Using Twitter Sentiments and Search Volumes Index To Predict Oil, Gold, Forex and Markets Indices

Tushar Rao, Netaji Subhas Institute of Technology, Delhi, India
Saket Srivastava, Indraprastha Institute of Information Technology, Delhi, India

Behavioral finance is an upcoming research field which is drawing a lot of attention of both academia and industry. With changing dynamics of internet behavior of millions across the globe, it provides opportunity to create a unified forecasting model comprising of large scale microblog discussions and search behavior for better understanding of market movements. In this work we used 2 million tweets and search volume index (SVI from Google) for a period of June 2010 to September 2011; studied causative relationships and developed a comprehensive and unified approach for a model for equity (Dow Jones Industrial Average-DJIA and NASDAQ-100), commodity markets (oil and gold) and Euro Forex rates. We investigate the lagged and statistically causative relations of Twitter sentiments developing prior during active trading days to market inactive days and search behavior of public before any change in the prices/ indices. Our results show extent of lagged significance with high correlation value upto 0.82 between search volumes and gold price in USD. We find weekly accuracy in direction (up and down prediction) uptil 94.3% for DJIA and 90% for NASDAQ-100 with significant reduction in mean average percentage error for all the forecasting models.

Categories and Subject Descriptors: A.1.1 [ ]:

General Terms: Opinion mining in Twitter, Sentiment Analysis, Behavioral Finance

Additional Key Words and Phrases: Stock market, sentiment analysis, Twitter, microblogging, social network analysis, oil, gold, Forex

1. INTRODUCTION

"Its not that people don't know about the probabilities, but decision revolves around how they distort it in their thinking."

- Daniel Kahneman

Most of the earlier works in computational finance comprise of *efficient market hypothesis* (EMH) that asserts market movements at present level are function of all the already existing news, hidden whispers and future valuation of dividends that particular equity shall yield as market comprises of investors who always seek out to maximize their earnings [14] and [20]. Research by Qian et al. comprise of comparative studies showing market is not fully efficient [22]. However behavioral finance [28] is rising research area that challenges the vary existence of efficient markets by placing the role of human sentiment and social mood as vital part of investment decisions [21]. It challenges the Efficient Market Hypothesis (EMH) by adding the notion of human emotion and macro-level mood play into investment decisions. For example for many investors, constantly rising stocks is indicative of sell to hold profits and subsequent portfolio adjustments towards options securities which in contrast at macro- level gives entirely altered picture of index/ price movements. Social mood indicates entirely abjured relations of mass opinion. For example if lots of people claim confidence to make buy decision for commodity, then price rises so fast that instead of stabilizing it falls; as we subsequently discuss how bullishness yield negative correlations with DJIA index.

Further accurate and comprehensive measurements of sentiment behavior at heart of the forecasting models and macro-level investment behavior.

Earlier in late 90s before the spread of social web, information regarding commodities/currency rates , direction and buy/sell sentiments took a long time (maybe even full day) to disseminate in the full community. Also, the companies and markets took a long time (weeks or months) to calm market rumors, news or false information. This era of web technology is marked high entropy of information as well as retrieval [5]. Recently scholars have made use of twitter feeds in predicting box office revenues [1], political game wagons [27], rate of flu spread [25] disaster news spread [9].

A large percentage of high frequency traders in US markets, have trained AI bots to capture buzzing trends in the social media feeds without learning dynamics of the sentiment and accurate context of the deeper information being diffused in the social networks. In social web mining context, distinctively there are two different approaches that researchers have taken for market prediction. First is through social media feeds and with recent volume of tweet dataset available makes it an important resource to measure investor mood at comprehensive scale. From earlier works through internet stock board [10] and quite recently through Livejournal, Facebook and Twitter are used to derive overall mood series of all the online activity and correlating it to make prediction models for DJIA index [10],[12], [24], [3; 18] and [15]. Secondly search volumes have been shown to give out predictive and causative relationships with market returns which varies for different equities and sectors standards [26].

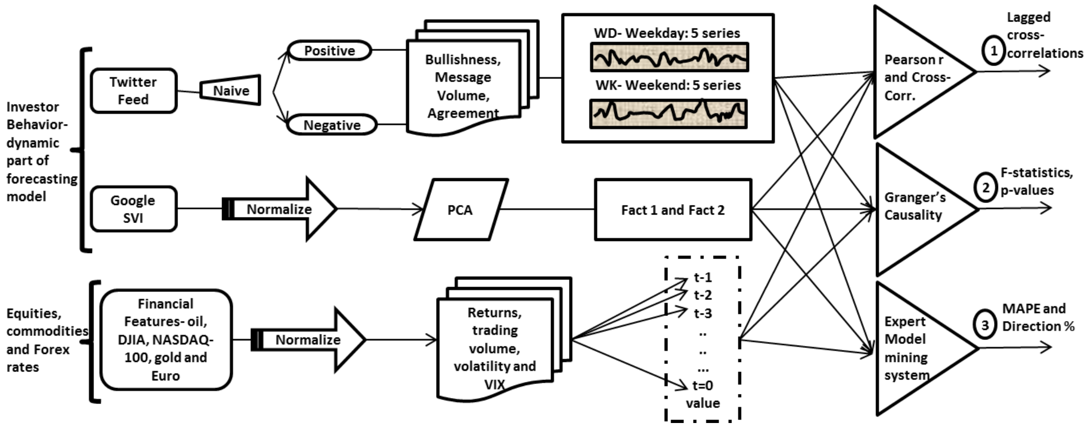

Investor Behavior-dynamic part of forecasting model
Twitter Feed
Naive
Positive
Negative
Bullishness, Message Volume, Agreement
WD- Weekday: 5 series
WK- Weekend: 5 series
Google SVI
Normalize
PCA
Fact 1 and Fact 2
Equities, commodities and Forex rates
Financial Features- oil, DJIA, NASDAQ-100, gold and Euro
Normalize
Returns, trading volume, volatility and VIX
t-1
t-2
t-3
..
..
...
t=0 value
Pearson r and Cross-Corr.
Granger's Causality
Expert Model mining system
1
Lagged cross-correlations
2
F-statistics, p-values
3
MAPE and Direction %

Fig. 1. Flowchart of the proposed methodology showing the various phases of sentimental analysis beginning with SVI/ Tweet collection to stock future prediction. In the final phase three set of results have been presented:(1) Correlation results for twitter sentiments and stock prices for different companies (2) Granger's casuality analysis to prove that the stock prices are affected in the short term by Twitter sentiments (3) Using EMMS for quantitative comparison in stock market prediction using tweet features

In this paper, we present a comprehensive study of relationships over wide range of market securities- commodities such as oil, gold, forex rates of Euro and equity markets such as DJIA and NASDAQ-100 with the dynamic features of the investor behavior as reflected in the opinions emerging on Twitter and trends in the search engine volumes for the given security. The summary of the whole study conducted in this paper is summarized in the figure 1. In section 3 we present data collection and prior processing explaining the terminologies used in the market securities and social mood series. Further in section 4 we present the statistical techniques implemented and discuss the results and draw conclusions. Future prospects of the work are given in section 6.

