**Stock Prices & Sales Prediction using Sentiment Analysis**

BYGB 7978 Web Analytics

Prof. Yilu Zhou

Dec 22, 2021

**By**

**Baris Baturay**

**Grant Preuninger**

**Sejal Panchal**

**Yuzhen Wu**

**Executive Summary**

Stock market prediction has been an active area of research for a long time. The Efficient Market Hypothesis (EMH), a hypothesis in financial economics, states that asset prices reflect all available information. In simple words, stock market prices react to new information and follow a random walk pattern. Over the last few decades, several hypotheses have been made to extract patterns in the way stock markets react and respond to external stimuli.

When it comes to sales of a newly launched product, marketing plays an important role. Time and again several studies have proven that public opinion about the products is crucial to determine the success of a new product in the long run. Analyzing the large volume of online reviews available would produce useful actionable knowledge that could be of economic values to businesses.

If external stimuli have a large impact on stock prices, can we predict stock prices using people's behaviour? Are sentiments expressed in online reviews correlated with the sales performance of products? These curious questions gave ideas to our project - ‘Stock Prices and Sales Correlation using Sentiment Analysis’.

In this project, we aimed to predict stock market prices and business sales based on the premise of behavioral economics for two large automobile companies: Ford Motor Company and Toyota Motor Corporation. We used online reviews of Ford’s and Toyota’s stakeholders’ to perform sentiment analysis; FAMA French Model to describe stock returns, linear regression analysis to predict continuity and correlation between stock, sales and sentiments.

**Business Goal Analysis**

The growth of social media platforms, advanced technology, and popularity of ecommerce have reduced the gap of interaction, shopping and voicing concerns. A research paper by Thorsten, Hennig-Thurau, Kevin P., Gwinner Gianfranco, and Walsh Dwayne D. Gremler, 2003 states that customer comments articulated via the Internet are available to a large number of audiences, and therefore can be expected to have a significant impact on the success of goods and services. Online reviews play a testament to consumer buying and communication behavior. The results of this research illustrates that consumers read online articulations mainly to save decision-making time and make better buying decisions. Structural equation modeling shows that their motives for retrieving online articulations strongly influence their behavior. As all this data is publicly available, web analytics provide an enormous scope and methods to collect this data, analyse it and make meaningful business decisions.

We performed sentiment analysis on publicly available online reviews of consumers and employees data to find the public mood and their opinion about the automobile brands - Ford & Toyota. Our study shows that people's behavior influences sales and stock market trends. By using regression analysis, we could numerically establish the correlation between sentiments and stock price and sentiments and sales growth rate.

This project will generate intuitive insights that will aid businesses in driving growth and optimizing strategy in terms of product development, launch intelligence, feature enhancement, marketing and distribution.

**Dataset Description and Crawling**

In this project, we collected data from online reviews from two main stakeholders viz. consumers and employees of Ford & Toyota. Python packages - Selenium and BeautifulSoup were used to mine data such as review text, date and rating from Glassdoor.com, and Indeed.com to collect employee reviews, and Influenster.com and EveryAuto.com to collect consumer reviews.

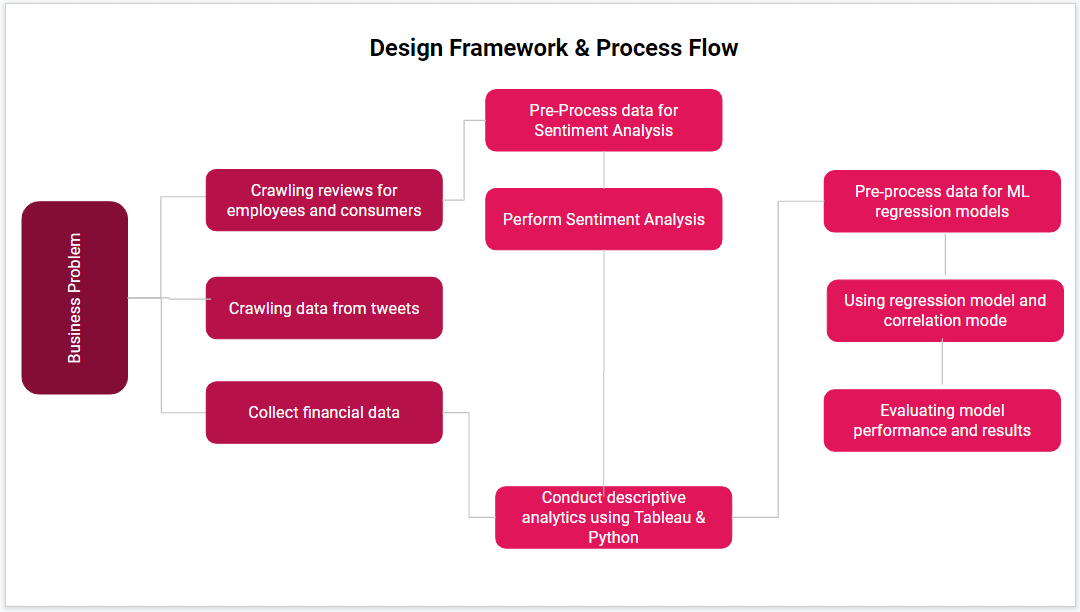
After preprocessing data, we got 37842 consumer reviews and 9757 employee reviews for Toyota and 35227 consumer reviews and 12418 employee reviews for Ford. Pandas and Numpy, which are powerful Python libraries, were used to clean data, and remove null values. Sentiment Intensity Analyzer from nltk library was used to give score values in the form of positive, negative and neutral for each review represented as a compound score.. We combined positive and negative review counts separately for monthly to spot trends and perform regression modelling.

Financial data was collected from Ford and Toyota’s publicly available financial reports from March 2000 - November 2021. Sales data for Ford and Toyota was collected from online open resources like Ychart.com. Both companies' stock market prices were collected by using Yahoo Finance API in Python. In order to run our models and to perform descriptive analytics we just used the closing prices as an indicator.

After doing the sentiment analysis we created 3 features from crawled data viz. percentage positives, average compound and average rate. Since, our data was crawled for each day, we had to group them monthly to do quarterly predictions. The total number of reviews changes significantly month by month so grabbing the total number of sentiments was not a good indicator. Instead of looking at total reviews we looked at each month what percent of the comments were positive and created ‘percentage positives’. Moreover, the compound score indicates how positive or negative each review is. For looking at them monthly we basically get the average compound score for all of the reviews monthly. We did the same for ratings to get average monthly ratings.

