Predicting stock market volatility, returns and trading volume of tech stocks using Semantic Vectors and Google Trends

Puneet Girdhar
School of Engineering and
Computer Science
University of Texas at Dallas
Email: pxg151330@utdallas.edu

Karan Shukla
School of Engineering and
Computer Science
University of Texas at Dallas
Email: kxs141930@utdallas.edu

Miti Desai
School of Engineering and
Computer Science
University of Texas at Dallas
Email: mkd150130@utdallas.edu

Sarvottam Singh
School of Engineering and
Computer Science
University of Texas at Dallas
Email: sxs155032@utdallas.edu

Tarun Kumar Sahu
School of Engineering and
Computer Science
University of Texas at Dallas
Email: tks160330@utdallas.edu

Himanshu Parashar
School of Engineering and
Computer Science
University of Texas at Dallas
Email: hxp151330@utdallas.edu

***Abstract*—Trying to foretell when the stock market is reaching its peak may be a fools errand. Even the most respected and famed investors and market pundits cant claim to know when its time to get outor for that matter into the market, no matter how savvy they may be on investing but that doesn't take away the fact of benefits of prediction where on one hand it can give someone opportunity to increase individual gains while on other hand it could also alert people before financial melt-down. Since Google Search reflects general trends in society be in financial, social or economical, we are using Google trends to predict financial market performance.**

The terms people search on Google have been used to forecast American Football winners, elections, travel plans etc. Since last few years, a vast majority of researchers are also looking into using Google trends to predict the stock market.This paper takes that research forward by investigating the efficacy of using Google Search Volume Index(SVI), a publicly available tool Google provides via Google trends, to predict the stock movements within the tech sector. The result of this experiment shows significant correlation between Google SVI and weekly trading volume as well as stock returns for last 5 years. We also studied linear regression model to predict stock prices based on search volume of closely related tech keywords.

Keywords*—*Stock Market, Google Trends, Machine Learning

I. Introduction

Current stock market trading strategies often depend on using the historic financial data for making predictions about volatility of the future market. However, the historic data reflects only the past actions and not any information about public thought i.e. a major predictor of the market's future. Following proven research that uses social data to predict other features of the market, we make a pitch for an attempt to predict stock market volatility, return and trading volume of tech stocks using Semantic Vectors and Google Trends, which reflects real-time popularity of search terms. Our approach comprises of two steps: First, used New York Times API to download 1 million documents and obtained a set of keywords related to the information technology domain using semantic vectors. Then, we train a linear regression model to predict the volatility of a future week using as input Google Trends data for those keywords in the previous week

II. Related Work

Already existing algorithms that predict the stock-market trends have proven that social media is useful. Recent proven strategies attempt to co-relate the "mood" of the audience of twitter and the daily changes in the closing values of the Dow Jones Industrial Average (Bollen et al. 2011) and other financial indicators (Zhang et al. 2011). Similarly, attempts at predicting the market volatility have majorly used historic data for making predictions (Fleming et al. 1998; Corrado et al. 2005), and these attempts have reported diverse levels of accomplishment. This approach aims to depict that social-media data can have powerful predictive power in modeling volatility, as it does for market movement. More recently, Preis et al. (2013) showed that an invented strategy based on Google Trends yields higher profits than traditional trading strategies like buy and hold. We add to their work in two significant ways. First, our final keyword set is obtained empirically rather than semi-automatically. Second, we attempt to predict a measure of market volatility rather than directly determine profits using a specific algorithm. This makes our method strategy-agnostic and solidifies the relationship between Google Trends and market movements.

III. Methodology

A. Search Terms

We started with 20 set of "seed words" (for example internet, computers, information etc.) related to technology to query the New York Times (NYT) Article Search API, which allows access to the abstract, title, and content of all articles in the NYT from 1851. We searched published articles from January 1st, 2012 and created a corpus of 1 million articles to be used for semantic vector generation using Google's Word2vec tool. We generated total list of 30287 words each having 300 dimension vector. We then clustered these vectors together into (k=120) classes to obtain similar set of words. Reason for choosing such a high ”K” number is to get less than 500 words per cluster. We then handpicked one cluster which contained all our desired seed words and formed a distinct set of keywords to be used in market prediction.