2. RELATED WORK

In the last decade several works related to web mining of data (blogposts, discussion boards and news) [10], [12] and to validate the significance of assessing behavioral changes in the public mood to track movements in stock markets. Bagnoli et al. have worked upon unofficial earnings forecast-earning whispers as more accurate proxies for market expectations than official first call forecast[2]. In the earlier works one of the titillating interest have been drawn to internet forums like stock message board like Yahoo! Finance and time-stamped blog archives. Earlier work by Das et al. mining information from investor communities to create speculation regarding private and forthcoming information and commentaries[16],[7]. Wysoci in 1998 found financial inputs significantly affected by stock message boards discussions[29]. Dewally in 2003 worked upon nave momentum strategy confirming recommended stocks through user ratings had significant prior performance in returns [8]. But these studies deal with the readily available quantitative information i.e. message volume and user ratings. But now with the pragmatic shift in the online habits of communities around the worlds, platforms like StockTwits [1] [30] and TweetTrader[2] have come up and their usage is virally spreading out. A recent work is done by Zhi et al. using search volume index (SVI) as indicator of investor demand for Russell 2000 stock for the time period from 2004 to 2008 [6]. Das and Chen made the initial attempts by using natural language processing algorithms classifying stock messages based on human trained samples [7]. However their result did not bring out statistically significant predictive relationship between message bullishness and index returns.

Gilbert and Karrie have used corpus from livejournal blogposts in assessing the bloggers sentiment in dimensions of fear , anxiety and worry making use of Monte Carlo simulation to reflect market movements in S&P 500 index [12]. Similar work is done by Bollen et al. who used dimensions of Google- Profile of Mood States to reflect changes in closing price of DJIA [3]. Another work by Mao et al. covers effect of search volumes data in description with the preliminary sentiment indices of entire twitter feed on stock market movements of DJIA and volatility index of commodities like gold [17]. Zhang et al also made have made use of dimensions in human behavior-fear and hope to show correlations with the stock market indicators [31]. However these approaches have been restricted to investor sentiment with only one perspective of macro-economics and are not complete and flexible in terms explaining complete dynamic system for individual stock index for companies. Also work done by Sprengers et al. analyzed individual stocks for S&P 100 companies and tried correlating tweet features such as bullishness, agreement and message volume for the Twitter messages about discussions of the stock discussions about the particular companies containing the Ticker symbol [24]. This study brings out a flexible new approach combining the search behavior along with the sentiment analysis that is scalable for both individual commodities stocks/ companies and can be exploited to make successful hedging strategies making *wisdom of the crowd usable even by a singular investor*.

3. DATA COLLECTION AND PROCESSING

In this section we discuss the collection of Twitter, Search volume indices and various financial data series being used in this work. This includes collecting search volumes indices for various search queries; mining of tweets- processing them for analysis and extracting tweet sentiment and normalizing them for forecasting.

[1] http://stocktwits.com/

[2] http://tweettrader.net/

3.1. Tweets Extraction and Processing

With the growing ubiquity of Twitter as a stable and widely recognized medium for discussion by financial community provides significant research opportunities to data analysts and investors to mine out voluminous amount of data produced every minute to track complex patterns in public sentiment causative of sensitive turbulence in various market securities. Trading is now completely social with rise of StockTwits and TweetTrader.net taking over previous means of generating mass opinions and discussions as investor forums and discussion boards, with now people sharing the losses they make, profit they rejoice and fear they have in various securities at impressive rate of as high as 250 million messages tweeted everyday (Techcrunch October 2011[3]). Tweets are made accessible through a simple search of keywords(different market securities in our case) through an application programming interface (API)[4].

In this paper, we have used tweets from period of 15 months and 10 days between June 2nd to 13th September 2011. During this period, by querying the Twitter search API for each of the market feature under study say Gold, Euro, Dow etc. we collected $1,964,044$ (by around 0.71M users) English language tweets Each tweet record contains (a) tweet identifier,(b) date/time of submission(in GMT), (c) language and (d)text. Subsequently the stop words and punctuation are removed and the tweets are grouped for each day (which is the highest time precision window in this study since we do not group tweets further based on hours/minutes).

3.1.1. Tweet Sentiment Extraction. In order to compute sentiment for any tweet we classify each incoming tweet everyday into *positive* or *negative* using nave classifier. For each day total number of positive tweets is aggregated as $Positive_{day}$ while total number of negative tweets as $Negative_{day}$. We have made use of JSON API from Twittersentiment [5], a service provided by Stanford NLP research group [13]. Online classifier has made use of Nave Bayesian classification method, which is one of the successful and highly researched algorithms for classification giving superior performance to other methods in context of tweets. These methods have high replicability and few arbitrary fine tuning elements. The training was done over a dataset of 1,600,000 tweets and classifier achieved an accuracy of about 82.7%.

In our dataset roughly 67.14% of the tweets are positive, while 32.86% of the tweets are negative for the market security under study. This result indicates stock/ commodity discussions to be much more balanced in terms of agreement than chat and internet board messages where the ratio of positive to negative from earlier works ranges from 7:1 [8] to 5:1 [10]. Balanced distribution of stock discussion provides us with more confidence to study information content of the positive and negative dimensions of discussion about the stock prices on microblogs.

3.1.2. Feature Extraction and Aggregation. Further positive and negative tweets from each day are aggregated to make weekly time domain indicators which is the time period under study. We selected weekly domain over daily, bi-daily, bi- weekly or monthly as it is the most balanced window to study effect of investor behavior between model performance accuracy keeping in-market monetization potential practically impeccable.

[3] http://techcrunch.com/2011/10/17/twitter-is-at-250-million-tweets-per-day/

[4] Twitter API is easily accessible at- https://dev.twitter.com/docs. Also Gnip - http://gnip.com/twitter, the premium platform available for purchasing historic and present public firehose of tweets has many investors as financial customers researching in the area, though due to confidentiality issues they are not explicitly named

[5] https://sites.google.com/site/twittersentimenthelp/

For every week, value of security (closing, volatility, volume, weekly returns for each index) is recorded at every Friday at closing time of the market trading hour $21:00$ UTC. To explore the relationships between weekly trading and on days when market remains closed (weekends, national holidays) we broadly focus on two domains of tweet sentiments- weekday indices and weekend indices, further referred as *WD* and *WK* respectively. We have carried forward work of Antweiler et al. for defining bullishness (B_t) for each time domain (time window is *WD* or *WK*) given by equation 1.

$$B_t = \ln \frac{1 + {M_t}^{Positive}}{1 + {M_t}^{Negative}} \tag{1}$$

Where ${M_t}^{Positive}$ and ${M_t}^{Negative}$ represent number of positive or negative tweets during particular time period *WD* or *WK*. Logarithm of bullishness measures the share of surplus positive signals and also gives more weight to larger number of messages in a specific sentiment (positive or negative). Message volume is simply defined as natural logarithm of total number of tweets per time domain for a specific security/index. And the agreement among positive and negative tweet messages is defined as:

$$A_t = 1 - \sqrt{(1 - \frac{(M_t^{Positive} - M_t^{Negative})}{(M_t^{Positive} + M_t^{Negative})}} \tag{2}$$

If all tweet messages about a particular company are bullish or bearish, agreement would be 1 in that case. Influence of silent tweets days in our study (trading days when no tweeting happens about particular company) is less than 0.1% which is significantly less than previous research [10; 24].

