Vysoká škola ekonomická v Praze

Národohospodářská fakulta

Hlavní specializace: Ekonomie

ONLINE TRENDS AND SENTIMENT AND THEIR POSSIBLE APPLICATION IN STOCK MARKET PREDICTION

Bachelor Thesis

Solver: Josef Stehno

Thesis supervisor: Mgr. Ing. Pavla Vozárová, Ph.D,

Rok: 2015

Hereby I declare that I have worked on this thesis on my own and to the best of my knowledge. It contains no material previously published or written by another person, except where due acknowledgment is made.

…….....

Josef Stehno

Prague, 15. 05. 2015

ACKNOWLEDGEMENTS

I wish to thank all who have helped me with this thesis. Firstly to my thesis supervisor Ms. Pavla Vozárová, for her much appreciated help and amazing patience she had when dealing with me. To Mr. Pierce Gibson Crosby, the Director of Business Development at StockTwits for willingness to provide me with the data and patience when dealing with my requests. Also I would like to thank to David Kolesa, account manager at Google for his willingness to help even though our data request was denied. Big thanks also belong to Martin Gavora for much needed help with programing the script for processing StockTwits data. Last but the most important Thank you, belongs to my mother Zuzana Stehnová, for giving me the incentive to study and helping me during studies.

University of Economics, Prague
Department of Economics

Faculty of Economics
Academic year: 2014/2015

BACHELOR THESIS TOPIC

Author of thesis : **Josef Stehno**

Study programme: Economics and Economic Administration

Field of study: Economics

Topic: **Online trends and sentiment and their possible application in stock market prediction**

Guides to writing a thesis:

1. New technologies and recent social development are creating a world which is globalized on such a scale that was unimaginable even few years ago. This recent trend of social networking and unprecedented availability of information through internet is also helping to literally transport our lives to zeros and ones. This has many consequences; one of them is the recent phenomenon of Big Data usage in business and investment decisions. But is it possible to use such data for prediction of stock prices or trade volumes development? The purpose of the thesis is to answer this question by studying whether search trends data and social network sentiment data can or cannot predict the evolution of stock markets.
2. The analysis will be based on econometric analysis of relationships between these trends or sentiments and various indicators of stock markets such as indexes, stock prices or trade volumes.
3. Data will be downloaded from Internet. Search trends data will be retrieved from Google trends (number of Google searches executed on a particular stock index or a firm), sentiment data will be retrived fromstocktwits.com (derivation of twitter focused directly on stock markets a community around them, where members of this community can express their sentiment by putting either bullish or bearish emoticon).

Length of thesis: 45 pages

Selected bibliography:

1. FAMA, E. F. Market efficiency, long-term returns, and behavioral finance. Journal of financial economics. 1998. Vol. 49, no. 3, pp. 283-306.
2. CHAN, E. Algorithmic trading: winning strategies and their rationale. John Wiley & Sons, 2013.
3. CHOI, H. – VARIAN, H. Predicting the Present with Google Trends [online]. 2009. Available at: http://static.googleusercontent.com/external _content/untrusted _dlcp/www.google.com/en/us/googleblogs/pdfs/google _predicting _the _present.pdf
4. LIU, B. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies.

   2012. Vol. 5, no. 1, pp. 1-167.
5. LOUGHLIN, Ch. – HARNISH, E. The Viability of StockTwits and Google Trends to Predict the Stock Market [online]. 2013. Available at: www.stocktwits.com/research/Viability-of-StockTwits-and-Google-Trends-Loughlin_Harnisch.pdf

Bachelor thesis topic submission date: February 2015

Deadline for submission of Bachelor thesis: May 2015

Josef Stehno

Solver

Mgr. Ing. Pavla Vozárová, Ph.D

Thesis supervisor

prof. Ing. Robert Holman, CSc.

Head of Institute

doc. Ing. Miroslav Ševčík, CSc.

Dean NF VŠE

Abstract

The goal of this thesis is to assess information contained in internet user's activity. I focus on two sources of data: Google Trends and sentiment contained in StockTwits posts. For both of them I examine the correlation of its percentage changes and percentage changes of variables describing the stock market development. Econometric testing consists of three phases, first is Least Squares Method, then ARIMA model, and lastly testing for Granger Causality. Conclusions are that activity of internet users does contain valuable information. The correlations are strongest for firms operating in IT business or generally focusing on modern technologies. Strong correlation is between trade volume or market volatility and Google Trends, whereas sentiment in post on StockTwits is statistically significant for stock price development.

Key words:

stock, sentiment, correlation, trade, volume, internet, price, prediction, market, development.

Abstrakt

Cílem této práce je zkoumání informací obsažených v aktivitě uživatelů na internetu. Soustředím se na dva zdroje dat, Google Trends a sentiment obsažený v příspěvcích na StockTwits. Pro tyto data zkoumám korelaci mezi procentuálními změnami a procentuálními změnami veličin popisujících vývoj akciového trhu. Ekonomické testování má tři fáze: prvně testování pomocí metody nejmenších čtverců, následně pomocí ARIMA modelu a nakonec testy na Granger kauzalitu. Závěr je takový, že tyto data vskutku obsahují cenné informace, korelace je nejsilnější pro firmy pohybující se na trhu s IT a moderními technologiemi. Silnou korelaci jsem nalezl mezi objemem obchodů nebo cenovou volatilitou a Google Trends, zatímco sentiment na sociální síti StockTwits je statisticky významný pro vývoj cen akcií.

Klíčová slova:

akcie, sentiment, korelace, internet, cena, obchod, objem, predikce, trh, vývoj

JEL classification:

G10, G12, G14, G17

TABLE OF CONTENTS

Introduction

Globalization and recent development of modern technologies are together changing world and creating opportunities which have not been available or even imaginable few years earlier. One of them are Big Data, a phenomenon of our time. Our life is more and more efficiently transmitted in to digital sphere and almost all aspect of our life and of our character are becoming quantifiable. And these data have already many times proved their information potential in many types of business. But are these data applicable also with connection to stock markets? Is there any viable information in activity of internet users, which can be extracted and used by investors for assessing the future development of stock market?

To assess this question I will go through the most know and accepted hypotheses of market development, The Efficient Market Hypothesis and The Adaptive Market Hypothesis and try to sum the recent state at which the research on this topic is. Sadly there is not so not so much papers and works focusing on this area. I will look at these market hypotheses in a light of conclusions of previous papers and works.

The practical part of the work will be an econometric analysis of both Google Trends and StockTwits data, where I will try to find a correlation between activity of internet users and stock market development both in respect of stock returns and trade volumes. Data concerning Google Trends will be downloaded from publicly available service of Google Trends, where Google provides records of searches executed on specific term on weekly basis. For StockTwits I was able to secure daily data directly from StockTwits which are not publicly available. The structure of testing will be following. Firstly I will test for correlation between variables using Least Squares Methods, then if the effect proves to be significant I will move to ARIMA test to soften the effect of autocorrelation, if needed and lastly I will test for Granger Causality to determine whether the sentiment on StockTwits and Google Trends are leading or lagging indicators of stock market.

For the testing I picked 18 stock market titles and one whole market index. Companies were selected in a way to provide somehow cross market overview on viability of these data for prediction of companies stock development. . For most of the companies I would expect at least correlation between changes in search volume on Google, e.g. Google Trends and changes of trade volume. As for the correlation between Google Trends and price movements, there I am mostly skeptical that Google Trends will contain

information precise enough to follow the movement of stock prices, but I believe that looking into absolute values of those changes, where the absolute values of changes represent stocks volatility. For the StockTwits I expect the highest correlation and viability in case of companies operating in IT business or generally in modern technology businesses such as Apple, Facebook or Tesla. Also interesting will be to examine stocks of companies which we could mark as trendy like Coca Cola or Michael Kors.

1. Theoretical part

1.1 Market hypotheses

Since the beginning of modern financial markets investors and speculators are trying to gain an edge on the other subjects on market. Even though traditional market theory and Efficient Market Hypothesis are saying that all effort contributed to predicting future movements of stock market is a wasted effort. Especially according Efficient Market Hypothesis by Eugene Fama for which he had been awarded with a Nobel Price for economic sciences in 2013, according which it is impossible to beat market as stocks are always trading at their fair value which is determined by all the available information. (Fama 1970) Meaning that it is impossible to legally gain edge on the other subjects on market because information is available for the whole market. This fact adjusts prices to its fair level at every moment making it impossible to buy undervalued stocks or sell overvalued ones. Beating the market and making a long run profit with trading strategies becomes theoretically impossible. Moreover any kind of predicting price based on past stock movements or on determining bargain stocks is impossible. This all essentially means that arbitrage opportunities to simultaneously buy and sell stocks on financial markets with making riskless profit are impossible to identify and exploit in long run view. The only way to make a higher return than is the average return on market is to buy more risky stocks. If this theory would be correct, all investors should basically stop trading

and speculating on stocks and focus on investing into index funds such as S&P500 and minimalizing their cost. That would, eventually, bring then the same profit as trading with stocks and their risk would be minimal. There is an old joke I came across, which is, in my view, perfectly summing up the logical process behind this theory, it goes like this:

“Two investors are going down the street. They come upon a $100 bill lying on the ground, and as one of them reaches down to pick it up, the other remarks, "Don't bother, if it was a genuine $100 bill, someone would have already picked it up."

Eventually Efficient Markets Theory was also empirically proved many times. For example as it is stated in the work Reflections on the Efficient Market Hypothesis: 30 Years Later (Malkiel, 2005) author concludes that evidence is clear that active portfolio management is in fact useless. He says that there is, based on hard data, evidence that switching from security to security accomplishes nothing, in fact, and even if markets

are not perfectly efficient, active portfolio management is likely to produce lower returns than just simple indexing and also is accompanied with increased transaction costs. Thus, both institutional as well as individual investors would be well served to use indexing investment strategies at least for the main part of their portfolio. He supports his claim by two quotes from legendary investors Benjamin Graham and his most famous student Warren Buffet, who is probably the most successful modern-day investor.

> *"I am no longer an advocate of elaborate techniques of security analysis in order to find superior value opportunities. This was a rewarding activity, say, 40 years ago, when Graham and Dodd was first published; but the situation has changed ... [Today] I doubt whether such extensive efforts will generate sufficiently superior selections to justify their cost ... . I'm on the side of the 'efficient market' school of thought."* (Benjamin Graham, 1976; cited by Malkiel 2005)

> *"Most investors, both institutional and individual, will find that the best way to own common stocks (shares') is through an index fund that charges minimal fees. Those following this path are sure to beat the net results (after fees and expenses) of the great majority of investment professionals."* (Warren Buffet, 1996; cited by Malkiel 2005)

The reality though often looks differently. The reason for that is that markets in fact may be efficient but they are not perfectly efficient or that they may often be efficient but they are not always. Main flaw in the Efficient Market Theory, in my opinion lies in the assumption of perfect information and its speed in which it is supposed to reach investors and it is also the main ground for critics of this theory. The point is that for markets it takes time to respond to new information, more importantly the information spread is not flat and so the information may reach different market subjects in different time, resulting in advantage for investors who get the information as first or for the ones who are able to react faster than other subjects. But that is not the only reason for which may be the Efficient Markets Theory incorrect. Next flaw in Efficient Market Hypothesis represents the fact that information is subjective, thus every investor may see the same information in a different light and it may cause a different reaction in stock evaluating process for every each of them, causing so deviations from situations described by Efficiency Market Hypothesis.

