2018

Analyzing Sentiments from Social Media: The Case of Financial Markets

Tianyou Hu
The University of Auckland, New Zealand, t.hu@auckland.ac.nz

Ananth Srinivasan
The University of Auckland, New Zealand, a.srinivasan@auckland.ac.nz

Arvind K. Tripathi
University of Auckland Business School, a.tripathi@auckland.ac.nz

ANALYZING SENTIMENTS FROM SOCIAL MEDIA: THE CASE OF FINANCIAL MARKETS

Research full-length paper

Track: Digital Markets

Hu, Tianyou, The University of Auckland, Auckland, New Zealand, t.hu@auckland.ac.nz

Srinivasan, Ananth, The University of Auckland, Auckland, New Zealand, a.srinivasan@auckland.ac.nz

Tripathi, Arvind, The University of Auckland, Auckland, New Zealand, a.tripathi@auckland.ac.nz

Abstract

Investors seeking quality information rely on market experts on financial news platforms such as Google Finance or Bloomberg. However, in recent years, stock discussion forums hosted on social media platforms are competing with financial news platforms and vying to become an important and credible source of information in this knowledge driven economy. Stock discussion forums are likely to attract retail investors who seek and share their opinions at no cost, and are competing with financial news platforms. This research compares the effect of information available on these two knowledge sources on stock returns. We use text mining methods to capture the sentiments revealed on a popular stock discussion forum and a news media platform and compare their ability to predict market returns. We find that sentiments from both social media and news media platforms predict future stock returns but the effect of social media appears to be stronger and more long lasting compared to news media.

Keywords: Social Media, Text Analytics, Financial Markets

1 Introduction

Social media platforms have democratized content generations by facilitating forums focussed on wide ranging topics such as politics, social issues and retail markets and investment. Though expertise of contributors on social media forum could be questionable, these forums still attract a large number of users seeking information. These users (content consumers) often appear to give more weight to the information coming from their peers than from experts on authenticated news media platforms. One of the notable discussion forums on social media are financial investment forums. We observe that instead of only focusing on experts' recommendations, retail investors increasingly turn to other fellow investors when looking for recommendations for investment.

Traditionally, financial analysis has been a domain of trained professional forecasters but now it is increasingly performed and broadcasted by retail investors (Chen et al. 2014). However, we are yet to understand, the value and effect of information generated on stock discussion forums by retail investors. While a few recent studies show significant positive relationships between number and sentiment of board messages and returns of underperforming small caps stocks (Leung and Ton, 2015), others find little or no evidence that investors' sentiment forecasts future stock returns (Kim and Kim, 2014). In this research, we investigate the impact of messages generated on HotCopper (HC), a prominent discussion forum in Australia, on stock returns. At this stage, we limit our investigation to large cap stocks.

Even with the proliferation of online stock discussion forums, financial experts on news media platforms are still an important source for investors for market forecasts. Tetlock (2007) collected data from Wall Street Journal and found that high media pessimism leads to downward pressure on market prices followed by a reversion to the fundamentals. At the same time, other studies (Fang and Peress, 2009) argue that stocks with no media coverage have higher returns than stocks with high media coverage. These studies highlight a possible role of social and news media platforms in financial markets. However, we are yet to understand how news media and social media effect financial markets.

Further, despite considerable research on the effect of social media and news media on stock markets, we are yet to understand how these two different information sources stack-up in predicting market trends. In this research, we compare the effect of content produced in stock discussion forums hosted on social media and news media in predicting the market returns of large cap stocks in the Australian market. Traditional news media, such as Google Finance or Bloomberg, are limited in their influence due to lack of information sharing, user interaction via comments and other tools available on social media platforms. In contrast, social media provides tools for information search and sharing which contributes to information diffusion (Westen, 2000; Rubin & Rubin, 2010).

We find a strong relationship between sentiments on stock discussion forum and individual stock returns. Bullish sentiments from stock discussion forum are related to higher stock returns with 1 to 3 days holding time. We also find that bullish sentiment from Google Finance is followed by higher returns with one holding day. Thus, sentiments on both social and news media platforms correlate with market returns. However, it appears that the sentiments on social media have a stronger and longer effect than news media.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature, section 3 describe the social media messages, news media articles, data collection strategy and summary statistics of our dataset. Later we discuss our research framework, methodology used and results. We conclude with discussion of our results and direction for future research.