**System Design**

Web analytics offers varied techniques to extract data that is publicly available, build data mining models to generate insights that power business decisions. By using appropriate programs, we can extract current, real-time as well as historical data. After doing individual research on potential topics for this project, we were certain to predict stock prices by using people’s sentiments. However, we also wanted to estimate the degree of correlation between sentiments and sales. After conducting industry research and brainstorming on domains, we decided to conduct our study on the automobile industry. It was easy to choose Ford and Toyota as they are the largest automobile companies and have wide ranges of automobiles for all demographics, thus, increasing the possibility of getting large datasets.

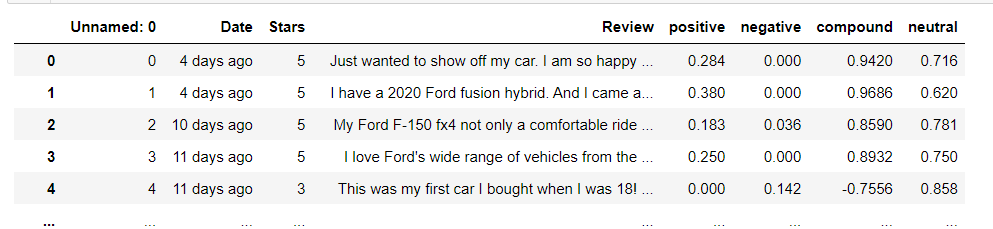
*Fig. 1 System Design Framework & Process Flow*

We were also curious if employee behavior affects stock prices, thus we decided to take both consumer and employee reviews in our analysis and find the highest correlation. By doing a thorough exploratory research on car review sites, we found that influenster.com and everyauto.com were the two best candidates for our study. The reason these two sites were chosen is because they had around 50,000 reviews for each car brand and reviews were not hidden in category layers such as date range, car model, car make etc. It was easy to crawl data from these sites by using the Python packages Beautiful Soup, HTTP Request and Selenium.

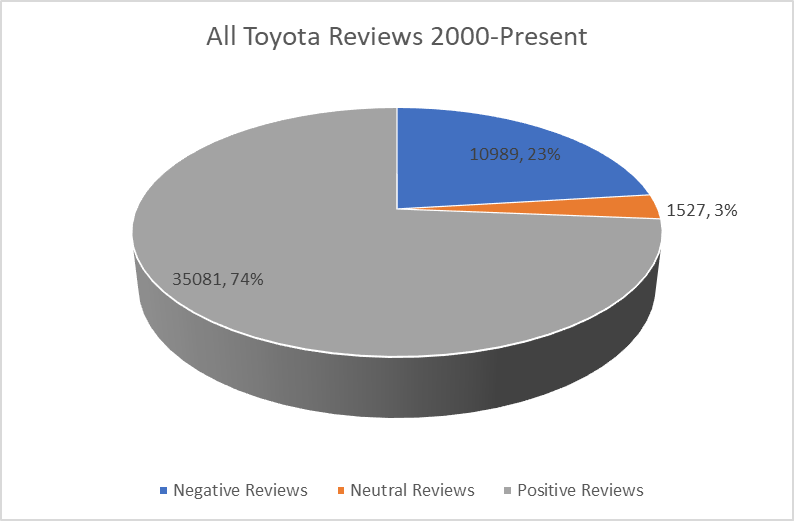
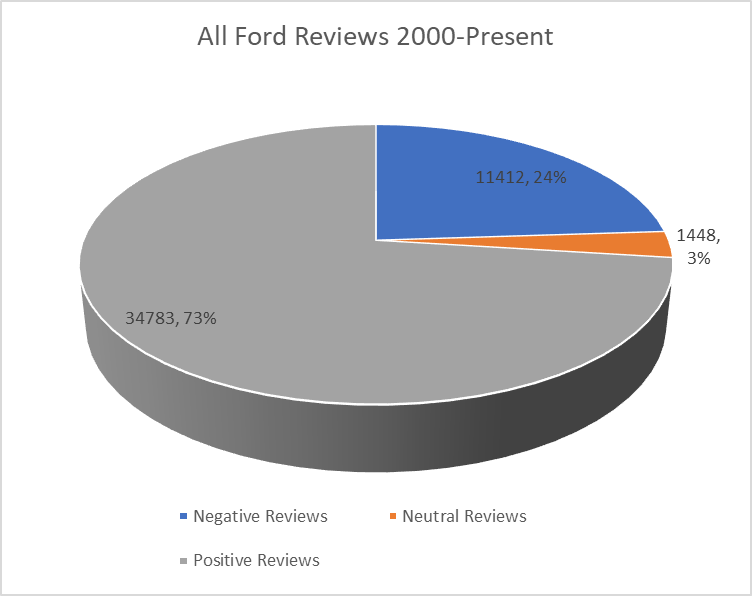
For the website Influenster.com, which is one continuous page of reviews that loads as you scroll, we first created a program with Selenium and Google Chrome driver to automatically scroll the pages to retrieve the reviews. However, in the chrome driver the pages turned out to stop loading after a short period, which was possibly due to security measures of the website. We instead opted to scroll down the webpage until we reached around 6500 reviews for each company and copied the outer html into a text file to use beautiful soup. We were able to easily crawl Indeed, and Every Auto using Beautiful Soup and HTTP requests. However, to crawl Glassdoor we used Selenium in order to log in to the website to see all the reviews.

