ARE SOME STOCKS MORE PREDICTABLE THAN OTHERS? A GOOGLE BASED INVESTOR SENTIMENT INVESTIGATION USING MACHINE LEARNING TECHNIQUES

Adeel Ali Qureshi[1]

[1]*Poznan University of Economics and Business, Poland*

adeel.ali-qureshi@phd.ue.poznan.pl

Abstract: Business finance, and common language dictionaries with sentiment directionality and a collection of 92000+ keywords are used to acquire 1150 random keywords (575 positive and 575 negative) to retrieve 10 years of Google SVI per keyword. Sorting by Google Search results, top 50 timeseries are forecasted 6 months in the future using Seasonal Naïve. We produce a sentiment index based on historical and forecasted data, to be used as an exogenous predictor to the forecasting of stock returns of 500 randomly selected US stocks from various industries. We utilize NBEATS based deep neural architecture and LSTM based recurring neural network along with endogenous technical variables, and further categorize stocks by characteristics, and generate two sub-portfolios–large, old, profitable, and dividend-yielding firms versus small, young, unprofitable, and non-dividend yielding firms–to find the former sub-portfolio with more accurate forecast, while latter to be more statistically significant, with aforementioned sentiment index.

***Keywords*:** stock market, sentiment, forecast, machine learning, attention

JEL: C22, C45, C53, G12, G41

INTRODUCTION

Internet has become one of the most used mediums of communication in our times, and Google has become synonymous for internet search. In 2023, 84.7% of all global internet searches were made on Google (Statista 2023). Both, literature and empirical evidence, suggest that Google's Search Volume Index can be a useful tool to capture investor attention. (Preis, Moat and Stanley 2013, Da, Engelberg and Gao 2011). Thus, we use Google Trends to retrieve Google Search Volume Index (SVI), and using that, generate a unique investor sentiment index to determine a quantifiable 'mood' of retail investors.

Assuming SVI for stock tickers itself remains indeterministic for investor attention, but research proves it useful for keywords, and in extension, to stock prices, and even stronger when business or finance related words are used instead of regular words or expressions. (Behrendt and Prange 2021, Dimpfl and Jank 2016, Preis, Moat and Stanley 2013, Zhang, Ren and Gao 2020, Da, Engelberg and Gao 2011). From the point of view of behavioral finance, loss aversion theories also reflect strongly with internet-based searches and coincide with previous research claiming negative keywords carrying stronger investor sentiment. (Tetlock 2007, Da, Engelberg and Gao 2014).

On our side, we focus primarily on positive and negative words, and disregard neutral words to form directionality. It must be noted that retrieved data is an index, therefore, zeroes denote no searches for that data point. Thus, a contrasting sentiment having zero would denote no sentiment, rather than lack thereof SVI. We use two dictionaries: Loughran–McDonald Dictionary of business vocabulary from University of Notre Dame, and common language and internet vocabulary dictionary from the University of Illinois Chicago.

Separately, we retrieve data for 500 US stocks which are randomly chosen with the sole criterion that their market capitalization per annum be in the range of 50 million to 1 trillion USD. We further flag each stock with characteristics, as such; size (large or small, concluded by market capitalization as being in the top 33% or bottom 33% of all stocks, respectively), age (young or old, concluded by founding year being after or before all stocks' median age, respectively), dividend yield (whether that stock yields dividend or not), and annual net income (whether that stocks' annual net income is positive or negative). Stocks filtered through each characteristic, we generate two sub-portfolios and expect one to be more sensitive to investor sentiment than the other:

i. Sentiment resistant companies (SRC): Large, old, dividend yielding, and positive net income stocks.
ii. Sentiment sensitive companies (SSC): Small, young, non-dividend yielding, and with negative annual net income stocks.

We use Google SVI based sentiment index, mentioned earlier, as an exogenous feature to forecast both sub-portfolios with additional variables such as VIX index as market volatility, and S&P 500 index as general market movements. We find sentiment index to be more statistically significantly related with SSC, while SRC to be more predictable using the same.