Every market index/ security thus have a total of 10 potentially causative time series from Twitter positive *WD*, negative *WD*, bullishness *WD*, message volume *WD*, agreement *WD* and from previous weekend we have positive *WK*, negative *WK*, bullishness *WK*, message volume *WK* and agreement *WK*.

3.2. Search Volume Index

Use of trends and patterns in the search engine volumes (SVI) has been used to indicate investor mood being extensively used to track movements in the financial markets by past researchers. To generate search engine lexicon for each of the five securities under study- Oil, DJIA, NASDAQ-100, Gold and Euro we tested by collecting weekly search volume for specific search terms related to respective sectors like- oil, GLD, Dow-30, nasdaq, oil price etc. as given in Table I from **Google Insights of Search**[6]. Google provides this service to provide search volume data at weekly minimum frequency since January 2004. Next we also take into account the top recommended relevant search terms by Google insights of search expanding the already existing number of search terms.

To further normalize and have ease in the computation by applying dimension reduction technique of principle component analysis; we are able to reduce the number of variables (uptil 50 for oil) from search domain by combining similarly behaving time series creating completely uncorrelated co-independent factors- Fact 1 and Fact 2 in our case. Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components which reveals underlying structure that is responsible for maximum variance. Table VIII, XI, X, IX and VII gives the extracted factors by varimax rotation technique to

[6] http://www.google.com/insights/search/

produce orthogonal factors. To identify the factors that cause maximum variance in retweets, we have used Kaiser criterion in which the factors with eigen values greater than 1 are extracted as given in the appendix.

Table I. Google search Terms for 5 Securities

Security or Commodity	Search Terms
US Oil Funds	oil commodity, crude oil, oil etfs, curde oil price, oil futures, oil quotes, oil price per barrel, oil prices bloomberg, wti crude oil, oil prices, how much of oil is left, crude oil ticker + 50 more similar terms etc.
DJIA	djia, dow jones industrial average, dow jones, dow, s&p 500, Stock Market, stock message board
Nasdaq-100	nasdaq up,djia today,dow futures quote,futures quote,djia quote,nylc,bank of america dividends
Gold	buy gold,invest in gold US data,invest in gold worldwide,dollar to pound exchange rate,dollar to pound exchange
Euro	exchange rates converter,dollar euro exchange rate history,rupee exchange rate,oanda currency,rupee exchange,dollar rupee exchange rate,bloomberg live tv,eurusd

3.3. Financial Market Data

We have done analysis in five different sectors broadly oil, DJIA , NASDAQ-100, gold and Euro. Most of the data including all the VIX indices and Euro to USD fedex rates used for analysis are collected from econometrics data from Federal Reserve Bank of St. Louis [7], gold prices are downloaded from World Gold Council [8] and weekly time series for US oil funds and weekly index movements in DJIA and NASDAQ-100 from Yahoo Finance! API[9].

The financial features (parameters) available from Yahoo finance under study are opening (O_t) and closing (C_t) value of the stock/index, highest (H_t), lowest (L_t) value and volume traded for the stock/index. In addition returns are defined as difference between the logarithm of closing values of the stock index between the week's Friday and previous week's Friday.

$$R_t = \{\ln Close_{(t)} - \ln Close_{(t-1)}\} \times 100 \tag{3}$$

Trading volume is the logarithm of number of traded shares every week. We estimate weekly volatility based on intra-day highs and lows using Garman and Klass volatility measures [11] given by the formula:

$$\sigma = \sqrt{\frac{1}{n}\sum \frac{1}{2}[\ln \frac{H_t}{L_t}]^2 - [2\ln 2 - 1][\ln \frac{C_t}{O_t}]^2} \tag{4}$$

Further in this section we will discuss the various security indices in each of the sector under study.

3.3.1. Oil. In this study we have taken USO- United States Oil Fund, an exchange traded fund (ETF) that is one of the highly traded security and strongly tracks movements of light, sweet crude oil purchased and sold at NYSE Arca. We have extracted weekly closing values, volatility and volume parameters from the lexicon. In addition

[7] Federal Reserve Economic Data: http://research.stlouisfed.org/fred2/

[8] http://www.gold.org/investment/statistics/goldpricechart/

[9] http://finance.yahoo.com/

to this we have also taken CBOE OIL volatility index [10] (further referred as VIX) which is index measure of market's expectation of 30-day volatility of crude oil prices.

3.3.2. DJIA. Its an aggregate of 30 traded and influential stock evenly distributed over all sectors. We have taken weekly returns, volatility and volume as parameters under the study. Further we have also extracted CBOE DJIA VIX which is indicative measure of fluctuation in 30-day future index sensitivities.

3.3.3. NASDAQ-100. Its an aggregate of the top 100 stocks from $NASDAQ$ exchange which indexes majority of the technological stocks in the market. For this as well we have taken weekly returns, volatility and volume as parameters under study. In addition we also extracted CBOE NASDAQ-100 VIX which is indicative measure of 30-day ahead index movements.

3.3.4. Gold. We have taken price in US dollar (USD) as its the most traded currency for gold in the world to accurately represent search volumes in each country and related twitter buzz for the precious metal. Further for Gold ETF VIX as well from CBOE as indicative of a month ahead fear-gauge in the price of the precious metal.

3.3.5. Euro. We have taken only two parameters- one Euro to USD (US dollar) conversion rates at closing of the market on Friday's eve for every week and CBOE Euro ETF VIX as measure of 30-day market fear for the same.

4. STATISTICAL TECHNIQUES AND RESULTS

In this section we begin statistical analysis and forecasting performance on each of the financial securities as discussed in section 3.3 from two dynamic components of the short term forecasting hedging model as 10 components from Twitter as discussed in section 3.1.2 and 1 or 2 principle factors from Google SVI as discussed in section 3.2. First we identify correlation patterns various time series at different lagged intervals, further testing the causative relationships of SVI and tweet features on the market securities using econometric technique of Granger's Casuality Analysis. Then we make use of expert model mining system (EMMS) to propose and test the forecasting model and draw performance conclusions.

4.1. Correlation and Cross-Correlation Analysis

We begin our study by identifying pairwise correlation metrics between 10 Twitter features for each security index given in section 3.1.2 and factors derived from SVI search factors given section 3.2.

4.1.1. Technique. As an evaluation of lagged response of relationships existing between financial features, Twitter sentiments and search volumes, we compute cross-correlation at lag of ± 7 week lag to show effectiveness in prediction and motivate us to look forward in making an accurate forecasting model by picking accurate regressor co-efficient.

