



Strathmore

UNIVERSITY

**ANALYZING THE INFLUENCE OF TWITTER HASHTAGS ON THE MOVEMENT
OF THE STOCK MARKET.**

NAME: WARIMAH JOY MUTHONI

NUMBER: 071862

**Submitted in partial fulfillment of the requirements for the Degree of
Bachelor of Business Science in Financial Economics at Strathmore University**

**School of Finance and Applied Economics
Strathmore University
Nairobi, Kenya**

November, 2015

This Research Project is available for Library use on the understanding that it is copyright material and that no quotation from the Research Project may be published without proper acknowledgement.

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the Research Project contains no material previously published or written by another person except where due reference is made in the Research Project itself.

© No part of this Research Project may be reproduced without the permission of the author and Strathmore University

JOY MUTHONI WARIMAH

J. Warimah [Signature]

13/11/2015 [Date]

This Research Project has been submitted for examination with my approval as the Supervisor.

..... [Name of Supervisor]

..... [Signature]

..... [Date]

School of Finance and Applied Economics

Strathmore University

Contents

ANALYSING INFLUENCE OF TWITTER HASHTAGS ON THE MOVEMENT OF THE STOCK MARKET	3
ABSTRACT	4
INTRODUCTION	4
Purpose of the Study	7
Research Objective	8
Research questions.....	8
Problem Statement.....	8
Significance of the Research.....	8
LITERATURE REVIEW	9
RESEARCH METHODOLOGY	16
Capital Market Data	16
Internet Data	16
METHODS	17
Sentiment Classification.....	17
Capital Data Classification.....	18
CORRELATION AND GRANGER CAUSALITY.....	18
SHORTCOMINGS	18
RESULTS AND INTERPRETATION	19
Pearson and Granger Causality Tests	19
Wald Test	19
Logit Regression.....	20
Table 1: Pearson Correlation.....	20
Table 2: Granger Causality	21
Table 3: Wald Test	22
Table 4: Logit Regression	23
DISCUSSION	24
CONCLUSION	25
APPENDICES.....	25
Table 5	25
Table 6	26

Table 7	29
REFERENCES	31
Bibliography	31

ANALYSING INFLUENCE OF TWITTER HASHTAGS ON THE MOVEMENT OF THE STOCK MARKET

ABSTRACT

This paper evaluates the existence of a relationship between a well-known micro-blogging platform, Twitter and the financial market. We particularly consider in a period of six months the daily collection of trending twitter hashtags and the daily movement of the NSE 20 share index

Granger causality, Correlation tests and logistic regression tests were used to examine the relationship. We find low Pearson correlation between the corresponding time series over the time period. We also found that none of the variables granger causes the other. We found no relationship between the sentiment as expressed from twitter and the movement of the stock market. As a series of other papers have already shown, there is a signal worth investigating which connects social media and market behavior. This opens the way, if not to forecasting, then at least to “now-casting” financial markets. This study is open to further and more advanced research using a longer time period as well and to employ sentiment classification techniques.

INTRODUCTION

Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making (Bollen, Mao, & Zeng). Does this also apply to societies at large, i.e. can societies experience mood states that affect their collective decision making? By extension is the public mood correlated or even predictive of economic indicators? Is what someone is talking about on twitter influential of his economic decision?

Behavioral finance focuses on the individual attributes, psychological or otherwise, that shape common financial and investment practices (Ritter, 2003). Social influence and interactions with other people will cause investors to behave irrationally. Investors may make common mistakes in a herd manner because of social influences and the force of media news. The media plays two roles; they set the stage for market moves and they also instigate the moves themselves (Shiller, 2000)

The media is a key factor in influencing decisions of individuals. The media actually plays two roles; they not only set the stage for market moves but they also instigate the moves themselves (Shiller, 2000) Shiller also documented that the media had a tendency to keep investors focused on specific stories for long periods of time called as an 'attention cascade'. In some cases, the attention cascade had contributed to speculative bubbles in stock markets. (Davis, 2006) confirmed the role of the media in the development of extreme market movements. The media were found to magnify market responses to news. At times of market crisis, the media had pushed trading activity to extremes. The media had triggered and reinforced opinions.

One such media platform that has seen an explosive rise in popularity is Twitter. Twitter is a free social networking microblogging service that allows registered members to broadcast short posts called tweets. Basically, you can use the service to post and receive messages to a network of contacts. Instead of sending a dozen e-mails or text messages, you send one message to your Twitter account, and the service distributes it to all your followers. (Strickland, 2007) Twitter members can broadcast tweets and follow other users' tweets by using multiple platforms and devices. The default settings for Twitter are public. To weave tweets into a conversation thread or connect them to a general topic, members can add hashtags to a keyword in their post. The hashtag, which acts like a meta tag, is expressed as #keyword. Tweets, which may include hyperlinks, are limited to 140 characters, due to the constraints of Twitter's Short Message Service (SMS) delivery system. Because tweets can be delivered to followers in real time, they might seem like instant messages to the novice user. Tweets are permanent, they are searchable and they are public. Anyone can search tweets on Twitter, whether they are a member or not. (Twitter)

Twitter has seen explosive growth over the past few years. As of March 31st 2015, twitter has 302 million monthly active users; 500 million tweets are sent per day. 80% of its active users use a mobile phone to access twitter. 77% of the accounts are outside the U.S. It supports 30 languages and has more than 35 offices around the world. (Twitter)

According to the 2014 second quarterly results, twitter grew by more than 16 million new users and now has around 271 million monthly active users. As of June/July 2014, Kenya has a confirmed more than 700,000 monthly active users (MAUs) on Twitter. Given that Twitter has 2

to 3 times its MAUs who do not login, this means that Kenya has in the region of 1.4 million to 2.1 million users in total. 80% of Twitter users in Kenya access the service on a daily basis. In terms of Daily Active Users (DAUs) therefore the number is approximately more than 570,000. Twitter growth has been doubling in Kenya year on year. This means that around this time next year there will be around 1.4 million Twitter MAUs in Kenya or anywhere between 2.8 million to 4.2 million users in total factoring in those who do not login. Whilst global mobile-based Twitter usage is approximately 75% or 203 million MAUs according to the second quarter of 2014 results. In Kenya its 93% which would translate to more than 651,000 MAUs on mobile and only 7% or more than 49,000 MAUs on desktop. (Kemibaro, 2014)

Kenya is the second most active country in Africa on the social networking site. South Africa led the pack with 5,030,226 tweets in the same period as revealed by the new study dubbed “How Africa Tweets” that analyzed over 11.5 million geo-located Tweets across the continent. Youth tweeting on mobile phones were also found to be the main drivers of Twitter’s growth in Africa, with 60 percent of the continent’s most active tweeters between the ages of 20 and 29.