METHODOLOGY

From the collection of 90,000+ words from both dictionaries, 1150 (575 positive and 575 negative) words are chosen by random. We further use Google Search's number of results as the so-called popularity to filter the top 50 (25 positive and 25 negative) keywords from the randomly chosen list. Ten years of Google SVI (2013 – 2023) is retrieved using Google Trends. It must be noted that Google produces weekly granularity if temporal range is greater than 9 months (less than 9 months generates daily, and greater than 5 years generates monthly). We use the following formula to produce sentiment index:

$$Sentiment_{\,t} = \frac{AVG_c(+SVI_{k,w})}{AVG_c(-SVI_{k,w})} - 1$$

Where:

- AVG_c = Conditional Average, Condition = SVI > 0
- w = Weekly iteration
- k = Keyword
- +/- SVI represent search volume index for positive and negative keywords.

We also generate two further variants, standardize this sentiment by reducing mean sentiment from each iteration and dividing it by standard deviation of all sentiment, and absolute standardized sentiment.

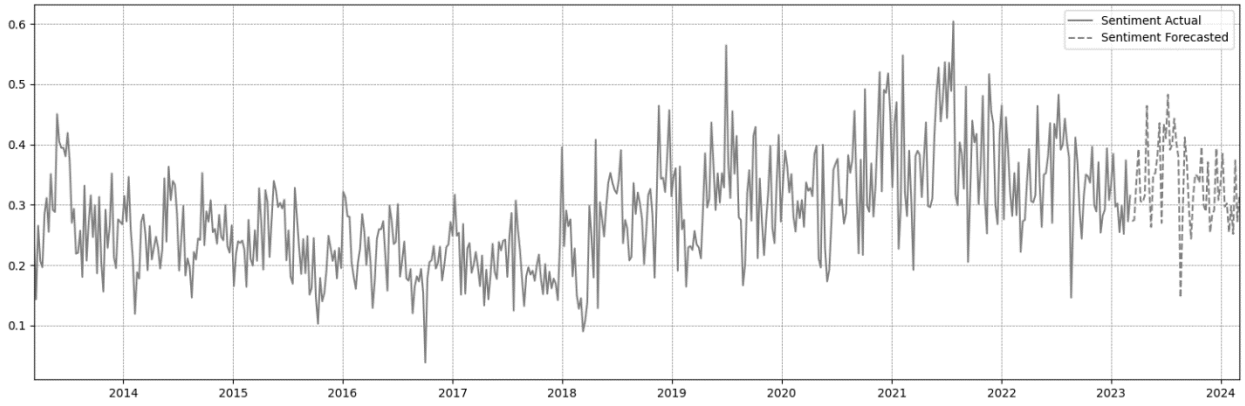

Figure 1. Calculated Actual Sentiment and Forecasted Sentiment

While Google SVI data is weekly, therefore, we resample stock market data from daily to weekly as well. We use Adjusted Close returns for all 500 stocks and calculate stock returns for each for the same ten years of temporal range. We filter SRC and SSC as mentioned in the previous section, and we deduce 97 SRC stocks and 69 SRC stocks. We calculate weekly average of their stock returns to produce each sub-portfolio's average stock returns. Similarly, we resample VIX index and S&P500 index to weekly and produce 5 timeseries. Separately, we also calculate endogenous variables such as relative strength index, fast, medium, and slow, moving averages.

We use investor sentiment as a feature to forecast both sub-portfolios separately. We use N-BEATS (Oreshkin, Carpov, Chapados, & Bengio, 2020) based deep learning model through Nixtla package in Python, that uses backward and forward residual links to generate the forecast. We use 24 previous iterations to forecast 12 upcoming ones. Separately, we use another method for forecasting purposes. We use long short-term memory based recurring neural network architecture (Hochreiter & Schmidhuber, 1997) through Keras and TensorFlow library packages in Python.

FINDINGS

To find correlation, we run regression models between the explanatory variables; Google SVI based investor sentiment (standardized and absolute standardized), VIX index, S&P500, and dependent variable; stock market returns of the sub-portfolio. We run regression models twice, for each sub-portfolio separately. We conclude that a significant relationship exists between sentiment and SSC returns, while it is not statistically significant with SRC return.

Table 1. Regression Results

	SRC	SSC
Const	-0.009	-0.01
Volume Lag 1	0.006	-0.024**
Standardized Sentiment Lag 1	0.001	0.007**
S&P500 Lag 1	0.102	-0.152
VIX Lag 1	0.001	0.002*
Returns Lag 1	-0.156	0.226
Returns Volatility Lag 1	-0.518	-0.208

This table provides regression results between SRC and SSC when regressed against given explanatory variables. Lag 1 denotes 1 iteration's lag. Each iteration denotes to 1 week in given temporal scope. The results show statistical significance between lagged volume of SSC, standardized sentiment of SSC and lagged VIX, whereas none for SRC.

