Computational Economic and Finance Gauges: Polls, Search, & Twitter

Huina Mao
Indiana University-Bloomington
Bloomington, Indiana,U.S
huinmao@indiana.edu

Scott Counts
Microsoft Research
Redmond,Washington,U.S
counts@microsoft.com

Johan Bollen
Indiana University-Bloomington
Bloomington, Indiana,U.S
jbollen@indiana.edu

ABSTRACT

We investigate whether the results of a variety of well-accepted socio-economic surveys, which are generally obtained from opinion polling, can be replicated (and predicted) by Computational Economic and Financial Gauges (CEFG) extracted from large-scale search engine and Twitter data. In particular, we examine the results of the Michigan Consumer Confidence Index, Gallup Economic Confidence Index, Unemployment Insurance Weekly Claims reported by U.S. Department of Labor as well as two investor sentiment surveys (i.e. weekly Investor Intelligence and Daily Sentiment Index). Our results show that CEFGs not only exhibit statistically significant correlations to many if not most existing socio-economic indices, but precede and thus predict survey data. We furthermore find that a CEFG of investor sentiment obtained from Twitter may be a leading indicator of the financial markets which existing surveys tend to lag.

Keywords

Consumer confidence index, unemployment rate, investor sentiment, Twitter and search engine.

1. INTRODUCTION

Predicting future economy and financial conditions is a matter of considerable interest to government and corporate entities. Prediction models, however, require information on past and current conditions, hence the use of economic and financial indicators. Some of the most well-known indicators include inflation and interests rates, GDP growth/decline, retail sales, consumer confidence, unemployment rates, and investor sentiment. In addition there exists a plethora of indicators of specific economic measures of inflation, interest rates and Gross Domestic Product. In this paper, we are specifically interested in a range of measures of consumer confidence, unemployment rates and investor sentiment. Because traditionally, surveys are the most direct method to capture these indicators, we aim to find a more efficient way to measure them.

The Michigan Consumer Confidence Index (CCI) is a widely accepted consumer confidence index. CCI is reported monthly by the Reuters/University of Michigan Surveys of Consumers[1]. This monthly poll is based on approximately 500 telephone interviews where adult men and women in the United States answer five questions about their evaluation on past and current financial conditions, and the expected financial outlook for their own household and the country. (See the five questions in the Appendix). Responses are combined into the index score which is intended to reflect how optimistic/pessimistic consumers feel about the current and future economic conditions.

The CCI is however recorded on a monthly basis. To make a refined comparison with our web data that is sampled more frequently (i.e. daily), we use another consumer confidence poll – daily Gallup Economic Confidence Index [2]. The Gallup ECI is based on telephone interviews with approximately 1,500 adults in the United States. Interviewees are asked to rate their opinion of current economic conditions and their overall economic outlook. The answers are combined and scaled to a range of -100 (negative) to 100 (positive).

Previous research has shown that consumer confidence is affected by many factors, such as inflation, interest and unemployment rates. Results from [1] indicate that awareness of unsatisfactory employment conditions affects economic expectations. Therefore, due to the influence of the unemployment rate on consumer confidence, as well as being an important economic indicator itself, we develop a computational gauge of unemployment and compare it with the weekly United States unemployment claims issued by the US Department of Labor [3].

Overall, consumer confidence and the unemployment rate are indicators of *economic* trends. Investor sentiment is a specific indicator to measure how investors tend to feel about near term prospects for the *stock market*. There are many polls on investor sentiment provided by different investment services, such as Merrill Lynch, Investor Intelligence, American Association of Individual Investors (AAII) (see details from [5]), and Daily Sentiment Index. These polls are well accepted in the finance field. Given their availability this paper, we use weekly Investor Intelligence (II) [4] and Daily Sentiment Index's (DSI)[5]. Since 1964, II has been published by an investment services company study, based on approximately hundred independent market newsletters. It assesses each author's current stance on the market, i.e. bullish, bearish or correction. In this paper, we use the bull percentage as II sentiment. DSI provides daily market sentiment readings on all active US markets daily since 1987. High bullish readings (i.e. 90% or higher) suggest that a short-term top is developing or has been made. Low bullish readings (i.e. 10% or lower) suggest that a short-term bottom is developing or has been made.

The main goal of doing these polls for consumer confidence, unemployment and investor sentiment is to model how well Americans feel about the economy and financial market, and thus predict the economic and financial climate to come. In particular, stock market prediction has attracted much attention from both academia and industry. Stock market prediction, although a particularly difficult problem, is a topic of tremendous commercial interest. However, history has shown us that investor sentiment polls are a trailing, rather than a leading, indicator for the stock market. By definition, a trailing indicator cannot

[1] http://www.sca.isr.umich.edu/

[2] http://www.gallup.com/poll/122840/Gallup-Daily-Economic-Indexes.aspx

[3] http://www.dol.gov/opa/media/press/eta/ui/current.htm

[4] http://www.investorsintelligence.com/x/us_advisors_sentiment.html

[5] http://www.trade-futures.com/dailyindex.php

foreshadow the future. Additionally, some think of investor sentiment from polls as a *contrarian* indicator. Quoted from the website of II [4], *"When the survey was developed by our founder, AW Cohen, he originally expected that the best time to be long the market was when most advisors were bullish. This proved to be far from the case – a majority of advisors and commentators were almost always wrong at market turning points"*

More recent efforts have turned to computationally extracting public opinion and sentiment from the tremendous amount of web data available online. Compared with polls, these computational measures can in principle offer considerable advantages: they rely on public data, and are recorded at very short time intervals and at a very large scale. In [2], the authors found a high correlation between sentiment word frequency from Twitter messages and surveys on consumer confidence and political opinion. A significant positive association is found between job-search variables and official unemployment data [3]. As a further step, [4] not only show a good correlation between Google search volume on "*jobs, welfare & unemployment*" and official unemployment data, but also found that he former helps predict the latter.

