Learning Stock Volatility Using Keyword Search Volume

Yan Yang

December 16, 2011

1 Introduction

As internet becomes ubiquitous, search engine has become a standard way for people to retrieve information. Naturally, interests arise to explore whether the search volume of particular keywords tells something about the collective "mood" of the general public. Such correlation has been explored to determine various facors affected by popular interest, from consumer reception to political optinion. Stock price, being prone to the general sentiment in the market, is an ideal candidate in such studies.

On the other hand, such search volume implied sentiment can be measured, thanks to two factors. Firstly, Google has consistently dominated a large market share in the global search engine market, capturing $60-75\%$ of all searches. This makes search volume from Google a reasonable measure of the overall interest on the internet. Secondly, Google has been helpful to researchers and data crunchers by providing a free service to publish its search volume data, firstly through Google Trends, and then through a more research-oriented Google Insights for Search.

2 Background & Inspiration

This study is largely inspired by two results. Firsly, Da, Z. (2011) reached the conclusion that search volume of stock ticker symbols can be used as a proxy to investor attention, and furthermore has correlation with abnormal price increases. The second result is the R-word index pioneered by The Economist. It tries to use the frequency of word "recession" on news articles to predict the onset of a real recession.

Hence I borrow from both results and try to find how the volatility of a stock correlates with the search volume of both the ticker symbol and the word "recession". Although Da, Z. (2001) ascertains that since it is easier for a layman investor to buy than to sell, a high investor attention normally corresponds to more buy actions than sell and can apply an upward force to the stock price, by intuition volatility clearly is a better match to investor attention compared with stock prices. Also, this does not affect the application of the study, while more sophisticated, investment strategies to exploit high volatility are easily crafted and widely used in the market.

3 Data & Model

The search volume data are extracted from Google Insights for Search. It is aggreagted weekly, from 05/01/2004 to 19/11/2011. Also, the search volume is standardized so the week with highest search volume for a particular term is always 100. This means search volume of two terms cannot be compared, but that has no bearing on my method. Also, the search area is restricted to be USA only. Furthermore, in addition to web search data, in cases where news search data is available weekly (many less common terms only have monthly news search data), they are also used.

The daily stock prices are extracted from Yahoo! Finance for the same period. A number of stocks are selected from three fairly representative sectors: Technology, Energy and Financial. All stocks have big capitalization and most are well known household names. When selecting the stock one catch is that the ticker symbol must not yield "noisy" search data. For example, ConocoPhilips is excluded since its ticker symbol "COP" clearly will give huge amounts of search data unrelated to the company. The stocks selected are listed in Table 1.

Sector	Stock	Symbol
Technology	Apple	AAPL
	Microsoft	MSFT
	Hewlett Packard	HPQ
	Activision	ATVI
	Applied Materials	AMAT
Financial	Berkshire Hathaway	BRK-A
	JP Morgan	JPM
	Wells Fargo	WFC
Energy	Exxon Mobil	XOM
	Chevron	CVX
	Praxair	PX

Table 1: Stocks Selected

The volatility of the stocks are computed using the standard formula

$$Y_n = \sigma_{t_n+1 \le t \le t_n+\delta}(\log(P_t/P_{t-1})) \tag{1}$$

where σ is the standard deviation over δ trading days after day t_n (last day of the week when search volume is registered), P_t is the price of the stock on day t. δ is determined as number of trading days in d weeks, and d serves as a parameter in my model.

Also, when an initial learning was conducted using the search volume index directly quoted from Google, the results are not very promising. As such a transformation is also made on the search volume. I compute the normalized version of the search volume, i.e.

$$X_n = \frac{SV_n - \overline{SV_n}}{\sigma(SV_n)} \tag{2}$$

where SV_n is the search volume index of the nth week. There are two such features, corresponding to the ticker symbol $X_{1,n}$ as well as the term "recession" $X_{2,n}$.

After these preprocessings, I try to apply the logistic regression and Gaussian discriminant analysis. The target variable is converted into a binary variable $1\{Y_n \ge \theta\}$, where θ is a threshold that is determined by the historical prices of the stock . This level is experimented and set as $\overline{Y_n} + \sigma(Y_n)/2$ so as to get sufficiant but not too many data points with a label of 1.