For any two series $x = \{x_1, ..., x_n\}$ and $y = \{y_1, ..., y_n\}$, the cross correlation lag γ at lag k is defined as:

$$\gamma = \frac{\sum_i (x_{i+k} - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_{i+k} - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}} \qquad (5)$$

In equation 5, $\overline{x}$ $\overline{y}$ are the mean sample values of x and y respectively. Cross-correlation function defined as short for ccf(x,y), is estimate of linear correlation be-

[10] http://www.cboe.com/micro/oilvix/introduction.aspx

tween x_{t+k} and y_t, which means keeping the time series y stationary, we move the time series y backward to forward in time by a lag of k i.e. k= [-7,7] for lags fo 7 weeks in positive and negative direction. Cross-correlation gives the measure of anticipated values of statistically significant relations in a time series x which can be made part of the forecasting equation discussed ahead.

4.1.2. Results. The Table II contains the summarized set of results for financial, Twitter and SVI time series after transformation to log scale.

For Twitter features we examine 5 series- positive, negative, bullishness (Bull), message volume (Msg Vol.) and agreement (Agrmnt) in two cases one as weekday(active market trading) and weekend (during market off days). We realize that the overall nature of relationship exhibit varying degree of association while the clear trend that we observe is that market-off days don't carry high weights when compared to overall data available on comparison to market active days but its still significantly correlated and can be successfully used in hedging strategies as discussed later in this paper for any week start investment. Weekday bullishness is one of the important feature out of all others to look out for any investment and show uniformly significant behavior in all the sectors with value of pearson 'r' as high as -0.73 for DJIA's weekly volatility. Another interesting trend to observe is returns in both DJIA and NASDAQ-100 show negative relationship of varying strength with both positive and negative feeds indicating heavy discussion which is more sensitive to message volume on Twitter before fall in the index, but significantly valuable relationship of 0.593 correlation of returns with our introduced feature term bullishness which is relative measure of positive to negative sentiment of investor community as explained in section 3.1.2. NASDAQ doesn't carry any relationship with weekends Twitter discussions on account fast and dispersive behavior of news memes among tech-savvy investors of technological stocks whom are expected to be faster response to news. For volatility indices (VIX) for various securities shows significant negative relation with weekday agreement index which is vector distance between positive and negative discussion about any security as measure of accurate picture of about to happen turbulence/ perceived market risk in the coming weeks except for DJIA which consists major 30 stocks only which are subjected to highly balanced consistent movements due to heavy trading activity across any time domain.

We find stronger correlation of the principle factors from SVI series uptil 0.826 for commodity funds like for oil, gold and Euro forex rates as compared to index movements of DJIA and NASDAQ-100 giving an impression that people tend to search more for commodity funds then stock equities giving a better control heuristics of actual market movements from investor behavior. VIX is important measure widely used while making investment decisions in futures and options and often referred as "investor gauge fear". From Table II we can see that the VIX is one of the highly correlated financial feature in all the 5 cases, thus maybe referred as a strong measure of investor behavior though computational gauge of "investor fear". Another important significant relation that we observed for NASDAQ-100 and DJIA is the negative correlation with returns in contrast to positive correlation of volatility, volume and VIX which is indicative of search behavior being caused by fall in the index values, increasing more volume in trading making the index movements more volatile.

As we can see in the figure 2 (a), highest correlation is exhibited by oil VIX and SVI which is roughly balanced on both the sides indicating a bi-causative relation in both the directions. Similar observations can be seen for oil fund and oil VIX to Twitter message volume. Tweet message volume have stagnant low slope on the negative lag side which indicates surge in oil related discussion on Twitter consistently prior to actual hike in the price. However interestingly oil fund to SVI is little bent on the

Table II. Pearson Correlation Coefficients Between Security Indices Vs Dynamic behavioral features from Twitter and SVI factors

Market Securities		WeekDay					WeekEnd					SVI	
		Positive	Negative	Bull	Msg Vol	Agrmnt	Positive	Negative	Bull	Msg Vol	Agrmnt	Fact 1	Fact 2
Oil	Close	.604**	.529**	-.304*	.584**	-0.27	.452**	.387*	-0.13	.448**	-0.05	.379*	0.244
	Volatility	-0.09	0.21	-.417**	0.08	-.345*	-0.13	-0.06	-0.07	-0.08	-0.12	.721**	-0.125
	Volume	.342*	.659**	-.734**	.527**	-.673**	0.24	.395**	-.409**	.384*	-.382*	.744**	0.332
	VIX	-0.08	0.22	-.440**	0.09	-.487**	-0.15	0.15	-.368*	0.02	-0.30	.720**	.464**
DJIA	Returns	-.393**	-.440**	.593**	-.470**	0.08	-0.27	-0.25	0.14	-0.27	0.17	-0.064	-0.086
	Volatility	.767**	.767**	-.730**	.814**	0.16	.440**	.615**	-.436**	.545**	-.308*	.450**	0.239
	Volume	-0.17	-0.17	.312*	-0.27	0.08	-0.21	-0.14	0.07	-0.22	0.02	0.024	-.335*
	VIX	.575**	.566**	-.573**	.628**	0.02	.417**	.553**	-.373*	.491**	-0.28	.438**	0.055
NDX[11]	Returns	-0.08	-.363*	.448**	-.333*	-.423**	0.06	0.12	-0.11	0.15	-0.12	-.411*	0.257
	Volatility	0.04	.607**	-.591**	.475**	.563**	-0.12	-0.15	0.07	-0.24	0.21	.505**	0.352
	Volume	0.09	0.25	-0.19	0.25	0.14	-0.17	0.07	-0.21	-0.02	-0.04	0.071	-0.082
	VIX	0.04	.561**	-.548**	.428**	.567**	0.04	-0.10	0.10	-0.07	0.02	.799**	.427*
Gold	USD	0.07	.530**	-0.20	0.23	-.381**	.294*	.574**	-0.19	.350**	-.349**	.826**	
	VIX	-0.19	0.16	-.292*	-0.14	-.321**	-0.22	-0.06	-0.22	-0.16	-0.23	-.370**	
Euro	EURUSD	0.18	0.06	.357**	.276*	0.01	0.12	0.21	0.18	.321**	-0.09	-.540**	.484**
	VIX	-0.09	.281*	-.473**	0.00	-.399**	-0.09	0.11	-.423**	-0.08	-.386**	.822**	0.030

[11] NDX is NASDAQ-100. (p - value $<$ 0.01:***, p - value $<$ 0.05:**, p - value $<$ 0.1:* .)