Existing work has also been done to create a proxy of investor sentiment. For example, in [6], negative economic search queries serve as a proxy of investor sentiment. In [7], news media content is taken as a proxy for investor sentiment or non-informational trading. In [9], a six-dimensional model of public mood (Calm, Alert, Sure, Vital, Kind and Happy) is extracted from Twitter and shown to be a strong predictor of market fluctuations. However, to the best of our knowledge, no previous research has compared traditional investor sentiment polls with computational measures.

In this paper, we extend prior work with the following contributions. First, we extract a series of Computational Economic and Financial Gauges (CEFG) from two popular web data sources: (1) Google, one of the most significant search engines and (2) Twitter, the most popular microblogging service. Second, we compare the resulting CEFGs with several polls on consumer confidence, unemployment and investor sentiment. Third, since investor sentiment surveys are typically a lagging/trailing indicator, we investigate whether any of our CEFGs can serve as a non-lagging indicator. Our research may help provide a new angle for modeling economic and financial relations from online data.

The rest of the paper is organized as follows. In Section 2, we conduct a pair-wise comparison among polls measuring these three economic and financial indicators (consumer confidence, unemployment and investor sentiment). In Section 3, we present results based on the search engine and poll data. In section 4, Twitter Investor Sentiment (TIS) is compared with the Daily Sentiment Index (DSI), we also compare both TIS and DSI with the stock market. Concluding remarks and future work are gathered in Section 5.

2. Intra-survey comparison

Our survey data includes monthly Michigan Consumer Confidence Index (CCI), daily Gallup Economic Confidence Index (ECI) and Job Creation Index (JCI), season-adjusted weekly initial unemployment claims (official unemployment rate) issued by the US Department of Labor as well as two investor sentiment polls: Daily Sentiment Index (DSI) and weekly Investor Intelligence (II). Here, we compare CCI with the official unemployment rate and investor sentiment from January 2004 to July 2011, as shown in Figure 1. Due to data availability issues, we only have Gallup data (ECI and JCI) from November 2008 onwards. Therefore, we can compare the Gallup ECI, JCI with the Michigan Consumer Confidence Index, Official Unemployment Rate and Investor Intelligence from November 2008 to July 2011 (see Figure 2 for results).

Because the data are collected at different time intervals (days, weeks or months), we convert the lower level (daily/weekly) to higher level (weekly/monthly) by taking the mean values in the same week or month. For example, in order to obtain the correlation coefficient between weekly Investor Intelligence and the monthly CCI, we take the mean value of these 4 weeks II in January as the II value in January, and so on for the other months.

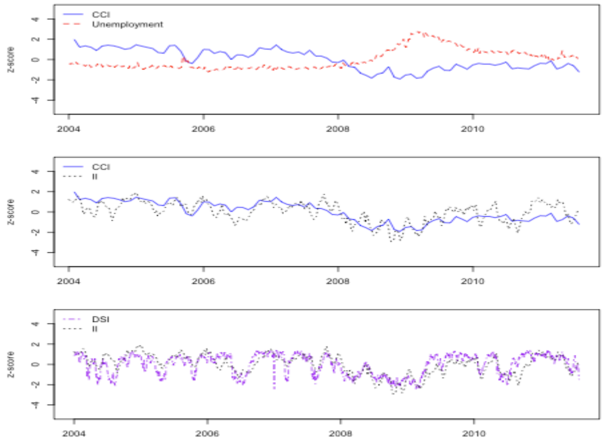

Figure 1: Comparison among Michigan Consumer Confidence Index (CCI), Unemployment, Investor Sentiment (II and DSI) from January 2004 to July 2011.

From Figure 1, we observe that changes in the Michigan consumer confidence index are related to changes in the employment rate. Indeed, the two series have a significant negative Pearson correlation correlation (γ = -0.73), indicating that the job market maybe one factor in explaining the change of consumer confidence. From the middle and bottom panels of Figure 1, we see a strong relation between the consumer confidence index and investor sentiment (γ=0.63). The two investor sentiment measures (II and DSI) have very similar trends and we found that γ = 0.72.