How Africa Tweets polled 500 of Africa’s most active tweeters, 81 of whom said they mainly used the social networking service to communicate with friends, while 60 percent said they used it to monitor news and another 22 percent used Twitter for employment opportunities.

The study was conducted by communications consultancy firm Portland Communications and UK based media platform Tweetminster carrying out a pan-African analysis of Twitter traffic over the last three months of 2011. (Portland Communications) (Capital Group)

The obvious popularity of Twitter has led many people and celebrities to join twitter in order to potentially connect with others and increase their popularity and awareness. With so many people tweeting about their various opinions about subjects ranging from toothpaste to the newest Apple products, Twitter is a rich source of real-time information regarding current societal trends and opinions. (Chang & Liu, 2011)

Stock market prediction has attracted much attention from academia as well as business. But can the stock market really be predicted? Early research on stock market prediction was based on

random walk theory and the Efficient Market Hypothesis. According to the EMH stock market prices are largely driven by new information, i.e. news, rather than present and past prices. Since news is unpredictable, stock market prices will follow random walk pattern and cannot be predicted with more than 50 percent accuracy. There are two problems with EMH. First, numerous studies show that stock market prices do not follow a random walk and can indeed to some degree be predicted thereby calling into question EMH's basic assumptions. Second, recent research suggests that news may be unpredictable but that very early indicators can be extracted from online social media (blogs, Twitter feeds, etc.) to predict changes in various economic and commercial indicators. This may conceivably also be the case for the stock market. For example, although news most certainly influences stock market prices, public mood states or sentiment may play an equally important role. We know from psychological research that emotions, in addition to information, play a significant role in human decision-making. Behavioral finance has provided further proof that financial decisions are significantly driven by emotion and mood. It is therefore reasonable to assume that the public mood and sentiment can drive stock market values as much as news. Over the past 5 years significant progress has been made in sentiment tracking techniques that extract indicators of public mood directly from social media content such as blog content and in particular large-scale Twitter feeds. Although each so-called tweet, i.e. an individual user post, is limited to only 140 characters, the aggregate of millions of tweets submitted to twitter at any given time may provide an accurate representation of public mood and sentiment. This has led to the development of real time sentiment-tracking indicators such as and "Pulse of Nation". Pulse of Nation illustrates the varying mood in the U.S., as inferred after analyzes over 300 million tweets that were created over the course of the day. Through a natural language processing algorithm called Sentiment Analysis, each tweet is assigned a mood score based on the number of positive or negative words it contained. Out of the resulting data, one then calculates the average mood score of all the users living in a state on a specific time basis. (Information Aesthetics)

Purpose of the Study

In this paper I analyze whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can be used to predict the stock market prices and volatility.

Research Objective

The objective of carrying out this research is to establish if there is a dependent relationship between social media, in this case twitter and the financial market by looking at the movement of the NSE 20 all share index.

This research could be helpful in predicting or influencing the behaviour of the stock market from the viewpoint of the investor.

Research questions

- i) Is there a relationship between twitter hashtags and the financial market
- ii) Can we use twitter hashtags to influence the behavior of the stock market

Problem Statement

Society can experience mood states that affect their collective decision-making. So, if each tweet is a condensed summary of a person's mood or opinion about a certain subject, then the aggregate of tweets about the subject should express the collective mood.

However, little or no research has been done to ascertain if this holds in the Kenyan market. This research purports to confirm or deny if there is a relationship between the mood as expressed from daily collections of the trending twitter hashtags in Kenya and the positive or negative movement of the Nairobi Stock Exchange.

Significance of the Research

A significant contribution of this paper is that the Twitter sentiment time series are made publicly available. They can be used not only to validate our results but also to carry out additional studies that do not necessarily follow the same methodology. The dataset allows one to study different sentiment aggregations, different events (points in time), and different post-event effects (such as drifts, reversals, and changes in volatility).

This paper also provides a significant platform to further this research without the time and data constraints that we experienced.

LITERATURE REVIEW

Financial Markets prediction on the basis of online sentiment tracking has drawn a lot of attention recently. The Efficient Market Hypothesis asserts that financial market valuations incorporate all existing, new and even hidden information since investors act as rational agents who seek to maximize profit. A market in which prices always “fully reflect “available information is called efficient. Fama (Fama, 1970) distinguished three nested information sets, namely past prices, publicly available information and all information including private information. (Kondak, 1997:36). The efficient market hypothesis is divided into three stages as the weak form, semi-strong form and strong form with respect to the availability of the three mentioned information sets. The strong form of efficiency states that market prices reflect all information including the past prices and all the publicly available information, plus all the private information. In such a market, prices would always be fair and any investor, including insider traders, cannot beat the market (Brealey, Myers, & Marcus, 1999) (Yalcin & Can, 2010) Behavioral Finance has challenged this notion by emphasizing the important role of behavioral and social factors, including social mood, in financial decision making. As a consequence, measuring investor and social mood has become a key research issue in financial prediction.

