Social media sentiment analysis of listed companies and its impact on Stock Prices

Francis J Kurian

12/12/2021

Abstract

Social networking sites and news feeds generate voluminous text information that could reveal market participant's sentiments about a business. Quantifying such sentiments and understanding the relevance of such information on a company's stock price is an area of interest for many researchers. Sentiment analysis, using extensive text mining, is technically challenging to build but services like IBM Watson Application Programming Interfaces (APIs) simplifies such tasks for researchers. This paper examines how stock price movements are correlated with the expressed public sentiments (quantified into a sentiment index leveraging Watson APIs) about certain companies. To achieve this goal, a tool that integrates relevant data sources and automates the analysis was developed in R. A multiple regression model was employed to quantify the relationship between stock prices and sentiment index. In addition to the sentiment index, exchange rates were introduced as explanatory variables to capture the international exposure of the companies. Validation of the model on two listed companies in the United States demonstrate significant explanatory power of sentiments index and thus supports the hypothesis that sentiment analysis is a useful predictor of stock price movement.

Introduction

Sentiment Analysis and Stock Prices

Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, i.e. can societies experience mood states that affect their collective decision making? [1] Efficient Market Hypothesis implies that stock market prices incorporate the new information available to the market participants instantaneously. News feeds and social media interactions generate large volume of content in the form of product reviews, service levels, management decisions and many other relevant information that are sensitive to the stock prices. For example, many customers posting negative reviews about a product in social media could impact the product demand and future profitability of a company. News about supply chain bottlenecks for certain electronics parts could result in on time product availability in the market and result in sales target missed. News about political instability and changes in international business environments could impact the foreign exchange rates and affect the profitability of companies exporting to those nations. Sentiment analysis of news feed and social media data could gather such information as it starts appearing in various platforms.

It's anticipated that sentiment analysis is relevant for understanding the stock price movement[2]. Quantification of human sentiments expressed in text documents like news feed or social media interaction is a computationally intensive task.This is where Watson API services help to look at these voluminous documents to capture feelings, such as anger, sadness, joy, fear, hesitation etc. Sentiment is a higher-level classifier that divides the spectrum of emotions into positive, negative, and neutral. Once such classfication is done it is possible to build an index to track the sentiments movement on a daily basis and analyse its sensitivity to stock prices. This system leverages on an index already built by TRaiCE Inc using Watson APIs and is available as a service. In following sections we will discuss the methodology, data sources and the analysis results.

Quantification of Stock Market Behavior using Sentiment Analysis

Several researchers empirically established the links between sentiments analysis and stock prices movements. Bollen et al. conducted a conclusive study using microblogs from twitter and established a causal relationship with sentiment score and Dow Jones Industrial Average. Seki et al. [3]. Zhang et al.[4] analyzed news text to build a business sentiment index to use it as an economic index in predicting stock prices. Yahya et al.[5] used a linear regression approach incorporating sentiment analysis as an explanatory variable and concluded that introducing sentiment analysis provided significant lift in predicting the stock price movement. On similar lines, we developed a rudimentary regression-based analysis system to understand the stock price movement with respect business sentiment index (BSI).

Objectives

This paper attempts to analyze stock price behavior of two companies listed in the US stock market with respect to changes in Business Sentiment index (BSI) . BSI quantifies the human feelings captured through the extensive text-based content from news feeds and social media interactions. In addition to BSI, foreign exchange rates are added as an additional explanatory variable to capture the international exposure of the companies with the assumption that exchange rates are a better proxy when news and social media interactions are conducted using a language other than English.

Methodology and Data

Analysis System and Linear Regression Model

An R based analysis system was developed to integrate stock prices, BSI and various exchange rates at company level. The design was in a such a way that analyst could make a function call with ticker symbol and produce data diagnostics and multiple regression analysis. This study leverages on that system.

