
ssr based F test:         F=1.4915  , p=0.2193  , df_denom=21, df_num=8
ssr based chi2 test:   chi2=21.5911 , p=0.0057  , df=8
likelihood ratio test: chi2=17.0970 , p=0.0291  , df=8
parameter F test:         F=1.4915  , p=0.2193  , df_denom=21, df_num=8

Granger Causality
('number of lags (no zero)', 9)
ssr based F test:         F=2.7064  , p=0.0345  , df_denom=18, df_num=9
ssr based chi2 test:   chi2=50.0675 , p=0.0000  , df=9
likelihood ratio test: chi2=31.6633 , p=0.0002  , df=9
parameter F test:         F=2.7064  , p=0.0345  , df_denom=18, df_num=9

Granger Causality
('number of lags (no zero)', 10)
ssr based F test:         F=2.1938  , p=0.0822  , df_denom=15, df_num=10
ssr based chi2 test:   chi2=52.6518 , p=0.0000  , df=10
likelihood ratio test: chi2=32.4431 , p=0.0003  , df=10
parameter F test:         F=2.1938  , p=0.0822  , df_denom=15, df_num=10
```

Fig. 4. Granger causality test results for Macy's number of Yelp 5 star ratings.

As we can see from the above figures, almost all p values are below 0.1 for a lag of 9 in both cases . This signifies that the number of both Yelp and Glassdoor 5 star ratings are most predictive of stock prices 9 periods into the future. Based on these tests we can reject the null hypothesis that the two features do not predict stock prices. The other features were also tested and we determined that they were less predictive and will therefore not be used in the model. Since each

```
Granger Causality
('number of lags (no zero)', 8)
ssr based F test:         F=1.7299  , p=0.1498  , df_denom=21, df_num=8
ssr based chi2 test:   chi2=25.0423 , p=0.0015  , df=8
likelihood ratio test: chi2=19.2363 , p=0.0136  , df=8
parameter F test:         F=1.7299  , p=0.1498  , df_denom=21, df_num=8

Granger Causality
('number of lags (no zero)', 9)
ssr based F test:         F=2.0488  , p=0.0934  , df_denom=18, df_num=9
ssr based chi2 test:   chi2=37.9022 , p=0.0000  , df=9
likelihood ratio test: chi2=26.0948 , p=0.0020  , df=9
parameter F test:         F=2.0488  , p=0.0934  , df_denom=18, df_num=9