Another issue on which was build reacting theory of Adaptive Markets Hypothesis is the fact that people trading on markets may make mistakes, may act irrationally and be subjected to herd behavior or be controlled by fears or greed instead of clear ratio. The Adaptive Markets Hypothesis connects implications of Efficient Market Hypothesis with behavioral science alternatives, specifically applying the principles of evolution, competition, adaptation and natural selection in terms of financial interactions. (Lo, 2014) The implications of this theory, which are going against Efficient Market Hypothesis are following:

Firstly that there is a proof that relation between risk and reward actually exists and it is unlikely to remain the same over time. Because this relation is determined by various factor which are changing over time. Such as preferences of population on market, relative sizes of groups with different preferences as well as market environment, created by laws and regulatory institutions. As all of these factors are variable over the time, the risk and reward relation is being affected and with that also the risk premium varies. The implication of this is that aggregated risk preference of the market is not and cannot be stable through different time periods. For example author provides an example regarding the turn of the millenniums and burst of the technology bubble. As there have been two completely different generations of investors. The one before the technology bubble burst which have never experienced genuine bear market and the population of investors active in years after the burst of technology bubble. According the author, in this context, it is a natural selection which had determined the market environment as the investors who have lived through the burst of technology bubble and suffered substantial losses most likely have exited market, creating so place for new generation of investors who in the light of recent market losses have a different relation between risk and reward, shifting so the aggregate preferences of the markets. (Lo, 2014)

Second implication which goes directly against classical Efficient Markets Hypothesis is that in fact arbitrage opportunities do from time to time exist. As author says from an evolutionary perspective, the existence of liquidity on markets implies that also profit opportunities must be present and they disappear as they are exploited. Hand in hand with that goes that new opportunities are continuously created. As in nature, while old species extinguish new are born. So in contrary to super efficiency predicted by EMH, the Adaptive Markets Hypothesis implies that markets are much more dynamic, ruled by panics, manias, bubbles and other phenomenon. (Lo, 2014)

Third implication is resulting from the first two. As the market environment is changing and the arbitrage opportunities are present and new appears as the old ones are exploited, investment strategies and its performance also varies in time. It may decline for time, but then, when market environment shifts they may become profitable again. Author again provides an example from the years after technology bubble burst when, in the period after the burst, the risk arbitrages declined significantly only to regain its place in a few years, when the activity of investment banking have risen back and number of M&A rose significantly. (Lo, 2014)

Final and main implication from Adaptive Market Hypothesis is that innovation is the key to survival and that survival is the only objective that matters. The classical Efficient Market Hypothesis states that higher levels of returns than average market return can be achieved only by bearing higher level of risk. On the contrary the Adaptive Market Hypothesis suggests that as the risk/reward relation changes in time with different market conditions, then profits higher than average profits on markets can be achieved simply by constant and swift adapting to changing market conditions. The surviving is the main and only objective that matters. As profit or utility maximization are both relevant factors in financial market evolution, the key and organizing principle is, the same as in nature, survival. (Lo, 2004)

Both of these theories now stand a challenge in a light of recent fast extension of the population with access to internet and expansion of the content available there. Which can in fact move the scales in favor of adaptive market theory. Nevertheless world of investment and trading had changed under these circumstances. With dramatic decrease of transaction costs, investing has been set free from local trading floors in to the world of internet, making possible to trade stocks and invest at any point of time, at any place, into stock all over the world. Eventually this massive use of internet and gradual transition of our lives in to the internet sphere resulted into birth and rapid extension of social media. This had fundamentally changed the way people communicate with each other and created new ways to instantly share their opinions, thoughts, and ideas with other people. In this way internet users creates, with their actions on the internet, unprecedented amount of information which has never been available before. Based on these sources of available data new ways of prediction had appeared and the access to these data together with their appropriate analysis are creating more sophisticated ways how to predict future on financial markets based on current mindset

of population, their preferences, opinions and beliefs. For the first time ever the fear and the optimism are becoming quantifiable.

This phenomenon of using information and data available on internet in form of user's sentiment for predicting future success or failure of businesses started a few years ago. One of them and probably the most known and cited is the paper Predicting the Future With Social Media in which authors, with regression model, proved strong correlation between future success of movies and sentiment of tweets on twitter concerning that movie or general topic of the movie. Even though that this study and other similar are not focused directly on financial markets their conclusions are still important and interesting for my work because they prove that there are hidden information on internet in form of a sentiment, which is with recent development of information technologies becoming available and quantifiable. Moreover it also demonstrates that information hidden in this data are actually useful and that there is a viability of social media in predictions for various topic.

1.2. Literature Review

Era of dot.com bubble marks the first signs of interest and the first attempts to collect and utilize data available on then rapidly developing web for the purpose of speculation and predicting on development on financial markets. One of the first papers on this subject: Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards focused on determining if the message-posting volume on stock message boards or investor chat rooms is just a noise or is somehow related to underlying firm´s characteristic and stock market activity. (Wysocki, 1998) In his work he comes to conclusion that message posting volume is, on average, higher for firms which are in some way not ordinary, for example they had extreme past returns or financial performance, or they had higher past volatility and trading volume or had been supported by analysis predicting higher future profits. Beside this the higher message posting volume is also connected with the healthcare or technology firms and IT business or generally with firms with the highest market capitalization. Also the time-series results show that daily posting volume increases during announcement, which seems logical as the atmosphere in investment world is getting more intense during these days lot of rumors may occur. Moreover there is an interesting conclusion that changes in overnight message posting volume predict next day trading volume and abnormal stock returns. The interpretations of these results may be that investors and individuals active on message boards focus on the firms with the greatest likelihood of generating future information flows, highest uncertainty and risk, so generally firms with the largest information asymmetry and the poorest accounting information. Another implication coming from the fact that the highest messages posting volume are as, have been mentioned, connected with firms which are in some way not ordinary, for example they had extreme past returns or financial performance, or they had higher past volatility and trading volume or had been supported by analysis predicting higher future profits or operating in the healthcare, technology and IT business, is that there could also be a behavioral effect in play and that is an irrational fixation on glamour or "cool and trendy" stocks.

Besides this paper there were few articles concerning possible links between message-posting volume and price moves if underlying stock as Traffic on financial web pages rises when the market falls (Bennett, 1998), or Gossip central - Internet message boards

can leave some stocks hanging by a thread and others (Batsell, 1991) coming into the same or very similar conclusion as Wysocki's work.

In 2001 Tumarkin and Whitelaw followed work of Wysockij using similar methods in their work News or Noise? Internet Postings and Stock Prices in which they focused directly on the web page RagingBull.com, at that time, a very popular page in investing community. Even though they proved a strong correlation between message-posting volume and trading volume during the next day, there was no correlation between message-posting and future development of market, so prediction viability for the way of stock price movements was denied. (Tumarkin and Whitelaw, 2001) Then when later in 2001 came burst of the dot-com bubble and interest about this area fell down and for a while this topic was left unattended.

In 2004 Antweiler and Frank again returned to the topic message-posting viability to prediction of movement on financial markets and in their paper: Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards again focused on how strongly are message boards related to stock markets. They come to conclusion that the stock messages reflect publicly available information very rapidly. Also the evidence clearly proves that this talk is not just a noise and there is relevant information for financial markets. And in some respects this information goes even beyond what can be found in newspapers or other news channels. Though this talk and information have proved to have viability for predicting stock returns, this effect is being pushed back with transaction costs. But viability of these information for predicting trade volume and volatility is relevant. (Antweiler and Frank, 2004)

Another improvement in this area was work of Sanjiv R. Das and Mike Y. Chen: Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. In their work they go beyond just analyzing message-posting volume, but for the first time there, they try to analyze the sentiment in those messages. For this purpose they had developed a method for extracting small investor's messages form message boards and analyzing sentiment in them with use of their own algorithm. The results were that there is no significant correlation between sentiment of the posts determined by their algorithm and specific stock price movements. Even though aggregated results for all 24 stocks together were much more promising. They attribute this to the fact that message boards contain large amount of noise, but when aggregated, this noise is reduced by larger amount of data. They also provide another explanation and that is the fact that on the message boards

usually meet smaller investors which do not hold such a market power to influence markets. Thus the sentiment on these board is not relevant factor determining market development. (Das and Chen, 2007)

Great change in this area is year came with era of social networks, especially with expansion of Twitter, since then the attention has shifted from message board and forums to this new social network as it has changed the way people interact, communicate and express their opinions or ideas. One of the best paper focused on predicting stock market movements with help of Twitter is work: Twitter mood predicts the stock market. (Bollen, Mao and Zeng, 2011) In which authors aggregated overall twitter chatter and evaluated individual posts according their mood. Based on this they came to conclusion that results indeed show that changes in the public moods can be tracked and extracted from the large scale content Twitter feeds by simple text processing methods. Moreover as there was scale of seven observed public moods the changes in some of them do correspond with the changes in Down Johns Index that occur three to four days later. What is even more surprising on this, is the fact that changes in DIJA values do not correspond with changes in mood dimensions labeled as positive to negative mood, but more with dimension labeled as calm. Suggesting so that the financial markets are not so much related to positive or negative moods of population as to general calmest or distress of population. (Bollen, Mao and Zeng, 2011)

Authors do note that their analysis is not designed for particular geographical location or subset of world's population base don language, as they were simply processing all the twitters feeds. They note that this may have caused some inaccuracy in their work, but as for the time of work the twitter users were almost exclusively residents of USA and there is valid expectation that only communication in English may be correlated with USA stock markets, the inaccuracy is not significant. But as Twitter become more and more international and English is becoming the most commonly used language in the world, future analysis will have to take these issues in account. (Bollen, Mao and Zeng, 2011)

This work again proves that there is viable information contained on internet. Here in form of a strong correlation between overall collective mood among population on Twitter and stock market movements and also proved this overall mood to be viable for predictions of upward and downward movements of the stock market.

As the paper Twitter mood predicts the stock market was based on assessing all twitter communication, the next work on this topic, The Information Content of Stock Microblogs focus directly on StockTwits as a platform for investors. (Sprenger et al., 2013) In his paper he tries to determine relations between message volume containing $S&P 100, sentiment of these messages and development of the index, coming to conclusion that it looks that online investor have matured during last ten years since the internet became globally available. The ratio between sell and buy signals is more stable and balance and also traders seem to be more stable and to don't follow naive strategies based on current trends, but even seem to recommend contrarian trading positions. Also the quality of the post seems to be more important than the number of post, as the sentiment is strongly related to stock returns then the message volumes are. Moreover for the importance of quality of posts speaks also the fact that users providing investment advice of higher quality tend to have much more followers and to have higher levels of retweets. Authors eventually conclude that stock microblogs do contain valuable information, which are yet not fully incorporated into trading strategies. Based on thesis information various indicators of future market development can be derived. This applies mainly for highly traded and trendy stocks, as those are one most heavily tweeted about. (Sprenger et al., 2013)

Another and probably one of the latest work on this topic is The Viability of StockTwits and Google Trends to Predict the Stock Market focused on determining the viability of Google Trends and Stocktwits sentiment for predicting stock returns. (Loughlin and Harnish, 2013) In this work I would like to continue and try to improve it. As Chris Loughlin and Erik Harnish have focused in their work only on IT sector, specifically on the biggest companies: Apple, Google, Facebook and Microsoft. Their concussion was that:

> *From our analysis, Google Trends data was not significant in predicting stock returns. But StockTwits data was significant in predicting Apple, Google, and Microsoft stock returns. When the data was lagged, the Bear and Bull indices were significant in predicting Apple and Microsoft stock returns. Because this data was lagged, StockTwits data is significant as a leading predictor of stock returns.* (Loughlin and Harnish, 2013, page: 17)

Other papers, even though not directly related to topic of my thesis, but never less very interesting are various works focused on analyses of Google Trends data for different

goals and different areas. Probably the most known are works of Seth Stephens-Davidowitz which I will mention more, when discussing Google Trends data and their possible applications.

As I have mentioned before, in my thesis I would like to revisit the work of Chris Loughlin and Erik Harnish, The Viability of StockTwits and Google Trends and look at it in a more complex way. Not only to measure the effect on returns of the stock, but also examine the viability of StockTwits and especially Google Trends data for determining future trade volume of particular stocks, where I would expect a strong correlation. Besides that I would like to look on other industries then just IT, even though I expect the correlations to be the strongest in the IT branch as investors focusing on IT are the ones most active on internet.

I will focus on information contained in the sentiment and try to prove a viability of sentiment on StockTwits, a derivation of twitter, focused directly to stock markets and Google Trends to predict future development of stock prices and trade volume. Since I believe that in the sentiment on social networks and sentiment expressed in volume of searches for particular expression is an information which is not yet fully incorporated in price development. The reason for this, I believe, is that investors creating the sentiments with their posts and search records are small investors, which have not enough power to influence the price on market, not even aggregated. Beside that small investors are also the ones who have the highest intention to react based on their sentiment. At least that is the conclusion of paper All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. (Barber and Odean, 2007) Even though they speak about the effect of attention and news on buying behavior, not the sentiment itself, I believe that as the sentiment is based mainly on these news, the effect would be very similar. In their work they are testing for attention-driven buying by matching stocks and events that are mostly likely to influence investors' mood or to ignite his attention. Matching is focused on abnormal trading volume, since extremely high trade volume must be attracting investors' attention, extreme one-day returns, as both positive and negative are likely to be coincided with events causing higher attention and also should be matched with periods with higher presence of particular firm in the news. Consistently with authors' expectations, individual investors show attention-driven behavior on markets. They are buying on high-volume days, follow both extremely negative and positive trends in one-day returns are

strongly influenced when the company is written about in news. As for institutional investors, they do not display these signs of attention-driven behavior. (Barber and Odean, 2007)

This results are applicable also in our case and also in order with our assumption that the sentiment on Google Trends and StockTwits is mainly created by small investors. This fact together with the lack of market power of small investors, large quantity of small investors with implication to Law of Large Numbers, which theoretically eliminates all bad judgments and gives somewhat of an average opinion aggregating the general idea of small investors about future development of stock market, providing so another channel of information which is not fully exploited.

For the purpose of my work I picked one whole market index and 18 different companies on which I will try to prove viability of sentiment included in StockTwits post and the sentiment of expressed in reach volume for particular stock on Google Trend. The work will be focused directly on US markets as I believe US market is one of the most advanced and progressive with most reliable data concerning Google Trends and for United States being country of residence for most of the users of StockTwits. I will try to cover more of the industrial sectors not just IT sector. Namely I have chosen Standard & Poor´s 500 as an index unifying 500 biggest publicly traded companies in US. Then for IT sector I have chosen Apple Inc., Facebook, Google Inc., Microsoft Corporation, Yahoo, Twitter and Blackberry as leaders of their markets and because they are strongly connected with internet by the sole definition of their business and that's why I expect the strongest correlation there. Outside of an IT sector I have chosen two of the firms leading current energy revolution: Tesla and Plug power, then I continued with General Motors as the biggest automotive company publicly traded in US, General Electric one of the most over-reaching corporations, Coca-Cola from foods industry, Delta Air Lines for transportation, Exxon for the oil and raw resources industry, Goldman Sachs for financial services, the Michael Kors Holdings Ltd. one of the most successful fashion companies of last years, Procter & Gamble, the biggest consumers goods company in US and Starbucks, nowadays one of the most successful and progressive representatives of franchise model. As mentioned above, I expect the strongest correlations between market development and sentiment on StockTwits or Google Trends at firms from IT sector, as for others I have mostly negative expectations because message volume concerning these companies on StockTwits is substantially lower, then in the case

of IT sector. But even here I would expect at least correlations between trade volume and Google Trends.