Traditionally, empiricists have taken two approaches to measuring investor sentiment. Under the first approach, empiricists proxy for investor sentiment with market-based measures such as trading volume, closed-end fund discount, IPO first-day returns, IPO volume, option implied or mutual fund flows (Malcom & Jeffrey, 2006)A particularly interesting thread of this literature examines sentiment following non-economic events such as sports (Edmans, Garcia, & Norli, 2007)), aviation disasters (Kaplanski & Levy, 2010)weather conditions (Hirshleifer & Shumway, 2003), seasonal active disorder (Kamstra, Kramer, & Levi, 2003)and shows these sentiment-changing events cause changes in asset prices. Although market-based measures have the advantage of being readily available at a relatively high frequency, they have the disadvantage of being the equilibrium outcome of many economic forces other than investor sentiment. Under the second approach, empiricists use survey-based indices such as the University of Michigan Consumer Sentiment Index or the UBS/GALLUP Index for Investor Optimism, (Lemmon & Portniaguina, 2006), and (Qiu & Welch, 2006).Compared to survey-based measures of investor sentiment, the twitter-hashtag-based sentiment measure we propose has several advantages. First, hashtags are available at a high frequency while survey measures are often available monthly or

quarterly. Second, using hashtags reveals attitudes rather than inquire about them. Although many people answer survey questions for altruistic reasons, there is often little incentive to answer survey questions carefully or truthfully, especially when questions are sensitive (Singer, 2002). Using the trending twitter hashtags has the potential to reveal more personal information where non-response rates in surveys are particularly high or the incentive for truth-telling is low. Finally, some economists have been skeptical about answers in survey data which are not cross verified with data on actual behavior observed by objective external measurement (Vissing-Jorgensten, 2003)

In recent years, researchers have explored a variety of methods to compute indicators of the public's sentiment and mood state from large-scale online data. In the past couple of years, public mood indicators extracted from social networks such as Facebook, Live journal and Twitter have been used to predict stock market fluctuations. (Mao, Counts, & Bollen, 2011) Accordingly, a large body of empirical literature attempts to measure investor sentiment and to assess its effects on stock market activity (e.g. (Neal & Wheatley, 1998) so far several different proxies have been employed to capture investor sentiment as the available measures have instability.

For instance, Neal and Wheatley (Neal & Wheatley, 1998) investigate the forecast power of three potential sentiment measures; discounts on closed-end funds, net mutual fund redemption, and the ratio of odd-lot sales to purchases. The results indicate that two of these three proxies, namely fund discounts and net redemptions display some ability to predict the size premium and the difference between small and large firm returns. However, (Neal & Wheatley, 1998) only document a very weak evidence that odd-lot ratio predicts market returns.

In another similar paper, Baker and Wurgler (Malcom & Jeffrey, 2006) examine the question of how sentiment affects the cross section of stock returns rather than its impact on the aggregate market returns. They introduce the concept of categorization of the market as opposed to looking at the market as an aggregate whole. They measure investor sentiment by a composite index of six commonly used sentiment variables. Malcom and Jeffrey (Malcom & Jeffrey, 2006) document that smaller stocks, high volatility stocks, unprofitable stocks, non-dividend-paying stocks, extreme-growth stocks, and distressed stocks earn high returns following the periods of low sentiment, and they earn low returns when the sentiment is high. This sentiment index is

likely to understate the predictive power of investor sentiment because their index is based on the first principal component of six sentiment proxies that may have a common noise component. (Huang, Jiang, Zhou, & Tu, 2013) Econometrically, the first principal component is the best combination of the six proxies that maximally represents the total variations of the six proxies. (Baker, Wurgler, & Yuan, 2012) provide further international evidence for the cross-section forecasting power of investor sentiment. In contrast, the significance of the predictability of investor sentiment on the aggregate stock market has not been established yet. For instance, Baker and Wurgler (2007) find that the predictability of investor sentiment on the aggregate stock market is statistically insignificant and is much weaker than the case with cross-sectional studies. If the impacts of investor sentiment were restricted only to the cross-section, its role in finance could be limited. Therefore, the open question is whether or not investor sentiment is only relevant at the cross-section or matters to the aggregate stock market as well.

To further modify what Baker and Wurgler said, Dashan Huang and friends propose a new investor sentiment index for explaining asset expected returns. It performs much better than any of the commonly used macroeconomic variables. Empirical results suggest that investor sentiment is important not only cross-sectionally, but also at the aggregate market level. Aligned investor sentiment can achieve substantial improvements in forecasting stock returns either at the aggregate level or cross-sectionally. The aligned investor sentiment is an improvement of the fundamental measure of Baker and Wurgler. (Huang, Jiang, Zhou, & Tu, 2013)

(Tetlock, Giving Content to Investor Sentiment, 2007) finds that high media pessimism predicts downward pressure on market prices followed by a reversion to fundamentals, and unusually high or low pessimism predicts high market trading volume. These and similar results are consistent with theoretical models of noise and liquidity traders, and are inconsistent with theories of media content as a proxy for new information about fundamental asset values, as a proxy for market volatility, or as a sideshow with no relationship to asset markets. Tetlock extends Baker's paper by looking at whole market systemic issues and categorizing them as either optimistic or pessimistic. In contrast to these studies that rely on indirect proxies of investor sentiment, this paper focuses on a direct and new measure of sentiment compiled from an online social network, and investigates the role of sentiment in NSE index movement. This paper uses direct proxies of investor sentiment because they tend to reveal attitudes as opposed

to inquire about them as well as whole market systemic issues by categorizing the trending hashtags as positive, negative or neutral economic news.

The breadth of information dissemination affects stock returns. Mass media outlets play an important role in disseminating information to a broad audience, especially to individual investors. Stocks not covered by the media earn significantly higher future returns than stocks that are heavily covered, even after accounting for widely accepted risk characteristics. The “investor recognition hypothesis” advanced by Merton (1987) posits that in informationally incomplete markets, investors are not aware of all securities. (Tetlock, Giving Content to Investor Sentiment, 2007) analyzes the linguistic content of the mass media and reports that media pessimism predicts downward pressure and a subsequent reversal. (Tetlock, Saar-Tsechansky, & Mackassy, More than Words: Quantifying Language to Measure Firm's Fundamentals, 2007) document that the fraction of negative words used in news stories predicts earnings and stock returns. These findings suggest that qualitative information embedded in news stories contributes to the efficiency of stock prices.