**Sentiment Analysis**

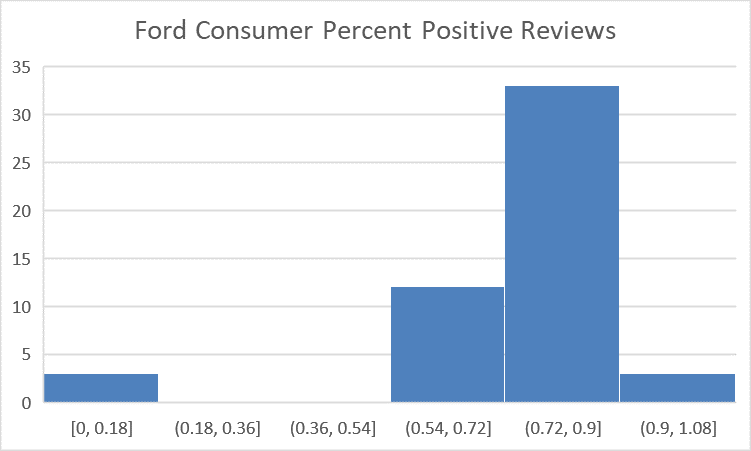
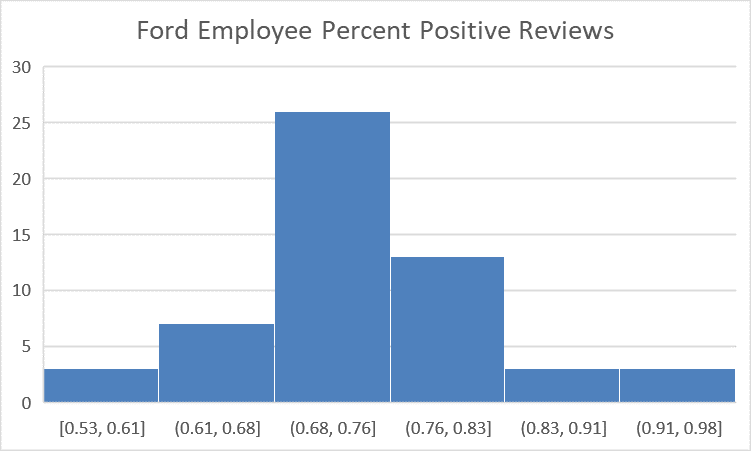
Sentiment Intensity Analyzer from nltk library was then used to understand people’s sentiments from reviews. Sentiment Intensity Analyzer gave scores for each review: positive, negative, neutral and compound. However, we only used compound score to determine positivity or negativity.

*Fig. 2 Snapshot of results of Sentiment Analysis*

From the pie chart, we can see the total positive, neutral and negative percentages are nearly similar for Ford and Toyota.

 *Fig 3.1 Sentiment Analysis for Ford Fig 3.2 Sentiment Analysis for Toyota*

The below bar graph shows that the average compound was less for employee reviews than consumers’.



*Fig 4.1 Ford Consumer Positive Reviews Fig 4.2 Ford Employee Positive Reviews*



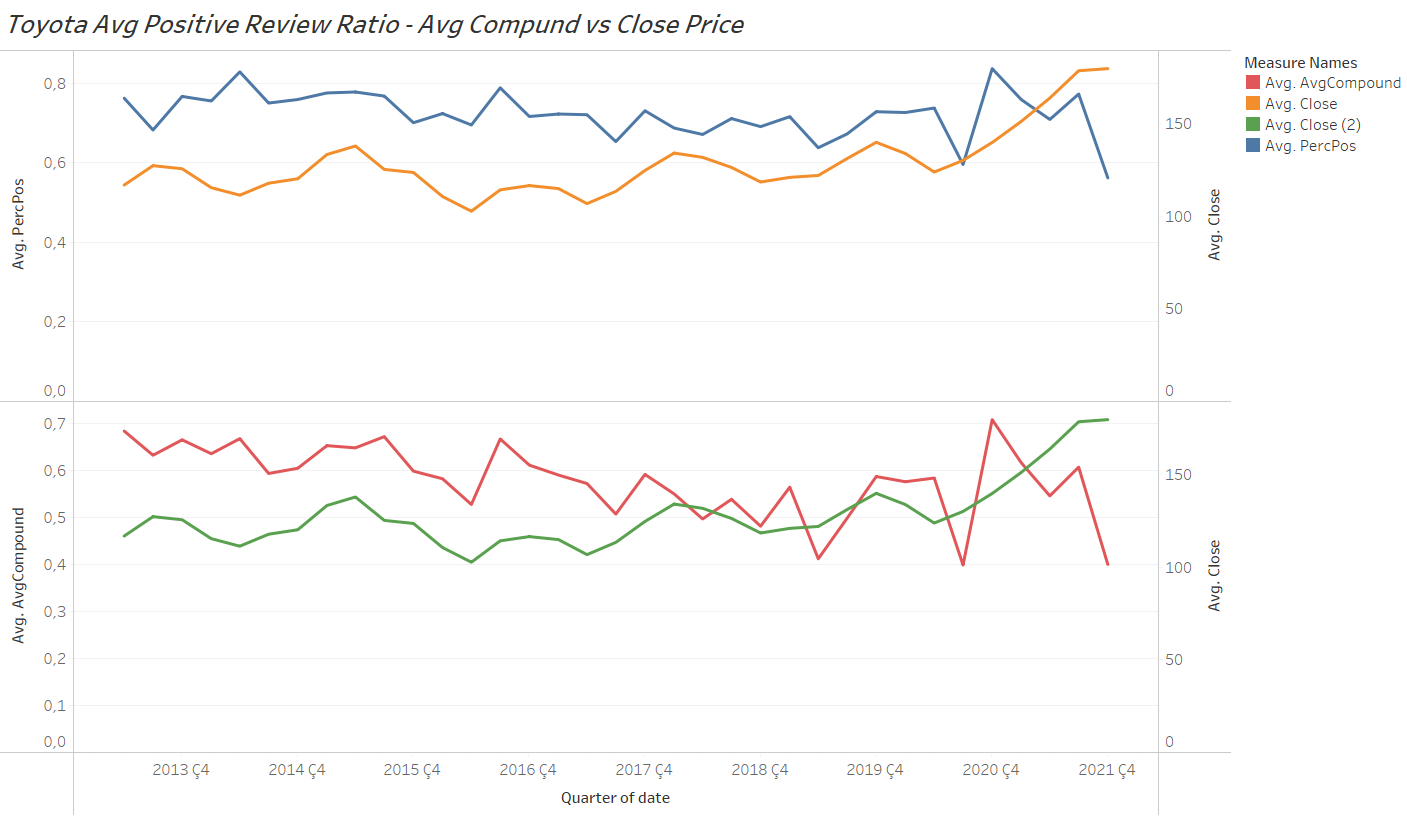
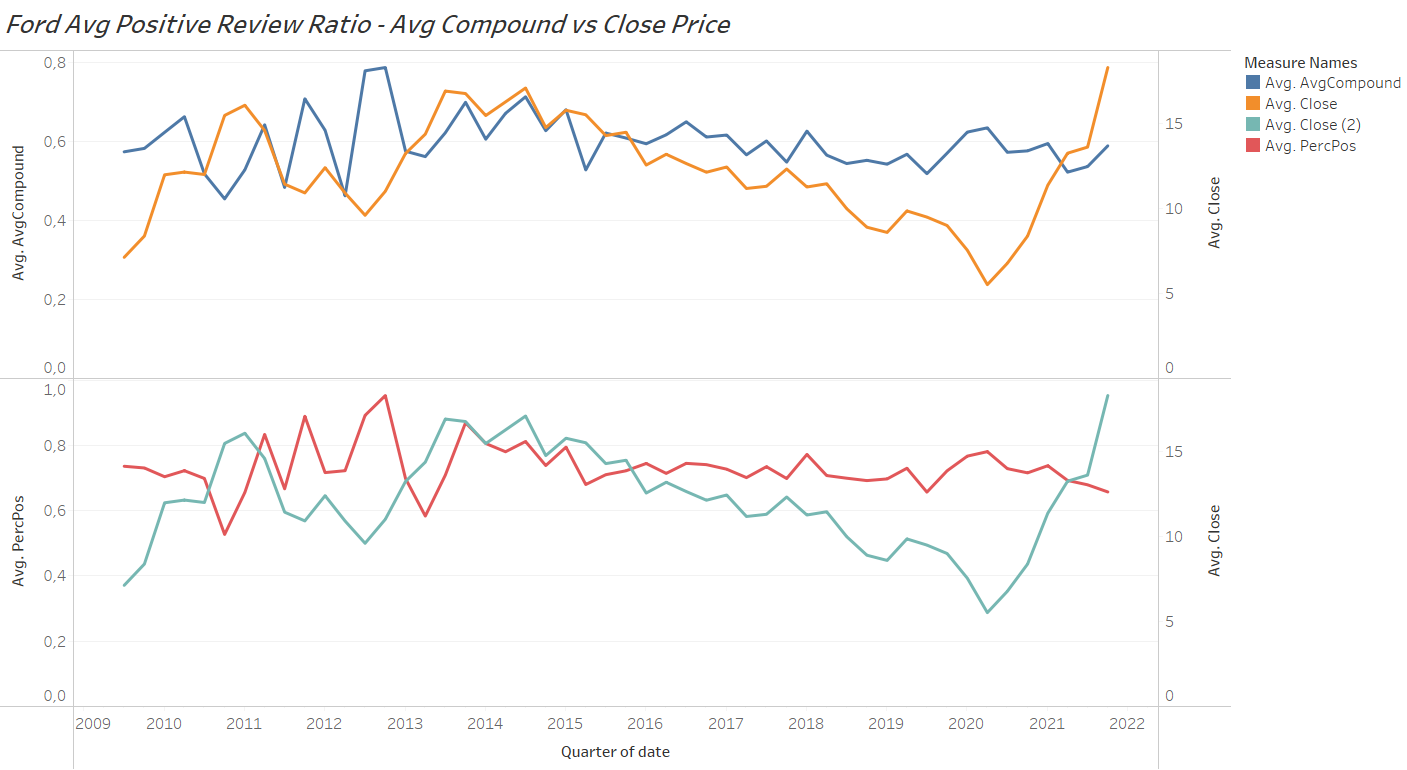


