Forecasting Stock Market Movements using Google Trend Searches

Melody Y. Huang, Randall R. Rojas, Patrick D. Convery

Department of Economics
University of California, Los Angeles
Los Angeles, CA 90095
m.huang@ucla.edu, rrojas@econ.ucla.edu, pconvery@ucla.edu

ABSTRACT
Previous studies have proposed using Google Trend data as buy and sell signals for trading. We find from backtesting that using these proposed trading strategies fails to consistently generate positive returns. Instead, we use a Granger causal framework to examine the potential predictive capabilities of using the relative volume of Google searches for different terms in forecasting the S&P 500. We apply Kaplan-Meier estimator to quantify the level of persistence in lagged correlation between the two series. We find that the stock market series and search trend data show high levels of persistence in their relationship with one another. We propose several generalized linear models for forecasting the probability of positive or negative directional movements in the S&P 500 using the relative search volume of different terms. We then generate directional forecasts within our backtesting period, with a success rate of 64%.

1. Introduction

There exists numerous inefficiencies within markets that result from human irrationalities. Being able to quantify human behavior–especially irrational human behavior–has become intrinsic to the goal of modeling and understanding stock markets. A popular method that has been used in past studies includes sentiment analysis across a variety of social media forms, such as Twitter and LiveJournal, in order to build predictive models [2, 6]. Other studies have focused on text mining methods, such as using topic modeling algorithms to explain market volatility, or using live news feeds to extract signals for stock movements [7, 8].

We look at large-scale user data in Google searches from the past decade to determine whether or not these patterns can be utilized in predicting market movements. Previous studies have looked at the forecastibility of unemployment, private consumption, and other economic indicators using Google search data [1, 11, 4]. However, not much has been done in the realm of forecasting financial markets using Google Trends, with the exception being Preis' 2013 study, in which he proposes utilizing search term fluctuations for a set of key words to create a portfolio trading strategy. Preis hypothesizes that the reason his proposed trading strategy yields positive returns (for some search terms) is because investor behavior is captured through the search trends of certain key terms [10].

However, we find that the proposed trading strategy fails to be plausible in implementation, and that there exists a lack of consistency in the performance of certain words across time. We utilize the same search terms proposed by Preis and find that the Granger causality between certain terms and the S&P 500 is relatively persis-

tent across time. We extract out Granger causal terms and fit a model utilizing only Granger causal terms to generate directional forecasts for the stock market.

2. Background

We begin by using the same trading strategy that Preis proposes–given an increase in search volume of a term, we choose to hold the S&P 500 index, while given a decrease in search volume, we choose to short the S&P 500 index. For each window of time, we sum up the total returns that would be obtained using this particular strategy and benchmark against the total returns that would be obtained by (1) blindly buying and selling at the beginning and end of each period, and (2) buying a share of the index at the beginning of our period and selling at the end of the period (a buy-and-hold strategy).

We scrape the weekly Google trend data and fix our time horizon from January 2006 to February 2017. We find that in agreement with the original study, blindly running this strategy through this period yields a reasonable number of terms that outperform both benchmarks (for a full list of terms that outperform the benchmark, see Table 4). However, actually implementing the strategy in real time yields several problems. For one, extending a given period for which we compute the cumulative returns from the strategy also shows that as we increase the time horizon, it becomes increasingly more difficult to outperform the market. An additional complexity is the potential to achieve high returns by reversing the position (i.e., given an increase in search volume, short the index, while given a decrease in search volume, hold the index). There fails to be a discernible relationship between terms that perform best with the stated strategy (i.e., 'dividend', 'gains', 'portfolio'), and terms that perform best with the reversed strategy ('stock market', 'dow jones').

More problematically, the strategy proves inconsistent in its performance across time. Taking a rolling window of 50 weeks across the 10 year time horizon, we simulate trades according to the proposed strategy. We rank each term by its relative performance with the rest of the terms in that given window. Therefore, a term given a rank of 1 will have performed best, while a term with a rank of 89 will have performed the worst (see Figure 1). Looking at the range of ranks that the terms could take on, we find that the average range of placements hovered around 74.4. Given that there are a total of 89 terms, this means that the average term would jump between anywhere from the bottom of the list in terms of performance to the very top, depending on which time period we were in. The general lack of consistency in terms of how the strategy performs, and the dependency on a term's performance upon the particular time period, bring into question how plausible it would be to actually implement this strategy.