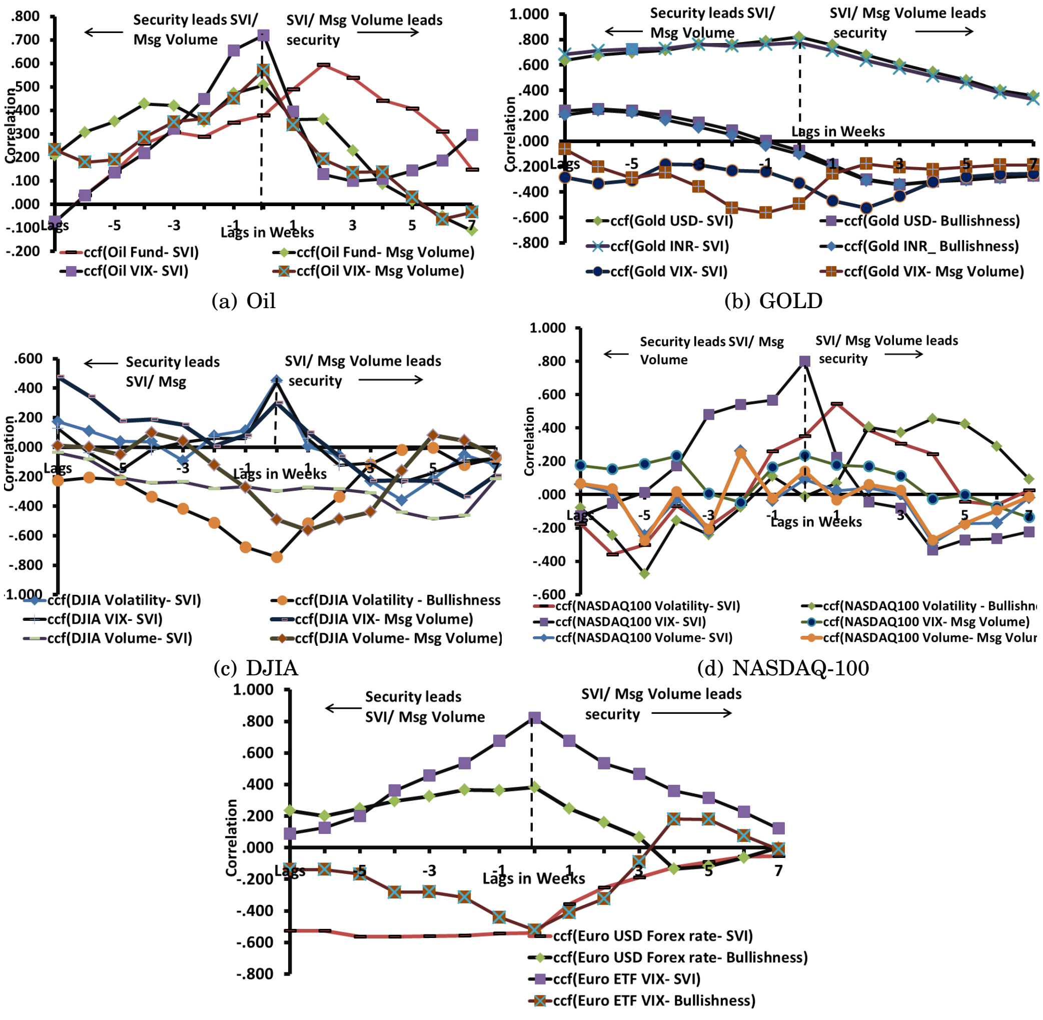
positive lag side with peak of approximately 0.6 at 2 week lag indicating peaks in search volumes for oil continuing upto two after the actual rise in the price.

For DJIA and NASDAQ-100 as observed from figure 2 (c) and (d), much balanced correlation factors can be observed for majority of the pairs in both the cases. However, for DJIA significant bend on the negative lag side is observed by volatility in the index for k=-1, indicating a fall of -0.8 correlation in tweet based bullishness atleast a week before the actual market trading and a similar effect is observed for search volumes uptil 4-5 weeks before the actual trading volume increases. NASDAQ-100's correlation activity doesn't give much insights into relationships between the features which maybe due to non- linear associations or significant relations hidden at smaller time domains frequencies as nature of tech-savvy investors of technological stocks. But we can see that bend on positive k lag side for volatility with search volumes for a week before and constant increase in bullishness prior 2 weeks before actual surge in volatility. However we leave this area for future exploration.

From figure 2 (b) and (e) we can see balanced correlation for gold prices and Euro conversion rates. However important conclusions comes when we see behavior of Gold ETF's VIX, which is negative correlation prior one to two weeks, indicating increase gold related tweet discussions before dip in VIX index occurs, but it shows negative correlation at positive lag with search volumes. In contrast indicating dip in VIX index (fear of buying gold) caused by increased discussion on Twitter as investors consider it as safe investment, hence the confounding effect further observed in the search volumes.

4.2. Granger Causality Analysis

GCA rests on the assumption that if a variable X causes Y then changes in X will be systematically occur before the changes in Y. We realize lagged values of X shall bear significant correlation with Y. However correlation is not necessarily behind causation. Like the earlier approaches by [3; 12] we have made use of GCA to investigate whether one time series is significant in predicting another time series. GCA is used not to establish statistical causality, but as an economist tool to investigate a statistical pattern of lagged correlation. A similar observation that smoking causes lung cancer is widely



(a) Oil (b) GOLD

(c) DJIA (d) NASDAQ-100

(e) EURO

Fig. 2. Cross Correlation of Twitter and SVI features vs commodities like oil (a) and gold (b); stock indices like DJIA (c) and NASDAQ-100 (d); and forex rate of Euro (e)

accepted; proving it contains carcinogens but itself may not be actual causative of the real event.

4.2.1. Technique. Let returns R_t be reflective of fast movements in the stock market. To verify the change in returns with the change in Twitter features we compare the variance given by following linear models in equation 6 and equation 7.

$$R_t = \alpha + \Sigma^n{}_{i=1}\beta_i D_{t-i} + \epsilon_t \tag{6}$$

$$R_t = \alpha + \Sigma^n{}_{i=1}\beta_i D_{t-i} + \Sigma^n{}_{i=1}\gamma_i X_{i-t} + \varepsilon_t \tag{7}$$

Equation 6 uses only 'n' lagged values of R_t , i.e. (R_{t-1}, . . .,R_{t-n}) for prediction, while equation 7 uses the n lagged values of both R_t and the tweet features time series given by $X_{t-1}, ..., X_{t-n}$. We have taken weekly time window to validate the casuality performance, hence the lag values [12]. will be calculated over the weekly intervals $1, 2, ..., 7$.

4.2.2. Results. From the Table III, we can reject the null hypothesis (H_o) that *the SVI and Twitter investor behavior do not affect returns in the financial markets* i.e. $\beta_{1,2,....,n} \neq 0$ with a high level of confidence (high p-values). However as we see the result applies to only specific negative and positive tweets (** for p-value < 0.05 and * for p-value < 0.1 which is 95% and 99% confidence interval respectively). Other features like agreement and message volume do not have significant casual relationship with the returns of a stock index (low p-values).

In Table III and IV we can see that at the lag of one week, almost all the features are significant in predicting changes in the financial features of oil, DJIA, NASDAQ-100, gold and Euro. However as we go in the positive lag direction from 1st to 4 weeks, the significance decreases showing Twitter and SVI mood series as Granger's causative of financial features. SVI shows uniform p values i.e. confidence of uptil 99% for almost all the sectors- both index (DJIA, NASDAQ-100) and commodities (gold, oil and forex rate of Euro). Twitter features specially for the indices- DJIA and NASDAQ-100 don't significance beyond 2-3 weeks, indicating the dispersive nature of information entropy on the social networks in contrast to the SVI factors.