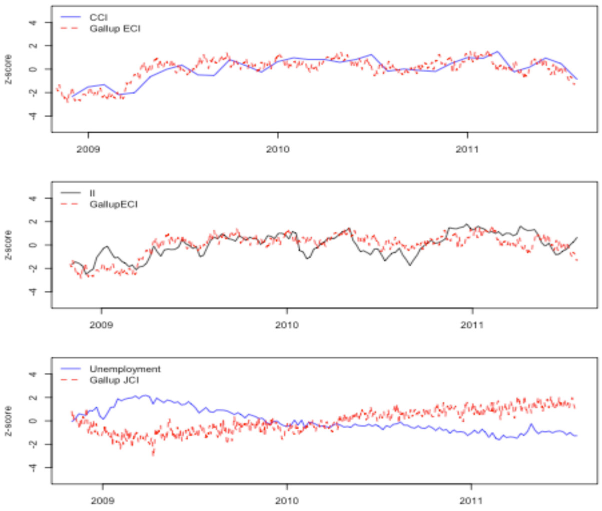

Figure 2: Comparison among Gallup Economic Confidence Index (ECI), Job Creation Index (JCI), Michigan CCI, and Unemployment Rate from November 2008 to July 2011.

The correlation between CCI and Gallup ECI is γ=0.88; the correlation between Gallup ECI and II is γ = 0.68, between unemployment data and job creation index, γ = -0.86. All the correlations are statistically significant, with *p-values* < 0.01.

In sum, we have found that two consumer confidence indexes (i.e. Michigan Consumer Confidence Index and Gallup Economic Confidence Index) have a high positive correlation (γ=0.88), and two investor sentiment indexes (Investor intelligence and Daily Investor Sentiment Index) have a high positive correlation (γ=0.72). Additionally, we found a negative correlation between unemployment and consumer confidence, implying that when the unemployment rate is high, people do not feel confident about future economic conditions. The high positive correlation between Investor sentiment and Consumer Confidence indexes indicates the close relation between the economy and financial markets.

In the next two sections, we aim to computationally obtain economic and financial market indicators, i.e. Consumer Confidence Index, Unemployment Rate, and Investor Sentiment from our two web data source, namely Google Insight Search and Twitter.

3. Analysis on Web Search and Surveys

3.1 Search Engine Data

Google released Google Insight Search (GIS), a product that provides temporal and spatial information for a given query on a weekly basis. The prediction power of Google query data has been explored in previous work, both in finance and other fields [6,8]. The 19 search queries we adopted are the Financial and Economic Attitudes Revealed by Search (FEARS) terms constructed in [6]. (See the Appendix for the word list). Due to different data availability of these two search engines, we used the weekly search volume of GIS from January 2004 to July 2011.

3.2 Does Google search help predict the polls?

3.2.1 Correlation between GIS and Consumer Confidence, Unemployment Rate, Investor Sentiment.

In this section, we compare the search volumes for the 19 FEARS terms with Michigan CCI, official unemployment rate and Investor Intelligence (II). The results are very good. Due to space limitations, we select the top 5 search queries according to their correlation coefficients. The list of their corresponding top five search queries are shown from (1) to (3) respectively as follows:

(1) For CCI: *recession, unemployment benefits, unemployment office, unemployment* and *credit card debt*;

(2) For Unemployment Rate: *unemployment office, unemployment benefits, unemployment, credit card debt* and *recession;*

(3) For II: *recession, credit debt, credit card debt, unemployment office* and *unemployment benefits.*

Their correlation coefficients are shown in Table 1.

Table 1: Top 5 Correlation coefficients between GIS and these three polls (CCI, Unemployment Rate (UR) and II)

Indicators	1	2	3	4	5
CCI	**-0.739**	**-0.711**	**-0.709**	-0.666	-0.562
UR	**0.882**	**0.878**	**0.843**	0.693	0.667
II	**-0.608**	-0.448	-0.432	-0.404	-0.294

We can see that among these search queries, "*recession*" has the highest correlation with CCI and II. The term also is also highly correlated with the unemployment rate. In Figure 3 we plot "*recession*" against these three indicators. As can be seen, the GIS trend is very positively correlated with unemployment rate. This confirms previous research [3,4] – web search can produce accurate, useful statistics about the unemployment rate. As an extension, in this paper we also compared *"recession"* with consumer confidence and investor sentiment, and found significantly high negative correlations, indicating that the more people search on negative economic words (e.g. *recession, unemployment benefits, unemployment office*, etc), the less they feel confident about the economy and the financial markets.

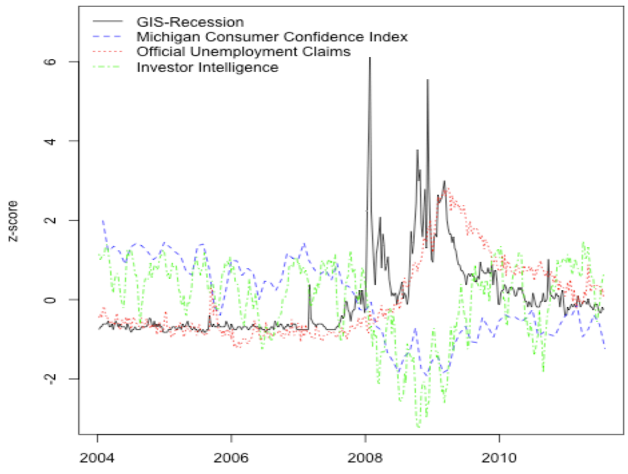

Figure 3: GIS search on "*Recession"* and Michigan CCI, official unemployment rate and II from Jan 2004 to Jul 2011.