Due to the fact that most of the data will be in form of times series, with a strong suspicion for autocorrelation I will have to use Autoregressive–moving-average model (ARMA model) in my thesis to eliminate this effect of autocorrelation. Once cleared I expect the model to prove that influence of Google Trends and sentiment on StockTwits is statistically significant.

1.3. Data

The main sources of data for my thesis will be Google Trends, StockTwits and Yahoo! Finance. In this chapter I will introduce each of them and say a little about how I will approach to collecting and analyzing data from them. Besides that I would like to thank once more all mentioned companies for providing all the data.

1.3.1. Google trends

Google Trends Interface

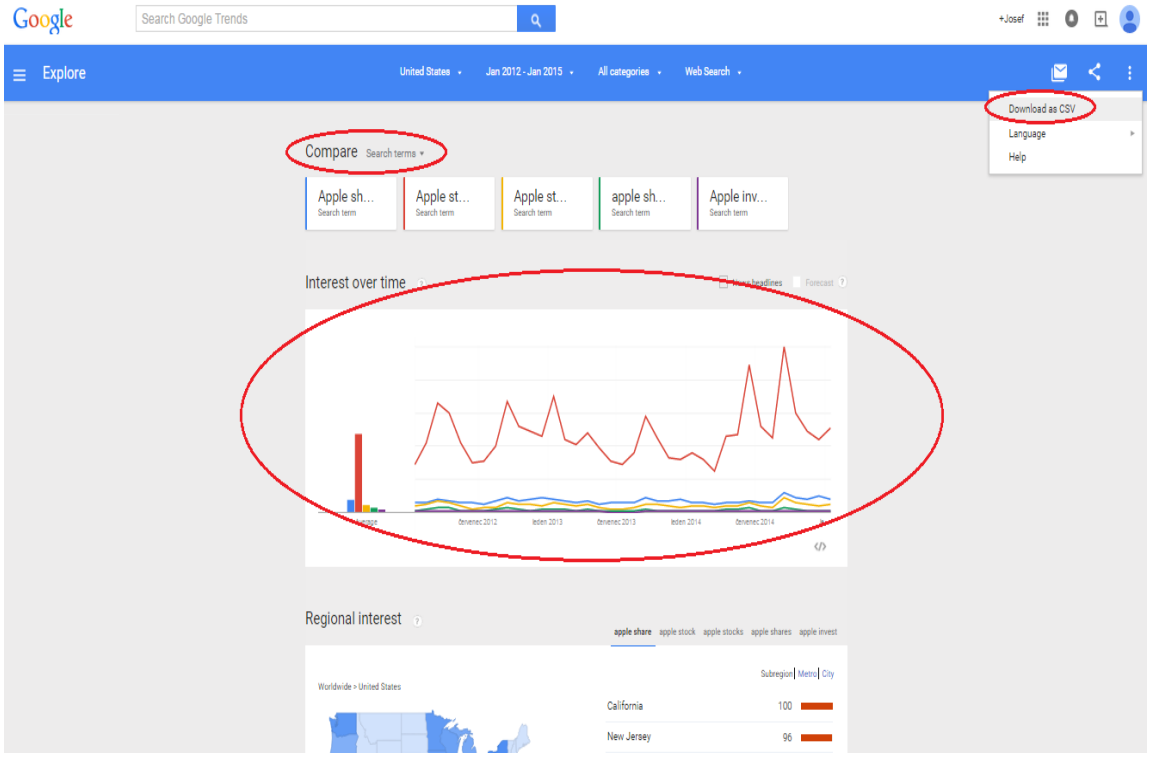

Fig. 1 Google Tends Interface (Google Trends, 2015)

In the Fig. 1 we can see example of Google Trends interface. Furthermore I have highlighted most important features as compare terms (up to five), graph expressing Google Trends and the most important: Download as CSV file, enabling to export data in to excel and their deeper analysis.

One of the main sources of data in my work will be Google Trends. Google Trends is a public web service of Google Inc. A history of Google Trends goes back to year 2006 when the first version of Google Trends was introduced in May with data going as far as to 01.01.2004. In year 2008, an extension of Google Trends called Insights for Search, was introduced allowing more detailed analysis of searches. In 2012 Google Trends and Insights fore Search were merged together to create Google Trends service, as we know it today. In last years, there were multiple attempts to determine possible applicability of data contained in Google Trends. For the example one of the most famous

and interesting papers on this topic is The Cost of Racial Animus on a Black Candidate: Evidence Using Google Data, in which author comes to conclusion that Obama lost approximately 4% of voles in 2008 election for the sole reason of being black. (Stephens-Davidowitz, 2014).

Seth Stephens-Davidowitz is one of the pioneers of usage of Google Trends data in this way, besides this mentioned paper, he is also author of many New York Times Columns, mainly focusing on popular topics as “What Do Pregnant Women Want,” The New York Times, Sunday Review, 5/17/2014, “Tell Me, Google. Is My Son a Genius,” The New York Times, Sunday Review, 1/18/2014 or “For the N.B.A., Zip Code Matters," The New York Times, Sunday Review, 11/3/2013 in which he makes conclusion based on analyses of these Google Trends data.

What are Google Trends in a technical view?

Google Trends are basically analyses of Google searches for a particular expression in given time period. Sadly Google does not provide data on daily basis, instead it is measuring on weekly basis. Also it does not provide us with an absolute value of searches executed, instead it determines a week with a normalized (Number of searches executed for one particular term in one area (for example US) is divided by total number of searches executed in that area. This ensure us that the trends for different areas are comparable no matter what population they contain. Maximal value of total reaches in given time period and then marks it as 100. This week is then used as a base and every other week is being compared to this one. Giving us an output in which every week, in a given period of time, is given a value from interval of 1 to 100, 100 marking a week with the highest number of searches and 1 being only 1 percent point of this maximum. Another important thing which Google managed to do, is that these data are cleared for duplicate searches, meaning that trends eliminate repeated searches from the same user which have been executed over a short period of time, leaving the final data more valid in view of general population.

For example let us see Google Trends for Apple Inc.

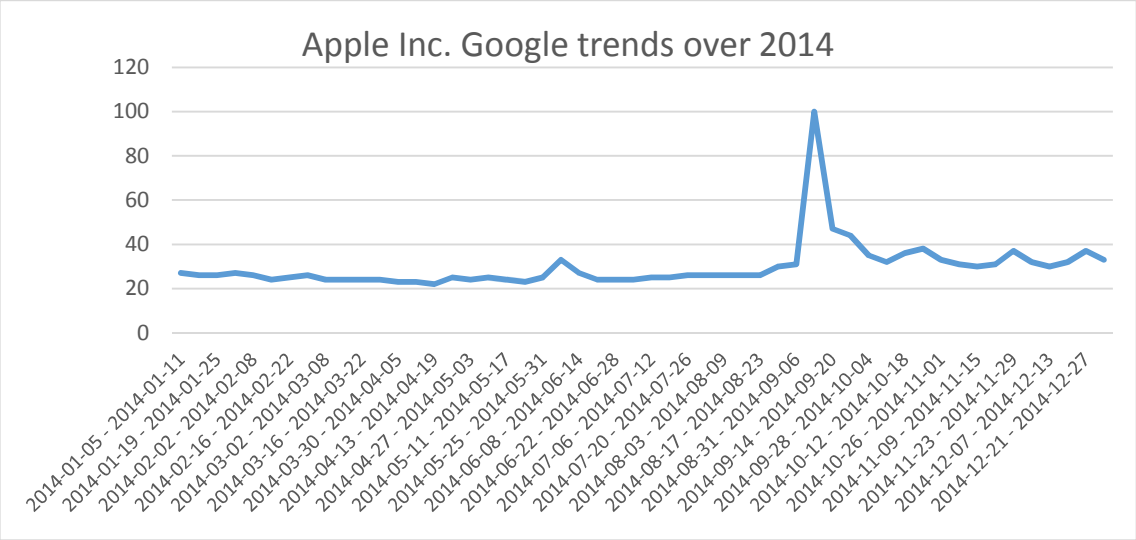

Fig. 2. Google trend for Apple Inc. for year 2014 (Data Source: Google Trends, 2015)
"Data Source: Google Trends (www.google.com/trends) "

Google trend in this case seems stable over the year with peak reaching its maximum in half of the September when new generation of iPhone was introduced.

How to type search item.

When working with Google trends, it is important to determine and then follow a way how we will type our search term. Google trends is using variety of operators to determine what exactly we want to search for. Using these different operators will influence the results of our Google Trends. In the following chart I will describe which operators are being used and how to work with them.

Search items (operator)	**What result you will see (function of operator)**
Apple shares	Results will include all searches including both terms apple and shares in any order. Result may also include other related searches like “sell apple shares”.
“Apple shares”	Results will only contain exact searches as included in quotation marks, the order of terms is also important.
Apple + shares	Results include searches containing words Apple OR shares
Apple - shares	Results will include searches containing word Apple but will exclude all searches including word shares.

Table 1: Search terms and operators functions in Google Trends. (Support.google.com, 2015)
Adjusted from https://support.google.com/trends/answer

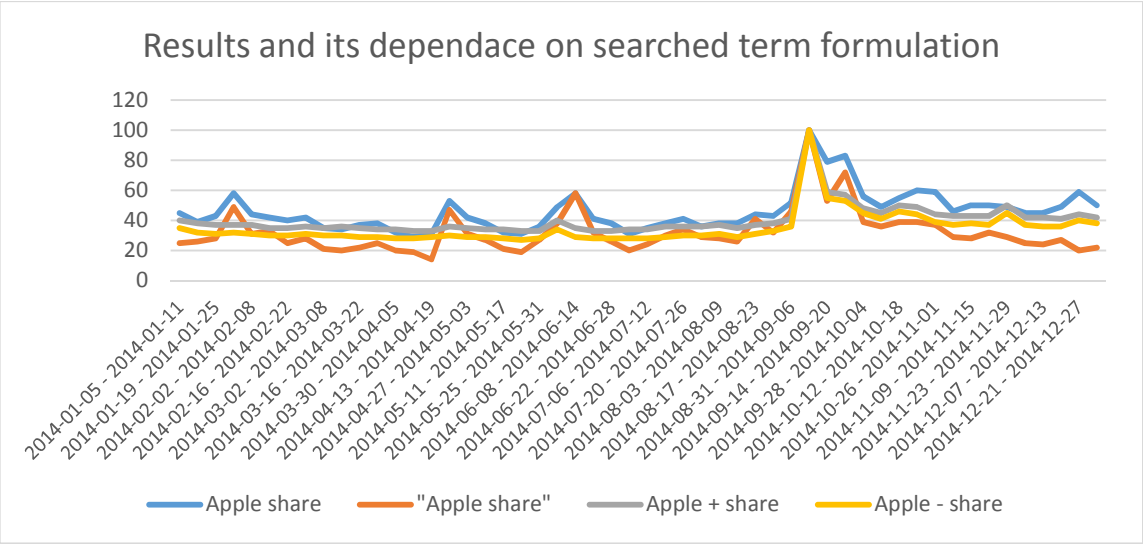

Fig. 3 Effect of different operators on Google Trends outcome. (Data Source: Google Trends, 2015)

Note that the Google trends data are normalized, so the peak for different operators is the same point. Besides that it clearly visible that Apple share (blue) has the highest volume as it covers more possibilities than other ones. Then “Apple share” which is the most exact expression has generally the lowest volume of searches, with peaks connected to important events.

Also no misspellings, synonyms, plural or singular of our search term are included. This represent a slight problem for us since even though we expect an educated user, we need to take in account possible misspelling, typos and to adjust our trends also for synonyms and plural possibilities. For this reason I will use a combination of inputs to cover synonyms and both plural and singular terms. For example in case of Apple Inc. it would be: apple share + apple shares + apple stock + invest apple + appl stock market acronym for Apple Inc.