Ordinary Least Square Regression:

X_i are the k independent variables and Y is a dependent variable, for each sample of n , the value of Y_n is:

$$Y_n = \sum_{i=0}^{k} \beta_i X_{ni} + \epsilon_i \tag{1}$$

ϵ_i =Random Error Term

The OLS model in this context is formulated as:

$$StockPrice = \beta_0 + \beta_1 BSI + \beta_2 USD.EUR + \beta_3 USD.YEN + \beta_4 USD.INR \tag{2}$$

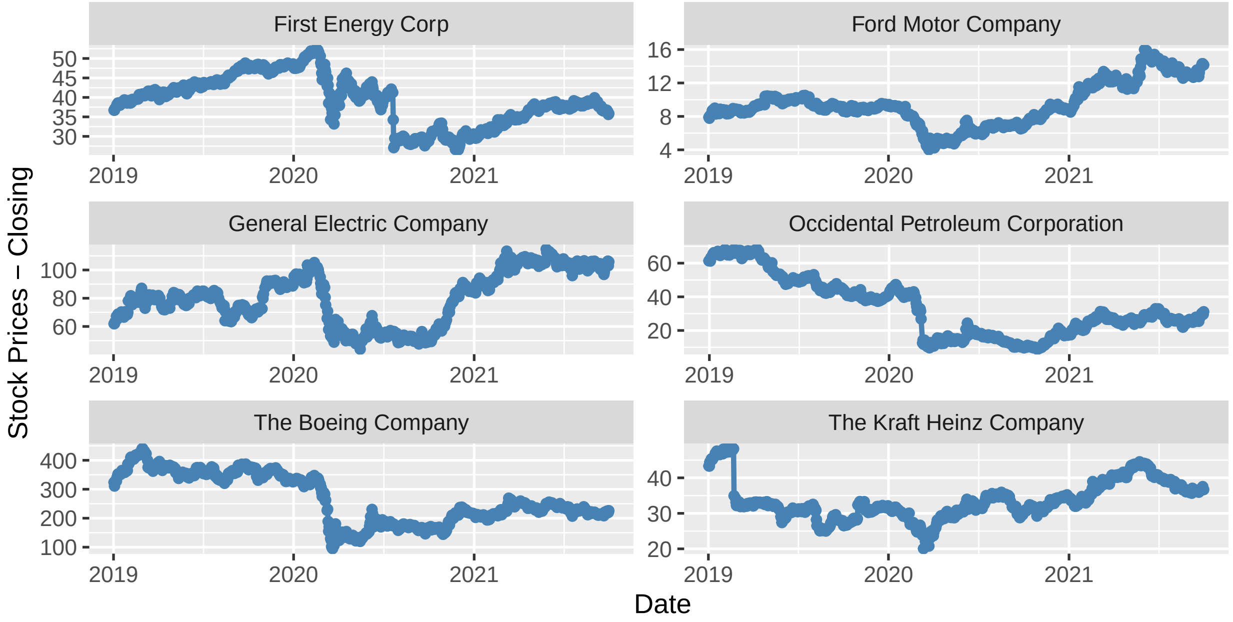
where:
StockPrice is Stock Price of the company;
BSI is Business Sentiment Index ;
USD.EUR is US Dollar vs Euro exchange rate;
USD.YEN is US Dollar vs Japanese Yen exchange rate;
USD.INR is US Dollar vs Indian Rupee exchange rate;
β_0 is the intercept, β_1,β_2,β_3,β_4 are the coefficients.

Data Collection and Profile

Daily data was collected for a period of two years (01-SEP-2019 to 31-AUG-2021) for the following data series. Daily stock prices[6], Business sentiments index[7] and foreign exchange rates[8]. The following graphs shows the data distribution for a select number of the companies including the ones we analyzed for the study.