4 Results

One sample result is as shown in Figure 1 where stock prices of Apple are taken and d is chosen to be 2. As seen from the plot, the high and low volatility points tend to have large overlap in the 2D space. This is possibly one reason that SVM attempted on the data does not give as good a result as GDA or logistics regression. Therefore, the method serves as a conservative strategy in terms of capturing high volatility events. Also, we can see that logistics regression performs less well in dtermining high volatility events compared to GDA. These trends largely hold for other stocks and d values as well.

Next, a 10-fold cross validation is performed using GDA method. The cross validation error serves as a proxy of the predictive power of my method. Firstly, to determine if d significantly affects the learning, cross

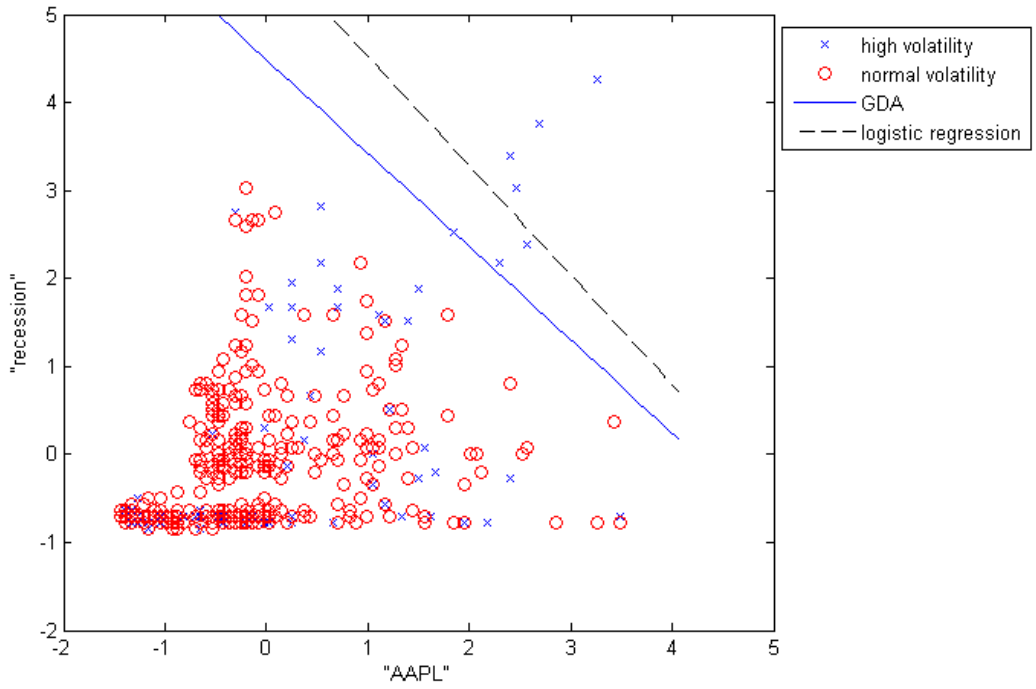

Figure 1: Sample Result for Apple and d=2

validation is performed on different values of d, as shown in Figure 2. As can be seen, the error drops as d increases for certain stocks like Apple and Activision, but increases for some other like JP Morgan. In below analysis I take a value of $d = 2$, which is a reasonable length and yield comparatively good results.

Once the method is fixed as GDA and the value of d fixed at 2, we can evaluate the performance of the method on various stocks. The results are compiled in Table 2. Here, Inst. Own stands for institutional ownership, while historical volatility is computed as the overall volatility of the stock over the entire timeline of the dataset. Furthermore, since the method is intended to identify high volatility events the normal notion of error being number of cases mis-classified may not be suitable. Hence another error measure, termed one-side error, is added which only measures number of events that are classified as high volatility but are actually not.

Also, for the six stocks with weekly news search volume data, I have carried out the same process on news search data rather than web search data. The results do not change and hence news search data, at least from Google, seems not to offer much improvement in performance relative to general web search.