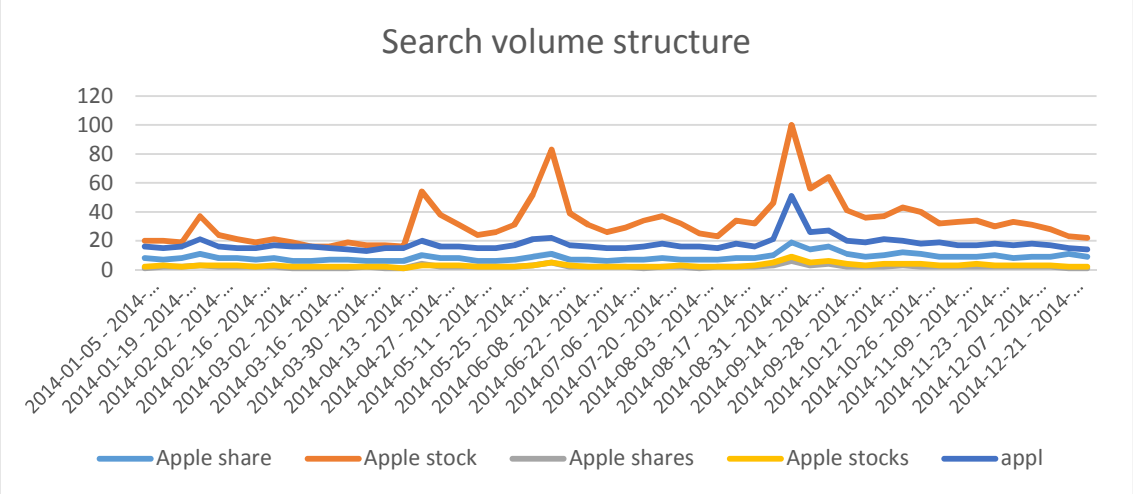

Fig. 4 Distribution on search volume between individual search terms of apple share + apple shares + apple stock + invest apple. (Data Source: Google Trends, 2015)
Data Source: Google Trends (www.google.com/trends).

As we can see most of the combined search volume of Apple share + Apple shares + Apple stock + invest Apple is contributed by the term Apple stock, second is Apple share, together they are counting for almost 98% of total search volume. Apple shares and apple invest hardly even visible in graph with Apple shares making approximately 2% and invest Apple even less than 1‰ of total search volume.

1.3.2. StockTwits

StockTwits Interface

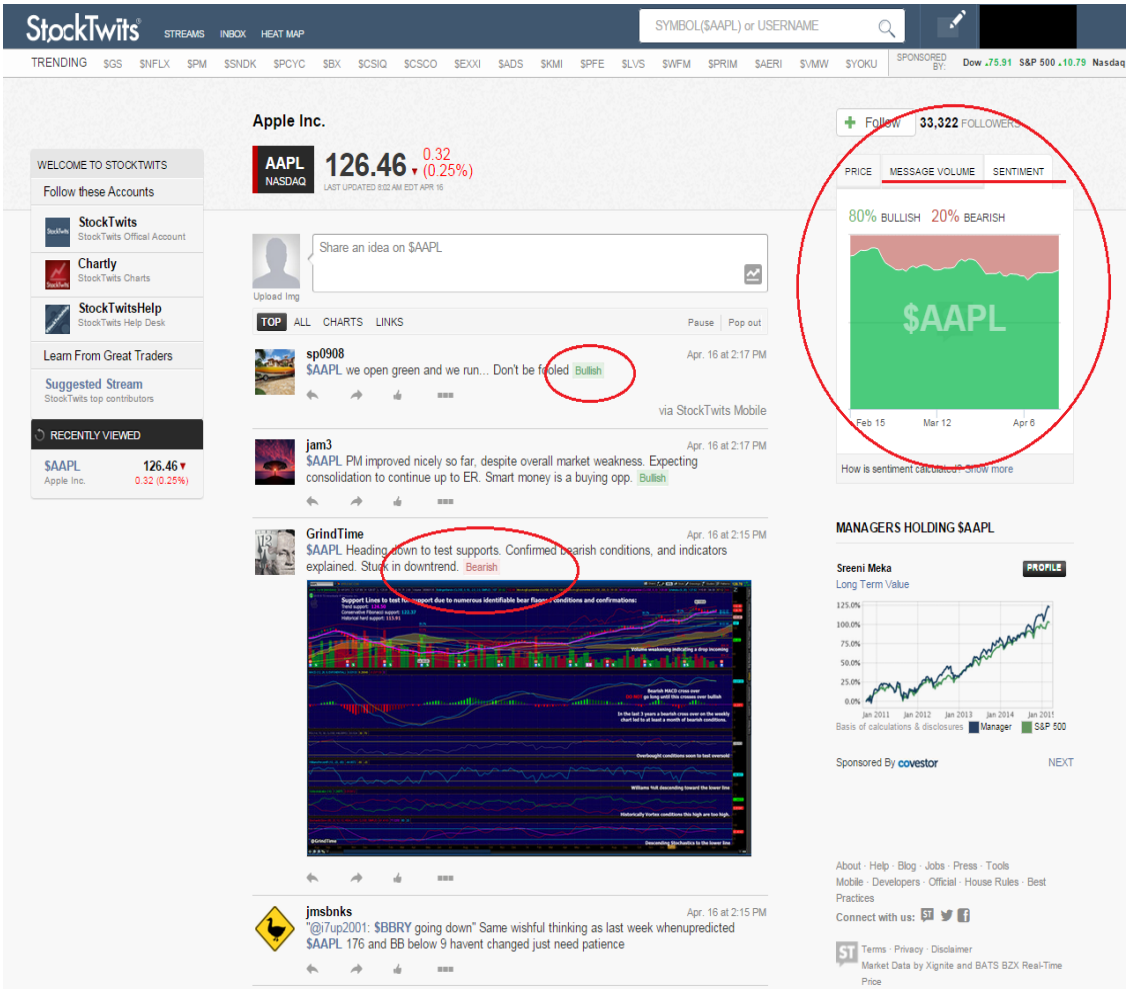

Fig. 5 StockTwits Interface (StockTwits, 2015)
Data Source: StockTwits, 2015, April 16th 8:00 am. (http://stocktwits.com/symbol/AAPL?q=AAPL)

Basic user interface on StockTwits.com. In the middle there is tweet channel. As u can see users are able to write short message, add a graph or picture explaining their opinion and most importantly for my work, accompany it with a tag expressing their sentiment as either “Bullish” or “Bearish”. On the right part of interface, user can choose see a graphs for ether, price development, message volume, and sentiment.

StockTwits Is a communications platform focused on financial markets and will be the source for my sentiment analysis data. StockTwits as a company was founded in 2008

and with creation of the $TICKER allowed organize a "Stream" of information concerning financial markets across web and social networks. This created a brand new form of insight and information source to be used by potential investors. It quickly became a center of many investors and as for today more then 300 000 investors, market professionals and public companies share their information and ideas about stock markets on StockTwits, creating so unprecedented source of information, not copying just one or few investors opinions but allowing to see a sentiment on financial markets as an aggregate mood of all subjects operating on market. This gives StockTwits an ability to become one most important supporting channels of information for investment decision.

"StockTwits streams consist of ideas, links, charts and other important financial data, summarized within 140 character messages. Users, which include analysts, media and investors of all types, as well as the public companies themselves, contribute to the stream. Investors, and others interested in stocks and markets, can easily follow individual stocks, specific contributors, as well as view the StockTwits stream across dozens of financial sites that integrate the stream including Yahoo! Finance, CNN Money, Reuters, TheStreet.com, Bing.com and The Globe and Mail." (StockTwits, 2015)

In my thesis I will focus on two of the tools available on StockTwits. First one is message volume, measuring the number of message on StockTwits stream for particular stock. For example in case of Apple the volume of messages on StockTwits stream looks like this:

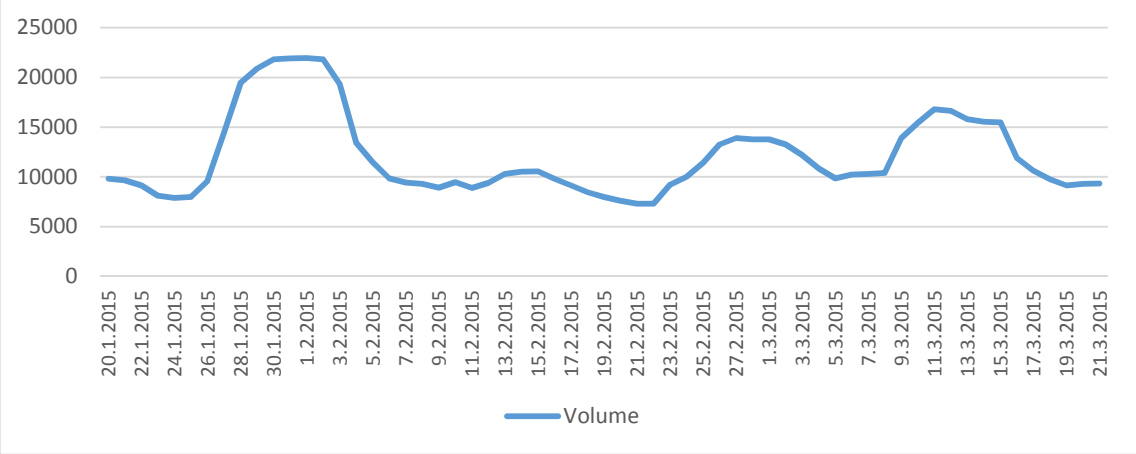

Fig. 6: Message volume development (From 20th of January till 21th of March) (StockTwits, 2015)

Data Source: StockTwits, 2015, April 16th 8:00 am. (http://stocktwits.com/symbol/AAPL?q=AAPL)
On the graph we can see the message volume for first three month of year 2015. Peaking at the end of January when the rumors around iWatch were the loudest.

Second tool is a sentiment. This works in a way that users can add to their post an emoticon or sticker expressing their mood about the current situation of particular stock on market. There are two options either Bullish, meaning that user is in believe that stock price will rise or Bearish meaning that user bets on decrease in stock price. StockTwits then aggregates these data into benchmark determining aggregate mood of users concerning that particular stock.

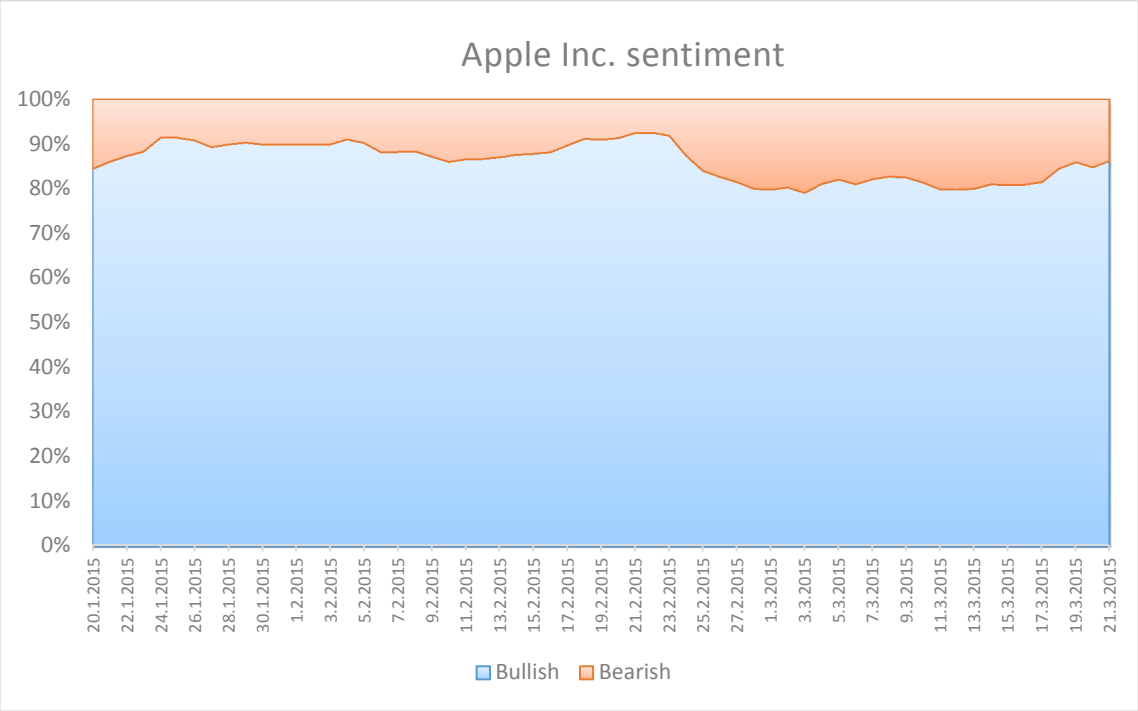

Fig. 7: Sentiment on StockTwits development (From 20th of January till 21th of March) (StockTwits, 2015) Data Source: StockTwits, 2015, April 16th 8:00 am. (http://stocktwits.com/symbol/AAPL?q=AAPL)

As we can see the sentiment on StockTwits concerning Apple stock was significantly bullish over first three months of year 2015. Firstly in a zone from 85% to 90% then with drop to levels around 80%, which was probably cause by skepticism about forthcoming introduction of iWatch.

The problem I encountered when dealing with StockTwits is that they do not provide data for free. Only form of public data available are graphs for last three months concerning message volume and market sentiment. Luckily after contacting them, they were willing to provide me with historical data. The issue was that as they do not process these data, they were only able to provide data in forms of log files in JavaScript from their website. Eventually with the size of dataset I came to need to somehow automatize the process of data mining from these logs. For this purpose, with big help of my friend, we wrote a C++ script for processing these logs into .csv files compatible with excel. Thanks to this script I was able to transpose GBs of data from .json file into excel spread sheets

and analyze them further more. All together for the year 2014 for which I was provided data I processed almost 11 million of message posts from which slightly less than two millions were directly linked to particular company and contained some form of sentiment and were used for further analysis, the rest was eliminated from further analyses, because even though they may also contain valuable information, they were not important for my work.