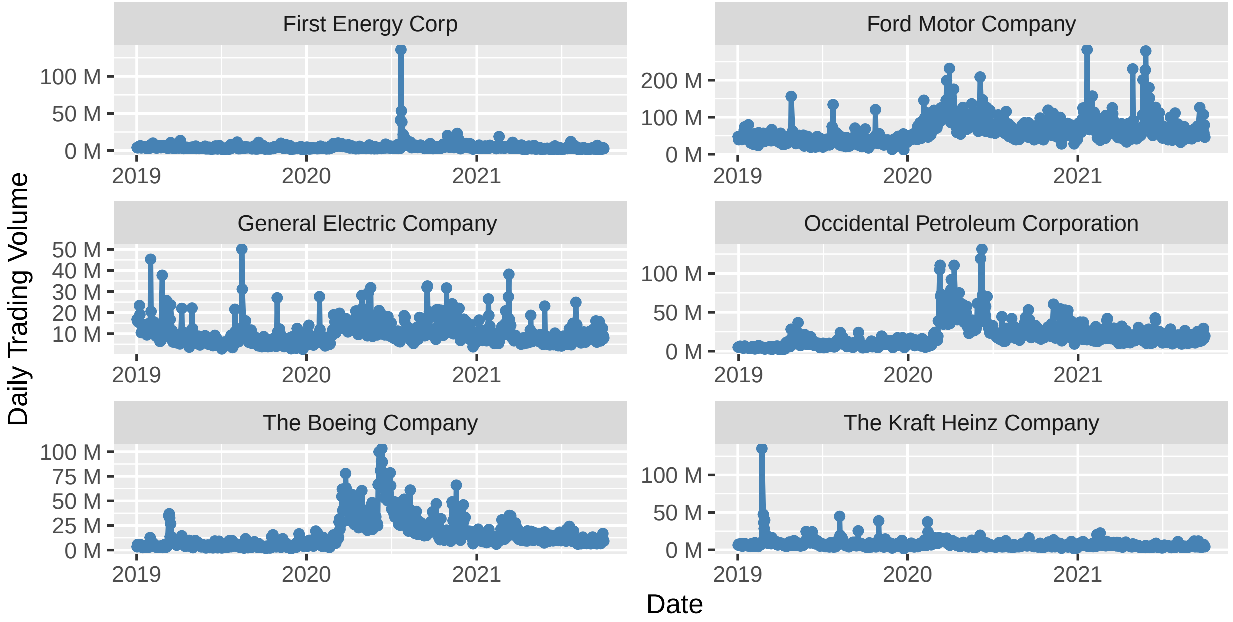
Time Series of Stock Prices by Company

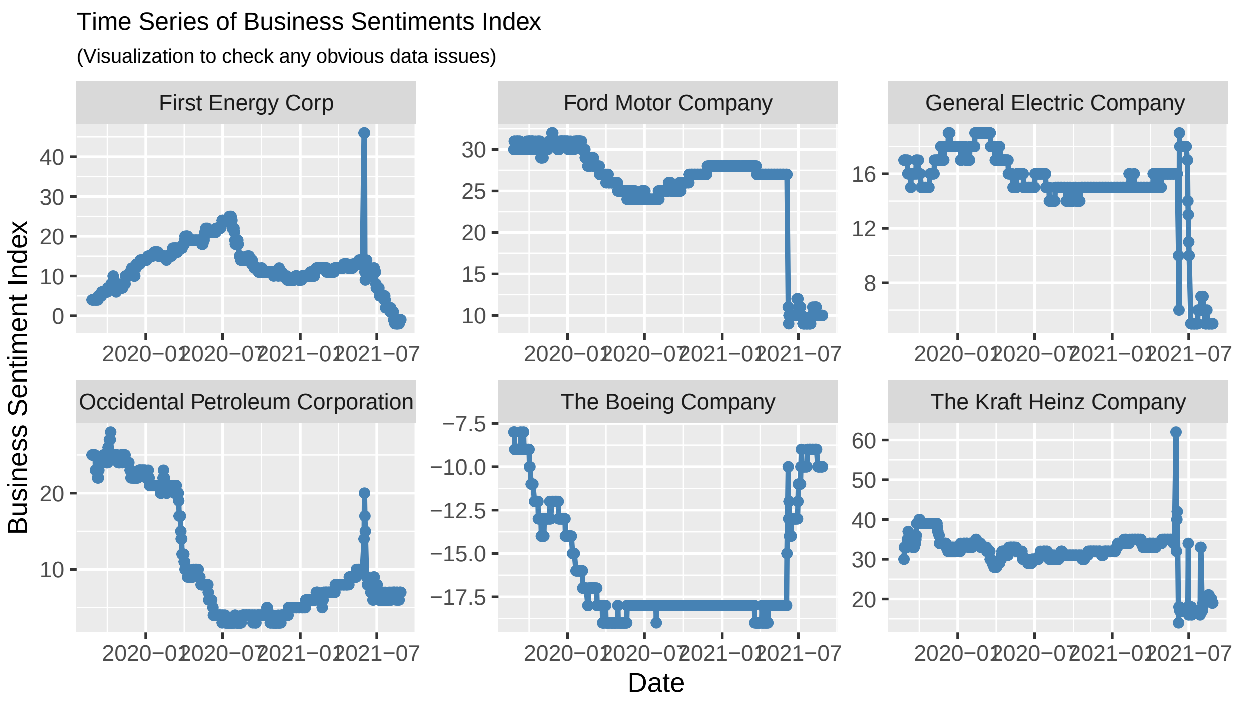
(Visualization to check any obvious data issues)