Fang and Peress (Peress & Fang, 2009) investigate cross-sectional link between media coverage and expected stock returns. Without classifying news reports into positive or negative information, they show that less-covered firms exhibit higher returns even after controlling for well-known risk factors. Although the internet has become a mainstream media, we are aware of only few studies that use internet activity as a proxy for public interest and thereby examine the implications for stock returns. This is an extension of Tetlock's paper to focus on mainstream media. Da et al (2009) introduced the search volume of stock tickers from google trends as a measure of investor attention. They show that internet search volume measures attention more timely than do other well-established attention variables and primarily captures the individual investor attention. Vlastakis also use data from google trends to approximate information demand and public interest at the market level for stocks traded on the New York Stock Exchange. The authors report a positive link between information demand and investor risk aversion. (Matthias Bank, 2011). We can therefore infer from this that we can use an indirect proxy-level of information demand for a particular stock as opposed to the type of information requested.

Twitter has a much younger demographic as compared to other forms of social media. Nearly half of all users are under the age of 34 and only 30% are over 45. Facebook users connect with friends, family, colleagues and other acquaintances; connections are founded in an offline relationship. While offline relationships are possible on Twitter, they aren't as common. Therefore, the interactions on this social network tend to be less personal and more business oriented. (Aedy, 2013) Twitter is used more as a listen exercise. Users listen in on what the target audience is discussing so as to stay up-to-date on industry trends and give users what they want. This social network is appealing to users because of the amazing search capabilities. Being able to segment content based on topic is a feature Facebook or other social media sites don't offer. Twitter users take advantage of hashtags, trending topics, and saved searches and this enables it to have a larger circle of influence. (Aedy, 2013)

Baker and Wurgler constructed an investment sentiment index estimated simply from data, and find that it has strong forecasting power for a large number of cross-sectional stock returns. To investigate this prediction empirically, and to get a more tangible sense of the intrinsically elusive concept of investor sentiment, we start with a summary of the rises and falls in U.S. market sentiment from 1961 through the Internet bubble. This summary is based on anecdotal accounts and thus by its nature can only be a suggestive, ex post characterization of fluctuations in sentiment. Nonetheless, its basic message appears broadly consistent with our theoretical predictive power of investor sentiment because their index is based on the first principal component of six pr. Because cross-sectional patterns of sentiment-driven mispricing would be difficult to identify directly, we examine whether cross-sectional predictability patterns in stock returns depend upon proxies for beginning-of-period sentiment. For example, low future returns on young firms relative to old firms, conditional on high values for proxies for beginning-of-period sentiment, would be consistent with the ex-ante relative overvaluation of young firms. As usual, we are mindful of the joint hypothesis problem that any predictability patterns we find actually reflect compensation for systematic risks. The first step is to gather proxies for investor sentiment that we can use as time-series conditioning variables. Since there are no perfect and/or uncontroversial proxies for investor sentiment, our approach is necessarily practical. Specifically, we consider a number of proxies suggested in recent work and form a composite sentiment index based on their first principal component. To reduce the likelihood that these

proxies are connected to systematic risk, we also form an index based on sentiment proxies that have been orthogonalized to several macroeconomic conditions. The sentiment indexes visibly line up with historical accounts of bubbles and crashes. We then test how the cross-section of subsequent stock returns varies with beginning-of-period sentiment. Using monthly stock returns between 1963 and 2001, we start by forming equal-weighted decile portfolios based on several firm characteristics. (Our theory predicts, and the empirical results confirm, that large firms will be less affected by sentiment, and hence value weighting will tend to obscure the relevant patterns.) We then look for patterns in the average returns across deciles conditional upon the beginning-of-period level of sentiment. We find that when sentiment is low (below sample average), small stocks earn particularly high subsequent returns, but when sentiment is high (above average), there is no size effect at all. Conditional patterns are even sharper when we sort on other firm characteristics. When sentiment is low, subsequent returns are higher on very young (newly listed) stocks than older stocks, high return volatility than low-return volatility stocks, unprofitable stocks than profitable ones, and nonpayers than dividend payers. When sentiment is high, Investor Sentiment and the Cross-Section of Stock Returns 1647 these patterns completely reverse. In other words, several characteristics that do not have any unconditional predictive power actually display sign-flipping predictive ability, in the hypothesized directions, once one condition on sentiment. These are our most striking findings. Although earlier data are not as rich, some of these patterns are also apparent in a sample that covers 1935 through 1961. (Matthias Bank, 2011) However, this method by Baker found that predictability of investor sentiment on the aggregate stock market is statistically insignificant and is much weaker than the use of cross-sectional studies. If the impacts of investor sentiment were restricted only to the cross section, its role in finance could be limited. Baker's method fails to answer an important question, how to measure investment sentiment and quantity effect.

The method proposed by (Huang, Jiang, Zhou, & Tu, 2013) widely used Baker and Wurgler sentiment index is likely to understate the predictive power of investor sentiment because their index is based on the first principal component of six proxies that may have a common noise component. They exploited information on Baker six sentiment proxies in an efficient manner to obtain a new index for the purpose of explaining expected return in the aggregate stock market. This paper proposes a new sentiment index that is aligned for explaining stock expected returns by eliminating the noise component. It finds that the aligned sentiment index has greater power

in predicting aggregate stock market than the Baker and Wurgler index. It increases the R^2 by more than five times both in-sample and out-of-sample, and outperforms any of the well-recognized macroeconomic variables. Predictability is statistically and economically significant. The new index substantially improves the forecasting power of the cross section of stock returns formed on industry, size, value and momentum. Statistically, the Partial Least Squares Method proposed by Wold (1966, 1975) and extended by Keily and Priutt (2012, 2013) does this job. It finds that Aligned Investor Sentiment can predict aggregate stock market well. Cross-sectionality, we compare Aligned Investor Sentiment index performs relative to Baker average sentiment index. It is useful for cross-sectionality as well.