We use Seasonal Naïve model to forecast the sentiment for the upcoming one year. With the two forecasting methods, for LSTM based neural network approach, we use sentiment as a regressor feature and produce technical indicators such as relative strength index, fast, medium, and slow, moving averages. We train on 90% of data and use 10% as training and validation and use 30 back candles for each epoch to forecast the upcoming iteration. It must be noted that before forecasting, we scale all variables (including exogeneous feature and endogenous technical indicators, and the average stock returns themselves) between 0 and 1, and after forecasting we reverse scale the same. For N-BEATS based deep learning approach, we use sentiment as a feature, and train on 495 weeks and test and validate on 26 weeks. We add a constant of 100 before forecasting and subtract afterwards to avoid any negative, zero, or infinity-

based error. To measure accuracy, for both approaches, we use mean absolute errors and mean squared errors. For N-BEATS, we also produce rooted mean squared error.

Table 2. Forecast Validation Results

Metric	Sub-Portfolio	Model Results	
		LSTM	N-BEATS
MAE	SRC	0.03	0.27
	SSC	0.08	0.32
MSE	SRC	0.01	0.10
	SSC	0.02	0.14
RMSE	SRC	-	0.32
	SSC	-	0.37

This table shows the results of the two forecasting models used. For each model, mean absolute error, mean squared error, and rooted mean squared error are displayed for each portfolio. Each model is run separately for each portfolio and with sentiment as feature.

Both models have separate prediction techniques and feature engineering is performed separately as well, nevertheless, both models produce same results. We notice that MAE, MSE, and RMSE, are always closer to zero for SRC, proving this category of stocks to be more predictable for the Google Trends based stocks.

CONCLUSIONS

In this research, we generate an investor sentiment based on Google SVI retrieved through Google Trends for ten years. We use business, finance, and common language-based dictionaries to randomly select keywords which are mentioned as negative or positive sentiment. We generate a weekly sentiment index and use it to regress against stock returns of companies filtered by their characteristics. We categorize them in two-sub-portfolios, small, young, dividend-yielding, and negative annual net income stocks, and vice versa for the opposite, labelling the former group as sentiment sensitive companies, and the latter as sentiment resistant companies. Through regression model, we find that SSC holds a statistically significant relationship with the said Google SVI-based sentiment index, while SRC does not. Quite the contrary for predictability, because through two separate methods (LSTM based neural network, and N-BEATS based deep learning) depicts SRC to be more predictable than SSC.

While reporting these results, we also generate opportunity for future research with enhanced feature engineering, and modelling approaches, in addition to, adding exogenous features to improve accuracies, for both portfolios.

REFERENCES

Behrendt, S., & Prange, P. (2021, January). What are you searching for? On the equivalence of proxies for online investor attention. *Finance Research Letters, 38*. doi:https://doi.org/10.1016/j.frl.2019.101401

Da, Z., Engelberg, J., & Gao, P. (2011). In Search of Attention. *Journal of Finance, The, 66*(5), 1461-1499. doi:https://doi.org/10.1111/j.1540-6261.2011.01679.x

Da, Z., Engelberg, J., & Gao, P. (2014). The Sum of All FEARS Investor Sentiment and Asset Prices. *The Review of Financial Studies*, 1-32. doi:https://doi.org/10.1093/rfs/hhu072

Dimpfl, T., & Jank, S. (2016). Can Internet Search Queries Help to Predict Stock Market Volatility? *European Financial Management*, 171-192. doi:http://dx.doi.org/10.2139/ssrn.1941680

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*. Retrieved from https://www.bioinf.jku.at/publications/older/2604.pdf

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2020). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *Cornell University*. doi:https://doi.org/10.48550/arXiv.1905.10437

Preis, T., Moat, H. S., & Stanley, E. (2013). Quantifying Trading Behavior in Financial Markets Using Google Trends. *Sci. Rep 3 1684*. doi:https://doi.org/10.1038/srep01684

Statista. (2023). *Statista - Online Search*. Retrieved December 2023, from https://www.statista.com/markets/424/topic/541/online-search/#overview

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *Journal of Finance, The*, 1139-1168. doi:https://doi.org/10.1111/j.1540-6261.2007.01232.x

Zhang, B., Ren, H., & Gao, Z. (2020). Googling Investor Sentiment around the World. *Journal of Financial and Quantitative Analysis*, 549-580. doi:10.1017/S0022109019000061