2 Literature Review

Advancement in technologies in the last decade has contributed to more connected and intertwined World via a wide range of web services. In particular, social media platforms have created a "web of communities" and has gained interest of researchers, businesses and policy makers.

Scholars have been trying to understand online information exchanges (Mudambi and Schuff 2010). Often these exchanges involve qualitative data, a classification approach is feasible and necessary to convert textual data into a categorical form, which helps in understanding the content and information filtering. However, categorization is often vulnerable and sensitive to misclassification. With a large number of messages being posted online, manual classification is impossible. Instead, machine learning techniques are quite feasible for this case. While classification approaches for long documents have made substantial progress, short and unstructured text classification approaches are still in early stages (Sriram et al. 2010; Sun 2012). Hence, it is not trivial to understand and examine the effect of social media stock discussion forums on financial markets.

Researchers have investigated links between opinions on message boards and stock returns but the results are mixed. For example, studies (Tumarkin and Whitelaw, 2001) have shown that while message board opinions and stock returns are linked on days of abnormal board activity, but there is no evidence that opinion predicts future returns. Similarly, Antweiler and Frank (2004) show that higher discussion forum posting volume will be followed by significant negative returns on the following day, but with small economic impact. Das and Chen (2007) argue that the combined high-tech sector sentiment is linked with high-tech sector index returns, but not for single stocks. Examining data from an equity review website (www.seekingalpha.com), Chen et al. (2014) demonstrate that views expressed in expert articles and users comments predict future stock returns and earnings. Leung and Ton (2015) study a stock message board (HC) and find that the number of messages and message sentiment have an effect on the contemporaneous stock returns. While these studies show that sentiments on social media platforms affect returns of small and large capitalization stocks, a few other studies argue the opposite. For example, a recent study argues that there is no evidence that investor sentiment forecasts future stock returns (Kim and Kim 2014). While this debate is still inconclusive, we aim to examine and compare the effect of social and new media platforms on market returns.

Another stream of research has examined why social media could influence the financial markets. For instance, Tumarkin and Whitelaw (2001) argue that company or sector professionals may want to disseminate value-relevant information on the internet, perhaps they have framed a long position in these stocks themselves. Boehme et al. (2009) show that online investors are more likely to disseminate information about stocks they are about to buy, instead of spreading false information to earn a profit because of the high cost of or the prohibitions on short selling.

Along with popular social media discussion forums, traditional news media such as newspaper, TV, and online news media also attracts a large number of investors (Fang and Peress 2009b). A few studies have examined the effect of news media on stock markets. For example, Barber and Loeffler (1993) analyse the Wall Street Journal column and observe average positive abnormal returns of 4 percent for the two days following the publication of the recommendation. Huberman and Regev (2001) study a Sunday New York Times article on a possible improvement of new cancer-curing drugs, which give rise to biotechnology stocks on the following Monday and in the three following weeks. Busse and Green (2002) focus on the Morning Call and Midday Call segments on CNBC TV and find that prices respond to reports within seconds of initial mention, with positive reports fully incorporated within one minute. Tetlock et al. (2008) find that more negative words in news focusing on specific firms predict low firm earnings. Dougal et al. (2012) show that financial journalists have the potential to impact investor behaviour, at least in a short term. Gurun and Butler (2012) demonstrate that local media uses fewer negative words when reporting local companies in comparison with reporting nonlocal companies. Abnormal positive local media slant is strongly linked with firm equity values.

Despite a large number of studies investigating the effects of social and news media on stock returns, there is little that compares the effect of these two media platforms on stock returns. We are yet to understand how these two media sources are different in terms of value and insights for investors. Does the high frequency of information dissemination and users' interaction through social media help generate more wisdom from crowds? In this research, we shed some light on these issues and examine the differences in effects of social and news media on stock returns. In particular, we examine how positive or negative sentiments expressed in news media and social media discussion forums impact stock returns.