1.3.3. Yahoo finance

The source of my data concerning the stock markets development and prices will be Yahoo! Finance, part of the Yahoo network. It provides financial news, reports, press releases and most importantly financial data available to download in form of excel spreadsheet. The output in spread sheet looks like this:

Date	Open	High	Low	Close	Volume	Adj Close
20.3.2015	2 090	2 114	2 090	2 108	5 554 120 000	2 108
19.3.2015	2 099	2 099	2 086	2 089	3 305 220 000	2 089

Table 2: Example of data table from Yahoo! Finance. (Data Source: Yahoo! Finance, 2015)

Yahoo! Finance provides data in form of csv file e.g. the data are downloaded in form of text string. Then I wrote macro in VBA to process these string in to table as you can see in example.

For my thesis I will use trade volume and adjusted closing price. It represents price at the end of daily trading adjusted for dividends and share splits. The data on Yahoo! Finance go as far as back to 1950, depending of course on each specific company. This makes Yahoo! Finance one of the largest and most robust financial market databases and a perfect source of data for my work.

1.3.4. Google Trends and Yahoo finance data.

When testing viability of Google Trends data I came across a problem that Google Trends provides us with weekly data, e.g. it show us an index for whole week not for each day individually, whereas Yahoo! Finance provides data for each day trading floors were open. As for this I need to adjust these data to make them comparable. There were two ways to do so. First option was to take the value of index for whole week and assign it to every day in the week. For example a Google Trends week starting on the 6th of January and ending on the 13th of January would have the same Google Trends value of 43 for all

trading days from 7th till 12th of January. Second option was to stay on weekly basis and create for each week average price of stock and average trade volume, meaning that for Google Trends week starting on the 6th of January and ending on the 13th of January sum adjusted closing prices and trade volumes for each trading day and then divide the sum with a number of trading days in particular week.

In the end I have chosen the second approach because I think that having the same value for multiple prices would collide in the model. In the Excel I developed a formula to transfer and adjust the stock price to week average. The formula had to be also able to take in to account cases when there were less than five trading days in a week. The same had to be done also in case of trade volume, e.g. to determine the average trade volume in a week given by Google Trends.

It will be shown on the example of S&P 500.

1A	B	C	D	E	F	G	H	I	J	K
2	S&P 500									
3	Google Trends			Yahoo! Finance			Calculation			
4	Start_of_period	End_of:Period	Google trend	Date	Volume	Adj Close Price	GT_Week	Week	Average Trade Volume	Average week price
5	6. leden 2008	12. leden 2008	46	7. leden 2008	4 221 260 000	1 416	1	1	4 788 802 000	1 407
6	13. leden 2008	19. leden 2008	58	8. leden 2008	4 705 390 000	1 390	2	1	5 006 464 000	1 366
7				9. leden 2008	5 351 030 000	1 409		1		
8				10. leden 2008	5 170 490 000	1 420		1		
9				11. leden 2008	4 495 840 000	1 401		1		
10				14. leden 2008	3 682 090 000	1 416		2		
11				15. leden 2008	4 601 640 000	1 381		2		
12				16. leden 2008	5 440 620 000	1 373		2		
13				17. leden 2008	5 303 130 000	1 333		2		
14				18. leden 2008	6 004 840 000	1 325		2		

Table 3: Example of processing Yahoo! Finance data in to weekly format as determined by Google Trends. (Data Sources: Google Trends, 2015, Yahoo! Finance, 2015)

G_T week sets us a number of Google trends week
Week sets in which Google Trends week a day of trading belongs.
Funtion: Week=*SLOOKUP(Date;$A$5:$G$6;7;TRUE)*
Average_Trade_Volume give us an average trade volume in GT_Week.
Function: Average_Trade_Volume = *SUMPRODUCT (($H$5:$H$14=G5)*($E$5:$E$14))/ COUNTIF ($H$5:$H$14;G5)*
Average_week_price gives and average stock price in GT_Week.
Function: Average_week_price = *SUMPRODUCT(($H$5:$H$14=G5)*($F$5:$F$14)) /COUNTIF($H$5:$H$14;G5)*

In this way we get weekly data for each stock, for weeks as set by Google Trends, e.g. starting on Sunday and ending on Saturday in a form that allows us to build models based on percentage change between weeks. I believe that using percentage changes instead of nominal values will be more precise and clear as Google Trend, Stock Prices and Trade volume are each in a different numerical order.

As for the time period, have chosen for my work is from year 2012, when the current version of Google Trends was started, to April 2015. It provides a time period of three year and three months, which should give testing sample big enough to test my hypotheses.

2. Practical part - Econometrical testing

2.1 Methodology

My dataset will consist of five variables, trend from Google Trend, adjusted closing price and trade volume from Yahoo! Finance, sentiment (expressed by bullish index) and message volume from StockTwits. Data will be used in form of percentage changes to eliminate differences in numerical orders, which are in some cases truly significant.

For Google Trends I will test for correlation between percentage changes in Google Trends and trade volume, Google Trends and price development and finally between absolute values of percentage changes in Google Trends and stock prices representing market volatility using Least Squares method. I will always show the equation only for the first category, for the following ones are same only with appropriately changed variables.

$$\text{ATV\_\%}\Delta_t = \beta_0 + \beta\text{TREND\_\%}\Delta_t + \varepsilon_t$$

If the effect of Google Trends for some stock proves to be statistically significant next step will be ARIMA models to soften the effect of autocorrelation in price development.

In the same way I will proceed with StockTwits data which I will test in four categories. Correlation between Bullish Index (ration of messages with bullish sentiment on total number of messages with sentiment stamp) and trade volume, bullish index and price, messages volume (number of messages concerning particular stock containing any kind of sentiment) and trade volume, and lastly message volume and price. For these ho will prove significant effect I will continue with ARIMA models and then with determining Granger Causality.

2.2 Google trends

2.2.1. Least Squares Method

Google Trends and trade volume (Least Squares method)

Dependent variable	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Google_ATV_%Δ	Google_TREND_%Δ	1,29458	0,47102	2,74845	**0,00860**	0,14374
Apple_ATV_%Δ	Apple_TREND_%Δ	0,56874	0,05197	10,94447	**0,00000**	0,41768
Tesla_ATV_%Δ	Tesla_TREND_%Δ	0,64102	0,08637	7,42165	**0,00000**	0,32967
Microsoft_ATV_%Δ	Microsoft_TREND_%Δ	0,60200	0,09262	6,49977	**0,00000**	0,20190
Coca-Cola_ATV_%Δ	Coca-Cola_TREND_%Δ	0,42187	0,10476	4,02695	**0,00010**	0,08851
Michael _Kors_ATV_%Δ	Michael _Kors_TREND_%Δ	0,11346	0,23954	0,47365	0,63710	0,00287
Facebook_ATV_%Δ	Facebook_TREND_%Δ	1,48611	0,18071	8,22378	**0,00000**	0,31364
General_Motors_ATV_%Δ	General_Motors_TREND_%Δ	0,21987	0,11802	1,86310	**0,06420**	0,02048
General_Electricts_ATV_%Δ	General_Electricts_TREND_%Δ	0,09716	0,08961	1,08428	0,27980	0,00699
Procter&Gamble_ATV_%Δ	Procter&Gamble_TREND_%Δ	0,16660	0,08670	1,92149	**0,05720**	0,03191
Starbucks_ATV_%Δ	Starbucks_TREND_%Δ	0,72762	0,08334	8,73061	**0,00000**	0,31339
SP_500 _ATV_%Δ	SP_500 _TREND_%Δ	0,21581	0,05169	4,17480	**0,00000**	0,09450
Delta_Airlines_ATV_%Δ	Delta_Airlines_TREND_%Δ	0,47650	0,14743	3,23210	**0,00150**	0,05887
Exxon_ATV_%Δ	Exxon_TREND_%Δ	0,35640	0,08283	4,30291	**0,00000**	0,09980
Goldman Sachs_ATV_%Δ	Goldman Sachs_TREND_%Δ	0,15492	0,07683	2,01644	**0,04540**	0,02377
Backberry_ATV_%Δ	Backberry_TREND_%Δ	0,80599	0,09416	8,56023	**0,00000**	0,30371
Yahoo_ATV_%Δ	Yahoo_TREND_%Δ	1,39381	0,22517	6,19005	**0,00000**	0,18572
Plug_Power_ATV_%Δ	Plug_Power_TREND_%Δ	0,11703	5,49825	0,02129	0,98300	0,00000
Twitter_ATV_%Δ	Twitter_TREND_%Δ	-0,05481	0,04794	-1,14331	0,25600	0,01448

Table 4: Results of Leas Squares Method for correlation between weekly ATV_%Δ and TREND_%Δ.

In the table we can see results of testing for correlation between Google Trends and trade volume. As mentioned in part dedicated to Data processing, for the reason that Google Trends provide data only on weekly basis, trade volume is taken as average for period given by Google Tends Week. The results clearly shows that the correlation is strongest in case of firms operating in IT business. Companies such as Google, Apple Inc., Microsoft, Facebook, Blackberry, Yahoo, but surprisingly not Twitter. Besides that we can see correlation also in case of companies which we could call trendy as Tesla, now one of the hot calls of investors, or Starbucks. The correlation is also clear for index S&P 500. and big companies in a position of leaders on their markets as Goldman Sachs, General Motors, Exxon, Delta Air Lines and Procter & Gamble, which generally has ones of the highest trade volumes. Furthermore I need to mention that the results for Plug Power are highly distorted because of the fact that as one of the search terms used for Google Trends analysis is firm's stock marker acronym, which for Plug Power is simply

plug, which has also other meanings and these other meaning clearly overweight the volume of searches for plug as a firm's symbol on stock market. Google Trend for Plug Power eventually looks like this:

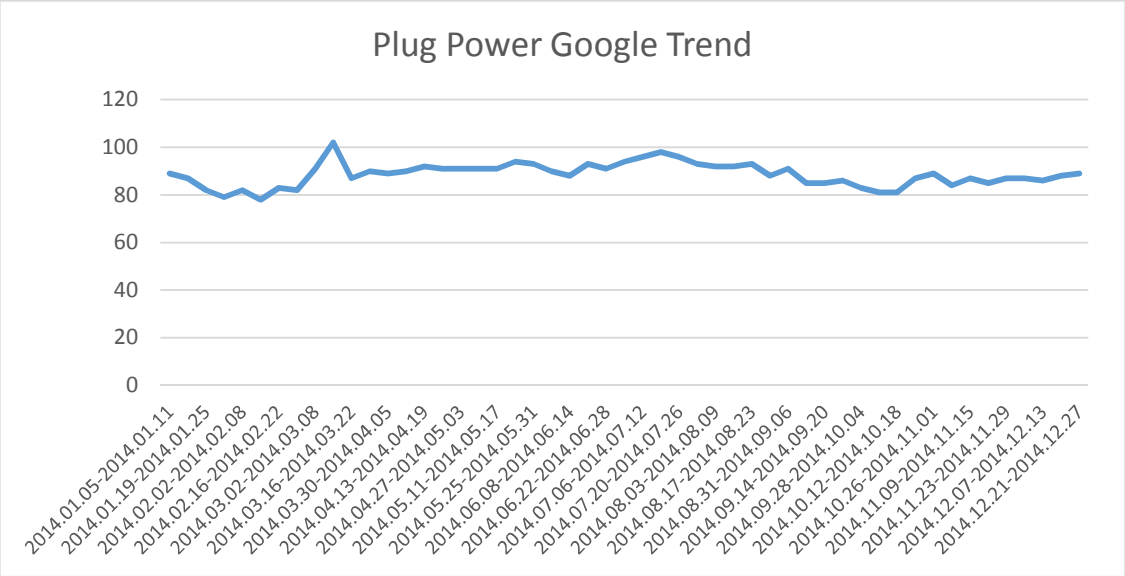

Fig. 8: Google Trends for terms Plug Power stock + Plug Power share + Plug Power shares + Plug Power stocks + plug. (Data Source: Google Trends, 2015)
Data Source: Google Trends (www.google.com/trends).

The volume of searches remains almost constant over three years and three months period and possible information hidden in search volume is hidden in more general searches for the term plug. This proves my theory that data considering Google Trends volume are distorted by the fact that one of the search terms was general word plug.

As for other companies in which cases results were no significant, like Michael Kors, General Electric and Twitter the explanation is not so clear. For Michael Kors could hold argument that it is small emission in a very specific industry so it does not attract so much attention from small investors. Nevertheless for General Electric this argument fades and so does for Twitter.