Therefore, our focus is to determine whether signals extracted from Google Trend data exhibit consistency overtime, and whether or not these signals can be then successfully utilized to forecast stock market movement in real time.

3. Forecasting the S&P 500

We focus on forecasting the S&P 500 index to see whether or not the signals from the Google Trend data is able to uncover information about the markets as a whole. We extract Granger causal terms and use those terms to fit a generalized linear model to

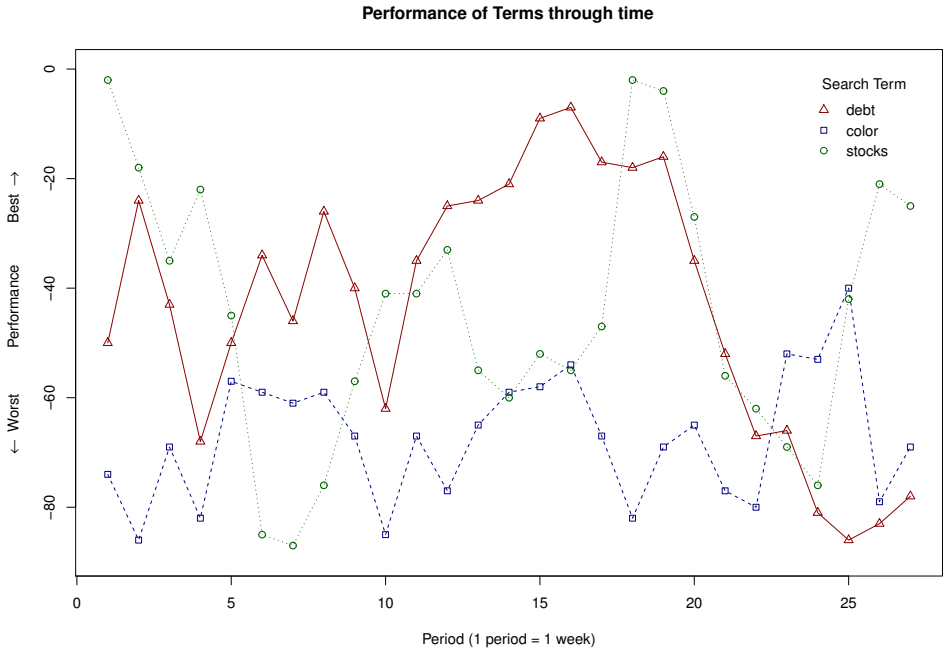

Figure 1. We use a rolling window of 50 weeks across a 10 year time horizon. For each window of time, we rank the terms by the performance achieved using the proposed strategy. A rank of 1 denotes that the term performed best. For ease of understanding in plotting, we use $-1 * rank_t$ for the y-axis values. We plot three sample terms, and see that there is a great deal of fluctuation in the performance of the trading strategy using various terms, depending on what period we are in.

forecast the directional movements of the S&P 500 at a weekly level.

3.1. *Persistence in Granger causality*

We begin by finding all Granger causal terms with adjusted closing prices of the S&P 500. (See Table 2 for a full list of statistically significant terms.) However, a concern that arises is that erms that appear Granger causal now may not necessarily have been Granger causal in the past, or be Granger causal in the future.

Therefore, we test for the persistence of Granger causality using both a moving window and a recursive backtesting scheme. In both, schemes over half the terms showed persistence in their significance levels. To measure the likelihood of a change in the Granger causality significance in a given term, we examine only the terms that showed fluctuations in significance, and use traditional survival analysis methods to quantify persistence in causality (or lack thereof). In particular, we use the Kaplan-Meier estimator, a non-parametric statistic used to estimate survival functions. Typical uses of the estimator involve estimating the survival rates of different groups of patients, or determining how long certain plants stay alive [3].