We have established that there exists a high correlation between GIS search and poll data. However, an even more important question is "*whether GIS precedes polls?*". To answer this question, we conduct the following study.

3.2.2 Cross correlation and Granger causality

To see whether GIS is a leading and useful indicator for poll prediction, we perform a cross-correlation and Granger causality analysis. From the above section, GIS has a better correlation with Unemployment Rate than with CCI and Investor Sentiment. Therefore, we take unemployment rate as an example.

Cross-correlation is a standard method of estimating the degree to which two series are correlated. Consider two series $x = \{x_i, \ldots, x_n\}$ and $y = \{y_i, \ldots, y_n\}$, the cross correlation r at lag k is then defined as:

$$r = \frac{\sum_i (x_{(i+k)} - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_{(i+k)} - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \qquad (1)$$

where $\bar{x}$ and $\bar{y}$ are the sample mean values of the x and y, respectively. We use the cross-correlation function provided in ccf, an R statistics package. For example, $ccf(x,y)$ estimates the correlation between $x[t+k]$ and $y[t]$. When $k > 0$, it means that y leads x, and vice versa. Here we are interested in the cross-correlation between GIS and Unemployment rate. i.e. ccf(Unemployment rate, GIS). The result is shown in Figure 4.

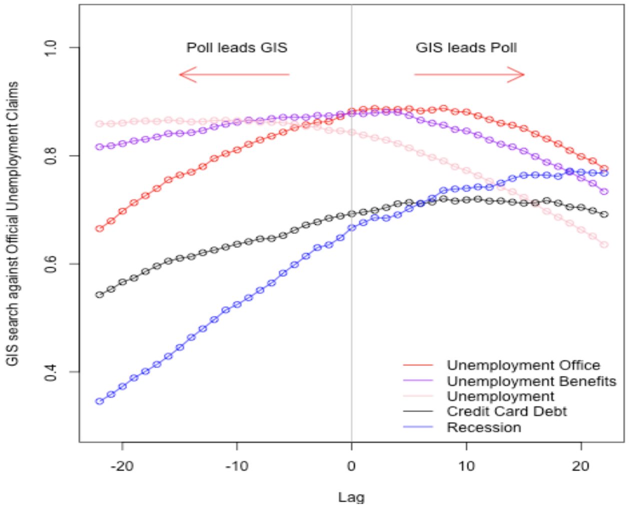

Figure 4: Cross-correlation between GIS search queries and Unemployment Rate (Separated by Lag = 0, the right side means GIS leads poll, while left side indicates that poll leads GIS)

In Figure 4, it can be observed that *unemployment and unemployment benefits* falls off faster for GIS-leads-Polls than Polls-leads-GIS. So, the polls seem to be leading indicators for these two queries. However, the correlations for *unemployment office*, *recession* and *credit card debt* are higher on the GIS-leads-Polls side than the other side. Especially, *recession* and *credit card debt* correlations increase on the right side while the lag increase. Thus GIS on *recession, credit card debt* and *unemployment office* appears to be a strong leading indicator.

Among these five search queries, *recession* shows the strongest leading property. Therefore, we choose *recession* as an example to test the *Granger Causality* significance of GIS search queries in forecasting the Official Unemployment Rate. Granger causality is usually used to find out whether changes in a variable will have an impact on changes other variables. It is assumed that if variable X causes Y then changes in X will systematically occur before changes in Y. An F-test is run on both the full model (with historical X values added) and the limited model (only the historical Y values). If the *p-value* < 0.05, X is statistically significant in predicting Y.

Based on the weekly GIS for "*recession*" and the Unemployment rate from Jan 2004 to July 2011, we conduct a Granger causality test and plot the *p-value* for both causal directions as shown in Figure 5. We can see that only when lag < 10 days, Official unemployment claims are a statistically significant predictor of GIS-*Recession*. However, it is always significant in the other direction (p << 0.05). Therefore, we conclude that GIS search is a useful indicator for unemployment rate prediction.

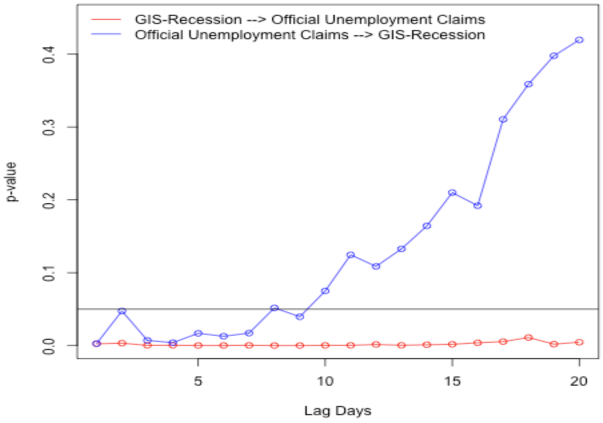

Figure 5: p-values of Granger Causality between GIS search - Recession and Official Unemployment Claims; (The vertical black line is where p-value = 0.05; when p < 0.05, it is statistically significant.)

For consumer confidence and investor intelligence, cross-correlation and granger causality analysis results have shown the similar leading property of GIS, though not as strong as for unemployment rate.