*Fig. 5.1 Word cloud of Ford consumer review Fig. 5.2 Word cloud of Toyota employee review*

**System Implementation**

To spot trends and correlations between sentiments - stock and sentiments - sales, we used Tableau to visualize results. Employee and consumer positive reviews were combined and used as one variable. Closing price, opening price and positive reviews and sales unit were plotted to analyze trends.

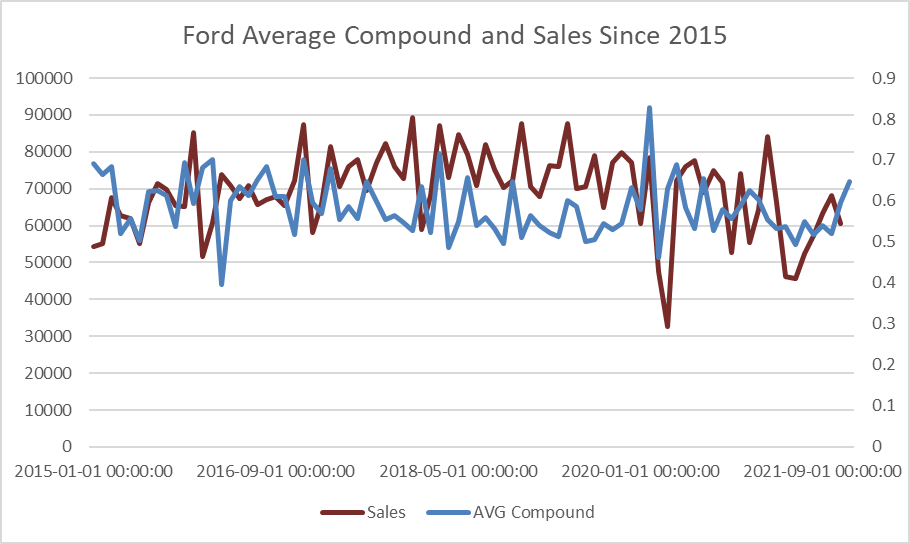
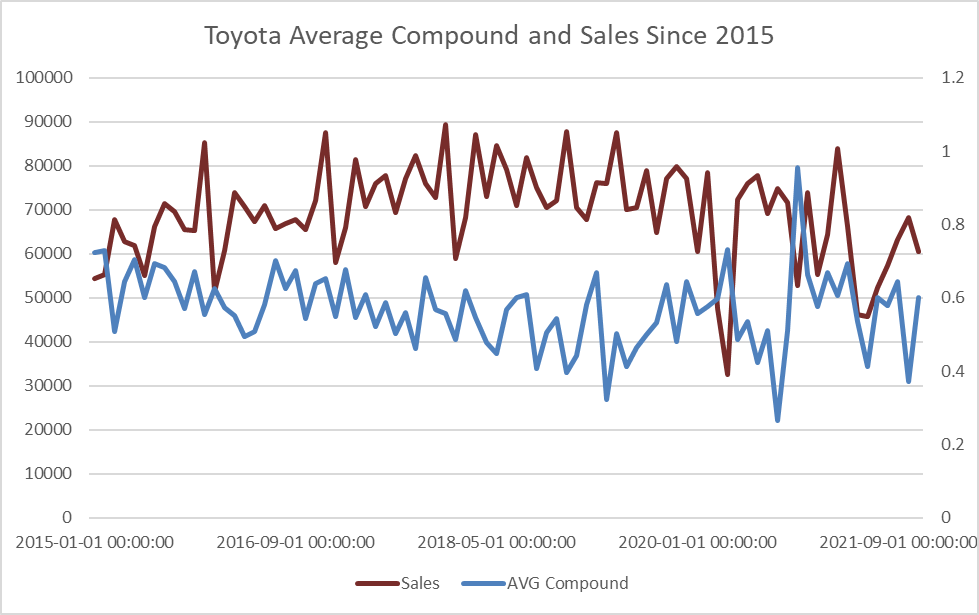
**Trend Analysis**  The analysis for Toyota between 2013-2020 revealed a fascinating pattern in its variables. When the average positive reviews dropped for a particular quarter, the closing stock price reached the apex and vice versa. This trend in sentiments and stock followed in all fourth quarters of 2013 and 2014. From 2015-2019, the average positive reviews and average closing stock started trending together i.e. stock prices were high when the number of positive reviews were more. This analysis effectively proves that stakeholders sentiments are highly correlated with quarterly stock prices for a company.