To understand the sentiment and opinions from media platforms, researchers have used different approaches. For example, a few studies (Antweiler and Frank, 2004) have used machine learning algorithms to classify social media posts and generate bullishness and agreement indexes. Antweiler and Frank (2004) use Naïve Bayes (NB) Algorithm to classify posts and propose a Bullishness index and an Agreement index. As the name suggests, Bullishness index measures the bullishness of the market and Agreement index measures the disagreement between positive and negative sentiments of the social media posts. Li (2008) uses NB to classify sentences from 10-K and 10-Q filings into different tone and content groups.

Kim and Kim (2014) use NB to compute bullishness index. Hu and Tripathi (2015) use NB and Support Vector Machine to compute bullishness and agreement indexes. Following the literature, we use Bernolli Naïve Bayes to classify messages and articles and use Bullishness and Agreement index to gauge the sentiments of social media and news media

3 Data

Data for this research has been collected from HotCopper (HC), the largest and most popular online stock discussion forum in the Australasian region. A web crawler was used to download and store messages in a database. We focus on 46 companies from the ASX 50 index from January 2014 to March 2015. ASX 50 represents 50 stocks but four stocks in ASX 50 underwent identity change during the sample period and have been removed from our dataset (Tirunillai and Tellis, 2012). The firms in our sample, in general, are large capitalization stocks. Our dataset contains 43375 messages from HotCopper (http://hotcopper.com.au). We selected HotCopper for this study because all the messages have self-disclosed sentiments by the authors, which made it good for training machine learning classifiers. Downloaded messages contain author sentiments ("None", "LT Buy", "ST Buy" "Buy", "Hold", "Sell", "ST Sell" and "LT Sell"), title, posting time, author, content, and ticker symbol of the firm. Here ST means short-term and LT means long-term. Table 1 presents the summary statistics of the posted messages. Among all the messages, 64.98% have revealed sentiment explicitly. In this research, we combine "LT Buy", "ST Buy", and "Buy" and term them as "Bullish" sentiment. Same is done for "Bearish" sentiments. We do this because we only need "Bullish" and "Bearish" polarity, not valence of sentiment to compute bullishness index (equation 1). Messages with "Hold" sentiment are discarded in computing the bullishness index following the literature (Antweiler & Frank, 2004). Following this approach, among the messages with the self-disclosed sentiment, 32.26% are "Bullish" and 9.62% are "Bearish" messages. We conjecture that retail investors are more likely to reveal "Bullish" instead of "Bearish" sentiment. This observation is consistent with other studies that investors try to use more positive words in messages (Boehme et al. 2009).

#Revealed Messages	%LT Buy	%ST Buy	%Buy	%Hold	%Sell	%ST Sell	%LT Sell	%None
43375	2.83%	0.54%	28.89%	23.10%	9.05%	0.45%	0.12%	35.02%

Table 1. Summary Statistics

Since not all the messages on this forum have revealed sentiments, we classify collected messages into "Bullish" and "Bearish" using Bernoulli Naïve Bayes (BNB) classifier. The training set is constructed using messages with self-disclosed sentiments.

In this research, we have developed an agent to collect news articles from Google Finance (GF) based on stock tickers. Google Finance covers a wide range of media sites. The summary of the collected messages and news articles is shown in Table 2. We classify these articles downloaded from Google Finance, into "Bullish" and "Bearish" using BNB classifier. This classifier performs best when compared with Multinomial Naïve Bayes, Linear Support Vector Machine and Support Vector Machine (SVM) with rdf kernel. The training set for classification of news articles is downloaded from well-known public training data set available at: Data for Everyone (https://www.crowdflower.com/data-for-everyone/). This training set is crowdsourced. Contributors viewed news article and rated the positivity of the article on a scale 1-9 with 1 being negative and 9 being positive. We classified all the articles with score of 1, 2 as negative, and 7, 8, 9 as positive (there are no news with score 10). As a result, we have 746 news articles in our training set, with about equal number of positive and negative articles. To make the results comparable and consistent, we also use BNB to classify messages from HC.