3.2.3 Forecasting Analysis.

As a further validation, we do forecasting analysis by adding the GIS data. We first smooth the weekly GIS to derive a more consistent signal based on the past *k* weeks data:

$$MA = \frac{1}{k}(X_{t-k+1} + X_{t-k+2} + \ldots + X_t) \qquad (2)$$

Here, we arbitrarily define *k* = 10. There are three linear models for forecasting, which are defined as follows:

$$\begin{aligned} &\textbf{M1:}\ \ Y_{t+L} = b_0 + b_1 Y_t + \varepsilon_t \\ &\textbf{M2:}\ \ Y_{t+L} = b_0 + aMA + \varepsilon_t \qquad (3) \\ &\textbf{M3:}\ \ Y_{t+L} = b_0 + b_1 Y_t + aMA + \varepsilon_t \end{aligned}$$

L indicates to predict the *L*-step value, i.e. to predict the unemployment rate in *t*+*L* week). Training data is through *t-1*, and we use the window ending on week *t* for forecasting. Model 1 and Model 2 means to predict Y (i.e. unemployment rate) only based on historical data Y or X. To see whether X can improve forecasting accuracy, we combine M1 and M2, which gives Model 3. During the period from Jan 2004 to July 2011, we correlate the forecasting results with the real values. The correlation coefficients are shown in Figure 6.

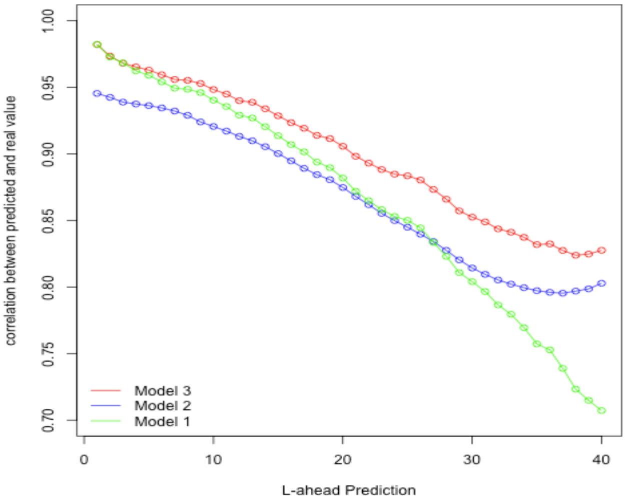

Figure 6: Correlation between the real values and predicted values obtained from M1, M2, M3) respectively.

From Figure 6, we can see that when predicting the unemployment rate on *L* <=3 weeks ahead, M3 and M1 have similar results. When *L* >=4, M3 starts performing slightly better and increases as L increases. When *L* > 18, M2 (i.e. only using GIS) can be equally good and even better than M1 as *L* continues to increase. Therefore, all our results demonstrate that GIS is very useful for predicting the unemployment rate, especially the long-term value (i.e. *L* >=3 weeks). For a clear view on the prediction results, we plot the real unemployment rate against the forecasted values obtained from these three models as shown in Figure 7. The correlation coefficients between real value and M1, M2 and M3 predicted values are 0.881, 0.875, and 0.906, respectively.

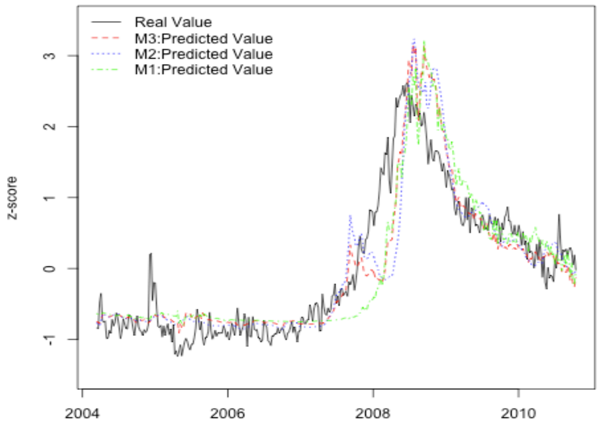

Figure 7: Real unemployment rate and the forecasted values from M1, M2 and M3; (L=20, k=10)

4. Investor Sentiment Analysis: Twitter Investor Sentiment (TIS) and Daily Sentiment Index (DSI).

4.1.1 Poll -- Daily Investor sentiment

In Section 2, we have seen that both measures of investor sentiment (i.e. Daily Sentiment Index and Investor Intelligence) are strongly correlated (γ = 0.72). For this study, we choose Daily Sentiment Index, because our Twitter data is sampled daily. Due to data availability, the study period is from July 1st 2010 to August 30th 2011, i.e. 426 days in total.

DSI was initiated in 1987 to gather the opinions of small retail traders on all active US futures markets, represented as the percentage of trader bullish. It has been used by top banks, money managers, brokerage firms, professional traders and speculators throughout the world. Similar to other market sentiment measures, DSI is also a contrary opinion indicator. It means that if a majority of traders agrees on the direction of a market move, then the odds are significant that prices will, in fact, move in the opposite direction.