*Fig. 6.1 Toyota Average Positive Review Ratio Fig. 6.2 Ford Average Positive Review Ratio - Average Compound vs Closing Stock Price Average Compound vs Closing Stock Price*

However, for Ford, a similar yet different trend was spotted. From 2015-2020, when the average positive reviews decreased, the average closing stock price was maximum for that quarter. However, there wasn’t high consistency in growth and fall trends. The initial years from 2009-2013 clearly indicate high correlation between variables: stock price and number of positive reviews. In 2012, when average positive reviews dropped at its lowest, Ford’s closing stock price marked the highest for that year. Thus, for Ford, people’s sentiments inversely affects its stock price.

We were also interested in comparing review sentiments to sales numbers for the auto companies. To observe correlation we created two line charts from 2015-present, with monthly sales and average compound. Average compound is the indicator of whether reviews were more or less positive during that month. Here, the average compound is the combination of both employee and consumer reviews. We can see correlation between these two variables, by observing the graphs below. The data suggests that the average compound of Ford reviews has a higher correlation with sales than Toyota's.

*Fig. 6.3 Line Graph: Ford Average Compound Fig. 6.4 Line Graph: Toyota Average Compound and Sales, Monthly, 2015-2021 and Sales, Monthly, 2015-2021*

**Correlation Analysis**

If we look at the correlations for closing price we can clearly see that sentiments are somewhat correlated with closing price. However, for sales we couldn’t prove correlation numerically with sentiments. We believe that because of the lack of data on some time intervals the features which belong to sentiments fluctuate more and it affects the total correlation. In the trend analysis part we prove quarterly correlation. Not many people gave reviews between 2010-2013, and it affected our correlation matrix.

*Fig. 7.1 Toyota Correlation Matrix*

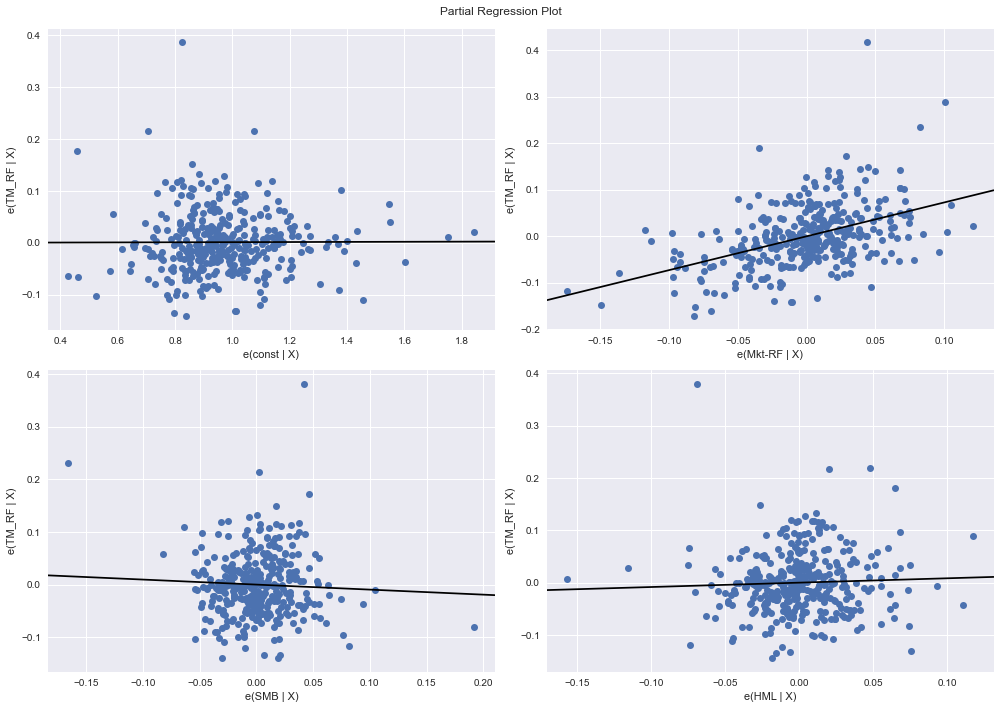
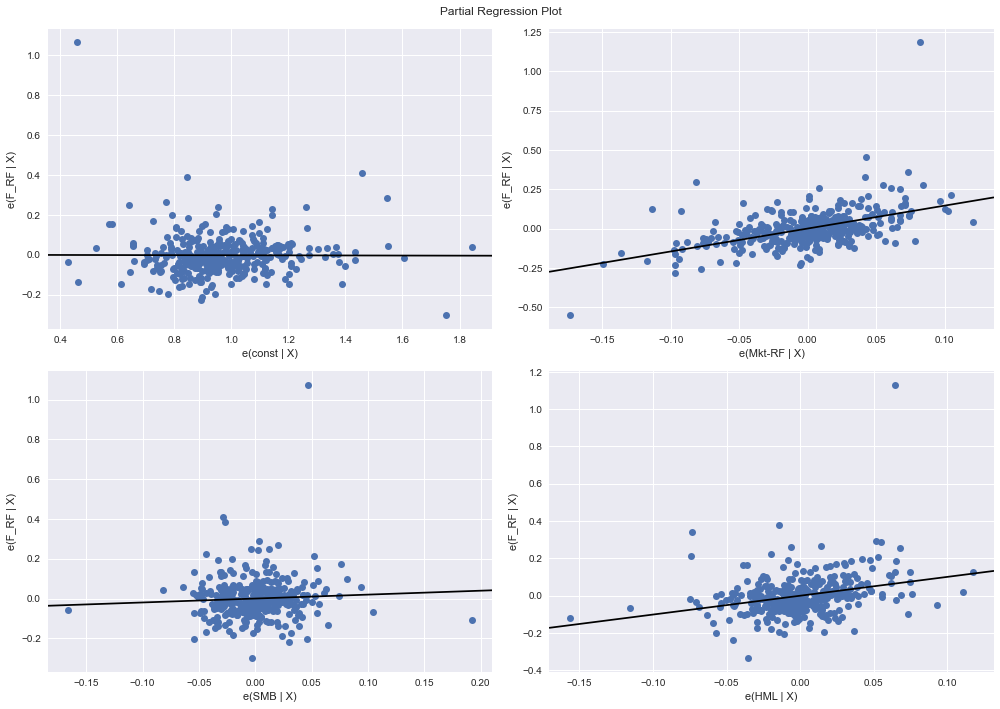
|  | **Closing Price** | **Sales** |
| --- | --- | --- |
| Previous Closing Price | 0.875 |  |
| Previous Sales | 0.492 | 0.544 |
| Mkt-RF | 0.142 |  |
| **AvgCompound** | **0.112** | **0.028** |
| **AvgRate** | **0.085** | **0.031** |
| **PercPos** | **0.082** | **0.027** |
| SMB | 0.024 |  |
| HML | 0.021 |  |