4.3. EMMS model for Forecasting Analysis of Financial features

In this section we work upon the perennial question of *how much? and how good?* are these features proposed in the earlier sections can be useful to make accurate forecasts of financial indicators as per their relative weights. For the same purpose we have used Expert Model Mining System (EMMS) which incorporates a set of competing methods such as Exponential Smoothing (ES), Auto Regressive Integrated Moving Average (ARIMA) and seasonal ARIMA models. These methods are widely used in financial modeling to predict the values of stocks/bonds/commodities/etc [19; 4]. These methods are suitable for constant level, additive trend or multiplicative trend and with either no seasonality, additive seasonality, or multiplicative seasonality.

4.3.1. Technique. Selection criterion for the EMMS is MAPE and stationary R squared which is measure of how good is the model under consideration then the baseline model [23]. The stationary R-squared can be negative with range $(-\infty, 1]$. A negative R-squared value means that the model under consideration is worse than the baseline model. Zero R-squared means that the model under consideration is as good or bad as the baseline model. Positive R-squared means that the model under consideration is better than the baseline model. Mean absolute percentage error (MAPE) is mean residuals (difference between fit value and observed value in percentage). To show the performance of tweet features in prediction model, we have applied the EMMS twice - first with SVI and Twitter sentiment features as independent predictor events and second time without them. This provides us with a quantitative comparison of improvement in the prediction using tweet features.

ARIMA (p,d,q) are in theory and practice, the most general class of models for forecasting a time series data, which is subsequently stationarized by series of transformation such as differencing or logging of the series Y_i. For a non-seasonal ARIMA (p,d,q) model- p is autoregressive term, d is number of non-seasonal differences and q is the number of lagged forecast errors in the predictive equation. A stationary time series ΔY differences d times has stochastic component

[12] *lag at k* for any parameter M at x_t week is the value of the parameter prior to x_{t-k} week. For example, value of returns for the month of April, at the lag of one month will be $return_{april-1}$ which will be $return_{march}$

Table III. Granger's Casuality Analysis- statistical significance (p values) at lags of 1,2,3 and 4 weeks between financial indicators and features of investor behavior

Securities		Lag				
			1	2	3	4
Oil	Close	Positive	.009**	0.1*	0.755	0.238
		Negative	.014**	0.352	0.666	0.204
		Bull	.014**	0.25	0.77	0.238
		Msg Vol	.018**	0.05**	0.911	0.397
		Agreement	0.061*	0.521	0.89	0.421
		SVI	0.038**	0.201	0.006**	0.001***
	VIX	Positive	0.048**	0.966	0.454	0.746
		Negative	0.6	0.683	0.303	0.621
		Bull	0.032*	0.819	0.364	0.742
		Msg Vol	0.078*	0.701	0.706	0.949
		Agreement	0.008**	0.804	0.411	0.957
		SVI	0.00002***	0.001***	0.037**	0.07*
DJIA	Return	Positive	0.675	0.601	0.986	0.266
		Negative	0.065*	0.056*	0.996	0.331
		Bull	0.38	0.442	0.991	0.305
		Msg Vol	0.052*	0.608	0.947	0.237
		Agreement	0.264	0.243	0.826	0.552
		SVI	0.021**	0.053*	0.021**	0.057*
	VIX	Positive	0.461	0.501	0.936	0.683
		Negative	0.024*	0.286	0.91	0.388
		Bull	0.38	0.527	0.672	0.583
		Msg Vol	0.033*	0.05*	0.666	0.97
		Agreement	0.427	0.436	0.616	0.752
		SVI	0.015**	0.06*	0.017**	0.03**
Nasdaq	Return	Positive	0.088*	0.017**	0.049**	0.1*
		Negative	0.737	0.017**	0.076*	0.064*
		Bull	0.061*	0.024**	0.213	0.136
		Msg Vol	0.253	0.218	0.043	0.473
		Agreement	0.091*	0.31	0.988	0.245
		SVI	0.091*	0.081*	0.064*	0.091*
	VIX	Positive	0.076*	0.086*	0.025**	0.042**
		Negative	0.001***	0.31	0.893	0.128
		Bull	0.043**	0.241	0.021**	0.04**
		Msg Vol	0.179	0.427	0.024**	0.148
		Agreement	0.019**	0.229	0.278	0.093*
		SVI	0.0002***	0.002***	0.02**	0.054*
Gold	Price USD	Positive	0.136	0.331	0.631	0.41
		Negative	0.56	0.712	0.807	0.66
		Bull	0.004***	0.023**	0.028**	0.058*
		Msg Vol	0.027**	0.1*	0.625	0.557
		Agreement	0.035**	0.009***	0.015**	0.045**
		SVI	0.0001***	0.00034***	0.00041***	0.001***
	VIX	Positive	0.083*	0.11*	0.192	0.05**
		Negative	0.1*	0.454	0.1*	0.033**
		Bull	0.385	0.641	0.509	0.755
		Msg Vol	0.793	0.1*	0.305	0.256
		Agreement	0.1*	0.385	0.493	0.184
		SVI	0.414	0.059*	0.057*	0.05**

(p - value $<$ 0.01:***, p - value $<$ 0.05:**, p - value $<$ 0.1:*.)

Table IV. (Continued from Table III)

Securities		Lag				
			1	2	3	4
Euro	EURUSD	Positive	**0.051***	**0.11***	**0.1***	0.336
		Negative	**0.043***	0.51	0.249	0.561
		Bull	**0.069***	0.754	0.521	0.497
		Msg	**0.1***	0.439	**0.1***	0.157
		Vol				
		Agreement	0.944	0.985	0.62	0.399
		SVI	**0.00001*****	**0.00006*****	**0.00008*****	**0.0001*****
	VIX	Positive	**0.1***	**0.085***	**0.092***	0.431
		Negative	**0.028****	**0.011****	**0.034****	0.068
		Bull	0.498	**0.1***	**0.1***	0.797
		Msg	0.443	0.256	0.987	0.213
		Vol				
		Agreement	0.384	0.587	0.55	0.557
		SVI	**0.091***	**0.0001*****	**0.002*****	**0.003*****

(p - value $<$ 0.01:***, p - value $<$ 0.05:**, p - value $<$ 0.1:* .)

Where μ_i and σ^2 are the mean and variance of normal distribution, respectively. The systematic component is modeled as:

$$\mu_i = \alpha_i \Delta Y_{i-1} + ... + \alpha_p \Delta Y_{i-p} + \theta_i \varepsilon_{i-1} + ... + \theta_i \varepsilon_{i-q} \tag{8}$$

Where, ΔY the lag-p observations from the stationary time series with associated parameter vector α and ϵ_i the lagged errors of order q, with associated parameter vector. The expected value is the mean of simulations from the stochastic component,

$$E(Y_(i)) = \mu_i = \alpha_i \Delta Y_{i-1} + ... + \alpha_p \Lambda Y_{i-p} + \theta_i \varepsilon_{i-1} + ... + \theta_i \varepsilon_{i-q}(2) \tag{9}$$

Seasonal ARIMA model is of form ARIMA (p ,d ,q) (P,D,Q) where P specifies the seasonal autoregressive order, D is the seasonal differencing order and Q is the moving average order. Another advantage of EMMS model is that it is a stepwise forecasting process which automatically selects the most significant predictors among all other Twitter sentiment series and SVI features.