Google Trends and stock price (Least Squares method)

Dependent variable	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Google_AWP_%Δ	Google_TREND_%Δ	-0,05239	0,03533	-1,48278	0,14510	0,04658
Apple_AWP_%Δ	Apple_TREND_%Δ	-0,00171	0,00675	-0,25366	0,80010	0,00039
Tesla_AWP_%Δ	Tesla_TREND_%Δ	0,03260	0,01145	2,84724	**0,00520**	0,06750
Microsoft_AWP_%Δ	Microsoft_TREND_%Δ	0,00322	0,00677	0,47520	0,63530	0,00135
Coca-Cola_AWP_%Δ	Coca-Cola_TREND_%Δ	-0,00685	0,00530	-1,29371	0,19760	0,00992
Michael Kors_AWP_%Δ	Michael Kors_TREND_%Δ	-0,00660	0,00971	-0,67913	0,49910	0,00588
Facebook_AWP_%Δ	Facebook_TREND_%Δ	0,06614	0,02041	3,24026	**0,00150**	0,06624
General Motors_AWP_%Δ	General Motors_TREND_%Δ	0,01607	0,00760	2,11274	**0,03610**	0,02619
General Electricts_AWP_%Δ	General Electricts_TREND_%Δ	0,00594	0,00434	1,36898	0,17280	0,01110
Procter&Gamble_AWP_%Δ	Procter&Gamble_TREND_%Δ	-0,00075	0,00491	-0,15306	0,87860	0,00021
Starbucks_AWP_%Δ	Starbucks_TREND_%Δ	0,00916	0,00653	1,40300	0,16250	0,01165
SP_500 _AWP_%Δ	SP_500 _TREND_%Δ	-0,00656	0,00454	-1,44709	0,14970	0,01238
Delta Airlines_AWP_%Δ	Delta Airlines_TREND_%Δ	0,03828	0,01123	3,40767	**0,00080**	0,08784
Exxon_AWP_%Δ	Exxon_TREND_%Δ	0,35640	0,08283	4,30291	**0,00000**	0,09980
Goldman Sachs_AWP_%Δ	Goldman Sachs_TREND_%Δ	-0,00066	0,00178	-0,37377	0,70900	0,00084
Backberry_AWP_%Δ	Backberry_TREND_%Δ	0,00799	0,01217	0,65644	0,51240	0,00256
Yahoo_AWP_%Δ	Yahoo_TREND_%Δ	0,01340	0,01562	0,85744	0,39240	0,00436
Plug_Power_AWP_%Δ	Plug_Power_TREND_%Δ	0,35984	0,31242	1,15177	0,25110	0,00783
Twitter_AWP_%Δ	Twitter_TREND_%Δ	-0,00128	0,00550	-0,23320	0,81610	0,00061

Table 5: Results of Leas Squares Method for correlation between weekly AWP_%Δ and TREND_%Δ.

Results for correlation between price and Google Trends are much less promising but also much more surprising. Contra dictionary to my expectations the correlation is not strongest in IT industry form which the effect of Google Trends was statistically significant only in case of Facebook. Besides this, results are significant in case of Tesla Motors, General Motors, Delta Air Lines and Exxon.

Also note that the coefficients are, when compared to coefficients for trade volume results, often very low. That is not because they would not have significant effect, but because the Google Trends are much more volatile than stock prices whereas the volatility of trade volume reaches even higher levels than the volatility of Google Trends.

I will show this relation graphically on case of Tesla Motors (fig. no.: 9). Where the average percentage change in Google Trends is 36% in trade volume 42% and in price it is only 5%. (Note that these values are changes for weekly data, e.g. average trade volume and average price).

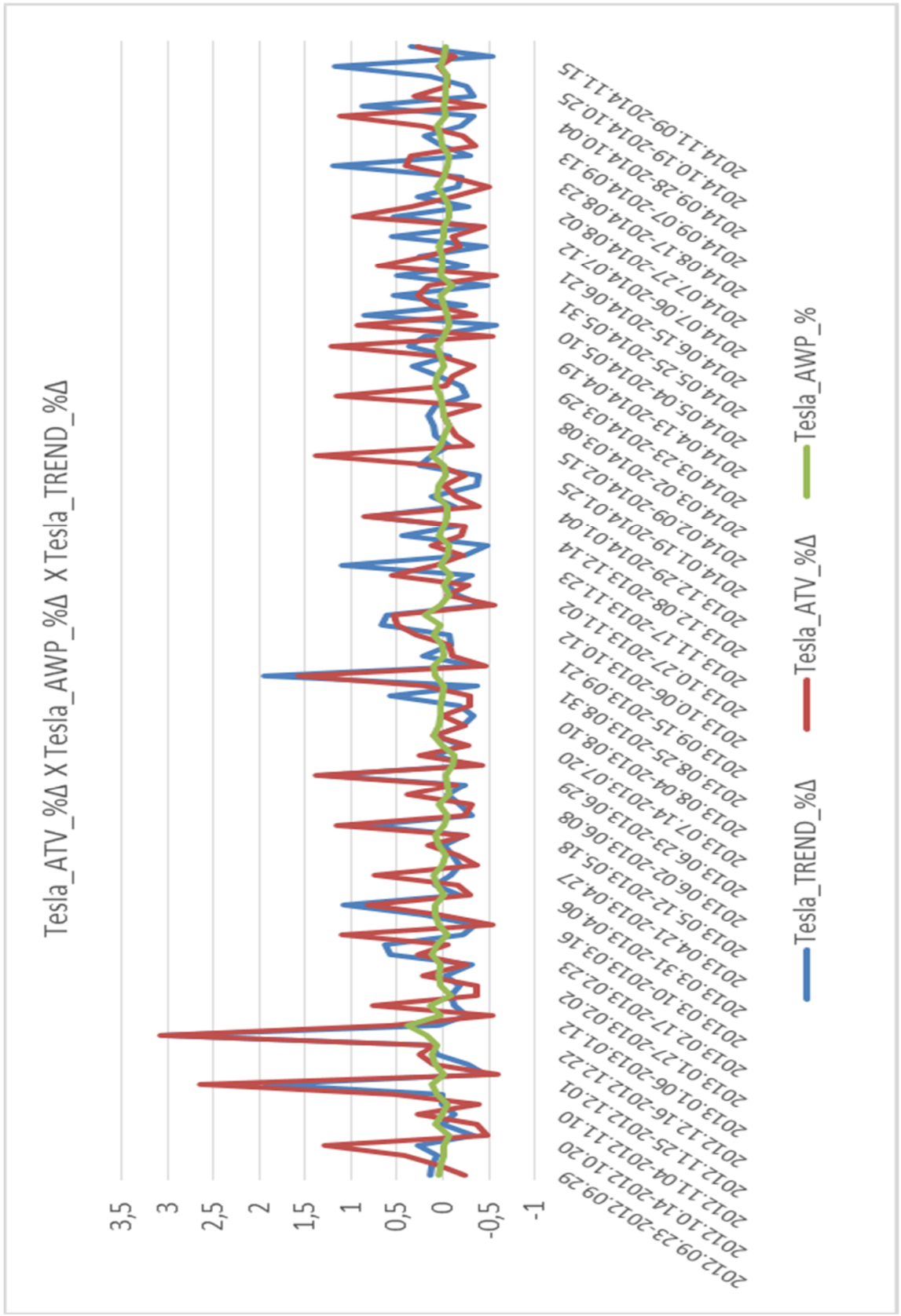

Fig. 9: Volatility of variables Tesla_ATV_%Δ X Tesla_AWP_%Δ X Tesla_TREND_%Δ

In the graph is clearly visible that the volatility of Tesla stock price is substantionally lower than volatility of trade volume and volatility of Google Trednds, whereas Google Trends and Trade volume shows approximately the same level of volatility.

Dependent variable	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Google_ABS_AWP_%Δ	Google_ABS_TREND_%Δ	0,02105	0,03212	0,65526	0,51560	0,00945
Apple_ABS_AWP_%Δ	Apple_ABS_TREND_%Δ	0,01155	0,00501	2,30487	**0,02240**	0,03083
Tesla_ABS_AWP_%Δ	Tesla_ABS_TREND_%Δ	0,01789	0,01077	1,66056	**0,09960**	0,02403
Microsoft_ABS_AWP_%Δ	Microsoft_ABS_TREND_%Δ	0,01144	0,00592	1,93093	**0,05520**	0,02184
Coca-Cola_ABS_AWP_%Δ	Coca-Cola_ABS_TREND_%Δ	0,00306	0,00506	0,60451	0,54630	0,00218
Michael Kors_ABS_AWP_%Δ	Michael Kors_ABS_TREND_%Δ	0,00476	0,00836	0,57009	0,57030	0,00415
Facebook_ABS_AWP_%Δ	Facebook_ABS_TREND_%Δ	0,05055	0,01973	2,56155	**0,01140**	0,04245
General Motors_ABS_AWP_%Δ	General Motors_ABS_TREND_%Δ	-0,00165	0,00623	-0,26540	0,79100	0,00042
General Electricts_ABS_AWP_%Δ	General Electricts_ABS_TREND_%Δ	0,00412	0,00435	0,94831	0,34430	0,00536
Procter&Gamble_ABS_AWP_%Δ	Procter&Gamble_ABS_TREND_%Δ	0,00510	0,00453	1,12705	0,26210	0,01121
Starbucks_ABS_AWP_%Δ	Starbucks_ABS_TREND_%Δ	-0,00215	0,00629	-0,34092	0,73360	0,00070
SP_500 _ABS_AWP_%Δ	SP_500 _ABS_TREND_%Δ	0,00462	0,00416	1,11056	0,26840	0,00733
Delta Airlines_ABS_AWP_%Δ	Delta Airlines_ABS_TREND_%Δ	0,03112	0,01202	2,58852	**0,01050**	0,03858
Exxon_ABS_AWP_%Δ	Exxon_ABS_TREND_%Δ	0,35640	0,08283	4,30291	**0,00000**	0,09980
Goldman Sachs_ABS_AWP_%Δ	Goldman Sachs_ABS_TREND_%Δ	0,00195	0,00146	1,33722	0,18300	0,00574
Backberry_ABS_AWP_%Δ	Backberry_ABS_TREND_%Δ	0,04364	0,00952	4,58461	**0,00000**	0,11120
Yahoo_ABS_AWP_%Δ	Yahoo_ABS_TREND_%Δ	-0,00924	0,01513	-0,61109	0,54200	0,00222
Plug_Power_ABS_AWP_%Δ	Plug_Power_ABS_TREND_%Δ	0,07754	0,39544	0,19608	0,84480	0,00023
Twitter_ABS_AWP_%Δ	Twitter_ABS_TREND_%Δ	-0,00304	0,00449	-0,67747	0,49990	0,00513

Table 6: Results of Leas Squares Method for correlation between weekly ABS_AWP_%Δ and ABS_TREND_%Δ.

As expected the results are very similar to previous results for the effect of changes in Google Trends on changes in stock prices. Absolute value of change in Google Trends has a statistically significant effect on volatility in case of Tesla Motors, Facebook, Delta Air Lines and Exxon, for which also the effect of Google Trends on stock price was significant. Besides these mentioned companies the effect of Google Trends is significant for Apple, Microsoft and Blackberry. That again, after surprising results in previous table for correlation between Google Trends and stock price, backs my expectation that the correlation would be strongest in IT industry.

The problem with using Least Squares method on these data is that there is strong autocorrelation expected in stock price development. To solve this problem. For all the stocks for which the effect of Google Trends proved to be significant for in Least Squares method I will do the testing again, now in model with ARIMA specification.

2.2.2. ARIMA models

To determine specification of ARIMA model I have followed Box-Jenkins methodology. Firstly I have ran Augmented Dickey-Fuller test to decline the hypothesis that the time series has a unit root, luckily with none of the time series had a unit root, which is most likely thanks to the fact that used variables in form of percentage changes.. Next step was to according correlogram of dependent variable and its autocorrelation and partial autocorrelation function determine ar and ma levels of ARIMA model. Then if residua were in a zone of white noise I have determined according Akaike, Schwartz and Hanna-Quinn criteria's the model with highest quality. (Box and Jenkins, 1970)

I need to note that I have tried to determine one model that would fit all. This means that the model can be in some cases far from an ideal one. I have decided to go this way as the main goal of my work is to asses and look for possible information hidden in online activity, not to deeply analyze all individual stocks and the effort needed to asses each stock individually would highly overreach the intended scope of my thesis. In the end I have decided for ARIMA with specifications ar(1) ar(2) ma(1) ma(2).

$$ATV\_\%\Delta_t = \beta_0 + \beta \mathrm{TREND}\_\%\Delta_t + \theta_1 y_{i-1} + \theta_2 y_{t-2} + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} + \varepsilon_t$$

And that will be specifications of all ARIMA models in my work. In the results I will note only coefficient for examined variable for the result tables to be more synoptic.

Google Trends and trade volume (ARIMA)

Dependent variable	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Google_ATV_%Δ	Google_TREND_%Δ	1,46431	0,37511	3,90370	**0,00040**	0,55547
Apple_ATV_%Δ	Apple_TREND_%Δ	0,55049	0,04755	11,57705	**0,00000**	0,53827
Tesla_ATV_%Δ	Tesla_TREND_%Δ	0,71636	0,08183	8,75434	**0,00000**	0,52868
Microsoft_ATV_%Δ	Microsoft_TREND_%Δ	0,62176	0,09376	6,63131	**0,00000**	0,33141
Coca-Cola_ATV_%Δ	Coca-Cola_TREND_%Δ	1,46431	0,37511	3,90370	**0,00040**	0,55547
Facebook_ATV_%Δ	Facebook_TREND_%Δ	1,66175	0,16746	9,92318	**0,00000**	0,43877
General_Motors_ATV_%Δ	General_Motors_TREND_%Δ	0,21631	0,12151	1,78021	**0,07690**	0,24181
Procter&Gamble_ATV_%Δ	Procter&Gamble_TREND_%Δ	0,15530	0,07862	1,97526	**0,05080**	0,33366
Starbucks_ATV_%Δ	Starbucks_TREND_%Δ	0,76148	0,08749	8,70358	**0,00000**	0,42153
SP_500 _ATV_%Δ	SP_500 _TREND_%Δ	0,24800	0,05294	4,68477	**0,00000**	0,29653
Delta_Airlines_ATV_%Δ	Delta_Airlines_TREND_%Δ	0,52297	0,14449	3,61949	**0,00040**	0,22601
Exxon_ATV_%Δ	Exxon_TREND_%Δ	0,39976	0,08030	4,97861	**0,00000**	0,34326
Goldman Sachs_ATV_%Δ	Goldman Sachs_TREND_%Δ	0,08322	0,07698	1,08106	0,28130	0,18207
Backberry_ATV_%Δ	Backberry_TREND_%Δ	0,74391	0,09223	8,06624	**0,00000**	0,34186
Yahoo_ATV_%Δ	Yahoo_TREND_%Δ	1,73418	0,19370	8,95303	**0,00000**	0,42323

Table 7: Results of ARIMA (2, 2) method for the effect of TREND_%Δ on ATV_%Δ.