Another method proposed is to use daily internet search volume from millions of households to reveal market-level-sentiment. By aggregating the volume of queries related to household concerns (e.g. unemployment, recession, bankruptcy) they constructed a Financial and Economic Attitudes Revealed by Search (FEARS) index as a new measure of investor sentiment. They find that FEARS can predict the short term return reversals; although increases in FEARS correspond with low market returns today, they predict high returns (reversal) over the next few days. It can also predict temporary increases in volatility as well as mutual fund flows out of equity funds and into bond funds. The results are consistent with theories of investor sentiment. This methodology is more transparent than what was proposed by Baker and predicts aggregate market returns.

This paper refocuses the issue of best variable. Although the internet has become a mainstream media, we are aware of only few studies that use internet activity as a proxy for publications and examine the implications of stock returns. Internet search volume measures attention more timely than do other well established attention variables and primarily captures individual investor attention. Therefore we seek to fill the gap of the best variable as a proxy for investor sentiment by using a daily collection of trending twitter hashtags as a proxy for investor sentiment.

RESEARCH METHODOLOGY

Capital Market Data

The data I employ is a daily collection of the NSE 20 stock price and stock movement for a period of six months between 13th May 2015 to 22nd October 2015. The data set consists of:

- i) Opening price for each day
- ii) Closing price for each day
- iii) Highest price of the day
- iv) Lowest price of the day
- v) Percentage change in price.

All this data is publicly available and can be downloaded from any source in the internet.

Internet Data

The data covers a period of approximately six months collected from 13th May 2015 to 22nd October 2015. The daily trending hashtags on Kenyans on twitter are collected on a twice daily basis and the data is analyzed and classified as negative, positive or neutral socio-economic news.

The resulting dataset is in the form of a time-series of negative, neutral and positive tweets for each day. We create the following time series for each day.

- i) Volume of trending hashtags/Trend count: total number of hashtags in a day.
- ii) The sentiment indicator: Each negative tweet is given a value of -1, positive tweets are each given a value of 1, and neutral tweets are each given a value of 0.
- iii) Sentiment average: The average of the sentiment indicator is obtained for each day.

The data was collected manually on a daily basis from twitter by checking the hashtags that were trending on that day. In the analysis we investigate the relation between sentiment and the movement of the stock market.

Our regression equation is

$$Return = \beta_0(Trending\ Hashtags) + \sum_{i=0}^n Control\ Variables + \mu$$

Where the control variables are the volume of tweets and the sentiment average.

METHODS

This section first describes the methodology used for sentiment classification. Then, it presents the methods used for the correlation analysis and Granger causality. We adopt the methodology as used in the paper ‘The Effects of Twitter Sentiment on Stock Price Return’ which uses similar data to establish if there is a presence of a relationship between the sentiment as observed from twitter and how the stock price return for a number of companies.

Sentiment Classification

The process of sentiment classification is not easy. Financial experts often disagree whether a given tweet represents a buy or a sell signal, and even individuals are not always consistent with themselves.

Our classification is very subjective. Our approach is that from each trending hashtag, we visit twitter and view the conversations that were taking place under that hashtag and see if it was positive, negative or neutral economic news. We then tag the hashtag as negative, neutral or positive.

A constraint of this method is that the classification is very subjective and the trending hashtags contain non-English non-verbal words that cannot be subjected to automatic sentiment classification software.

We then introduced a sentiment indicator whereby a positive sentiment is indicated by a 1, negative sentiment is indicated by -1 and neutral sentiment is indicated by 0. We then obtain the sentiment average for each day and also use an indicator for the average whereby if the average is less than one we use 0 and greater than 0 we use 1. We also obtained the trend count for each day.

Capital Data Classification

We use an indicator for the percent change in the movement of the stock market whereby a percentage of less than 1 is indicated with a 0 and one that is greater than one is indicated with a value of 1. Similarly, we tag the change as negative or positive depending on the same classification as mentioned above.

CORRELATION AND GRANGER CAUSALITY

For an initial investigation of the relation between the Twitter sentiment and stock prices, we apply the Pearson correlation and Granger causality tests. We use the Pearson correlation to measure the linear dependence. Given two time series, X_t and Y_t , the Pearson's correlation coefficient is calculated as:

$$\rho(X, Y) = \frac{X_t Y_t - (X_t)(Y_t)}{\sqrt{[(X_t^2) - (X_t)^2][(Y_t^2) - (Y_t)^2]}}$$

We also perform the Granger causality test to check if the Twitter variables help in the prediction of the price returns.

The steps of the procedure applied are summarized as follows:

- i) Determine if the two time series are non-stationary, by the Augmented Dickey-Fuller (ADF) test.
- ii) Build a Vector Autoregressive (VAR) -ARDL model and determine its optimal order by considering four measures: AIC, BIC, FPE, HQIC.
- iii) Fit the VAR model with the selected order from the previous step.
- iv) Perform the Granger causality test.
- v) Perform the Ljung-box test for no autocorrelation in the residuals of the fit.
- vi) Perform a logit regression.

SHORTCOMINGS

We did not account for abnormal stock movements and volume of tweets.

RESULTS AND INTERPRETATION

This section first presents the correlation and Granger causality results over the entire time period.

Pearson and Granger Causality Tests.

Pearson correlation: Table 1 shows the computed Pearson correlations, as defined in the methods section. We find a low Pearson correlation between the time series over the entire time period with correlation values all being below 0.2.

Granger causality. The results show the results of the causality test in both directions: from the Twitter variables to the market variables and vice versa. The table gives the Granger causality links between

- a) Sentiment average and volume of tweets
- b) The volume of tweets and the movement of the stock market
- c) The sentiment average and the movement of the stock market

The conclusions that can be drawn are that neither the volume of tweets nor the sentiment of the market as indicated by twitter influence how the stock market behaves.

Wald Test

Having seen that the two variables; volume of tweets and the sentiment average are individually insignificant because of their high p-values, we test if they have joint significance. From the table 3, we see that the coefficients have high p-values and therefore this shows that the volume of tweets as well as the sentiment average jointly does not have an influence on the movement of the stock market.