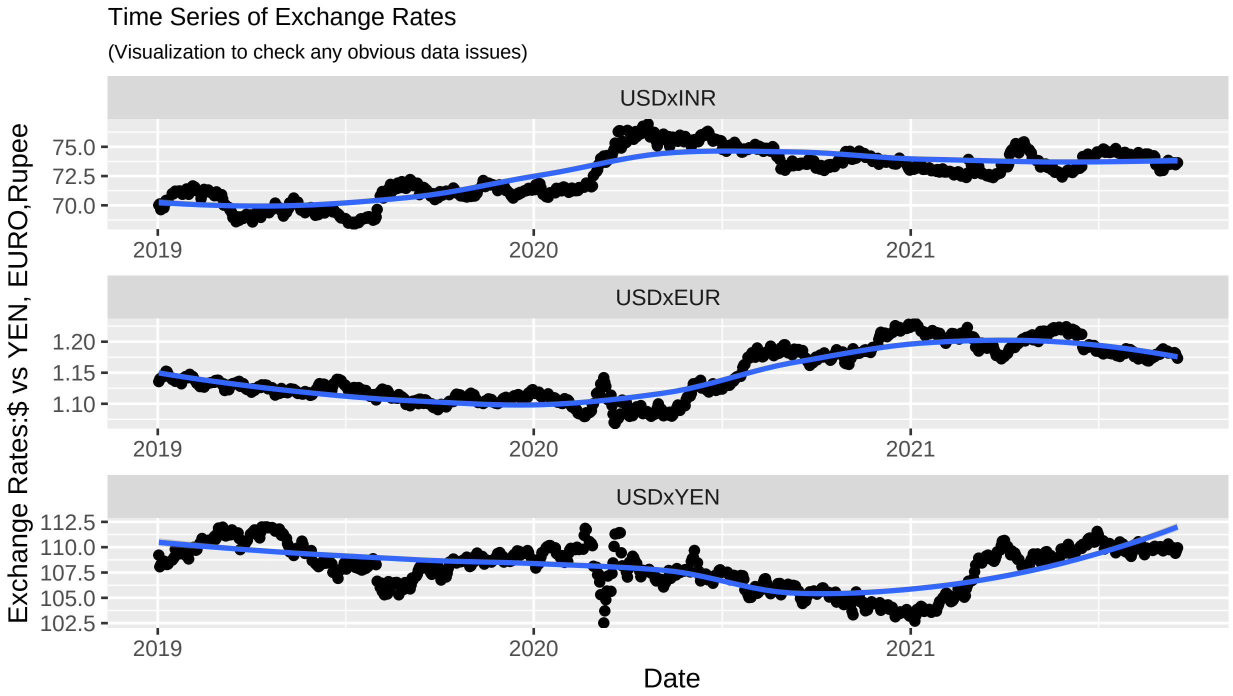
Time Series of Stock trading volume

(Visualization to check any obvious data issues)