4.3.2. Results. Model equation for two cases are given below as equation 10 for forecasting without predictors and equation 11 for forecasting with predictors. In these equations Y is the financial feature- oil, gold, DJIA etc. and X represents the investor mood series from SVI and Twitter features.

$$WithoutPredictors : Y_t = \alpha + \Sigma^n{}_{i=1} \beta_i Y_{t=i} + \epsilon_t \tag{10}$$

$$WithPredictors : Y_t = \alpha + \Sigma^n{}_{i=1} \beta_i Y_{t=i} + \Sigma^n{}_{i=1} \gamma_i X_{t=i} + \epsilon_t \tag{11}$$

In the dataset we have time series for a total of 66 weeks, out of which we use approximately 76% i.e. 50 weeks for the training both the models given in equation 11 and 10 for the time period 2nd June 2010 to 27th May 2011. Further we verify the model performance as one step ahead forecast over the testing period of 16 weeks from May 30th to 13 September 2011 which count for wide and robust range of market conditions. Forecasting accuracy in the testing period is compared for both the models

Table V. Forecasting results for the financial securities

Market Securities		Predictors	MAPE	Direction
US Oil Funds	Index	Yes	2.3202	75
		No	2.4203	62.5
	VIX	Yes	4.5592	75
		No	5.1218	56.3
DJIA	Index	Yes	0.8557	94.3
		No	1.1698	60
	VIX	Yes	5.3017	82.9
		No	5.6943	62.9
NASDAQ-100	Index	Yes	1.3235	90
		No	1.3585	50
	VIX	Yes	3.2415	83.3
		No	5.7268	50
Gold	USD	Yes	1.5245	78.6
		No	1.5555	64.3
	VIX	Yes	0.2534	71.9
		No	5.2724	56.1
Euro	EURUSD	Yes	2.6224	74.1
		No	4.3541	58.6
	VIX	Yes	4.4124	69
		No	4.7878	53.4

in each case in terms of mean absolute percentage error (MAPE) and the direction accuracy. MAPE is given by the equation 12, where $\hat{y_i}$ is the predicted value and y_i is the actual value.

$$MAPE = \frac{\Sigma^{n}{}_{i}|\frac{y_i - \hat{y_i}}{y_i}|}{n} \times 100 \tag{12}$$

While direction accuracy is measure of how accurately market or commodity up/ down movement is predicted by the model, which is technically defined as logical values for $(y_{i,\hat{t}+1} - y_{i,t}) \times (y_{i,t+1} - y_{i,t}) > 0$ respectively. This is of prime importance to the high frequency traders and investors who hedge their investment in derivative markets as lots of prices (option premium, bonds etc.) are solely determined by the direction of the moving index or price.

As given in Table V we observe that the there is significant reduction in the values of MAPE for all the sectors for the forecasting model with the use of predictor sentiment and SVI series than the predictor model without the use of the these predictor series. Also for index values of DJIA direction accuracy of uptil 94.3% is achieved, while it is for 90% for NASDAQ-100. SVI and measure of wisdom of crowd on Twitter gives quite a robust picture of how changing dynamics of the public opinion can be reflective of the market movements that would happen in near future.

5. DISCUSSIONS

From Table VI, we can see that earlier works in the area of behavioral finance were limited to profile of mood states and dimensions of public mood in context of investing. In simple words it focusses on finding when is general public more likely to invest and feeling positive or negative about the bullish or bearish market performance. Besides general trends like thanksgiving, Christmas etc. which are recurrent and pre known to the professionals, it also captures out of the box market situations like recession, natural calamity or tense political environment. However these approaches are not scalable to individual commodity or underlying security. Sensitivity between different market instruments show wide variance in response to investor behavior which is prime importance for portfolio managers with investments across diverse set of securities and commodities. Henceforth this research dives into identifying the pattern of market sentiment over market active and non-trading days individually for different securities or commodities.

Table VI. Prior research in sentiment analysis for predicting sentiment analysis

Previous Approaches	Bollen et al. [3; 18] and Gilbert et al. [12]	Sprenger et al. [24]	Our Approach
Approach	Mood of complete Twitter feed	Stock Discussion with ticker $ on Twitter	Combining Twitter sentiment + Google search volumes
Dataset	28th Feb 2008 to 19th Dec 2008, 9M tweets sampled as 1.5% of Twitter feed	1st Jan 2010 to 30th June 2010- 0.24M tweets	2nd June 2010 to 13th Sept 2011- 1.9M tweets through search API
Techniques	SOFNN, Grangers and linear models	OLS Regression and Correlation	Cross- Corr, GCA, Expert Model Mining System (EMMS)
Results	86.7% directional accuracy for DJIA	Corr values uptil 0.41 for S&P 100 stocks	Corr uptil 0.82 for OIL, DJIA, NASDAQ-100, Gold & Euro. Directional accuracy uptil 94%
Feedback/ Drawbacks	Individual modeling for stocks not feasible	News not taken into account, very less tweet volumes	Comprehensive and customizable approach

Primary objective is to bring out a uniform model combining search volume behavior along with *how people are speaking and about what? on Twitter* and observe how severe or accurate these effects get over the increasing time lag. Tracking sentiments of discussions revolving around the key security or commodity instead of just stock discussions gives more comprehensive measure of public demand and market fear. Use of bullishness, agreement and message volume provides additional features to measure sentiment in a subjective manner. As seen in Table V, we observe one of the most significant improvements can be observed for NASDAQ-100's VIX (MAPE- 3.2415) and Gold VIX (MAPE- 0.2534), indicating the tech savvy investors holding significant power for the index movements. Comparing general prediction performance of behavior features (SVI + Twitter sentiment series) for market indices and commodity prices over the VIX index; better performance can be observed for VIX accounting for the fact that these behavior features are better indicative of investor fear before the actual price movement occurs in the stock. However for the forex price of Euro, investor sentiment is more centralized factor in controlling the price movement as compared to the VIX index. Modeling market sentiment is luring area that investors are looking forward for use in hedging the investment instruments. The above discussed model can be a successful in decision making events like conversion of risk into a safely hedged security during oncoming bearish market conditions.

6. CONCLUSION AND FUTURE WORK

Proposed approach unifying Twitter sentiment with search volumes is scalable, with the decision-taking around the breaking news dynamics powered by Twitter sentiments incorporated into the prediction process. It combines the advantage of sophisticated statistical and linguistics summarization techniques. We are able to capture a good picture of both the changing rates and rising & falling probabilities of both commodities and stocks. It is no surprise that this approach is far more robust than its predecessors in tracking complexities comprising of search behavior combined with the sentiment changes across the microblogs. Moreover, as far as practical implementation is concerned, our approach not only helps to improve index movements but also the present volatility and the VIX index which is the measure of the 30-days ahead market fear. More importantly it can be also used to determine risk to security conversion decisions for hedging the portfolio with greater confidence. Future exploration in this area might compromise of changing tracks and dynamics of blogging taking location specific data along with more refinement with new approaches and perspectives into examining local exchanges.