The Kaplan-Meier estimate of a survival function is as follows:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1, \\ \prod_{t_i \leq t}[\, 1 - \frac{d_i}{Y_i}] & \text{if } t_1 \leq t. \end{cases}$$

Y_i represents the total number of individuals at risk at time t_i (defined more formally by the number of individuals who 'die' at any time beyond time t_i), while d_i represents the total number of individuals that have died off. Within our framework, individuals correspond to terms, and persistence in Granger causality represents survival. More specifically, we use a dummy variable to indicate when the series encounters a change in significance level, such that all series begin with an assigned value of 1 (irregardless of significance). When the term either becomes Granger causal at any given point in time, or becomes no longer significantly Granger causal, we assign a value of 0. In this manner, we can measure the "survival" of a term's significance persistence through time.

We find that the estimated survival rate of the terms appears to be relatively persistent for about 100 periods, which translates into 100 weeks (roughly 2 years worth of time) (see Figure 2). While we have no means to ensure that the terms that are Granger causal today will persist to be Granger causal next week, this is an observably long persistence that can be measured from historical data, and we therefore continue with our analysis.

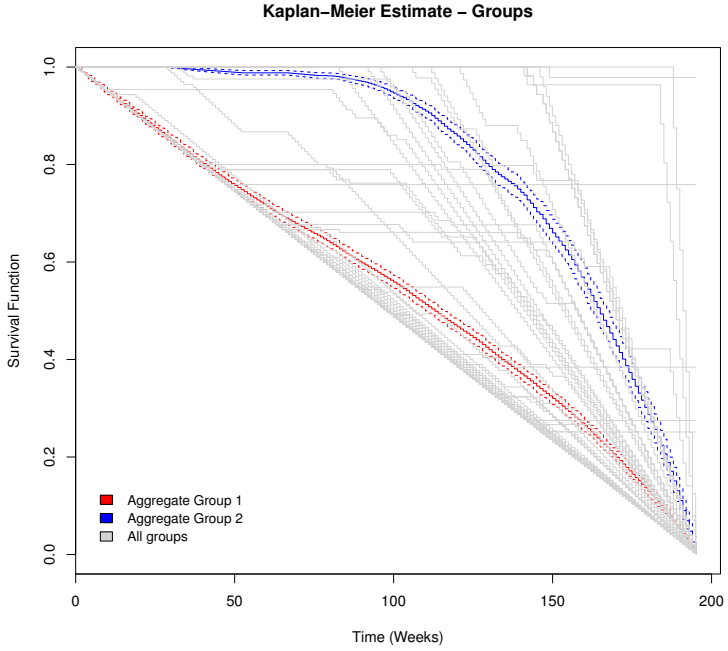

Figure 2. We use the Kaplan-Meier estimate to gauge 'survival' of a term's Granger causality, and the persistence in its statistical significance. Above, Aggregate Group 1 refers to the group of terms that begin the period as Granger causal with the series, while Aggregate Group 2 refers to the group of terms that begin the period as not Granger causal. We find that there exists more persistence in the significance of causality for terms that begin the period as Granger causal.

3.2. *Directional Movements*

To check the capability of the Google search terms to forecast directional movements, we decompose the S&P 500 index into the individual movements between periods. Each period is equivalent to one week, so we are essentially looking at whether or not

the index has moved up or down, relative to where it was at seven days ago. We use a dummy variable (δ_t) to indicate this upward or downward movement:

$$\delta_t = \begin{cases} 1 & \text{if } sp500_t > sp500_{t-1} \\ 0 & \text{if } sp500_t \leq sp500_{t-1} \end{cases} \tag{1}$$

We test for Granger causality between all 89 terms and this constructed series. Subsetting terms that are significant only at the 5% level, we are left with 10 remaining terms (see Table 3 for full list).

Using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), we fit a vector autoregressive (VAR) model of lag four:

$$P(sp500_t > sp500_{t-1}) = \alpha_t + \sum_{j=1}^{10} \Big(\sum_{i=1}^{n} \varphi_i^{(j)} term_{t-i}^{(j)} \Big) + \sum_{i=1}^{n} \gamma_i \delta_{t-i} + \varepsilon_t, \tag{2}$$

where $n = 4$, and $term_t^{(j)}$ represents the time series consisting of the relative search volume for the j-th term.