4.1.2 Extract investor sentiment from Twitter.

Twitter is the most popular Microblogging service. Users post messages with less than 140 characters, averaging 11 words per message. We randomly collect 30% of all the public tweets on every day for our analysis. Then, we derive daily sentiment scores by counting bullish and bearish messages. A message is defined as bullish if it contains the term "*bullish"*, and bearish if it contains the "*bearish"*. We then define the investor sentiment score, Twitter Investor Sentiment (TIS) on day ***t*** as the ratio of bullish tweets over all the bullish and bearish tweets on that day:

$$TIS = \frac{N_{Bull}}{N_{Bull} + N_{Bear}} \qquad (3)$$

N_{Bull} is the number of bullish tweets on day ***t*** ; N_{Bear} is the number of bearish tweets on day ***t*** .

4.1.3 Correlation analysis between TIS and DSI.

During this period, the correlation coefficient between TIS and DSI is *r* = 0.348 (p-value = 1.468e-13), thus indicating a highly significant correlation between these two series. Because the daily TIS is very volatile, we smooth the raw TIS based on its historical values (see Equation (2)). When the length of the smoothing window is 10, 30, 50, 60, 90 days, their corresponding correlation coefficient with DSI is 0.563, 0.612, 0.670, 0.773, respectively. So, the longer the smoothing window, the higher the correlation is. But of course, too much smoothing makes it impossible to see fine-grained changes of the Twitter Investor Sentiment. Figure 8 has shown the time series of DSI, TIS raw and 30-day and 90-day moving average.

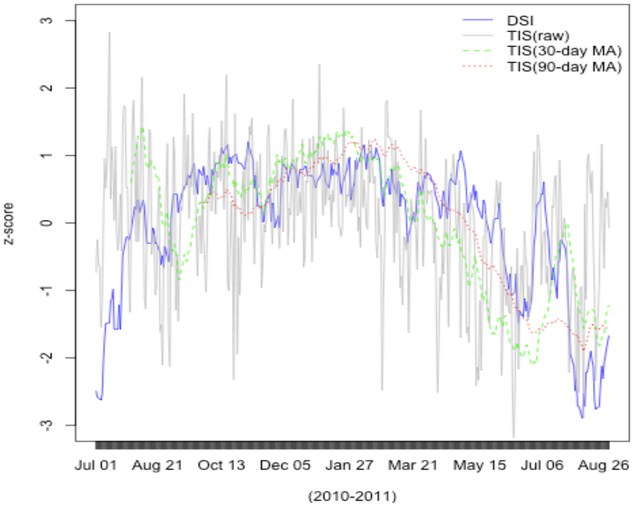

Figure 8: Daily Sentiment Index (DSI) and raw & smoothed Twitter Investor Sentiment (TIS).

4.1.4 Cross-correlation and Granger-Causality analysis between TIS and DSI.

Similarly with Section 3, in order to see the lead/lag relation between TIS and DSI, we do the cross-correlation and Granger-causality analysis. The cross-correlation results are shown in Figure 9. As we can see, the correlation coefficient on DSI-leads-TIS side falls off faster and lower than on the TIS-leads-DSI side. So, TIS seems to be a leading indicator for DSI. The Granger-Causality analysis also shows that TIS is useful for predicting DSI. Meanwhile, DSI is also a useful indicator for TIS prediction.

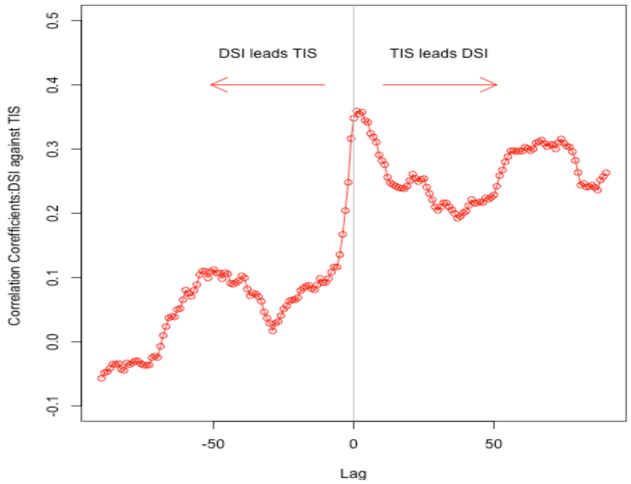

Figure 9: Cross-correlation between Daily Sentiment Index and Twitter Investor Sentiment

However, for the investor sentiment study, the problem that we are most concerned is not whether one measure can lead and predict the other one, but which one can be more useful for *stock market* prediction. Therefore, in the next section, we perform an analysis to compare DSI &TIS with the Dow Jones Industrial Average (DJIA).

4.1.5 Is TIS a trailing indicator for financial market?

We have obtained the daily DJIA data from Yahoo Finance. Then, we convert the raw closing price to log return. The "log return" of stock price $S(t)$ over a time interval Δt is defined as:

$$R_{\Delta t} = \log S(t+\Delta t) - \log S(t) \quad (4)$$

Here $\Delta t = 1$.

For log return, when lag = -1, 0, 1, the cross-correlation coefficient between log return and TIS is: 0.463, 0.226 and 0.191; between log return and DSI is: 0.165, 0.160 and 0.028. As we can see, there are higher correlation between log return and TIS than with DSI. For a better view, we plot the graph for log return against TIS/DSI on one day previous from Jan to Aug 2011 as shown in Figure 10.