The threads from HC can run for weeks/months, especially if the threads are listed in "TOP RATED POSTS" by HC. In comparison, news reports related to financial markets normally do not have such a long life span.

3.1 Computation of Investor Sentiment

The standardised bullishness index $Bullishness_{i,t}$ (Antweiler & Frank, 2004) for stock i at time t can be calculated as following:

$$Bullishness_{i,t} = \frac{M_{i,t}^{Bullish} - M_{i,t}^{Bearish}}{M_{i,t}} * ln(1 + M_{i,t}) \qquad (1)$$

$M_{i,t}^{Bullish}$ is the number of messages/articles with "Bullish" sentiment, $M_{i,t}^{Bearish}$ is the number of messages/articles with "Bearish" sentiment. Here $M_{i,t} = M_{i,t}^{Bullish} + M_{i,t}^{Bearish}$ is the total number of relevant messages.

	HC Messages	GF Articles
Total#Messages/Articles	43375	65658
Avg.# WordsPerMessage	64.79	407.2
StDev.#WordsPerMessage	100.74	725.9

Table 2. Summary for the collected messages from HotCopper (HC) and Google Finance (GF)

Prior research has argued that the disagreement among trading population drives trading volume and intensity (Antweiler and Frank 2004). Disagreement on stock discussion forums can be measured by looking at the volume and intensity of competing arguments. Disagreement has been measured via an agreement index *Agreement*$_{i,t}$ (Antweiler and Frank 2004). This index measures the disagreement between the sentiments of messages. Literature has shown somewhat controversial and conflicting effect of "disagreement" on the trading volume. For example, Harris and Raviv (1993) showed that "disagreement" could increase trading volume while Milgrom and Stokey (1982) demonstrated that "disagreement" gives rise to 'no trade' behaviour in the financial markets. Antweiler and Frank (2004) proposed a proxy to capture disagreement among message posters, which is given by:

$$Agreement_{i,t} = 1 - \sqrt{1 - \left(\frac{M_{i,t}^{Bullish} - M_{i,t}^{Bearish}}{M_{i,t}}\right)^2} \in [0,1] \quad (2)$$

where 0 represents complete disagreement. They find that greater agreement on a given day is followed by more trades on the next day. Their findings were in contrast to the literature which found that greater disagreement induces more trades on the next day.

We report the summary statistics of the sentiment measures in Table 3, where *GF_Bullishness*$_{i,t}$ is the bullishness index (equation 1) from all the news articles that appeared on Google Finance for stock *i* on day *t*, *GF_Agreement*$_{i,t}$ is the agreement index (equation 2) from Google Finance for stock *i* on day *t*, *HC_Bullishness*$_{i,t}$ is the Bullishness index for HotCopper, *HC_Agreement*$_{i,t}$ is Agreement index for HotCopper.

	Min	Max	Mean	Median	S.D.
$GF\_Bullishness_{i,t}$	-2.639	3.933	0.112	0	1.118
$GF\_Agreement_{i,t}$	0	1	0.604	1	0.454
$HC\_Bullishness_{i,t}$	-2.773	3.754	0.648	0.693	1.002
$HC\_Agreement_{i,t}$	0	1	0.783	1.000	0.391

Table 3. Min, Max, Mean, Standard Deviations

4 Methodology

Previous studies have focused on a general relationship between social media and financial market activities (Antweiler & Frank, 2004; Das & Chen, 2007; Kim & Kim, 2014; Leung & Ton, 2015) or between news media and market returns (Tetlock, 2007; Tetlock et al., 2008; Chen et al., 2014). Researchers have employed the contemporaneous regression with one holding day (Kim & Kim, 2014; Leung & Ton, 2015), and one-day or two-days lead-lag (Antweiler & Frank, 2004; Chen et al., 2014; Leung & Ton, 2015). Scholars disagree on optimal number of holding days. For example, Tetlock et al. (2008) used one holding day, which is a short holding time, while, others have used longer holding times, such as one month to 36 months holding time (Chen et al. 2014).