From the impulse response function of the VAR model, we isolate out the response from our directional series to a unit shock from each individual search term series. We find that unit shocks in the volume of various search terms will lead to a spike or decrease in the probability of the S&P 500 index moving upward in future periods (see Figure 3). Additionally, for the terms 'cash', 'dividends', 'revenue', 'transaction', and 'gain', there exists a positive overall response in the probability of an upward movement in the index, while for the terms 'bubble' and 'war', there exists a negative response in the probability of an upward movement in the index. The terms 'stocks', 'returns', and 'society' show some fluctuations in terms of whether the response is positive or negative. The first set of terms ('cash', 'dividends', 'revenue', 'transaction', and 'gain') all have a positive connotation behind them, while the second set ('bubble' and 'war') are much more negative in nature. The last set are all relatively neutral. This is consistent with what we would intuitively hypothesize to occur. As the search volume for terms with positive sentiment towards markets increases, the likelihood of the S&P 500 increasing should also increase. Likewise, as the search volume for terms with negative sentiment towards markets increases, we would expect the markets to respond accordingly. To examine whether this could be the result of the markets driving public sentiment (i.e., the volume of searches for various terms is a response to the markets changing), we look at the results of the Granger causality test, for which we check whether the markets Granger cause search volume changes. There exists no significance at $\alpha = 0.05$, and we therefore conclude that market movements do not Granger cause search volume changes for these ten terms.

We fit a series of generalized linear models, using penalized maximum likelihood, with several different regularization paths. In particular, we looked at a ridge regression in hopes of minimizing coefficients of correlated predictors, a lasso regression, which picks one predictor and 'discards' the rest, and then an elastic-net regression, which serves as a combination of the two [5]. For simplicity, we choose $\alpha = 0.5$ for the elastic-net regression, as we assume there to be existing correlations between our groups of terms. Lambda is set at a value that yields the minimum mean cross-validated error, and coordinate descent is utilized to solve the objective function.

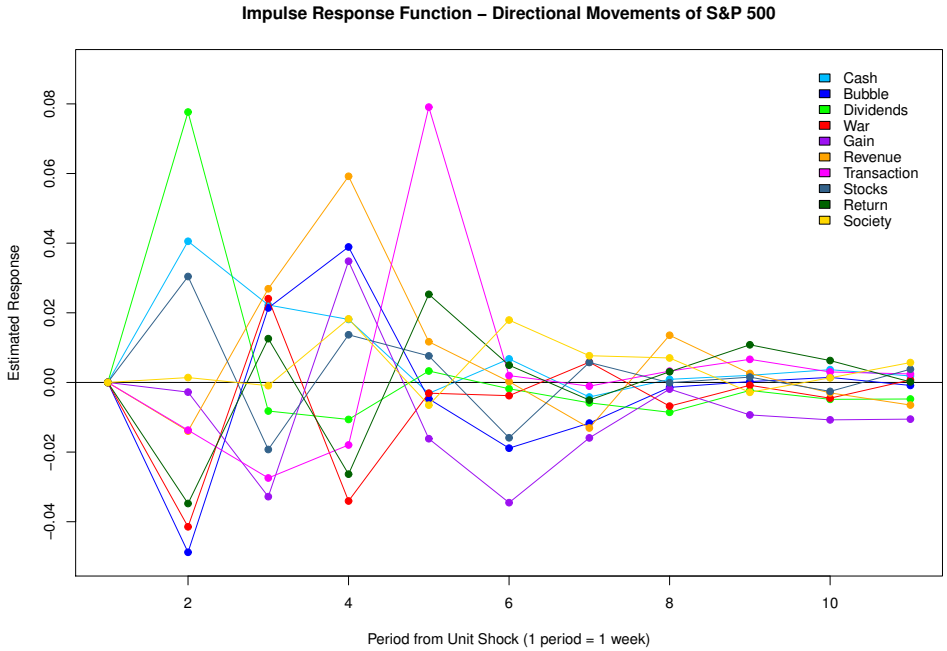

Figure 3. We estimate the effect of a unit shock to the volume of searches for each individual term on the probability of an upward movement in the S&P 500 index. For the most part, after 6 periods (which translates into 6 weeks of time), the effect dissipates. However, prior to then, there appears to a great deal of fluctuation in the directional series in response to an increase in search volume. Unit shocks to different terms yield different effects that peak at different future periods.