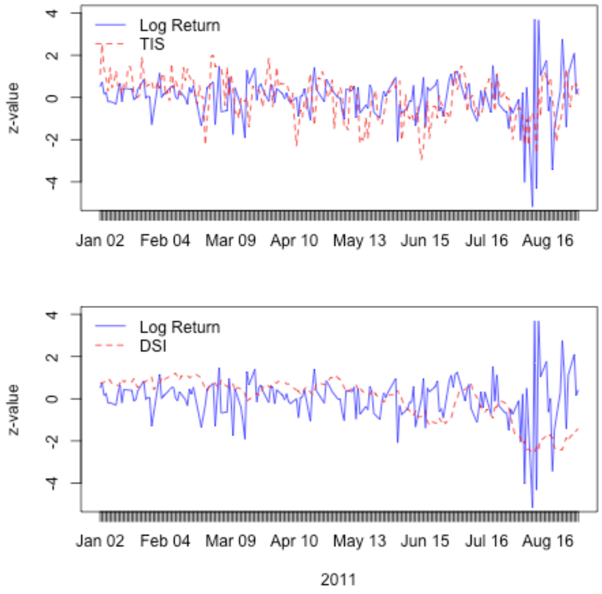

Figure 10: DJIA log return against one day previous 's DSI and TIS.

From Figure 10, it can be seen that DSI is relatively flat and not tracking the change of market. However, the daily TIS is volatile and therefore tracks the volatile market better.

We perform a Granger Causality analysis to see whether TIS and DSI is useful for stock market return prediction. Here we choose lags from 1 to 7 days. Results are listed in Table 2.

Table 2. p-value of Granger Causality between Daily Sentiment Index (DSI),Twitter Investor Sentiment (TIS) and log market return (R_1). (Note: (DSI, R_1) means DSI cause log return)

Lag	DSI, R_1	R_1,DSI	TIS, R_1	R_1,TIS
1	0.878	0.348	0.0005***	<<<0.001***
2	0.188	0.070 *	0.020 **	<<<0.001***
3	0.312	0.679	0.061 *	<<<0.001***
4	0.405	0.625	0.066 *	<<<0.001***
5	0.672	0.441	0.064 *	<<<0.001***
6	0.235	0.345	0.013 **	<<<0.001***
7	0.333	0.380	0.055 *	<<<0.001***

*: significant at the p < 0.1 level; ** significant at p < 0.05 level.

We observe the following. First, historical DSI values (past week) do not seem to be useful for log return prediction. Second, the historical market return is not very useful for DSI prediction, except when lag = 2, although the p-value indicates a marginal statistical significance of 0.1. So, DSI may not be highly Granger causative of previous market changes. This is interesting, but may confirm what DSI's founder, Jake Bernstein, claims "*DSI is a valuable tool that is not derived directly from price but from trader perception of price.*" Whether the trader impression of price can help the stock market prediction is definitely a big question. Third, the investor sentiment extracted from Twitter (TIS) is very useful in predicting the market return from lag 1 to 7 days, especially the 1-day earlier TIS with p = 0.0005. Fourth, different from DSI, adding the historical market return is highly useful for the market return prediction.

To view the trend similarity between the investor sentiment indexes and the DJIA, we plot their time series as presented in Figure 10. Here we smoothed both TIS and DSI based on its past 1 month historical value (i.e. k=30 for Equation 1) to make the time series look smoother.

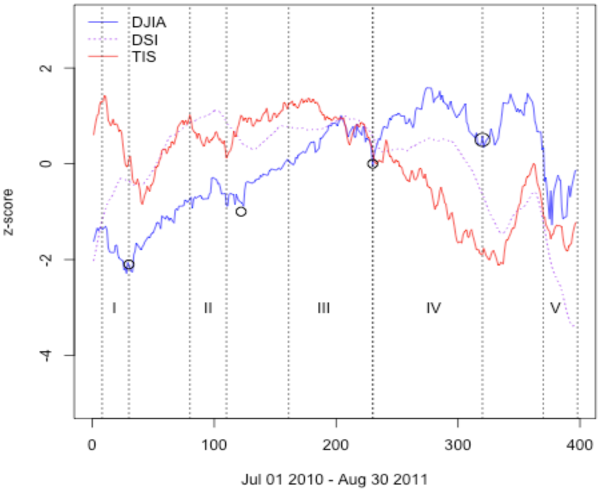

Figure 11: Raw daily Dow Jones Industrial Average (DJIA), 30 day-smoothed Investor sentiment poll (DSI) and 30 day-smoothed Twitter Investor Sentiment (TIS) from July 2010 to August 2011.

In Figure 11, we marked four time points by round circles when the DJIA is already in or is followed by a down trend. In period I (Aug 6th 2010 to Aug 28th 2010), we found that TIS declines with DJIA, while DSI keep increase though. In period II (Oct 17th 2010 to Nov 16th 2010), DSI went up with DJIA. TIS started declining which foreshadow the fast decline of DJIA in the end of Nov 2010. In period III (Jan 6th 2011 to Mar 16th 2011), both DSI and TIS declines before the market highly went down in early March 2011 (as marked by the 3rd circle). However, TIS started declining before DSI did. In period IV (Mar 16th 2011 to Jun 14th 2011), there is a big divergence between TIS and DJIA. TIS keeps declining after Feb 2011, even when the DJIA went up for about two months until May 2011. So, again TIS gives a right signal of down market, but DSI fails in prediction in this period. As we can see DSI only declines when the market seems to go down, i.e. DSI lag behind the market. In the last period V (Aug 3rd 2011 to Aug 30th 2011), TIS reacts positively to the upswing of DJIA, while DSI still keeps declining, indicating that DSI may have not recovered from the down trend of market from the end of July.