In this research, we examine how sentiments expressed on online stock discussion forum and news media affect stock returns with three different holding periods (one, two or three days). We control for market index and firms' characteristics.

$$Ret_{i,t,t+2} = LN\left(\frac{P_{t+2}}{P_{t-1}}\right) \qquad Ret_{i,t,t+1} = LN\left(\frac{P_{t+1}}{P_{t-1}}\right) \qquad Ret_{i,t} = LN\left(\frac{P_t}{P_{t-1}}\right) \tag{3}$$

$$Ret_{i,t,t+2} = \alpha + \beta_1 GF\_Bullishness_{i,t} + \beta_2 GF\_Agreement_{i,t} + \beta_3 Log(MarketCap)_{i,t} + \beta_4 Log(StockIndex)_{i,t} + \beta_5 Ret_{i,t-1} + \varepsilon \quad (4)$$

$$Ret_{i,t,t+2} = \alpha + \beta_1 HC\_Bullishness_{i,t} + \beta_2 HC\_Agreement_{i,t} + \beta_3 Log(MarketCap)_{i,t} + \beta_4 Log(StockIndex)_{i,t} + \beta_5 Ret_{i,t-1} + \varepsilon \quad (5)$$

$$Ret_{i,t,t+2} = \alpha + \beta_1 HC\_Bullishness_{i,t} + \beta_2 HC\_Agreement_{i,t} + \beta_3 GF\_Bullishness_{i,t} + \beta_4 GF\_Agreement_{i,t} + \beta_5 Log(MarketCap)_{i,t} + \beta_6 Log(StockIndex)_{i,t} + \beta_7 Ret_{i,t-1} + \varepsilon \ (6)$$

Equation 3 shows our approach of calculating raw returns with different number of holding days. $Ret_{i,t,t+2}$ is the raw return of stock *i* with holding time of three days from day *t* to day *t+2*. Similarly we calculate $Ret_{i,t,t+1}$ (holding two days from day *t* to *t+1*) and $Ret_{i,t}$ (holding one day for day *t*). Here, *t* represents the day on which the message appeared on HC (or GF) or the following trading day if the message/article was posted on a non-trading day (e.g., Sunday). Previous research has used raw return, which is the natural logarithm of the last holding day's adjusted close price divided by the adjusted close price on day t-1 (Kim & Kim, 2014; Leung & Ton, 2015). P_t is the adjusted close price of stock *i* on day *t*. $Log(MarketCap)_{i,t}$ is the log of market capitalization of stock i on day t, and $Log(StockIndex)_{i,t}$ is the log of ASX 50 stock index, $Ret_{i,t-1}$ is the raw return on day *t-1*.

Equation 4, 5 and 6 use the bullishness and agreement indexes to capture the sentiments from social media and news media. All of them have used control variables for market capitalization and stock index and one-day lagged returns to control for possible autocorrelation (Sabherwal et al. 2011).

5 Results and Contributions

Table 4 shows the summary of results focusing on three regressions (equation 4, 5 and 6). Coefficients are standardised. In equation 4, we use bullishness and agreement index to capture the sentiments of news media. These results show that sentiments expressed on news and social media platforms positively correlate with raw returns. Our results show that if the standard deviation (SD) of Bullishness index increases by one, then the return will increase by 0.088 for the same day. Prior studies have shown that high media pessimism predicts downward pressure on market prices (Tetlock, 2007) and abnormal positive local media slant is strongly linked with positive increase in firm equity values (Gurun and Butler, 2012). Our findings are consistent with previous results.

	$Ret_{i,t}$			$Ret_{i,t,t+1}$			$Ret_{i,t,t+2}$		
$GF\_Bullishness_{i,t}$	0.088 * (0.039)		0.084 * (0.038)	0.040 (0.039)		0.035 (0.038)	0.059 (0.039)		0.054 (0.038)
$GF\_Agreement_{i,t}$	0.041 (0.038)		0.041 (0.038)	0.034 (0.039)		0.035 (0.038)	0.023 (0.039)		0.024 (0.038)
$HC\_Bullishness_{i,t}$		0.130 ** (0.040)	0.127 ** (0.040)		0.164 *** (0.040)	0.163 *** (0.040)		0.155 *** (0.040)	0.153 *** (0.040)
$HC\_Agreement_{i,t}$		0.0190 (0.039)	0.017 (0.039)		0.007 (0.039)	0.006 (0.039)		-0.010 (0.039)	-0.010 (0.039)