The general objective function takes on the following form:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} -\left[\frac{1}{N}\sum_{i=1}^{N}\delta_i\cdot(\beta_0+x_i^T\beta)-\log(1+e^{(\beta_0+x_i^T\beta)})\right]+\lambda\big[(1-\alpha)||\beta||_2^2/2+\alpha||\beta||_1\big], \tag{3}$$

where x_i are our independent variables (more specifically, lagged values of the search patterns of certain search terms). To identify the optimal lambda value, we perform k-fold cross-validation, using $k = 10$ and binomial deviance as the loss function.

We first test each regression on the full set of lagged variables that were found to be Granger-causal with directional movements, with the maximum lag set to 4. We also limit our set of variables in two different ways in order to minimize the risk of collinearity. The first way uses only the statistically significant lagged values from our initial VAR model. The second way uses Mallows C_p to determine the included variables. The first method leaves us with 15 variables, while the second leaves us with 9. There exists some variability in the results because the k-folds within the k-fold cross validation are chosen randomly; therefore, we perform 1000 runs and average out the errors across these 1000 iterations per model. For each iteration, we compute the percentage of accurately classified periods within our testing set, as well as the area under the receiver operating characteristic curve (AUC) in order to estimate the expected rate of true positives in our out-of-sample testing (see Table 4 for full results).

We find that utilizing a ridge regression with the truncated data set including lags determined to be statistically significant from our initial VAR model achieves the

highest accuracy rate, with 63.75% accuracy, and an AUC of 58.214%. Therefore, we expect the proportion of true positives using this model to be around 58% (Figure 4).

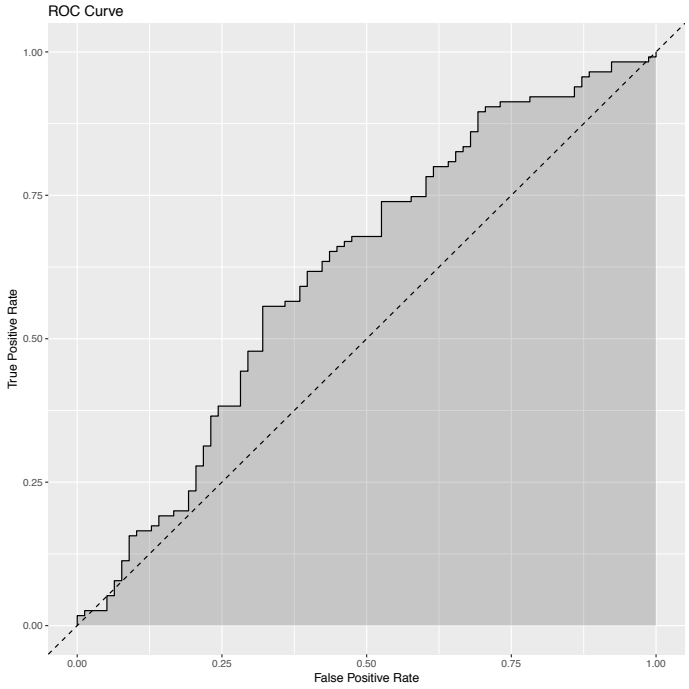

Figure 4. We graph the ROC curve for our ridge regression, using only values deemed statistically significant from our initial VAR model. We estimate the expected proportion of true positives using our proposed ridge regression over 1000 iterations. We find that the average AUC within these1000 trials was around 58.214%.

We also test the viability of using a support vector machine (SVM) to categorize directional movements. However, while performance using the SVM was often times comparable to the linear models when including all variables, it failed to outperform the aforementioned ridge regression.