Time Series of Business Sentiments Index

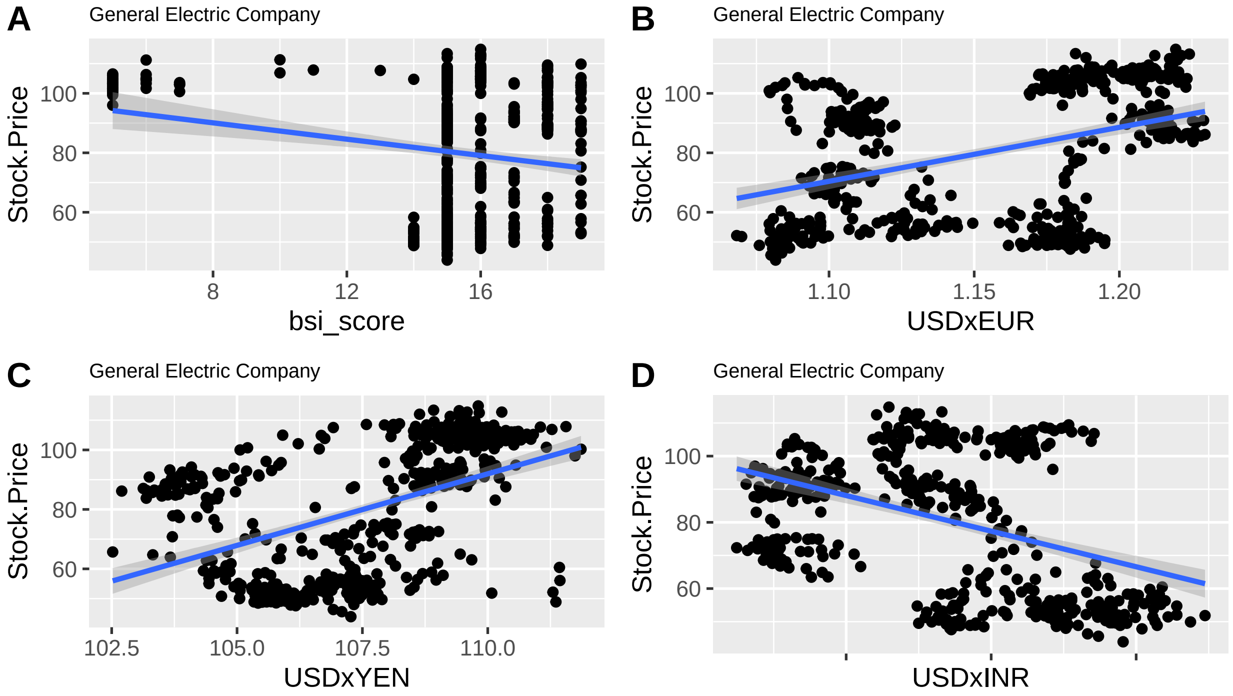
(Visualization to check any obvious data issues)


Time Series of Exchange Rates

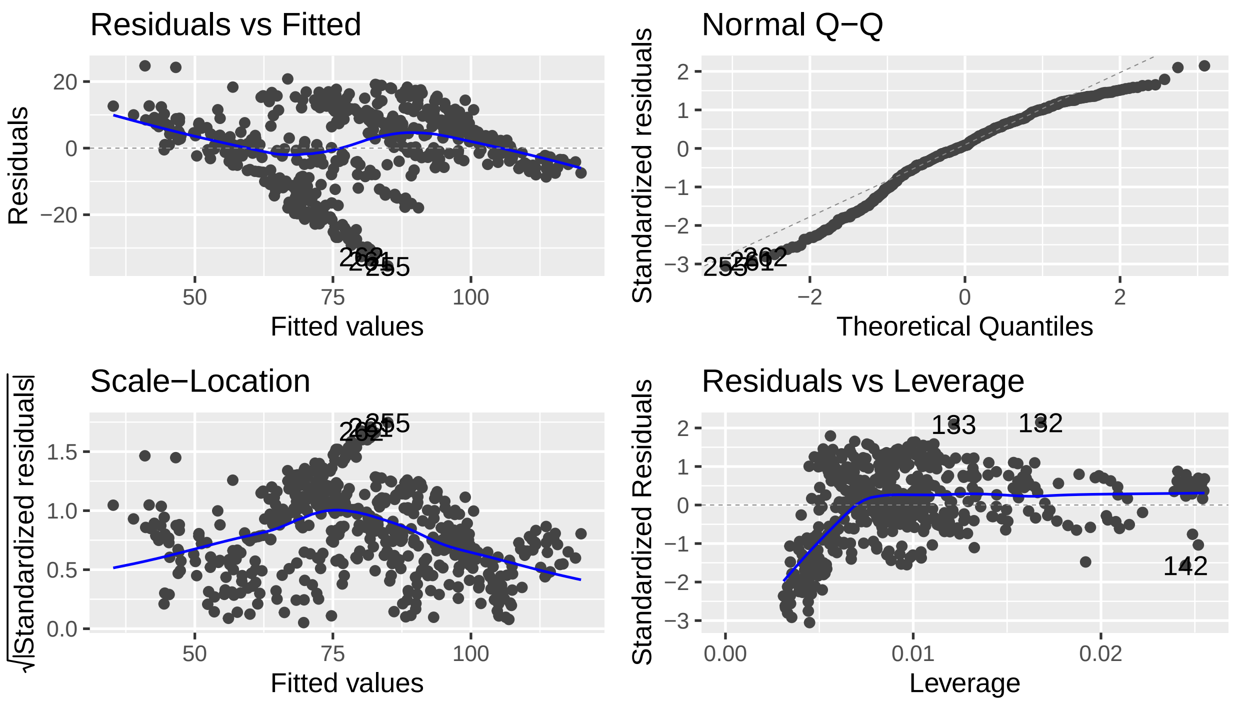
(Visualization to check any obvious data issues)