Logit Regression

From the results as seen in the table, we see that an increase of tweet count by one increases the movement indicator by 0.25 *ceteris paribus*. We therefore conclude that there is a weak relationship between the volume of tweets and the movement of the stock market. This relationship is however insignificant.

Table 1: Pearson Correlation

	MOVEMENT T_INDICAT OR	TWEET_AV ERAGE	TWEET_CO UNT
MOVEMENT _INDICATO R	1.000000	-0.053801	0.115157
TWEET_AV ERAGE	-0.053801	1.000000	-0.019019
TWEET_CO UNT	0.115157	-0.019019	1.000000

Table 2: Granger Causality

Pairwise Granger Causality Tests

Date: 10/30/15 Time: 10:37

Sample: 5/13/2015 10/22/2015

Lags: 2

Null Hypothesis:	Obs	F-Statistic	Prob.
TWEET_AVERAGE does not Granger Cause TWEET_COUNT	113	0.08960	0.9144
TWEET_COUNT does not Granger Cause TWEET_AVERAGE		0.45250	0.6372
MOVEMENT_INDICATOR does not Granger Cause TWEET_COUNT	113	0.73804	0.4804
TWEET_COUNT does not Granger Cause MOVEMENT_INDICATOR		1.18392	0.3100
MOVEMENT_INDICATOR does not Granger Cause TWEET_AVERAGE	113	0.01403	0.9861
TWEET_AVERAGE does not Granger Cause		0.63013	0.5345

MOVEMENT_INDICATOR

Table 3: Wald Test

Wald Test:

Equation: Untitled

Test Statistic	Value	df	Probability
F-statistic	1.655357	(2, 105)	0.1960
Chi-square	3.310714	2	0.1910

Null Hypothesis: $C(5)=C(6)=0$

Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(5)	-0.016663	0.010572
C(6)	-0.000357	0.010482

Restrictions are linear in coefficients.

Table 4: Logit Regression

Dependent Variable: MOVEMENT_INDICATOR

Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)

Date: 10/27/15 Time: 16:34

Sample: 5/13/2015 10/22/2015

Included observations: 115

Convergence achieved after 2 iterations

Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-1.185629	0.479245	-2.473952	0.0134
TWEET_COUNT	0.051479	0.042633	1.207492	0.2272
TWEET_AVERAG				
E	-0.854649	1.522482	-0.561353	0.5746

McFadden R-

squared 0.012118 Mean dependent var 0.347826

S.D. dependent var	0.478365	S.E. of regression	0.479272
Akaike info criterion	1.328696	Sum squared resid	25.72663
Schwarz criterion	1.400303	Log likelihood	-73.40001
Hannan-Quinn			
critere.	1.357761	Deviance	146.8000
Restr. deviance	148.6008	Restr. log likelihood	-74.30041
LR statistic	1.800800	Avg. log likelihood	-0.638261
Prob(LR statistic)	0.406407		
<hr/>			
Obs with Dep=0	75	Total obs	115
Obs with Dep=1	40		
<hr/>			

DISCUSSION

In this work we present no significant evidence of dependence between the movement of the stock market and the twitter hashtags. Our results are not in agreement with previous results of similar research that has been carried out which have found that there is a relationship between public sentiment as obtained from twitter and other types of social media and the behavior of the stock market as is the paper on The effects of twitter sentiment on stock price return. (Ranco, Aleksovski, Caldarelli, Grčar, & Mozetic, 2015) We attribute this difference in results to the fact that the Kenyan market is not yet developed enough to a point where they can use social media to influence the activity of the financial market as with the case with other countries which are developed.

Similarly, this difference could also be due to the shortcomings in our model in that:

- i) The internet data that we collected in the form of hashtags was not economic data. It was just general social data.
- ii) We did not subject the hashtags to a sentiment tracking software. The tagging of the hashtags as positive, negative or neutral was very subjective

iii) The research study was subjected to time constraints therefore the data that we had covered a very short period of time seeing that the data was manually collected on a daily basis.

CONCLUSION

As a series of other papers have already shown, there is a signal worth investigating which connects social media and market behavior. This opens the way, if not to forecasting, then at least to “now-casting” financial markets. This study is open to further and more advanced research which is not subjected to time or data constraints as was the case with this paper.

Studies as this one could be well used in order to establish a direct relation between social networks and market behavior. A specific application could, for example, detect and possibly mitigate panic diffusion in the market from social network analysis. To such purpose there is some additional research to be done in the future. One possible direction is to test the presence of forecasting power of the sentiment time series.

APPENDICES

Table 5

Dependent Variable: MOVEMENT_INDICATOR

Method: Least Squares (Gauss-Newton / Marquardt steps)

Date: 10/30/15 Time: 10:09

Sample (adjusted): 5/18/2015 10/22/2015

Included observations: 112 after adjustments

MOVEMENT_INDICATOR = C(1)*MOVEMENT_INDICATOR(-1) + C(2)

*MOVEMENT_INDICATOR(-2) + C(3)*TWEET_AVERAGE(-1) + C(4)

$$*TWEET_AVERAGE(-2) + C(5)*DTWEET_COUNT(-1) + C(6)$$

$$*DTWEET_COUNT(-2) + C(7)$$

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	0.195022	0.097945	1.991142	0.0491
C(2)	-0.013819	0.096859	-0.142671	0.8868
C(3)	-0.349282	0.365322	-0.956092	0.3412
C(4)	0.170963	0.360928	0.473675	0.6367
C(5)	-0.016663	0.010572	-1.576061	0.1180
C(6)	-0.000357	0.010482	-0.034039	0.9729
C(7)	0.276002	0.064455	4.282121	0.0000
R-squared	0.067305	Mean dependent var		0.348214
Adjusted R-squared	0.014008	S.D. dependent var		0.478545
S.E. of regression	0.475182	Akaike info criterion		1.410224
Sum squared resid	23.70879	Schwarz criterion		1.580130
Log likelihood	-71.97254	Hannan-Quinn criter.		1.479160
F-statistic	1.262823	Durbin-Watson stat		1.934079
Prob(F-statistic)	0.280916			