Log(MarketCap)$_{i,t}$	-0.051 (0.039)	-0.0036 (0.038)	-0.041 (0.038)	-0.042 (0.039)	-0.030 (0.038)	-0.032 (0.038)	-0.036 (0.039)	-0.024 (0.038)	-0.028 (0.039)
Log(StockIndex)$_{i,t}$	0.102 ** (0.039)	0.077 ** (0.039)	0.080 * (0.039)	0.115 ** (0.039)	0.087 * (0.039)	0.089 * (0.039)	0.102 ** (0.039)	0.075 . (0.039)	0.077 * (0.039)
Ret$_{i,t-1}$	0.079 * (0.039)	0.064 (0.039)	0.067 (0.039)	0.002 (0.039)	-0.014 (0.038)	-0.012 (0.039)	-0.035 (0.039)	-0.049 (0.039)	-0.047 (0.039)

Standard errors in parentheses; . p<0.1, * p<0.05, ** p<0.01, *** p<0.001

Table 4. Summary of Results

In equation 5, we use Bullishness and Agreement index to capture the effect of sentiments expressed of social media on raw returns. We demonstrate that raw returns are significantly positively related to the Bullishness from HotCopper, which is consistent with previous result (Leung and Ton 2015). Our results show that if the standard deviation (SD) of bullishness index increases by one, then the return will increase by 0.127 for the same day and 0.163 for two or 0.155 for three holding days. These results confirm that sentiment from social media is reflected in market price quite quickly.

In equation 6, we test the effect of Bullish sentiments expressed in Hot Copper stock discussion forums and in articles published in Google Finance using *GF_Bullishness$_{i,t}$*, *HC_Bullishness$_{i,t}$*.We find that effect of sentiments expressed on social media remains significant on market returns for at least 3 days, while the effect of sentiments expressed on news media on market returns last only for one day. By using one holding day, we show that the parameter for HC (0.127) is larger than the parameter for GF (0.084) after fitting the model for equation 6. As a result, we demonstrate that sentiments from HC has a stronger effect on market returns than the sentiments from GF. These results are shown in table 4.

We conjecture that HC is a specialized retail investor forum where passionate investors contribute based on their own experience and are likely to influence other inexperienced and naïve investors. Users in social media forums such as HC, are quick to respond and echo sentiments from community leaders quickly leading to bearish or bullish sentiments on the stock discussion forum (HotCopper). Further, short messages on HC have clear positive and negative sentiments compared to news articles, which tend to be more balanced. In Table 2, our summary statistics show that average number of words in Google Finance articles are almost six times to messages posted on social media forums such as Hot Copper.

In contrast to investor forums, articles published on Google Finance are written by industry experts and therefore these articles tend to discuss both pros and cons of the market and firms.

Prior research (Chen et al., 2014) has shown effect of sentiments on social media platforms on market returns. Our research confirms and adds nuances to the extant literature. Focusing on articles written by experts on an equity review website (www.seekingalpha.com), Chen et al. (2014) established the impact of sentiments in expert articles on market returns. In this research, we extended that line of research to examine the impact of sentiments generated in retail investor forums on market returns. Retail investor forums, such as HotCopper used in this study, attract hobbyist and amateur investors and often compete with expert forums, such as SeekingAlpha.com, in terms of value of information. Note that experts are paid to contribute on equity research websites (e.g., seekingAlpha), whereas hobbyists provide information on HotCopper for free. Further, articles written by experts are structured and follow accepted financial jargons; therefore, sentiments of these articles can be analyzed using standard dictionary tools such as L&M dictionary. In contrast, short text messages posted on amateur investor forums that are used in this study, are unstructured, therefore require machine-learning approaches for sentiment analysis. Chen et al. (2014) use Dow Jones News Service, while we use Google Finance news service. Google Finance aggregates a wide coverage of news from many different news media platforms. Fourth, we use all large cap stocks, whereas Chen et al. (2014) use all stocks that have been discussed on their chosen social media. Stocks with large and small caps will be influenced by social media differently (Leung and Ton 2015). Thus, by only focusing on large cap stocks, we control for the heterogeneity emanating due to market cap size.