B. *Choice of tech stocks*

To minimize the uncertainties of stock market prediction, we chose 8 stocks from NASDAQ 100 with active high trading volume. These stocks generally guarantee a sizeable pool of interested invidividual retail investors that are likely to seek information to these stocks. This provides us with a good sample size to observe variations in investor interest.

Company Name	Ticker	Dollar Volume (Million)
Apple Inc.	AAPL	$1,644.87
Amazon Inc.	AMZN	$86.28
Baidu.com,Inc.	BIDU	$232.99
Cisco Systems Inc.	CSCO	$126.39
Intel Corporation	INTC	$185.13
Microsoft Corporation	MSFT	$314.96
Netflix, Inc.	NFLX	$1023.83
Qualcomm, Inc.	QCOM	$759.55

We selected technology related keyword

C. *Features*

1) SVI: We obtained Weekly search volume index(SVI) for each word in our keyword dataset by querying a Google Trends API. SVI shows often a specific search term is searched relative to the total search volume across the world, over a defined date range that the user inputs. Here is the mathematical formulation:

$$SVI = \frac{\#\ queries\ for\ specific\ keyword}{Total\ Google\ search\ queries}$$

For each data point, the SVI of previous week is also recorded as SVI_pre in order to correlate changes in SVI with stock movements in the subsequent week.

2) Weekly Stock Returns: Weekly returns on a stock are measured by taking the natural log of the ratio of the closing price of the current week to the closing price of the week before.

$$WEEKLY\ RETURNS = r_{t,w} = \log \frac{P_{close,w}}{P_{close,w-1}}$$

where $r_{t,w}$ is the weekly returns for day t of week w and $P_{close,w}$ is the closing price of week w.

3) Stock Volatility: Stock volatility is a statistical measure of the dispersion of returns for a given stock or market index. It is a great way to detect trends in stock prices. To calculate, for each week we measure the percentage increase in the stock price as compared to the opening price of that stock in the same week. Here is the mathematical formulation:

$$STOCK\ VOLATILITY = SV_w = 100 * \frac{high - low}{open}$$

IV. Data Analysis

A. *The basic ARIMA model analysis of historical stock prices*

To perform the basic ARIMA time series analysis on the historical stock prices, we first make a plot of the raw data, i.e. the weekly close prices of each company over time. As an example plot of AAPLE is shown below.

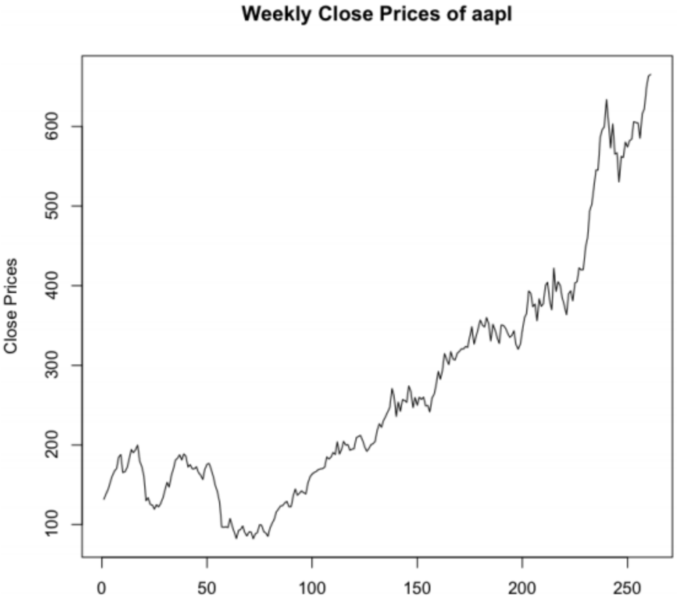

Fig. 1. Weekly Closing price of AAPL

This plot shows that the close price of aapl increases in general over the past five years. However, there is no apparent pattern in the movement of the stock price. The variance of stock price seems to increase slightly with time.

We then calculate weekly log returns stock volatility from stock price data.