5 Analysis & Conclusion

Institutional ownership is an important measure since the rational is that stocks with low instituional ownership should by right be more subject to general public sentiment. This seems to be not much a factor here. Also, the one-side error offers a very different trend as the overall error. Apple has a particularly low one-side error, and HP's higher than rest overall error shrinks to a lower than rest one-side error. Besides the two stocks, the one-side error of the other stocks all roughly float around $2 - 3\%$

Overall, the errors seem not to have much correlation with the two mesures listed here. In future, to further expand this study, several things can be improved. Firstly, a much larger set of stocks and a longer period (ideally one not covering a real recession) should be used. While the former is easily done but requires

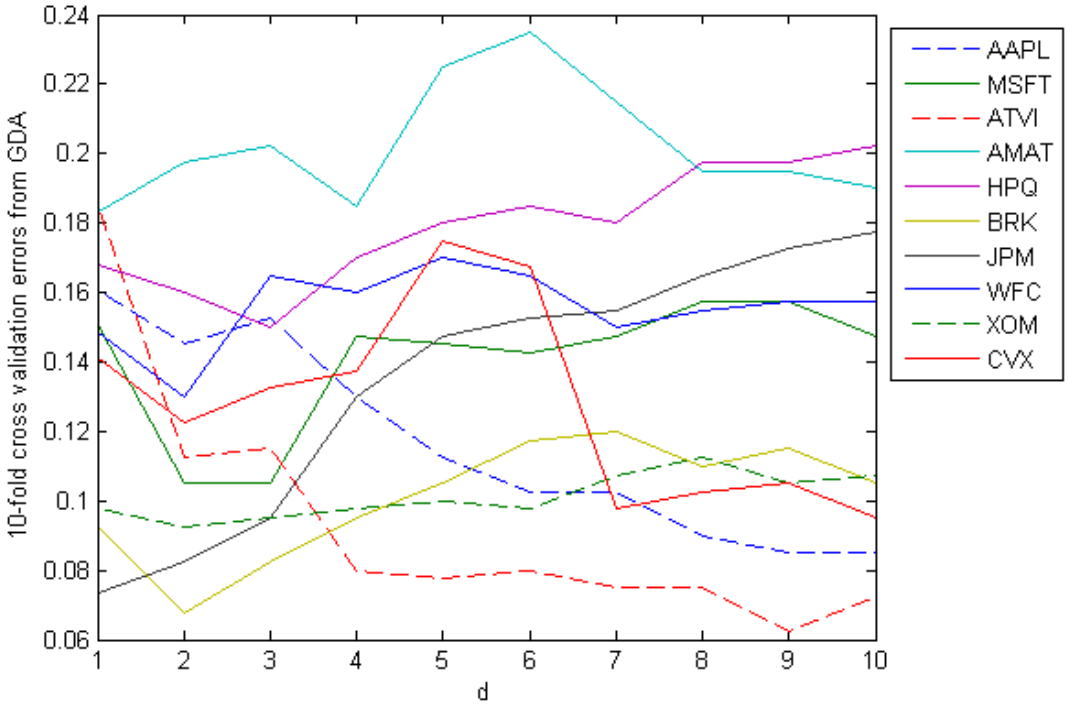

Figure 2: Plot of 10-fold cross validation error from GDA vs. d

	Stock	Overall Error	One-side Error	Inst. Own	Historical Volatility
Technology	Apple	0.145	0.005	70.80%	0.036
	Microsoft	0.105	0.023	64%	0.011
	HP	0.198	0.013	74.80%	0.018
	Activision	0.113	0.020	33.90%	0.031
	Applied Materials	0.113	0.020	79.30%	0.019
Financial	Berkshire Hathaway	0.160	0.035	48.10%	0.009
	JP Morgan	0.068	0.025	73.70%	0.010
	Wells Fargo	0.083	0.035	77%	0.008
Energy	Exxon Mobil	0.130	0.023	49.30%	0.012
	Chevron	0.093	0.020	63.20%	0.027
	Praxair	0.123	0.023	0.90%	0.013

Table 2: Cross validation errors of GDA compared to certain stock attributes

more time and effort, the latter is largely limited by reality. Secondly, I would like to design an investment strategy and test it on historical data to find the expected return. However, the main difficulty here is that the simplest startegies that exploit high volatility, such as a straddle, depend on historical option prices which I cannot get hold of. Thirdly, more features in addition to ticker symbol and "recession" can be used and supposedly more sophsticated correlation among stock volatility and search volume may be uncovered.

In all, this method is straightforward and simple-to-use, it also seems to be quite capable of capturing particular weeks when individual stocks register higher than average volatility.

References

[1] Da, Z., Engelberg, J. & Gao, P. (2011), In Search of Attention. *Journal of Finance*, **66**: 1461-1499.

[2] "The recession index: Words that can harm you" The Economist. Nov 21, 2002. ⟨`http://www.economist.com/node/1455116?story_id=E1_TQVVTTG`⟩