**Fama and French Three Factor Model**

**Parameters**

The Fama and French three-factor model is one of the CAPM(Capital Asset Pricing Model) for analyzing the relationship between expected return and risks. We choose Fama and French three factors to explore whether Ford or Toyota stock are worthwhile to invest in. We were concerned with three elements: Market Risk Premium, SMB(Small minus big), and HML(High minus low) to analyze the stock performance. SMB represents the companies with small market caps which generate higher returns. On the other hand, HML is the value premium, concentrating on the spread between the high book-to-market value ratio companies and low book-to-market value ratio companies.

As for doing the Fama French model, we used python to extract data from yahoo finance. We first selected the Adj closing for each stock to compare with the market index from the Ken French Data Library. Observing the data interpretation, we could notice that the β*sml* is lower than 0 (-0.0954), which means that the Toyota portfolio is predominantly large-cap stocks. Compared with the Ford portfolio, the β*sml* is higher than 0, which is 0.1975 means that the Ford portfolio is predominantly small-cap stocks. As for the HML beta coefficient, we found that both Ford and Toyota have positive values, but Ford is higher than Toyota. This situation means that both portfolios positively correlate with the value premium; however, Ford has a better performance than Toyota.



*Fig. 8.1 Fama Model Regression Line for Ford Fig. 8.2 Fama Model Regression Line for Toyota*

**Stock Price and Sales Prediction with ML**

**Preprocessing**

For the regression modeling we decided to use average ratings, average compound, sales, percentage of positive comments, previous closing price, and Fama model tree features for predicting stock prices. However, for sales we only used previous sales, average ratings, average compound and percentage of positive comments for a time interval. In order to get more accuracy we scaled the data by using MinMaxScaler() Python function between 0 and 1. Since we are working with time series, we shifted the data quarterly in order to train the model with the actual values after 4 months and called ‘predictions’. So, every future is used to predict the price or sales for 4 months later. Performed a train, test split with random test sample of 0.15.

**Model Comparison**

We compared four different models including; neural network regression(sequential model), linear regression model, decision tree regression model and random forest regression model. We compared our results by looking at mean squared error, mean absolute error and r2 values. By looking at these values the neural network model and random forest regression model performed better than the others.

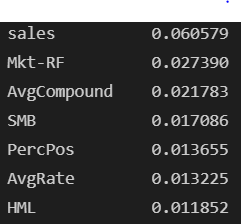
Random forest and neural network models have a similar 0.012 mean squared error and 0.089 mean absolute error with a 0.69 r squared value for stock price prediction for Ford. Also, Toyota stock price prediction errors and .r2 scores have similar values which indicates the importance of the sentiments while running the regression models.

For sales prediction we couldn’t provide strong correlation between sales and sentiments so our models accuracy significantly dropped compared to close price prediction. Our mean squared errors and mean absolute errors are raised to 0.0109 and 0.0999. Also, r^2 value has dropped to 0.0785 which indicates a worse model compared to stock price.

In summary, we managed to predict stock prices more accurately than sales for both Ford and Toyota. Also, by comparing the errors and r2 value we have decided to use random forest regression for stock price and neural networks for sales prediction.