The significance remained as it was according Least Squares method and also the coefficient remained approximately the same. Whereas the R-Squared rose significantly as the effect of autocorrelation was softened at some cases even overreaching 50%.

Google Trends and stock price (ARIMA)

Dependent Variable:	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Tesla_AWP_%Δ	Tesla_TREND_%Δ	0,02488	0,00962	2,58585	**0,01110**	0,19422
Facebook_AWP_%Δ	Facebook_TREND_%Δ	0,07500	0,01891	3,96560	**0,00010**	0,19587
General_Motors_AWP_%Δ	General_Motors_TREND_%Δ	0,01460	0,00731	1,99730	**0,04750**	0,13376
Delta_Airlines_AWP_%Δ	Delta_Airlines_TREND_%Δ	0,03340	0,00977	3,42029	**0,00080**	0,13201
Exxon_AWP_%Δ	Exxon_TREND_%Δ	-0,00639	0,00539	-1,18579	0,23750	0,08941

Table 8: Results of ARIMA (2, 2) method for the effect of TREND_%Δ on AWP_%Δ.

Dependent Variable:	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Tesla_ABS_AWP_%Δ	Tesla_TREND_%Δ	0,04253	0,00601	7,07488	**0,00000**	0,13557
Facebook_ABS_AWP_%Δ	Facebook_TREND_%Δ	0,01893	0,00431	4,39580	**0,00000**	0,23205
General_Motors_ABS_AWP_%Δ	General_Motors_TREND_%Δ	0,00065	0,00647	0,09993	0,92050	0,04011
Delta_Airlines_ABS_AWP_%Δ	Delta_Airlines_TREND_%Δ	0,02861	0,00897	3,18851	**0,00170**	0,11079
Exxon_ABS_AWP_%Δ	Exxon_TREND_%Δ	0,01616	0,00505	3,20323	**0,00160**	0,12823
Apple_ABS_AWP_%Δ	Apple_TREND_%Δ	0,00861	0,00496	1,73585	**0,08450**	0,06847
Microsoft_ABS_AWP_%Δ	Microsoft_TREND_%Δ	0,01357	0,00596	2,27816	**0,02400**	0,07266

Table 9: Results of ARIMA (2, 2) method for the effect of ABS_TREND_%Δ on ABS_AWP_%Δ.

The same applies when testing for the ARIMA models for changes in stock price and also for absolute values of this changes. If the effect was significant according to Least Squares Method it remained significant also in ARIMA models with the exception of Exxon for percentage changes in price and General Motors for absolute values of percentage changes. Coefficients remained approximately the same and R-squared have risen significantly.

2.2.3. Granger causality

Last tests concerning Google Trends will be devoted to Granger Causality to determine if Google Trends are leading or lagging indicator of stock market development. The test will concern only those companies for which the change of Google Trends have proved to be statistically significant for changes of trade volume or its stock prices. When determining if Google Trends are leading indicator of stock market movements I will add lags in to the same ARIMA (2, 2) model used before.

$$\text{ATV\_\%}\Delta_t = \beta_0 + \beta_1 \text{TREND\_\%}\Delta_{t-1} + \beta_2 \text{TREND\_\%}\Delta_{t-2} + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2}$$

And for determining importance of Google Trends as lagging indicator I will use simple Least Squares method as the autocorrelation of Google Trends is no so strong as in a case of price development.

$$\text{TREND\_\%}\Delta_t = \beta_0 + \beta_1 \text{ATV\_\%}\Delta_{t-1} + \beta_2 \text{ATV\_\%}\Delta_{t-2} + \varepsilon_t$$

As for lagging I will test only for first two lags because model are based on weekly data. Also I have run the model with both lags, first and second I realize that there is a high possibility of autocorrelation, nevertheless when I back test model by creating a model for each lag separately the result were approximately the same at least for the statistical significance of lags, so again as the goal of my thesis is not deeper analysis of specific stocks but only to assess the information and determining if there is viable information hidden I believe that this way of testing will be sufficient enough.

Dependent Variable:	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Google_ATV_%Δ	Google_I,_LAG_TREND_%Δ	-1,37541	0,73493	-1,87148	**0,06250**	0,37771
	Google_II,_LAG_TREND_%Δ	-0,90923	0,73778	-1,23238	0,21900	
Apple_ATV_%Δ	Apple_I,_LAG_TREND_%Δ	-0,12672	0,07233	-1,75191	**0,08170**	0,19289
	Apple_II,_LAG_TREND_%Δ	-0,06444	0,07171	-0,89860	0,37020	
Tesla_ATV_%Δ	Tesla_I,_LAG_TREND_%Δ	0,16532	0,14465	1,14285	0,25570	0,09251
	Tesla_II,_LAG_TREND_%Δ	0,02683	0,13587	0,19746	0,84390	
Micrososoft_ATV_%Δ	Micrososoft_I,_LAG_TREND_%Δ	0,00894	0,11545	0,07746	0,93840	0,17218
	Micrososoft_II,_LAG_TREND_%Δ	0,00063	0,11514	0,00548	0,99560	
Coca-Cola_ATV_%Δ	Coca-Cola_I,_LAG_TREND_%Δ	0,16904	0,56854	0,29732	0,76790	0,37771
	Coca-Cola_II,_LAG_TREND_%Δ	-0,61010	0,55985	-1,08976	0,28310	
Facebook_ATV_%Δ	Facebook_I,_LAG_TREND_%Δ	-0,10253	0,29504	-0,34750	0,72870	0,09624
	Facebook_II,_LAG_TREND_%Δ	-0,22211	0,26224	-0,84695	0,39850	
General_Motors_ATV_%Δ	General_Motors_I,_LAG_TREND_%Δ	0,00600	0,12679	0,04733	0,96230	0,22776
	General_Motors_II,_LAG_TREND_%Δ	0,00901	0,12611	0,07143	0,94310	
Procter&Gamble_ATV_%Δ	Procter&Gamble_I,_LAG_TREND_%Δ	-0,00640	0,09289	-0,06884	0,94520	0,29463
	Procter&Gamble_II,_LAG_TREND_%Δ	0,06937	0,08837	0,78499	0,43430	
Starbucks_ATV_%Δ	Starbucks_I,_LAG_TREND_%Δ	-0,24359	0,12863	-1,89370	**0,06010**	0,16042
	Starbucks_II,_LAG_TREND_%Δ	-0,04763	0,12083	-0,39415	0,69400	
S&P500_ATV_%Δ	S&P500_I,_LAG_TREND_%Δ	0,07807	0,05652	1,38145	0,16910	0,25309
	S&P500_II,_LAG_TREND_%Δ	-0,04952	0,05300	-0,93426	0,35160	
Delta_Air_Lines_ATV_%Δ	Delta_Air_Lines_I,_LAG_TREND_%Δ	0,20319	0,15638	1,29938	0,19570	0,15917
	Delta_Air_Lines_II,_LAG_TREND_%Δ	-0,02365	0,15584	-0,15175	0,87960	
Exxon_ATV_%Δ	Exxon_I,_LAG_TREND_%Δ	0,02625	0,09425	0,27855	0,78100	0,25328
	Exxon_II,_LAG_TREND_%Δ	-0,14965	0,09348	-1,60085	0,11140	
Goldman Sach_ATV_%Δ	Goldman Sach_I,_LAG_TREND_%Δ	-0,04056	0,07600	-0,53368	0,59430	0,24373
	Goldman Sach_II,_LAG_TREND_%Δ	0,17719	0,07547	2,34796	**0,02010**	
Blackberry_ATV_%Δ	Blackberry_I,_LAG_TREND_%Δ	-0,09703	0,12638	-0,76778	0,44380	0,14362
	Blackberry_II,_LAG_TREND_%Δ	-0,10345	0,12238	-0,84539	0,39920	
Yahoo_ATV_%Δ	Yahoo_I,_LAG_TREND_%Δ	0,70374	0,27657	2,54448	**0,01190**	0,22787
	Yahoo_II,_LAG_TREND_%Δ	0,02903	0,27205	0,10671	0,91520	

Table 10: Results of lagged ARIMA (2, 2) method for the effect of lagged TREND_%Δ on ATV_%Δ.

Dependent Variable:	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Tesla_ATV_%Δ	Tesla_I,_LAG_TREND_%Δ	0,01098	0,01246	0,88137	0,38020	0,17493
	Tesla_II,_LAG_TREND_%Δ	-0,00877	0,01234	-0,71098	0,47870	
Facebook_ATV_%Δ	Facebook_I,_LAG_TREND_%Δ	0,02995	0,02186	1,37025	0,17280	0,10373
	Facebook_II,_LAG_TREND_%Δ	-0,00249	0,02160	-0,11518	0,90850	
General_Motors_ATV_%Δ	General_Motors_I,_LAG_TREND_%Δ	-0,01142	0,00854	-1,33724	0,18310	0,08608
	General_Motors_II,_LAG_TREND_%Δ	0,00641	0,00852	0,75233	0,45300	
Delta_Air_Lines_ATV_%Δ	Delta_Air_Lines_I,_LAG_TREND_%Δ	-0,01709	0,01325	-1,28967	0,19910	0,06266
	Delta_Air_Lines_II,_LAG_TREND_%Δ	-0,00894	0,01312	-0,68158	0,49650	
Exxon_ATV_%Δ	Exxon_I,_LAG_TREND_%Δ	0,00613	0,00622	0,98591	0,32570	0,11451
	Exxon_II,_LAG_TREND_%Δ	0,01616	0,00619	2,61015	**0,00990**	

Table 11: Results of lagged ARIMA (2, 2) method for the effect of lagged TREND_%Δ on AWP_%Δ.

Dependent Variable:	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Tesla_ABS_ATV_%Δ	Tesla_I,_LAG_TREND_%Δ	0,04321	0,00985	4,38806	**0,00000**	0,26798
	Tesla_II,_LAG_TREND_%Δ	-0,00984	0,00978	-1,00658	0,31650	
Facebook_ABS_ATV_%Δ	Facebook_I,_LAG_TREND_%Δ	0,10233	0,01749	5,85119	**0,00000**	0,31981
	Facebook_II,_LAG_TREND_%Δ	-0,02059	0,01761	-1,16925	0,24430	
General_ABS_Motors_ATV_%Δ	General_Motors_I,_LAG_TREND_%Δ	0,00267	0,00674	0,39633	0,69240	0,03483
	General_Motors_II,_LAG_TREND_%Δ	-0,00066	0,00674	-0,09721	0,92270	
Delta_Air_Lines_ABS_ATV_%Δ	Delta_Air_Lines_I,_LAG_TREND_%Δ	0,01619	0,01208	1,34097	0,18190	0,04224
	Delta_Air_Lines_II,_LAG_TREND_%Δ	-0,00382	0,01207	-0,31694	0,75170	
Exxon_ABS_ATV_%Δ	Exxon_I,_LAG_TREND_%Δ	0,00319	0,00547	0,58216	0,56130	0,07097
	Exxon_II,_LAG_TREND_%Δ	0,00611	0,00547	1,11771	0,26540	
Apple_ABS_ATV_%Δ	Apple_I,_LAG_TREND_%Δ	0,01533	0,00492	3,11587	**0,00220**	0,10331
	Apple_II,_LAG_TREND_%Δ	0,00014	0,00487	0,02859	0,97720	
Microsoft_ABS_ATV_%Δ	Microsoft_I,_LAG_TREND_%Δ	0,00187	0,00622	0,30049	0,76420	0,03525
	Microsoft_II,_LAG_TREND_%Δ	-0,00136	0,00625	-0,21740	0,82820	
Backberry_ABS_ATV_%Δ	Backberry_I,_LAG_TREND_%Δ	0,02827	0,01030	2,74519	**0,00670**	0,08710
	Backberry_II,_LAG_TREND_%Δ	-0,00976	0,01024	-0,95277	0,34220	

Table 12: Results of lagged ARIMA (2, 2) method for the effect of lagged ABS_TREND_%Δ on ABS_AWP_%Δ.

Here it is safe conclude that Google Trends on weekly basis are not very reliable leading indicator for stock market development. Viability for predicting next week trade volume was proven only in case of Google, Apple, Starbuck and Yahoo. Where for first three the coefficient is negative whereas for Yahoo is positive. As for price development the result are significant only in case of Exxon and only the second lag, which could be denoted as a simple coincidence. Nevertheless the results for absolute values of percentage changes in Google Trends and stock prices are more promising. First lags of absolute values of percentage changes in Google Trends are statistically significant for determining next week volatility in cases of Tesla, Facebook, Apple and Blackberry suggesting that Google Trends are to some extent a leading indicator of next week volatility on stock market, at least for firms operating in IT and modern technology business. When higher search volume in one week marks higher volatility in the next week. Again backing up my expectations that correlations will be strongest in those areas.