Table 6

System: UNTITLED

Estimation Method: Least Squares

Date: 10/30/15 Time: 06:39

Sample: 5/18/2015 10/22/2015

Included observations: 112

Total system (balanced) observations 336

Coefficient	Std. Error	t-Statistic	Prob.
-------------	------------	-------------	-------

C(1)	0.195022	0.097945	1.991142	0.0473
C(2)	-0.013819	0.096859	-0.142671	0.8866
C(3)	-0.349282	0.365322	-0.956092	0.3398
C(4)	0.170963	0.360928	0.473675	0.6361
C(5)	-0.016663	0.010572	-1.576061	0.1160
C(6)	-0.000357	0.010482	-0.034039	0.9729
C(7)	0.276002	0.064455	4.282121	0.0000
C(8)	-0.002101	0.026174	-0.080268	0.9361
C(9)	0.005303	0.025884	0.204882	0.8378
C(10)	0.066370	0.097627	0.679839	0.4971
C(11)	0.090132	0.096452	0.934471	0.3508
C(12)	-0.000216	0.002825	-0.076587	0.9390
C(13)	-0.000916	0.002801	-0.327084	0.7438
C(14)	-0.036836	0.017224	-2.138587	0.0332
C(15)	0.267166	0.846581	0.315582	0.7525
C(16)	0.202955	0.837199	0.242421	0.8086
C(17)	-0.848775	3.157650	-0.268800	0.7883
C(18)	-2.626774	3.119664	-0.842005	0.4004
C(19)	-0.678411	0.091381	-7.423961	0.0000
C(20)	-0.340383	0.090601	-3.756932	0.0002
C(21)	-0.416883	0.557110	-0.748294	0.4548
Determinant residual covariance		0.049567		

Equation: MOVEMENT_INDICATOR =
C(1)*MOVEMENT_INDICATOR(-1) +
C(2)*MOVEMENT_INDICATOR(-2) +
C(3)*TWEET_AVERAGE(-1) + C(4)
*TWEET_AVERAGE(-2) + C(5)*DTWEET_COUNT(-1) +
C(6)

$$*DTWEET_COUNT(-2) + C(7)$$

Observations: 112

R-squared	0.067305	Mean dependent var	0.348214
Adjusted R-squared	0.014008	S.D. dependent var	0.478545
S.E. of regression	0.475182	Sum squared resid	23.70879
Durbin-Watson stat	1.934079		

Equation: TWEET_AVERAGE =

$$C(8)*MOVEMENT_INDICATOR(-1) + C(9)$$

$$*MOVEMENT_INDICATOR(-2) +$$

$$C(10)*TWEET_AVERAGE(-1) + C(11)$$

$$*TWEET_AVERAGE(-2) + C(12)*DTWEET_COUNT(-1) +$$

$$C(13)$$

$$*DTWEET_COUNT(-2) + C(14)$$

Observations: 112

R-squared	0.015825	Mean dependent var	-0.042797
Adjusted R-squared	-0.040413	S.D. dependent var	0.124494
S.E. of regression	0.126985	Sum squared resid	1.693144
Durbin-Watson stat	2.000475		

Equation: DTWEET_COUNT =

$$C(15)*MOVEMENT_INDICATOR(-1) + C(16)$$

$$*MOVEMENT_INDICATOR(-2) +$$

$$C(17)*TWEET_AVERAGE(-1) + C(18)$$

$$*TWEET_AVERAGE(-2) + C(19)*DTWEET_COUNT(-1) +$$

$$C(20)$$

$$*DTWEET_COUNT(-2) + C(21)$$

Observations: 112

R-squared	0.348310	Mean dependent var	-0.017857
-----------	----------	--------------------	-----------

Adjusted R-squared	0.311071	S.D. dependent var	4.948349
S.E. of regression	4.107218	Sum squared resid	1771.270
Durbin-Watson stat	2.158484		

Table 7

Sample: 5/13/2015 10/22/2015

Included observations: 114

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob	
**** .	**** .	1	-0.513	-0.513	30.810	0.000
. .	** .	2	0.013	-0.340	30.830	0.000
. .	* .	3	0.037	-0.197	30.990	0.000
. .	* .	4	0.012	-0.089	31.007	0.000
. .	. .	5	-0.018	-0.051	31.046	0.000
* .	** .	6	-0.112	-0.216	32.585	0.000
. **	. *	7	0.258	0.109	40.789	0.000

**		*		8	-0.291	-0.162	51.376	0.000
*		*		9	0.101	-0.157	52.661	0.000
.		*		10	0.034	-0.100	52.810	0.000
*		*		11	0.086	0.115	53.766	0.000
**		*		12	-0.218	-0.146	59.903	0.000
*		*		13	0.084	-0.137	60.826	0.000
*		*		14	0.184	0.093	65.319	0.000
**		.		15	-0.215	0.041	71.490	0.000
*		.		16	0.109	0.043	73.093	0.000
.		.		17	-0.041	-0.005	73.320	0.000
.		*		18	-0.008	-0.083	73.330	0.000
.		.		19	0.004	0.068	73.332	0.000
.		.		20	0.028	0.022	73.440	0.000
.		.		21	0.054	0.060	73.856	0.000
*		*		22	-0.120	0.079	75.913	0.000
.		.		23	0.007	-0.046	75.921	0.000
*		.		24	0.084	-0.007	76.955	0.000
.		.		25	-0.036	0.008	77.148	0.000
.		.		26	-0.026	0.040	77.247	0.000
.		.		27	0.042	0.048	77.522	0.000
*		*		28	-0.066	-0.096	78.195	0.000
.		.		29	0.060	0.021	78.748	0.000
.		.		30	-0.021	-0.018	78.817	0.000
.		.		31	0.002	-0.028	78.817	0.000
.		.		32	-0.025	-0.048	78.921	0.000
.		.		33	0.005	-0.047	78.925	0.000
.		*		34	-0.035	-0.167	79.131	0.000
*		.		35	0.083	-0.065	80.292	0.000
.		.		36	-0.049	-0.057	80.699	0.000