Granger Causality
('number of lags (no zero)', 10)
ssr based F test:         F=1.6970  , p=0.1717  , df_denom=15, df_num=10
ssr based chi2 test:   chi2=40.7290 , p=0.0000  , df=10
likelihood ratio test: chi2=27.2434 , p=0.0024  , df=10
parameter F test:         F=1.6970  , p=0.1717  , df_denom=15, df_num=10
```

Fig. 5. Granger causality test results for Home Depot's number of Glassdoor 5 star ratings.

period is of 3 months, the two features (number of Yelp and Glassdoor 5 star ratings) seem to be useful in forecasting stock prices around 27 months into the future. This makes sense as customer and employee ratings can tell us how the stock price of a certain company would move in the long term and should not be indicative of short term fluctuations.

In Fig. 6 and 7 we can see how the two features that are determined to be most predictive are correlated with the stock price.

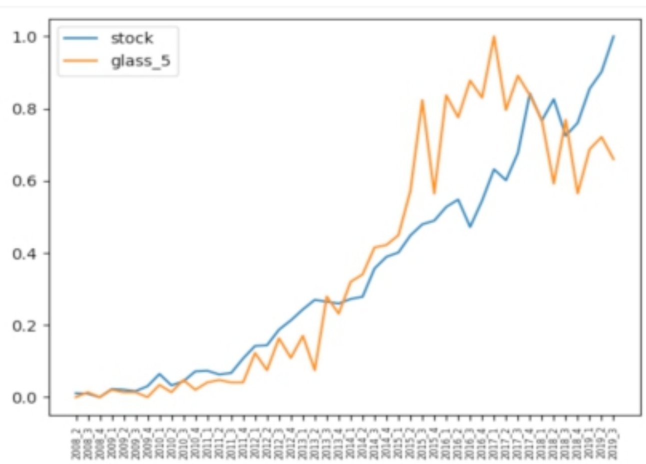

Fig. 6. Correlation of Home Depot's number of 5 star Glassdoor ratings with its stock price.

B. Long Short Term Memory

Long short-term memory (LSTM) uses an artificial recurrent neural network architecture with feedback connections. It can process entire sequences of data. We used a Stacked LSTM model to model our data. A Stacked LSTM has multiple hidden LSTM layers where each layer contains multiple memory cells. We have selected 5 as the length of each sequence to be

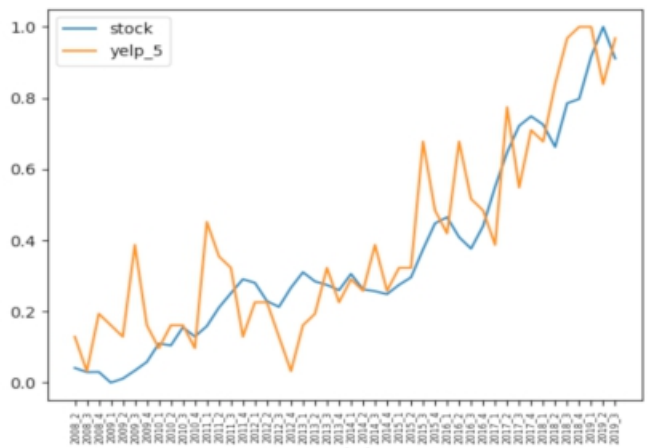

Fig. 7. Correlation of McDonald's number of 5 star Yelp ratings with its stock price.

Company	RMSE		Accuracy(%)	
	All	Only Stock	All	Only Stock
Macys	0.109	0.414	66.67	0
McDonalds	0.106543	0.469657	100	25
Nike	0.2796	0.3539	100	66.67
Apple	0.220086	0.2450142	100	100
Home Depot	0.086977	0.103624	100	100
Gap	0.172304	0.35267	25	25

TABLE I
RMSE AND ACCURACY OF PREDICTIONS.

fed to the model. A lag of 9 is selected for all companies. This value is kept constant to avoid over-fitting for each company.

The model is trained on 90% of the data and tested on the remaining 10%. At first we tried to predict stock prices 9 periods (27 months) into the future using past values of only the stock price. Figure 8 is the result we obtained. As we can see from the plot, the model does not work well when it is trained on only the stock price. There are two ways we have evaluated the result - root mean squared error (RMSE) and prediction accuracy, where prediction accuracy is the accuracy with which we have classified the future stock price as having increased or decreased from the current price. In our first test of the model, using only past stock prices, the RMSE is 0.414 and the prediction accuracy is 0%. This result is not considered to be absurd because past values in the stock market are only slightly indicative of future values. This is where alternative data comes into play.