6 Conclusions

Advances in Internet technologies in general and social media in particular have led to proliferation of content creation and sharing platforms. These platforms have democratized content creation and as a result of that, there is plenty of information on any topic coming from a wide range of sources. The onus is on content consumer to determine the quality of the information and credibility of the sources. We posit that the challenge and risk associated with finding quality information becomes higher in financial markets. Therefore, it is imperative to understand the role and effect of these varied information sources in financial markets. This paper examines and compares the effect of messages in stock discussion forums hosted on social media and expert articles on news media on individual stock returns.

This research examines the impact of sentiments expressed in social media and news media on market returns. We have collected data from an online discussion forum, HotCopper (HC), and investigated the impact of users' sentiments expressed via messages/posts on this forum on the stocks listed in ASX 50. Further, news articles related to these stocks were collected from Google Finance. Messages and articles are classified using Bernoulli Naïve Bayes classifier and sentiments are analysed using the *Bullishness* index (Antweiler and Frank 2004).

Our findings show a significant effect of sentiments from social media and news media on market returns. We find that sentiments on social media have a longer lasting and stronger effect on market returns than the sentiments on news media. Stock discussion forums such as the one used in this study typically attract retail or hobbyist investors. In essence, our results show that sentiments expressed in these forums have a stronger and longer lasting effect compared to effect of financial news articles written from market experts. These results certainly need more investigation. However, we believe that these results give early indications about how various information sources could affect financial markets. First, while market research articles on Google finance tend to be more balanced (positive and negative), opinion posts and comments on stock discussion forums tend be more biased and therefore are likely to attract a higher number of users. Second, users appear to have a higher trust in information coming from their peers in stock discussion forums than from unknown market experts. And finally, due to herding effect, users tend to lean or rely more on information or sentiment that has been echoed by many community/forum members. Combination of these effects may result in higher trading volume based on these sentiments.

These results suggest that instead of only relying on news reports, retail investors should also dig into social media forums to get a better understanding of sentiments of potential investors. For researchers, we recommend that studies should control the effect of news media sentiments when investigating the effect of social media sentiments on financial markets.

This research has some limitations that should be addressed as the work progresses. First, we only focus on large cap stocks and therefore would caution extending our results to small caps. Second, our data comes from Australian market which is smaller compared to US market with many stock discussion forums and larger number of finance articles on Google finance or Bloomberg. Despite these limitations, we believe that our results provide echo many findings in the literature and provide new insights into how various information sources could affect stock returns.

References

Antweiler, W., and Frank, M. Z. 2004. "Is all that talk just noise? The information content of Internet stock message boards," Journal of Finance (59:3), pp. 1259–1294 (doi: 10.1111/j.1540-6261.2004.00662.x).

Barber, B. M., and Loeffler, D. 1993. "The 'Dartboard' Column: Second-Hand Information and Price Pressure," The Journal of Financial and Quantitative Analysis (28:2), Cambridge University Press, p. 273 (doi: 10.2307/2331290).

Boehme, R. D., Danielsen, B. R., and Sorescu, S. M. 2009. "Short-Sale Constraints, Differences of Opinion, and Overvaluation," Journal of Financial and Quantitative Analysis (41:02), Cambridge University Press, p. 455 (doi: 10.1017/S0022109000002143).

Busse, J. A., and Clifton Green, T. 2002. "Market efficiency in real time," Journal of Financial Economics (65:3), pp. 415–437 (doi: 10.1016/S0304-405X(02)00148-4).

Chen, H., De, P., Hu, Y. Y. (Jeffrey), and Hwang, B.-H. B.-H. 2014. "Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media," Rev. Financ. Stud. (27), p. hhu001– (doi: 10.1093/rfs/hhu001).

Das, S. R., and Chen, M. Y. 2007. "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web," Management Science (53:9), INFORMS, pp. 1375–1388 (available at http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1070.0704).

Dougal, C., Engelberg, J., Garcia, D., and Parsons, C. A. 2012. "Journalists and the Stock Market," Review of Financial Studies (25:3), pp. 639–679 (doi: 10.1093/rfs/hhr133).