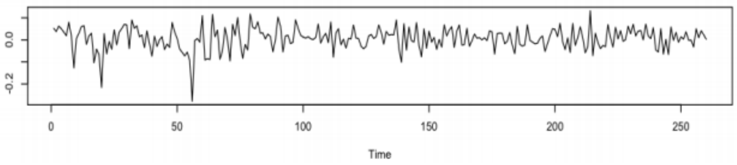

Fig. 2. Weekly Log Returns

As we can see this is stable across different timeframe and hence exhibits less correlation with time. It can thus further be used for prediction model. Similar observation can be made for Stock volatility.

To confirm whether we removed any correlation in the data from previous data points, we plotted weekly returns data against historical data. As shown in Fig4, we could observe here that except previous week data, all correlations were removed.

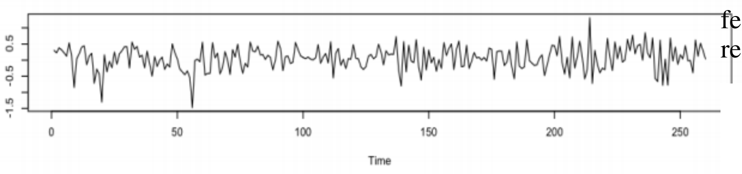

Fig. 3. Stock volatility

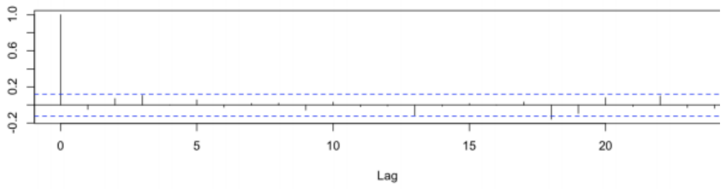

Fig. 4. Autocorrelation Function of Weekly log returns

B. Prediction Model

In order to study the effect of Google Search Volume interest over the stock market, we built a linear regression model to predict stock volatility and log-returns. Formally, for each week w in our dataset, we define $x \in \mathbb{R}^n$ to be the featureset for w containing the SVI values from week $w-1$ for all keywords in the keyword set K ($|K| = n$). We also define our response variables $y \in \mathbb{R}$ and thus our training dataset consists of m training examples $(x^i, y^i); i = 1, ...m$.

Following multivariate regressions were conducted. Correlations were drawn between SVI, returns and volatility for each week. Regressions were drawn for all 8 stocks as an aggregate and subsequently for each stock to investigate differences in relationships between SVI and stock movements between 8 stocks.

1) Correlating SVI and returns: Weekly returns are regressed against previous week's SVI values and previous week return. Idea is to predict the future return if we know the present condition. Here is the mathematical formulation:

$$returns_w = W_0 + W_k * SVI + W_i * returns_{w-1}$$

2) Correlating SVI and volatility: Stock volatility is regressed against previous week's SVI values and previous week volatility data. Idea is similar to above where we are predicting future based on prsent condition.

$$volatility_w = W_0 + W_k * SVI + W_i * volatility_{w-1}$$

C. Experiments and results

In financial trading, any prediction model which often has low error but occasionally makes large error proves to be loss-making. To strongly penalize larger errors, we use Mean Squared Error (MSE) as our metric. We also performed dimensional reduction since set of words were quite huge and we were over fitting the training samples.

1) Results: We experimented with Linear regression, Decision Tree regression and XGDBoost in our study and compared their output. We also performed lasso to remove certain features which are not important in prediction. Here are some results