**Important Features**

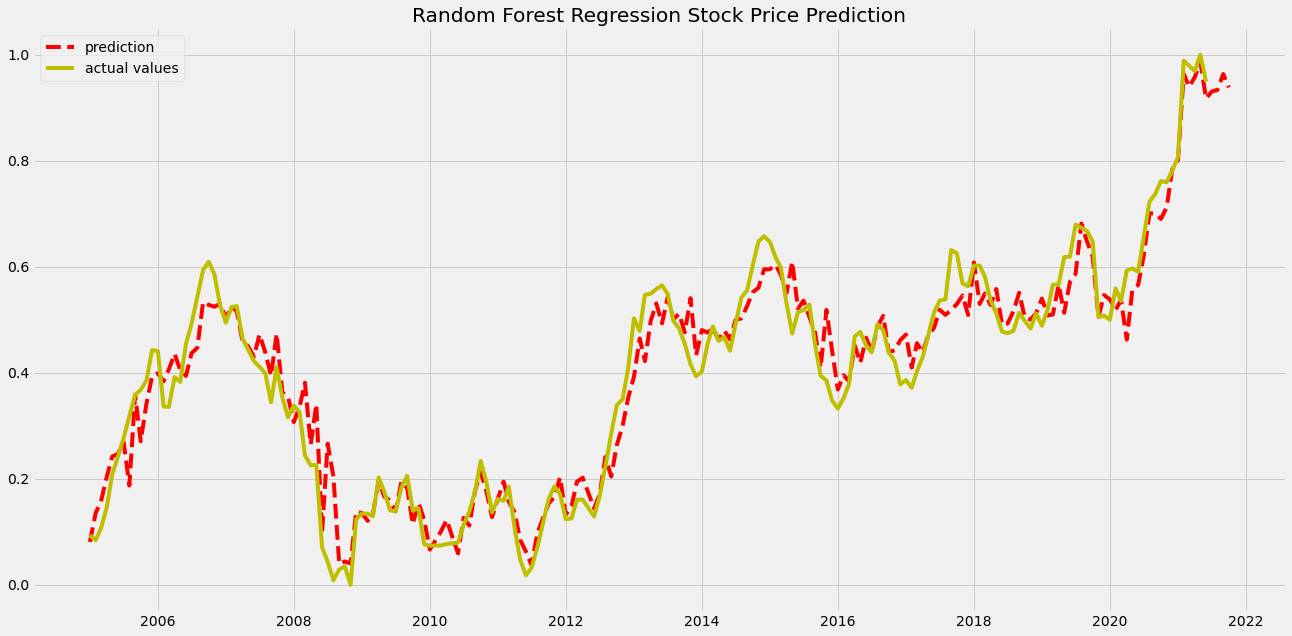
To prove the sentiment's predictive strength we looked at the future importance algorithm provided by random forests model in python Sci-kit learn library. If we look at the features sorted by importance for the Toyota stock price prediction we can say that average compounds play a significant role in prediction. SMB and HML which are fama model features are way poorer options for predicting the Toyota stock price.