Overall, various time series data demonstrate volatility especially around the early 2020. This could be attributed to the Covid-19 uncertainties in the market at the time. BSI data clearly show some outliers that need a review and probably a correction. With these points noted, the data was used for the further analysis.

Analysis of General Electric Corporation



```
## MULTIPLE REGRESSION ANALYSIS RESULTS
## Call:
## lm(formula = Stock.Price ~ bsi_score + USDxEUR + USDxYEN + USDxINR,
##     data = df_all)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.418  -6.164   0.936   8.423  24.720
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -577.9548    50.9904 -11.335   <2e-16 ***
## bsi_score      0.1159     0.1905   0.608    0.543
## USDxEUR      300.7839    13.0462  23.055   <2e-16 ***
## USDxYEN        6.4028     0.2780  23.028   <2e-16 ***
## USDxINR       -5.1636     0.3490 -14.797   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.63 on 493 degrees of freedom
## Multiple R-squared:  0.7071, Adjusted R-squared:  0.7047
## F-statistic: 297.6 on 4 and 493 DF,  p-value: < 2.2e-16
```



```
## [1] TRUE
```

The circulatory model consisted of a racetrack that was effectively made rigid through the use of tether points with an inner lumen, two straight sections connected by two curved regions, and a moving region at the bottom of the racetrack, representing the heart tube that moved with a preferred motion. The racetrack design was used to stay consistent with past designs for easier comparison to other analyses [**Waldrop:peristalsis?**].

The elastic region had a 4:1 length:diameter ratio with the inner 3/4 of the tube length consisting of points tethered to target points, which drove the preferred peristaltic motion (Fig. **??**). The rest of the racetrack were tethered to target points which remained still throughout the simulations. Target point stiffness (k_{targ}) was chosen as 30.0 to remain consistent with the model in [**Waldrop:peristalsis?**].

The force equation used to drive peristalsis in the model is:

$$\mathbf{f}(r,t) = k_{targ}(\mathbf{Y}(r,t) - \mathbf{X}(r,t)) \tag{3}$$

where $\mathbf{Y}(r,t)$ is the preferred position of the boundary. Only the preferred motion of the boundary in each model of peristalsis differed. Each model of driving peristalsis is described below.

Opposing sine-wave peristalsis model The sine-wave model defines the motion of the boundary as two opposing sine waves:

$$y_{top,bot} = R_{top,bot} \pm A\sin(2\pi ft + 2\pi cx_t) \tag{4}$$

where f is the compression frequency, c is the compression-wave speed (held constant throughout the study at a non-dimensional speed of 3.0), A is the amplitude of the contraction, and x_t is the horizontal distance from the beginning of the prescribed motion section. The compression ratio gives the percent occlusion and is equal to $2A$. The peristaltic waves created by Eq.~4 propagated from left to right, therefore driving fluid flow counter-clockwise in the lumen of the racetrack. The stiffness of the boundary and target point stiffness ($k_{targ} = 30.0$) allowed for very little independent elastic motion in the peristaltic region of the tube.