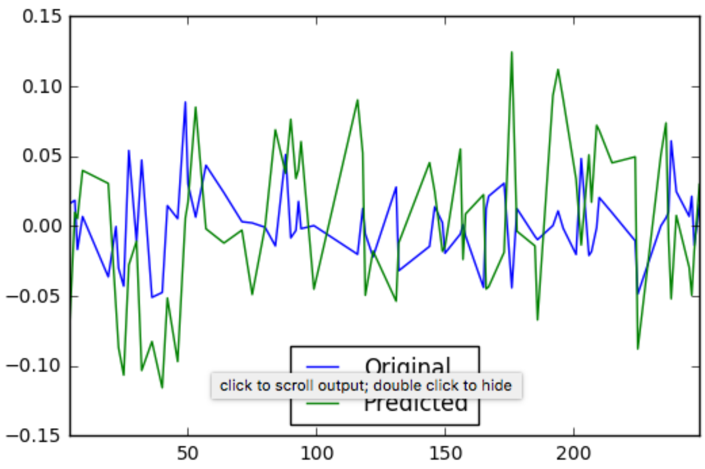

Fig. 5. Fitted vs Original Plot of stock returns

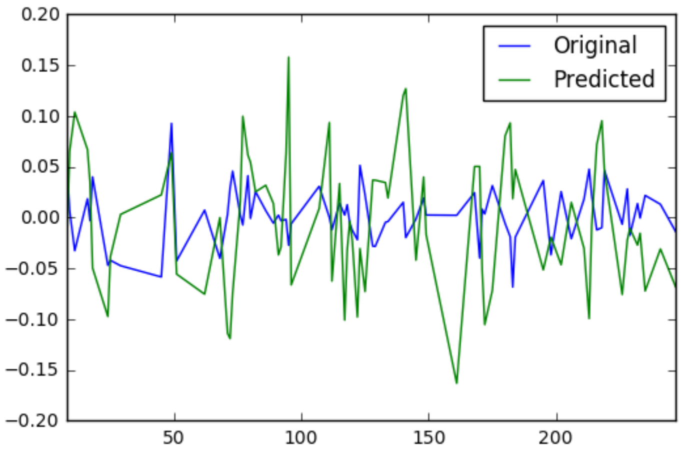

Fig. 6. Fitted vs Original Plot of stock volatility

As we can see that predictions closely follow the original pattern but there is quite a noise in prediction. Below are Root mean squared values obtained from three different models.

Linear Regression	0.006
Decision Tree	0.004
XGD Boost	0.003

However RMSE doesn't seem to be high enough but please note that we are predicting log of returns which itself is a very small quantity. With our experiments, XGD Boost performed better than other models but not to high extent. However, this cannot be generalized for all time periods since hidden factors effecting stock price may come into picture.

Appendix A
Conclusion

This study introduces a novel approach to selecting stocks for studies on search volume, as it uses active trading volume and relevant keywords from New York Times API. This serves to maximize the accuracy of using Google SVI as measure of investor interest. This study also offers evidence on positive relationship between Google SVI and weekly traded volume, realized volatility and weekly close price for specifically traded stocks in tech sector. The positive relationship between Google SVI and weekly returns is shown to be slightly more prevalent amongst the tech stocks since they are more active online. Furthermore, this study presents new evidence that Google SVI has become an increasingly significant predictor of realized weekly volatility in the stock market over the years. Result also suggest most significant ”herding” behavior than before.

On intution level, taking a non-contemporaneous approach with weekly stock data and SVI is non-optimal since investors are unlikely to wait a week between researching and making investment decisions. However public data of google trends is only available in week granularity from May 2012 and hence greater granuarlity in such data is therefore needed to improve the predictive power of Google SVI. This would allow us to more closely study the effect of SVI over stock market prediction. Can this be used in actual real stock trading, we will leave that question to the reader. If you earn some money using above techniques, please do write to us and share your experience.

Acknowledgment

The authors would like to thank Dr. Latifur Khan for teaching the class and sharing with us his knowledge of the foundations of Big Data architecture, its algorithms, and its newest technologies. The authors would also like to thank Mr. Mohiuddin Solaimani and Mr. Zhuoyi Wang for aiding in the class’s learning and generously taking their time to provide individual help and ensure student success.

References

[1] Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2(1):18.*

[2] Corrado, C. J., Miller Jr, T. W., et al. (2005). The forecast quality of cboe implied volatility indexes *Journal of Futures Markets, 25(4):339373.*

[3] Fleming, J. (1998). The quality of market volatility forecasts implied by sp 100 index option prices. *Journal of empirical finance, 5(4):317345.*

[4] Preis, T., Moat, H. S., and Stanley, H. E. (2013). Quantifying trading behavior in financial markets using google trends.*Scientific reports, 3.*

[5] Zhang, X., Fuehres, H., and Gloor, P. A. (2011). Predicting stock market indicators through twitter *Procedia-Social and Behavioral Sciences, 26:5562.*