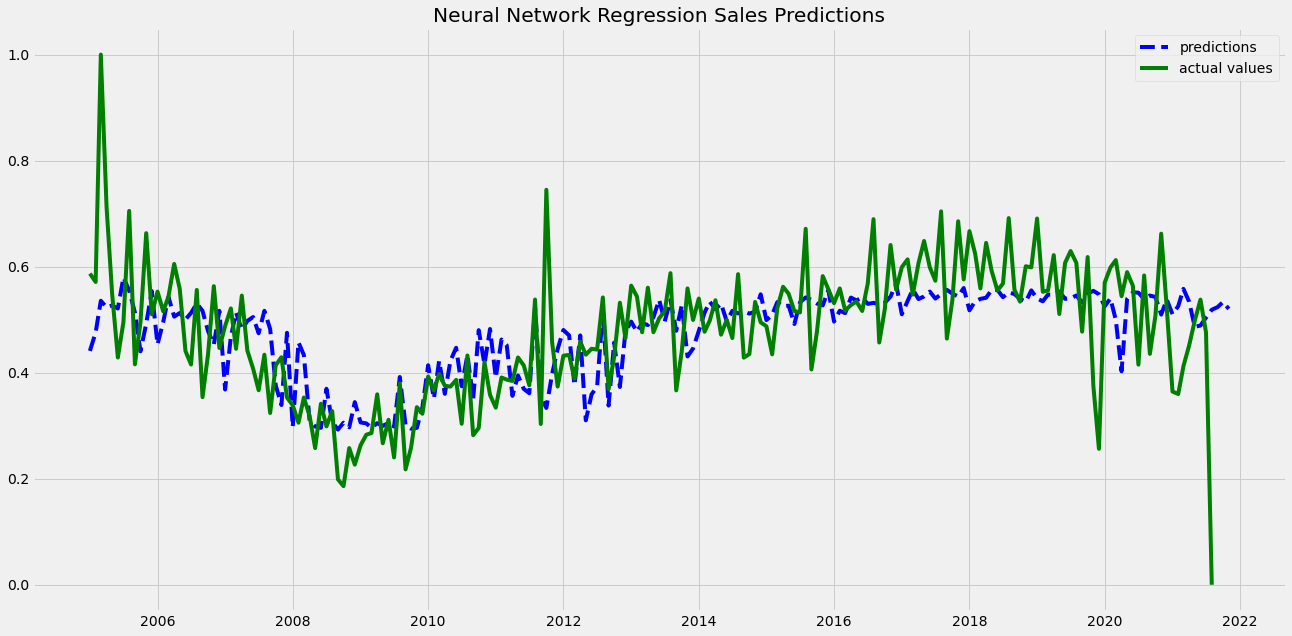
*Fig. 9 Important Features*

**Model Prediction Results**

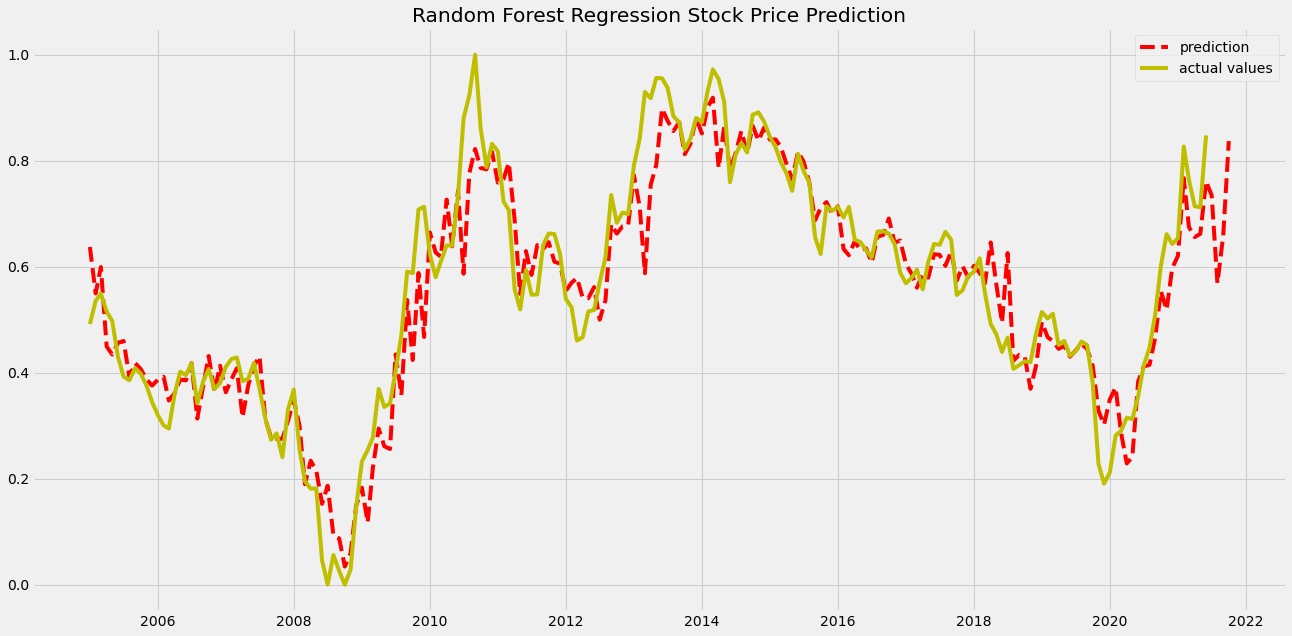
To visualize, we trained our model with the 0.85 of the data and tested with 0.15 % of it and got our mean squared errors, mean absolute errors, and r2 values. However, to visualize we make the model to do predictions for the whole data to see what is the difference between the predicted results and the actual values. Also, we made 4 month predictions for each model. For sales we used the same method to visualize, however, as we mentioned before the accuracy for sales dropped significantly.

****

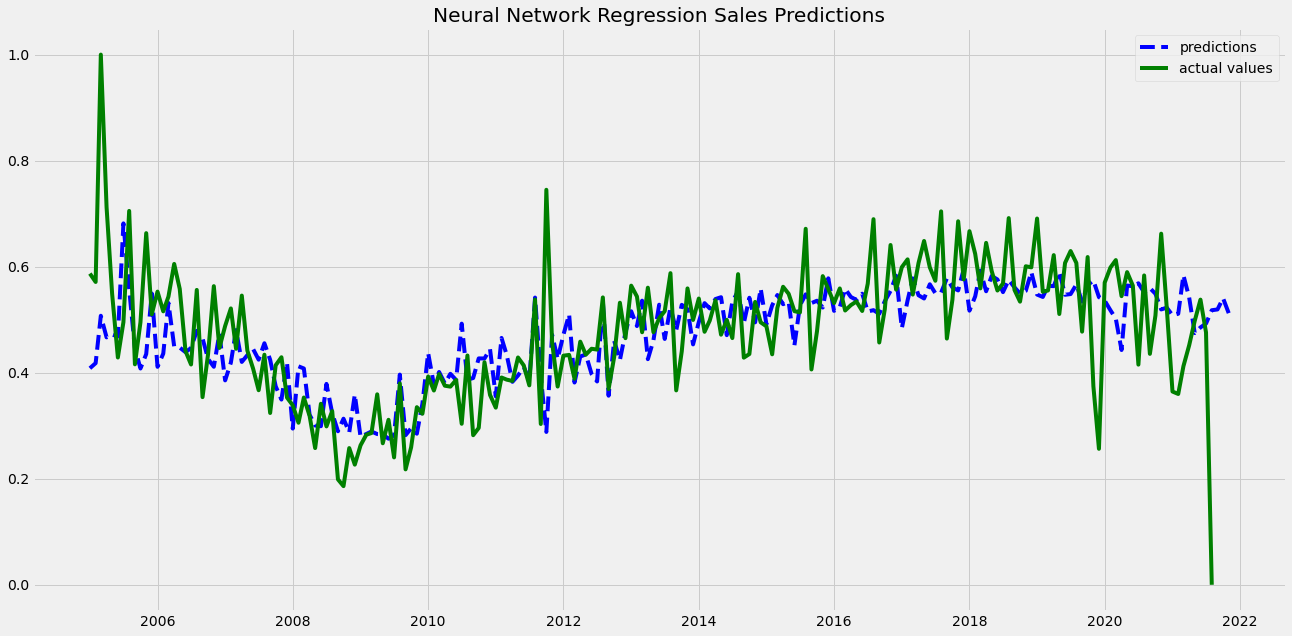
*Fig. 10.1 Random Forest Regression Model - Ford Stock Price Prediction*

****

*Fig. 10.2 Neural Network Model - Ford Sales Prediction*

****

*Fig. 10.3 Random Forest Regression Model - Toyota Stock Price Prediction*

****

*Fig. 10.2 Neural Network Model - Toyota Sales Prediction*

**Evaluation**

For Ford and Toyota, the results of our study shows a causative relationship between sentiments - stock prices and sentiments - sales unit. To analyze reviews, average compound, compound positive and negative, is an important feature. Ford and Toyota can improve their vehicle features using people’s sentiments. Negative reviews can help improve car and truck performance while positive reviews can be leveraged to further improve sales.

**Conclusion & Future Direction**

To conclude, we could predict stock market prices for both Ford and Toyota. We predict a decent increase in stock price if positive reviews increase. Our analysis indirectly proves that employees' confidence in a company affects the company’s financial portfolio. Lastly, it is worth mentioning that we did not take into consideration several external factors that directly or indirectly affect stock and sales performance. With our analysis, it can be concluded that people’s opinion about a product, service or company indeed affect the company’s financial performance. There is no direct correlation between people who work or own a Ford or Toyota vehicle are the people investing in these companies' stocks. We recommend that investors investing in the stock market should study whether the market is driven by people’s emotions to invest smartly and get higher returns.

**References**

Papers

Gonsalves, O. (2020, December 11). *A study of sentiment analysis and sales prediction: Tourism domain*. International Journal of Engineering Research & Technology. Retrieved from <https://www.ijert.org/a-study-of-sentiment-analysis-and-sales-prediction-tourism-domain>

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). *Electronic word-of-mouth via consumer-opinion platforms ...* Wiley InterScience. Retrieved from <https://www.marketingcenter.de/sites/mcm/files/downloads/research/lmm/literature/hennig-thurau_et_al._2004_jim_electronic_word-of-mouth_via_consumer-opninion_platforms_what_motivates_consumers_to_articulate_themselves_on_the_internet.pdf>