We benchmark the performance of our ridge regression to a univariate, basic autoregressive integrated moving average (ARIMA) model fit to the series containing the directional movements of the S&P 500 index. Using the same partition of training and testing set, we find that the model achieves 59% accuracy in the 193 testing set; however, this is somewhat skewed, as looking at the forecasted values, the ARIMA model predicts an upward movement in the S&P 500 index at every period. Because there are 115 total actual upward movements, the model forecasts all 115 of these movements correctly. However, this is not particularly useful from a forecasting perspective, because the model fails to account for any potential directional changes in the index.

Therefore, a more accurate measure for the model's success would be to count the number of "shifts" (i.e., changes from 1 to 0, or vice versa, from 0 to 1) and see how many times the forecasting model is able to account for them. From our testing set, there are a total of 101 shifts. The ARIMA model accounts for none of the shifts. However, the generalized linear model is able to forecast 36 different shifts, 21 of which are forecasted correctly. There is obviously improvement needed to be made, but there is demonstrable potential for utilizing a model similar to this to forecast

shifts in movements within the index in real time.

4. Discussion

In conclusion, we find that previously proposed trade strategies using Google search data fails to continually achieve positive returns within our backtesting period. However, we estimate a large degree of persistence in the Granger causality between certain terms and movements in the market.

Our study goes beyond previous work done in this area by proposing an actual model for which we can forecast directional movements in the S&P 500. We find that using relatively simple linear models can allow us to successfully forecast directional movements in the S&P 500 with over 60% accuracy. Additionally, the model can correctly account for some of the shifts within the index. Therefore, we conclude that Google search data can be used as potential signals for stock market movement.

A natural extension of this study would be to increase the number of terms we are including within our model. Currently, we are using 89 terms taken from Preis' original study. However, there is no guarantee that these terms are the 'best' terms include within our model. Some data scraping shows that we can easily extract out the top five related search terms to each of our 89 terms, for a total of 445 new terms. Hypothetically speaking, the inclusion of more terms may help the performance of the various linear models proposed, as some of the dynamics within the stock market movements may not be fully captured by our existing terms. Additionally, much of the existing work done in this area has kept behavioral data from various sources relatively discrete. An interesting study may be to pool sentiment analysis data from Twitter and the existing Google Trend data to create a model that captures different dynamics within the behavioral inefficiencies.

Other extensions of this study include changing the frequency of the data. Google Trend data actually exists at a monthly, weekly, daily, and even hourly rate. Thus far, we have only tested weekly data. Our preliminary analysis shows that utilizing Google Trend data at a daily week yields very few, if not zero, terms that Granger cause the stock market (and vice versa). This decoupling in relationship at a finer aggregation level of data may be interesting to look into.

Lastly, future analysis should be done to test different trading strategies based on the forecasts of our proposed model. Depending on the forecast produced by the regression, one could hypothetically buy or sell depending on whether the forecast for the week ahead shows an upward or downward movement in the index.

Search Terms Outperforming Benchmark - Preis

Search Term	Return	Excess Return	Absolute Excess Return
dividend	2.11	1.28	1.28
gains	1.42	0.59	0.59
portfolio	1.35	0.51	0.51
fond	1.34	0.50	0.50
rare earths	1.11	0.27	0.27
risk	0.95	0.11	0.11
cash	0.93	0.09	0.09
bubble	-0.86	-1.70	0.02
bonds	-0.88	-1.72	0.04
hedge	-0.96	-1.79	0.12
housing	-0.96	-1.80	0.12
derivatives	-0.97	-1.81	0.13
consume	-1.05	-1.89	0.21
energy	-1.05	-1.89	0.21
fun	-1.11	-1.95	0.27
kitchen	-1.11	-1.95	0.27
stock market	-1.12	-1.96	0.28
dow jones	-1.46	-2.29	0.62

Table 1. List of terms that yield significantly positive or negative results under Preis' proposed trading strategy on our period of January 2006 to February 2017. Only terms that outperform the market during this period are included. Additionally, terms that would have outperformed the market should the strategy be flipped are also included. Excess return is computed by subtracting the return from the benchmark (i.e., a buy-and-hold strategy over the last decade period), and absolute excess return takes the absolute value of the return achieved using the strategy and subtracting that from the benchmark.