REFERENCES

Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. *Computing*, 25(1):492499, 2010.

Mark Bagnoli, Messod D. Beneish, and Susan G. Watts. Whisper forecasts of quarterly earnings per share. *Journal of Accounting and Economics*, 28(1):27 – 50, 1999.

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *Computer*, 1010(3003v1):1–8, 2010.

George Edward Pelham Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990.

danah m. boyd and Nicole B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

Zhi Da, Joseph Engelberg, and Pengjie Gao. In search of attention. *Russell The Journal Of The Bertrand Russell Archives*, (919), 2010.

Sanjiv R. Das and Mike Y. Chen. Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web. *SSRN eLibrary*, 2001.

Michal Dewally. Internet investment advice: Investing with a rock of salt. *Financial Analysts Journal*, 59(4):65–77, 2003.

S. Doan, B.-K. H. Vo, and N. Collier. An analysis of Twitter messages in the 2011 Tohoku Earthquake. *ArXiv e-prints*, September 2011.

Murray Z. Frank and Werner Antweiler. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *SSRN eLibrary*, 2001.

Mark B. Garman and Michael J. Klass. On the estimation of security price volatilities from historical data. *The Journal of Business*, 53(1):67–78, 1980.

Eric Gilbert and Karrie Karahalios. Widespread worry and the stock market. *Artificial Intelligence*, pages 58–65, 2010.

Alec Go, Richa Bhayani, and Lei Huang. Twitter Sentiment Classification using Distant Supervision.

Erkam Guresen, Gulgun Kayakutlu, and Tugrul U. Daim. Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, 38(8):10389 – 10397, 2011.

Yigitcan Karabulut. Can Facebook Predict Stock Market Activity? *SSRN eLibrary*, 2011.

Alina Lerman. Individual Investors' Attention to Accounting Information: Message Board Discussions. *SSRN eLibrary*, 2011.

H. Mao, S. Counts, and J. Bollen. Predicting Financial Markets: Comparing Survey,News, Twitter and Search Engine Data. *ArXiv e-prints*, December 2011.

Huina Mao, Scott Counts, and Johan Bollen. Predicting financial markets: Comparing survey,news, twitter and search engine data. Quantitative Finance Papers 1112.1051, arXiv.org, December 2011.

Garth P. McCormick. Communications to the editorexponential forecasting: Some new variations. *Management Science*, 15(5):311–320, 1969.

Hong Miao, Sanjay Ramchander, and J. K. Zumwalt. Information Driven Price Jumps and Trading Strategy: Evidence from Stock Index Futures. *SSRN eLibrary*, 2011.

John R. Nofsinger. Social mood and financial economics. *Journal of Behavioral Finance*, 6(3):144–160, 2005.

Bo Qian and Khaled Rasheed. Stock market prediction with multiple classifiers. *Applied Intelligence*, 26:25–33, February 2007.

Laura Sabani. Forecasting structural time series models and the kalman filter, a. c. harvey. cambridge university press, cambridge, 1989. isbn 0-521-32196-4, cloth, 55.00 pp. xvi + 554. *Journal of Applied Econometrics*, 6(3):329–331, 1991.

Timm O. Sprenger and Isabell M. Welpe. Tweets and Trades: The Information Content of Stock Microblogs. *SSRN eLibrary*, 2010.

Martin Szomszor, Patty Kostkova, and Ed De Quincey. swineflu : Twitter predicts swine flu outbreak in 2009. *3rd International ICST Conference on Electronic Healthcare for the 21st Century eHealth2010*, (December), 2009.

PAUL C. TETLOCK. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *International AAAI Conference on Weblogs and Social Media Washington DC*, pages 178–185, 2010.

Amos Tversky and Daniel Kahneman. Prospect theory: An analysis of decision under risk. Technical report, 1979.

Peter Wysocki. Cheap talk on the web: The determinants of postings on stock message boards. *Working Paper*, 1998.

Max Zeledon. Stocktwits may change how you trade, 2009. This is an electronic document. Date of publication: [2009]. Date retrieved: September 01, 2011. Date last modified: [Date unavailable].

Xue Zhang, Hauke Fuehres, and Peter A Gloor. Predicting stock market indicators through twitter i hope it is not as bad as i fear. *Anxiety*, pages 1–8, 2009.

7. APPENDIX

The PCA component matrixes for Oil, DJIA, NASDAQ-100, Gold and Euro are given in Tables VII, VIII, IX, X and XI respectively. Feature reduction is an important step before development of any model so as to increase predictive accuracy, simplicity and comprehensibility of the mined results. Effect of so many search terms can be concisely mapped to double or single factors i.e. original high-dimensional data onto a lower dimensional space. The new PCA factors are uncorrelated, and are ordered by the fraction of the total information each retains and filtered out on the basis of Kaiser criterion, that is the factors with the eigen value of greater than 1. Each of the search term factors (Fact 1 and Fact 2) explain significant amount of variance as in the original feature set of search keywords given in the Tables below.

Table VII. Vector Matrix for Oil SVI Factors

Search Terms	Oil SVI search term Factors	
	Fact 1	Fact 2
oil commodity	.671	.688
crude oil etf	.319	.897
oil funds	.304	.900
oil etf	.303	.898
oil quotes	.593	.599
oil prices per barrel	.853	.338
spot oil prices	.727	.533
wti crude	.361	.504
how much oil is left	.452	.446
futures price	.521	.759
how to buy oil	.298	.918
oil ticker	.634	.655
current oil	.741	.021
crude oil futures	.441	.639
crude oil price	.442	.700

Table VIII. Vector Matrix for DJIA SVI Factors

Search Terms	Dow SVI search term Factors	
	Fact 1	Fact 2
djia	.931	.038
dow jones industrial average	.811	.495
dow jones	.929	.116
dow	.966	.010
s&p 500	.635	-.032
Stock Market	.689	-.337
stock message board	-.012	.936

Table IX. Vector Matrix for Nasdaq SVI Factors

Search Terms	Nasdaq SVI search term Factors	
	Fact 1	Fact 2
nasdaq today	.833	.398
dow futures quote	.888	.239
futures quote	.673	-.004
NASDAQ quote	.821	.326
nylc	.257	.936
bank of america dividends	.161	.947

Table X. Vector Matrix for Gold SVI Factors

Search Term	Gold SVI search term factors Fact 1
buy gold	.575
invest in gold US data	.942
invest in gold worldwide	.884
dollar to pound exchange rate	.905
dollar to pound exchange	.904

We obtain only 1 factor for gold as most of the search terms for gold fall on the same dimension plane in the feature vector map.

Table XI. Vector Matrix for Euro SVI Factors

Search Terms	Euro SVI search term factors	
	Fact 1	Fact 2
exchange rates converter	.530	-.383
dollar euro exchange rate history	.798	.079
rupee exchange rate	-.047	.063
oanda currency	.105	.810
rupee exchange	.929	-.102
dollar rupee exchange rate	.938	.043
bloomberg live tv	.828	-.228
eurusd	-.237	.741