Fang, L., and Peress, J. 2009a. "Media Coverage and the Cross-section of Stock Returns," The Journal of Finance (64:5), pp. 2023–2052 (doi: 10.1111/j.1540-6261.2009.01493.x).

Fang, L., and Peress, J. 2009b. "Media Coverage and the Cross-section of Stock Returns," The Journal of Finance (64:5), pp. 2023–2052 (doi: 10.1111/j.1540-6261.2009.01493.x).

Gurun, U. G., and Butler, A. W. 2012. "Don't Believe the Hype: Local Media Slant, Local Advertising, and Firm Value," The Journal of Finance (67:2), pp. 561–598 (doi: 10.1111/j.1540-6261.2012.01725.x).

Harris, M., and Raviv, A. 1993. "Differences of Opinion Make a Horse Race," Review of Financial Studies (6:3), pp. 473–506 (doi: 10.1093/rfs/5.3.473).

Hu, T., and Tripathi, A. K. 2015. “The Effect of Social Media on Market Liquidity,” SSRN Electronic Journal (doi: 10.2139/ssrn.2661099).
Huberman, G., and Regev, T. 2001. “Contagious Speculation and a Cure for Cancer: A Nonevent that Made Stock Prices Soar,” The Journal of Finance (56:1), pp. 387–396 (doi: 10.1111/0022-1082.00330).
Kim, S.-H., and Kim, D. 2014. “Investor sentiment from internet message postings and the predictability of stock returns,” Journal of Economic Behavior & Organization (107), Elsevier B.V., pp. 708–729 (doi: 10.1016/j.jebo.2014.04.015).
Leung, H., and Ton, T. 2015. “The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks,” Journal of Banking & Finance (55:July 1998), Elsevier B.V., pp. 37–55 (doi: 10.1016/j.jbankfin.2015.01.009).
Li, F. 2008. “The Determinants and Information Content of the Forward-looking Statements in Corporate Filings - A Naive Bayesian Machine Learning Approach,” SSRN Electronic Journal (doi: 10.2139/ssrn.1267235).
Milgrom, P., and Stokey, N. 1982. “Information, trade and common knowledge,” Journal of Economic Theory (26:1), pp. 17–27 (doi: 10.1016/0022-0531(82)90046-1).
Rubin, A., and Rubin, E. 2010. “Informed Investors and the Internet,” Journal of Business Finance & Accounting (37:7-8), pp. 841–865 (doi: 10.1111/j.1468-5957.2010.02187.x).
Sabherwal, S., Sarkar, S. K., and Zhang, Y. 2011. “Do Internet Stock Message Boards Influence Trading? Evidence from Heavily Discussed Stocks with No Fundamental News,” Journal of Business Finance & Accounting (38:9-10), pp. 1209–1237 (doi: 10.1111/j.1468-5957.2011.02258.x).
Tetlock, P. C. 2007. “Giving content to investor sentiment: The role of media in the stock market,” Journal of Finance (62:3), pp. 1139–1168 (doi: 10.1111/j.1540-6261.2007.01232.x).
Tetlock, P. C., Saar-Tsechansky, M., and MacSkassy, S. 2008. “More than words: Quantifying language to measure firms’ fundamentals,” Journal of Finance (63:3), pp. 1437–1467 (doi: 10.1111/j.1540-6261.2008.01362.x).
Tirunillai, S., and Tellis, G. J. 2012. “Does Chatter Really Matter? Dynamics of User-Generated Content and Stock Performance,” Marketing Science (31:November 2014), pp. 198–215 (doi: 10.1287/mksc.1110.0682).
Tumarkin, R., and Whitelaw, R. F. 2001. “News or Noise? Internet Postings and Stock Prices,” Financial Analysts Journal (57:3), Association for Investment Management and Research, pp. 41–51 (doi: 10.2469/faj.v57.n3.2449).
Westen, T. 2000. “E-Democracy: Ready or Not, Here It Comes,” National Civic Review (89:3), pp. 217–228 (doi: 10.1002/ncr.89304)