Granger Causal Terms with S&P 500 Price Series

Terms	P-Values	Lags
bubble	0.01	2.00
consumption	0.00	2.00
debt	0.00	2.00
hedge	0.01	2.00
revenue	0.02	2.00
finance	0.03	3.00
gains	0.04	3.00
headlines	0.03	3.00
returns	0.02	3.00
bonds	0.00	4.00
stock market	0.01	4.00
dividend	0.01	5.00
markets	0.03	5.00
transaction	0.02	5.00
dow jones	0.01	6.00
nasdaq	0.04	6.00
money	0.01	9.00

Table 2. Results from testing Granger causality between the search trends of the 89 different terms. We provide a list of the terms that are statistically significant Granger causal with the S&P 500 price series at an $\alpha = 5\%$ level. In other words, lagged values of these terms may be potentially helpful in explaining the S&P 500 prices in the present.

Granger Causal terms with S&P 500 Directional Series

Terms	P-values	Lags
cash	0.01	1.00
bubble	0.03	2.00
return	0.02	2.00
stocks	0.03	3.00
gain	0.02	4.00
transaction	0.04	5.00
dividend	0.00	6.00
revenue	0.02	7.00
war	0.04	7.00
society	0.04	10.00

Table 3. Results from testing Granger causality between the 89 terms and the directional series of the S&P 500. Extracting only terms that are Granger causal with the directional movement of the index, we are left with 10 terms. $\alpha = 0.05$ is used as our significance level.

Regression Models and Corresponding Results

	Percentage Accurately Classified			AUC		
	(1)	(2)	(3)	(1)	(2)	(3)
SVM	59.58%	55.95%	57.06%	51.789%	53.712%	51.54%
Ridge Regression	59.59%	63.75%*	59.14%	57.572%	58.214%	53.57%
LASSO	53.50%	60.096%	53.84%	50.915%	51.586%	53.84%
Elastic Net	54.97%	60.25%	51.65%	51.789%	51.685%	51.65%

Table 4. Results from out-of-sample backtesting. (1) Full Data Set, (2) VAR selected variables, (3) Mallows C_p selected variables. The following models were tested over 1000 iterations, and the accuracy rates and AUC's computed at each iteration were averaged to find the best performing model. The best performing model is the ridge regression, fit to the VAR selected variables, and achieves 63.75% accuracy.

References

[1] Askitas, N. & Zimmermann, K. F. "Google Econometrics and Unemployment Forecasting" *Applied Economics Quarterly: Vol. 55, No. 2, pp. 107-120,* 2009

[2] Bollen, J., Mao, H., & Zeng, X. "Twitter mood predicts the stock market" *Journal of Computational Science,* 2011

[3] Borgan, O. "Kaplan-Meier Estimator" *Encyclopedia of Biostatistics,* 2005

[4] Choi, H. & Varian, H. "Predicting the Present with Google Trends" *Economic Record, Vol. 88, Issue S1,* 2012

[5] Friedman, J., Hastie, T., & Tibshirani, R. "Regularization Paths for Generalized Linear Models via Coordinate Descent" *Journal of Statistical Software, Vol. 33(1),* 2010

[6] Gilbert, E. & Karahalios, K. "Widespread Worry and the Stock Market" *Proceedings of the Fourth International AAI Conference on Weblogs and Social Media,* 2009

[7] Hisano, R., Sornette, D., Mizuno, T., & Ohnishi, T. "High quality topic extraction from business news explains abnormal financial market volatility," *PLoS ONE 8(6): e64846,* (2013)

[8] Ingle V. & Deshmukh, S. "Live New Streams Extraction for Visualization of Stock Market Trends" *Lecture Notes in Electrical Engineering, Vol. 395,* 2016

[9] Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. "Quantifying Wikipedia Usage Patterns Before Stock Market Moves" *Scientific Reports, 3, Article no. 1801,* 2013

[10] Preis, T., Moat, H. S., & Stanley, H. E. "Quantifying Trading Behavior in Financial Market using Google Trends" *Scientific Reports, 3, Article no. 1684,* 2013

[11] Vosen, S. & Schmidt, T. "Forecasting private consumption: survey-based indicators v. Google Trends" *Journal of Forecasting,* 2011