5. Conclusion

In this paper, we have widely studied three economic and financial indicators – consumer confidence, unemployment rate and investor sentiment. First, several polls for measuring these three indicators are collected, which include the Michigan Consumer Confidence Index, Gallup Economic Confidence, Gallup Job Creation, Unemployment rate claimed by the US labor office, and two investor sentiment measures – Investor Intelligence and Daily Sentiment Index. By comparing these surveys, we found a strong negative correlation between consumer confidence and unemployment rate (γ=-0.73), indicating that unemployment rate maybe one factor in explaining the change of consumer confidence. A high positive correlation between consumer confidence and investor sentiment, indicates that although the two indices are designed to measure different things (economy vs. financial market), they are highly related with each other. Different polls for the same indicator (e.g. Michigan Consumer Confidence Index vs Gallup Economic Confidence, Investor Intelligence vs Daily Sentiment Index) are high similar.

Second, from Google Insight Search (GIS), we obtained the weekly search volume trend from Jan 2004 to July 2011. The search queries are 19 negative economic words, such as *recession, unemployment, bankruptcy, etc*. Results have shown that there is strong correlation between these search queries with unemployment rate, consumer confidence and investor sentiment. Moreover, certain GIS search queries such as *recession, unemployment office,* and *credit card debt* can precede and help predict the unemployment rate.

Third, we have defined a computational investor sentiment index --- Twitter Investor Sentiment (TIS). It is extracted from Twitter from July 01 2010 to August 30 2011, by counting the “bullish” and “bearish” Tweet messages posted every day. We found a statistically significant positive correlation between TIS and Daily sentiment index (DSI). More importantly, unlike polls, TIS may be not a trailing indicator for stock market.

From these results we conclude that the automatic and computational extraction of economic and financial indicators from Search Engine and Twitter is a very promising research direction. Where traditional polls frequently lag economic and financial indicators, an analysis of the information provided by online source can provide a rapid and effective channel to gauge the public’s mood state and opinions in real-time. These new information may provide a new angle for research that aims to address the relations between various economic and financial phenomena.

6. ACKNOWLEDGMENTS

We thank Dave Allman from Socionomics Institution for sharing the data of weekly Investor Intelligence and Daily Sentiment Index.

7. REFERENCES

[1] Mueller,E. 1966. The Impact of Unemployment on Consumer Confidence. The Public Opinion Quarterly, 30 (1): 19-32.

[2] Brendan O’Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the International AAAI, Conference on Weblogs and Social Media.

[3] Ettredge,M., Gerdes, J and Karuga,G. 2005. Using web-based search data to predict macroeconomic statistics. Communications of the ACM, 48(11): 87–92.

[4] Choi H, Varian H. 2009. Predicting initial claims for unemployment benefits. Available at http://research.google.com/archive/papers/initialclaimsUS.pdf.

[5] Fisher,K and Statman,M., 2000. Investor sentiment and stock returns. Financial Analysis Journal, 56 (2): 16-23.

[6] Da,Z, Engelberg,J and Gao,P. The Sum of All FEARS: Investor Sentiment and Asset Prices. Available at http://ssrn.com/abstract=1509162

[7] Tetlock,P. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance. 62: 1139-1168.

[8] Preis T, Reith D, Stanley HE. 2010. Complex dynamics of our economic life on different scales: insights from search engine query data. Philosophical Transactions of the Royal Society A 368:5707–5719.

[9] Bollen,J, Mao H, Zeng,X. 2011. Twitter mood predicts the stock market. Journal of Computational Science. 2 (1): 1-8.

Appendix

1. Five survey questions of Michigan Consumer Confidence Index

(1) "We are interested in how people are getting along financially these days. Would you say that you (and your family living there) are better off or worse off financially than you were a year ago?"
(2) "Now looking ahead—do you think that a year from now you (and your family living there) will be better off financially, or worse off, or just about the same as now?"
(3) "Now turning to business conditions in the country as a whole—do you think that during the next twelve months we'll have good times financially, or bad times, or what?"
(4) "Looking ahead, which would you say is more likely—that in the country as a whole we'll have continuous good times during the next five years or so, or that we will have periods of widespread unemployment or depression, or what?"
(5) "About the big things people buy for their homes—such as furniture, a refrigerator, stove, television, and things like that. Generally speaking, do you think now is a good or bad time for people to buy major household items?"

2. 19 search queries we adopted are the Financial and Economic Attitudes Revealed by Search (FEARS)

the depression, the great depression, great depression, job search, job openings, job opportunity, bankruptcy,bankruptcy court,job bank, unemployment office, unemployment, insurance, unemployment benefits, credit debt, credit card debt, debt consolidation, inflation,inflation rate,recession