REFERENCES

Bibliography

- Aedy, S. (2013, March 15). *Facebook vs Twitter: Know whom you are writing for*. Retrieved June 29, 2015, from Socialnomics: [www.socialnomics.net/2013/03/2015/facebook-s-twitter-know-whom-you're-writing-for/](http://www.socialnomics.net/2013/03/2015/facebook-s-twitter-know-whom-you-re-writing-for/)
- Baker, M. P., Wurgler, J., & Yuan, Y. (2012). Global, Local and Contagious Investor Sentiment. *Journal of Financial Economics*.
- Bollen, J., Mao, H., & Zeng, X.-J. (n.d.). *Twitter mood predicts the stock market*.
- Brealey, R. A., Myers, S. C., & Marcus, A. J. (1999). Market Rationality: Efficient Market Hypothesis. *Fundamentals of Corporate Finance*.
- Capital Group. (n.d.). *Lifestyle magazine*. Retrieved June 19th, 2015, from Capital Fm: <http://www.capitalfm.co.ke/lifestyle/2012/01/26/kenya-is-africas-2nd-most-active-on-twitter/>
- Chang, S., & Liu, S. (2011). An analysis of the predictive powers of real-time social media. *Predicting Stock Market Fluctuations from Twitter*.
- Da, Z., Engelberg, J., & Gao, P. (2009). The sum of all FEARS: Investor Sentiment and Asset Prices.

- Davis, A. (2006). The role of mass media in investor relations. *Journal of Communication Management*.
- Edmans, A., Garcia, D., & Norli, O. (2007). Sport Sentiment. *Journal Of Finance*.
- Fama, E. (1970). Efficient Capital Markets. *Journal of Finance*.
- Hirshleifer, D., & Shumway, T. (2003). Good Day Sunshine: Stock Returns and the Weather. *Journal of Finance*, 1009-1032.
- Huang, D., Jiang, F., Zhou, G., & Tu, J. (2013). Investor sentiment aligned: A powerful predictor of stock returns. *Research Collection Lee Kong Chian School of Business*.
- Information Aesthetics. (n.d.). *Pulse of the Nation: US Mood throughout the day inferred from twitter*. Retrieved June 19, 2015, from infosthetics.com:
http://infosthetics.com/archives/2010/07/pulse_of_the_nation_us_mood_throughout_the_day_inferred_from_twitter.html
- Kamstra, M. J., Kramer, L. A., & Levi, M. D. (2003). *Winter Blues: A SAD Stock Market Cycle*.
- Kaplanski, G., & Levy, H. (2010). Sentiment and Stock Prices: The case of aviation disasters. *Journal of Financial Economics*.
- Kemibaro, M. (2014, August). *Sizing up twitter in Kenya*. Retrieved June 17, 2015, from Moses Kemibaro: Pure Digital Passion: <http://www.moseskemibaro.com/2014/08/01/sizing-up-twitter-in-kenya/>
- Lemmon, M., & Portniaguina, E. (2006). Consumer Confidence and Asset Prices: Some empirical evidence. *Oxford Journals: Review of Financial Studies*, 1499-1529.
- Malcom, B., & Jeffrey, W. (2006). Investor Sentiment and the Cross-Section of Stock Return. *Journal of Finance*.
- Mao, H., Counts, S., & Bollen, J. (2011). *Predicting financial markets: Comparing surveys, news, twitter and search engine data*.
- Matthias Bank, M. L. (2011). Google Search Volume and Its Influence on liquidity and returns of German Stocks. In Bank, Matthias and Larch, Martin and Peter, Georg, *Google Search Volume and Financial Markets and Portfolio Management* (pp. 239-264).
- Neal, R., & Wheatley, S. (1998). Do Measures of Investor Sentiment Predict Stock Returns. *Journal of Financial and Quantitative Analysis*, 523-547.
- Peress, J., & Fang, L. (2009). Media Coverage and the Cross-Section of Stock Returns. *Journal of Finance*.
- Portland Communications. (n.d.). *Portland*. Retrieved June 19th, 2015, from [www.portland-communications.com: http://www.portland-communications.com/publications/how-africa-tweets-2014/#sthash.KSgTdVTz.dpuf](http://www.portland-communications.com/publications/how-africa-tweets-2014/#sthash.KSgTdVTz.dpuf)

- Qiu, L. X., & Welch, I. (2006). *Investor Sentiment Measures*.
- Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetic, I. (2015). *The Effects Of Twitter Sentiment On Stock Price Return*. Warwick.
- Shiller, R. J. (2000). The News Media. In R. J. Shiller, *Irrational Exuberance* (pp. 71-94). New Jersey: Princeton University.
- Singer, E. (2002). *The Use of Incentives To Reduce Nonresponse in Household Surveys*.
- Strickland, J. a. (2007, December 17). *How Twitter Works*. Retrieved June 2015, 2015, from HowStuffWorks.com: <<http://computer.howstuffworks.com/internet/social-networking/networks/twitter.htm>>
- Tetlock, P. C. (2007). Giving Content to Investor Sentiment. *Journal of Finance*.
- Tetlock, P. C., Saar-Tsechansky, M., & Mackassy, S. (2007). More than Words: Quantifying Language to Measure Firm's Fundamentals. *9th Annual Texas Finance Festival*.
- Twitter. (n.d.). Retrieved June 17, 2015, from WhatIs.com: whatistechtarget.com/definition/Twitter
- Twitter. (n.d.). Retrieved June 17, 2015, from www.twitter.com: <http://about.twitter.com/company>
- Vissing-Jorgensten, A. (2003). *Perspectives on Behavioural Finance: Does irrationality disappear with wealth*.
- Yalcin, & Can, K. (2010). Market Rationality: Efficient Market Hypothesis vs Market Anomalies. *European Journal of Economic and Political Studies*.