For additional details on the opposing sine-wave peristalsis model, see [**Waldrop:peristalsis?**].

Opposing Gaussian-peak peristalsis model The pinch model defines the motion of the boundary as two sharp, Gaussian peaks, with the remainder of the boundary being free to flex with little restriction by the target points:

$$y_{top,bot} = R_{top,bot} \pm A\exp((-0.5(x_t - \gamma)/\sigma)^2) \tag{5}$$

Where γ is the position of the pinch on the x-axis of the center of the tube and σ is the width of the pinch. The pinch was advanced by altering γ depending on the time step of the simulation. For the points within the region of the Gaussian wave, the target point stiffness was chosen to be extremely stiff ($k_{targ} = 2500$) so that the target points adhered closely to the programmed waveform. Outside the peak region, the target points were tethered very loosely ($k_{targ} = 0.7$) with a spring constant about two orders of magnitude stiffer to allow for elastic interactions between fluid and the heart tube.

Analysis of Flow and Pressure Fields

Several calculations of non-dimensional fluid flow and pressure were made for each simulation in VisIt 2.9.1 [**HPV:VisIt?**] and *R* [9], similar to the analyses in [**Waldrop:peristalsis?**]. Positive flow speeds indicate fluid motion in the counter-clockwise direction in the racetrack, the same direction as the traveling peristaltic wave. All values presented in the analysis are dimensionless, and more information about nondimensionalizing values can be found in the supplemental information to this paper.

At each time step in the simulation, the magnitude of dimensionless fluid velocity was recorded and then spatially averaged across each area to find $|\mathbf{u'}|$ across four rigid sections of the racetrack: the upper position, a connecting vertical position, the inflow region (vena cava) and outflow region (aorta). The mean speeds $|\mathbf{u'}|$ were then temporally averaged to find the average flow speed across each simulation (U_{avg}). The maximum value of flow speed, $\mathbf{u'_m}$, was also taken at each time step, and the maximum of these in a simulation represents the peak flow speed (U_{peak}).

Non-dimensional pressure was also recorded for each time step of the simulation and spatially averaged at each time step near the inflow area (vena cava position) and the outflow area (aorta position) of the elastic region. For each simulation, the vena cava and aorta positions' pressures were averaged temporally to find p_{in} and p_{out}, respectively. Each inflow pressure was subtracted from the outflow pressure at each time step to find their difference, and these differences were averaged over simulation time to find ΔP.

Volume flow rate was calculated using the velocity profile across the upper position of the racetrack for each simulation. At each time step during a simulation, the velocities were sampled across the diameter of the tube to create a velocity profile across the tube. Each value was then used to calculate the volume of a concentric ring of fluid that passed through the tube during the time step based on the velocity at that position in the tube. These rings were then summed to find the volume flow rate at that time step, then these volume flow rates were averaged temporally to find the average volume flow rate of the simulation, Q.

Results

References

1. Bollen J, Mao H, Zeng X-J. 2011 Twitter mood predicts the stock market. *Journal of computational science* **2**, 1–8.
2. Mehta P, SharnilPandya, Kotecha K. 2021 Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Computer Science* **13**.
3. Seki K, Ikuta Y, Matsubayashi Y. 2021 News-based business sentiment and its properties as an economic index. *Information Processing & Management* **59**.
4. Zhang Y, Shirakawa M, Hara T. 2021 Predicting temporary deal success with social media timing signals. *Journal of Intelligent Information Systems*
5. Cakra YE, Trisedya BD. 2015 Stock price prediction using linear regression based on sentiment analysis. *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*
6. *Daily historical data.* https://finance.yahoo.com. Yahoo Finance.

7. Business sentiments index (BSI). *https://www.traice.io*

8. *Foregn exchange rate: Country data*. https://www.federalreserve.gov. Federal Reserve.

9. Team RDC. 2011 *R: A language and environment for statistical computing.* http://www.R-project.org/. Vienna, Austria: R Foundation for Statistical Computing.