When we added the two features - number of Yelp and Glassdoor 5 star ratings to the model, we obtained the result shown in figure 9. The RMSE is only 0.109 and we classified the increase or decrease in the price with an accuracy of 66.67.

The addition of customer and employee reviews provided the model with a considerable amount of additional predictive power as expected. This has been seen to be the case across all companies as observed from Table 1.

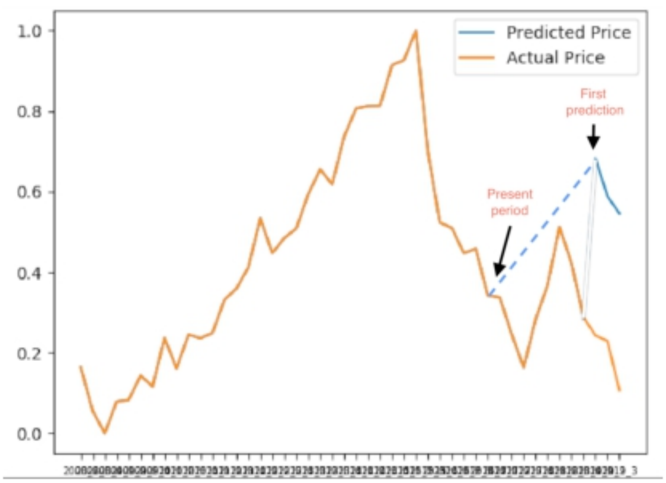

Fig. 8. Forecast of Macy's future (27 month's out) stock price using only past stock price values.

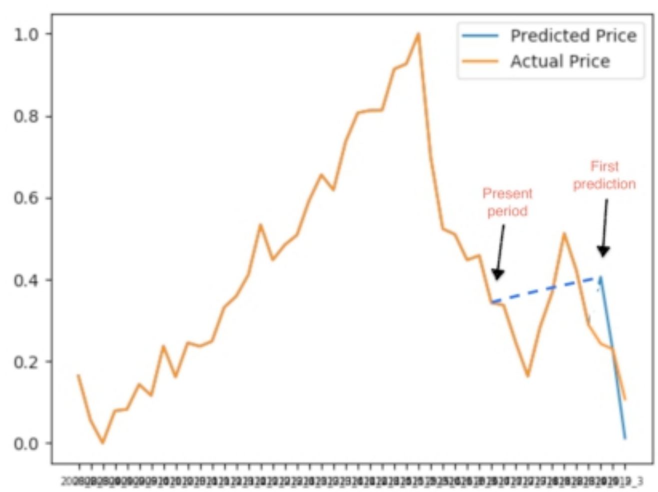

Fig. 9. Forecast of Macy's future (27 month's out) stock price by combining past values of number of Yelp and Glassdoor 5 star ratings and stock prices.

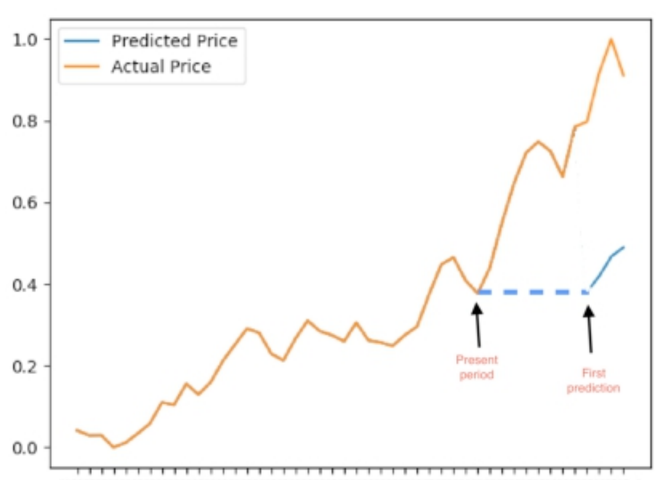

Fig. 10. Forecast of McDonald's future (27 month's out) stock price using only past stock price values.

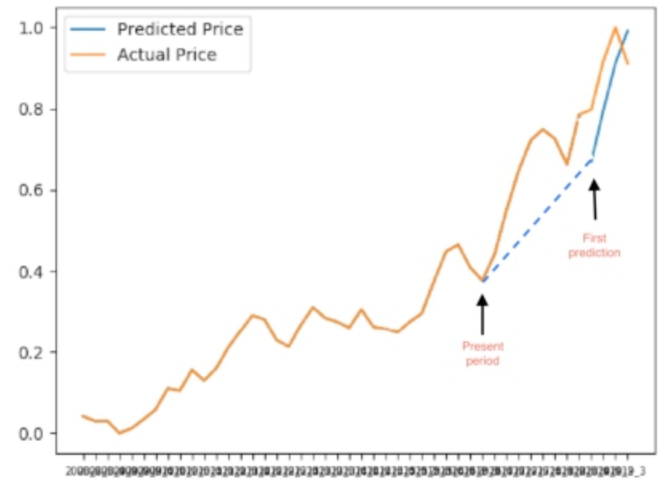

Fig. 11. Forecast of McDonald's future (27 month's out) stock price by combining past values of number of Yelp and Glassdoor 5 star ratings and stock prices.

VI. Conclusion

From our results we can conclude that consumer and employee ratings are highly predictive of the movements in stock prices in the long term. While in most cases only past values of stock price fail to predict which direction the price would move in, the patterns in past alternative data, namely Yelp and Glassdoor ratings, have been proven to hold a considerable amount of predictive power towards long term movements in the stock price of a company. Stock prices can double, triple or become a fraction of the current value in the span of 2 years. The power to forecast the direction in which prices would move in the long term can be a very valuable asset while making investment decisions or adjustments to portfolios. Further work would include finding a way to combine the training data of all companies without losing the sequential nature of training data that the LSTM model requires. This would result in a much larger training dataset and might improve the performance of the model. Also, including ratings from websites other than Yelp and Glassdoor could provide us with additional predictive power.

References

[1] Dr. Andrew Chamberlain, "Do People Drive Profit? Glassdoor Ratings Linked to Company Performance in the United Kingdom", March 2018.
[2] Susanna Ganjalyan, "Star Rating remains the most important part of a Review", Feb 2018.
[3] Efthymia Symitsi, Panagiotis Stamolampros,George Daskalakis and Nikolaos Korfiatis, "Employee Satisfaction and Corporate Performance in the UK" in SSRN,Feb 2018.
[4] Chris Campbell, "Star Ratings Matter Just as Much as (If Not More Than) Online Reviews" ,October 2015.