Dependent Variable:	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Google_TREND_%Δ	Google_I,_LAG_ATV_%Δ	-0,06159	0,04541	-1,35649	0,18220	0,04536
	Google_II,_LAG_ATV_%Δ	-0,03165	0,04504	-0,70269	0,48610	
Apple_TREND_%Δ	Apple_I,_LAG_ATV_%Δ	-0,04736	0,09199	-0,51484	0,60740	0,00674
	Apple_II,_LAG_ATV_%Δ	-0,09462	0,09213	-1,02700	0,30590	
Tesla_TREND_%Δ	Tesla_I,_LAG_ATV_%Δ	-0,05734	0,08658	-0,66224	0,50920	0,02455
	Tesla_II,_LAG_ATV_%Δ	-0,14001	0,08650	-1,61854	0,10840	
Micrososoft_TREND_%Δ	Micrososoft_I,_LAG_ATV_%Δ	-0,10715	0,06120	-1,75090	**0,08180**	0,02311
	Micrososoft_II,_LAG_ATV_%Δ	-0,08793	0,06121	-1,43649	0,15280	
Coca-Cola_TREND_%Δ	Coca-Cola_I,_LAG_ATV_%Δ	-0,16651	0,05780	-2,88078	**0,00450**	0,05249
	Coca-Cola_II,_LAG_ATV_%Δ	-0,02442	0,05791	-0,42170	0,67380	
Facebook_TREND_%Δ	Facebook_I,_LAG_ATV_%Δ	0,01059	0,03157	0,33541	0,73780	0,00327
	Facebook_II,_LAG_ATV_%Δ	-0,01618	0,03127	-0,51722	0,60580	
General_Motors_TREND_%Δ	General_Motors_I,_LAG_ATV_%Δ	-0,05779	0,05598	-1,03236	0,30340	0,00732
	General_Motors_II,_LAG_ATV_%Δ	-0,00673	0,05590	-0,12044	0,90430	
Procter&Gamble_TREND_%Δ	Procter&Gamble_I,_LAG_ATV_%Δ	-0,04434	0,10798	-0,41064	0,68210	0,01895
	Procter&Gamble_II,_LAG_ATV_%Δ	-0,15836	0,10934	-1,44826	0,15040	
Starbucks_TREND_%Δ	Starbucks_I,_LAG_ATV_%Δ	-0,02325	0,06130	-0,37923	0,70500	0,00604
	Starbucks_II,_LAG_ATV_%Δ	-0,06061	0,06139	-0,98726	0,32500	
S&P500_TREND_%Δ	S&P500_I,_LAG_ATV_%Δ	-0,32216	0,11543	-2,79087	**0,00590**	0,04569
	S&P500_II,_LAG_ATV_%Δ	-0,14152	0,11585	-1,22165	0,22360	
Delta_Air_Lines_TREND_%Δ	Delta_Air_Lines_I,_LAG_ATV_%Δ	-0,02548	0,04105	-0,62071	0,53570	0,00240
	Delta_Air_Lines_II,_LAG_ATV_%Δ	-0,00298	0,04111	-0,07247	0,94230	
Exxon_TREND_%Δ	Exxon_I,_LAG_ATV_%Δ	-0,04744	0,07359	-0,64461	0,52010	0,00844
	Exxon_II,_LAG_ATV_%Δ	-0,08533	0,07385	-1,15543	0,24960	
Goldman Sach_TREND_%Δ	Goldman Sach_I,_LAG_ATV_%Δ	-0,03362	0,07325	-0,45895	0,64690	0,01803
	Goldman Sach_II,_LAG_ATV_%Δ	-0,12589	0,07254	-1,73542	**0,08450**	
Blackberry_TREND_%Δ	Blackberry_I,_LAG_ATV_%Δ	-0,04001	0,05381	-0,74365	0,45810	0,01135
	Blackberry_II,_LAG_ATV_%Δ	-0,06864	0,05399	-1,27139	0,20540	
Yahoo_TREND_%Δ	Yahoo_I,_LAG_ATV_%Δ	0,01126	0,02430	0,46338	0,64370	0,01433
	Yahoo_II,_LAG_ATV_%Δ	-0,03329	0,02431	-1,36965	0,17270	

Table 13: Results of lagged Least Squares method for the effect of lagged ATV_%Δ on TREND_%Δ.

Dependent Variable:	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Tesla_TREND_%Δ	Tesla_I,_LAG_AWP_%Δ	-0,57409	0,79813	-0,71929	0,47350	0,00476
	Tesla_II,_LAG_AWP_%Δ	0,21181	0,79835	0,26531	0,79130	
Facebook_TREND_%Δ	Facebook_I,_LAG_AWP_%Δ	0,12440	0,34157	0,36421	0,71620	0,00565
	Facebook_II,_LAG_AWP_%Δ	-0,29952	0,33163	-0,90317	0,36790	
General_Motors_TREND_%Δ	General_Motors_I,_LAG_AWP_%Δ	-0,50166	0,79663	-0,62974	0,52970	0,00936
	General_Motors_II,_LAG_AWP_%Δ	-0,72874	0,79647	-0,91497	0,36160	
Delta_Air_Lines_TREND_%Δ	Delta_Air_Lines_I,_LAG_AWP_%Δ	-0,96751	0,52229	-1,85243	**0,06580**	0,02059
	Delta_Air_Lines_II,_LAG_AWP_%Δ	0,09228	0,52259	0,17659	0,86000	
Exxon_TREND_%Δ	Exxon_I,_LAG_AWP_%Δ	0,29350	1,00821	0,29111	0,77130	0,02115
	Exxon_II,_LAG_AWP_%Δ	1,81621	1,00780	1,80215	**0,07340**	

Table 14: Results of lagged Least Squares method for the effect of lagged AWP_%Δ on TREND_%Δ.

Dependent Variable:	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Tesla_TREND_%Δ	Tesla_I,_LAG_ABS_AWP_%Δ	-0,14412	1,07604	-0,13394	0,89370	0,00157
	Tesla_II,_LAG_ABS_AWP_%Δ	-0,40560	1,07867	-0,37602	0,70760	
Facebook_TREND_%Δ	Facebook_I,_LAG_ABS_AWP_%Δ	-1,05011	0,45730	-2,29631	**0,02310**	0,04190
	Facebook_II,_LAG_ABS_AWP_%Δ	-0,22204	0,44349	-0,50066	0,61740	
General_Motors_TREND_%Δ	General_Motors_I,_LAG_ABS_AWP_%Δ	-1,34809	1,34996	-0,99862	0,31950	0,00625
	General_Motors_II,_LAG_ABS_AWP_%Δ	-0,09690	1,34903	-0,07183	0,94280	
Delta_Air_Lines_TREND_%Δ	Delta_Air_Lines_I,_LAG_ABS_AWP_%Δ	-2,35237	0,75458	-3,11746	**0,00220**	0,05651
	Delta_Air_Lines_II,_LAG_ABS_AWP_%Δ	0,27898	0,75386	0,37007	0,71180	
Exxon_TREND_%Δ	Exxon_I,_LAG_ABS_AWP_%Δ	0,57779	1,63292	0,35384	0,72390	0,00317
	Exxon_II,_LAG_ABS_AWP_%Δ	-1,04471	1,63581	-0,63865	0,52390	
Apple_TREND_%Δ	Apple_I,_LAG_ABS_AWP_%Δ	-3,17450	1,49816	-2,11894	**0,03560**	0,03409
	Apple_II,_LAG_ABS_AWP_%Δ	-1,64227	1,49777	-1,09648	0,27450	
Microsoft_TREND_%Δ	Microsoft_I,_LAG_ABS_AWP_%Δ	-2,69986	1,33141	-2,02783	**0,04420**	0,03647
	Microsoft_II,_LAG_ABS_AWP_%Δ	-1,94552	1,33320	-1,45928	0,14640	
Backberry_TREND_%Δ	Backberry_I,_LAG_ABS_AWP_%Δ	-0,41848	0,72466	-0,57749	0,56440	0,00539
	Backberry_II,_LAG_ABS_AWP_%Δ	0,55995	0,72276	0,77474	0,43960	

Table 15: Results of lagged Least Squares method for the effect of lagged ABS_AWP_%Δ on ABS_TREND_%Δ.

Result for Google Trends as lagging indicator of stock market are also not very promising. As week trade volume is concerned we can see statistical significance of the first lag in case of firms like Microsoft, Coca Cola, S&P index and statistical significance of second lag in case of Goldman Sachs. Coefficients are in all cases negative, suggesting that higher trade volume in one week is followed by lower search volume in following period. A for the effect of price development on next period search volume it seems to have only small statistical significance and that is in case of Delta Air Lines where the first lag shows statistical significance with negative coefficient, suggesting so that the rise in stock price marks next week decrease in search volume. Statistically significant is also the effect of second lag in case of Exxon, now with positive coefficient. As for volatility of market here the effect on Google Trends is again significant mostly for firms operating in IT and modern technology business, namely: Microsoft, Apple, Facebook and also for Delta Air Lines. Coefficient is negative in all cases, suggesting that higher volatility is followed by lower search volumes in next week.

2.3. StockTwits

For analysis of StockTwits data I will use the same approach as I did for Google Trends. First I will test for correlation with Least Squares Method, then for those cases where the effect of StockTwits data will be significant I will follow with ARIMA model to solve for autocorrelation. (Again with ARIMA (2, 2) specification). Then I will try to determine if StockTwits data are leading or lagging indicator of stock market development.

Based on StockTwits data I have decided to test in four categories and that is: correlation between Bullish Index (ration of messages with bullish sentiment on total number of messages with sentiment stamp) and trade volume, bullish index and price, messages volume (number of messages concerning particular stock containing any kind of sentiment) and trade volume, and lastly message volume and price.

2.3.1. Least Squares Method

$$\text{ATV\_\%}\Delta_t = \beta_0 + \beta \text{TREND\_\%}\Delta_t + \varepsilon_t$$

Dependent Variable	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Google_TV_%Δ	Google_BI_%Δ	-0,61821	0,62579	-0,98788	0,32420	0,00390
Apple_TV_%Δ	Apple_BI_%Δ	0,09766	0,17385	0,56174	0,57480	0,00126
Tesla_TV_%Δ	Tesla_BI_%Δ	0,18976	0,18780	1,01041	0,31330	0,00408
Microsoft_TV_%Δ	Microsoft_BI_%Δ	-0,10592	0,10020	-1,05711	0,29150	0,00450
Coca-Cola_TV_%Δ	Coca-Cola_BI_%Δ	-0,04022	0,04320	-0,93110	0,35270	0,00346
Michael Kors_TV_%Δ	Michael Kors_BI_%Δ	0,13729	0,10880	1,26185	0,20820	0,00633
Facebook_TV_%Δ	Facebook_BI_%Δ	-0,06197	0,14343	-0,43208	0,66610	0,00075
General Motors_TV_%Δ	General Motors_BI_%Δ	-0,00073	0,02955	-0,02474	0,98030	0,00000
General Electricts_TV_%Δ	General Electricts_BI_%Δ	-0,02890	0,04245	-0,68067	0,49670	0,00185
Procter&Gamble_TV_%Δ	Procter&Gamble_BI_%Δ	-0,01942	0,03722	-0,52192	0,60220	0,00109
Starbucks_TV_%Δ	Starbucks_BI_%Δ	-0,02983	0,03018	-0,98828	0,32400	0,00389
SP_500 _TV_%Δ	SP_500 _BI_%Δ	-0,01732	0,02075	-0,83468	0,40470	0,00278
Delta Airlines_TV_%Δ	Delta Airlines_BI_%Δ	-0,06982	0,05979	-1,16774	0,24400	0,00543
Exxon_TV_%Δ	Exxon_BI_%Δ	-0,01663	0,03338	-0,49810	0,61890	0,00099
Goldman Sachs_TV_%Δ	Goldman Sachs_BI_%Δ	0,04781	0,03582	1,33458	0,18320	0,00707
Backberry_TV_%Δ	Backberry_BI_%Δ	-0,33131	0,37065	-0,89385	0,37230	0,00319
Yahoo_TV_%Δ	Yahoo_BI_%Δ	-0,02879	0,12757	-0,22564	0,82170	0,00020
Plug_Power_TV_%Δ	Plug_Power_BI_%Δ	-0,18418	0,48860	-0,37696	0,70650	0,00057
Twitter_TV_%Δ	Twitter_BI_%Δ	0,36058	0,18307	1,96964	**0,05000**	0,01528

Table 16: Results of Leas Squares for correlation between BI_%Δ and TV_%Δ.

Dependent Variable	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Google_Price_%Δ	Google_BI_%Δ	0,00079	0,01300	0,06089	0,95150	0,00002
Apple_Price_%Δ	Apple_BI_%Δ	0,19061	0,02277	8,37195	**0,00000**	0,21897
Tesla_Price_%Δ	Tesla_BI_%Δ	0,07207	0,00824	8,74379	**0,00000**	0,23492
Microsoft_Price_%Δ	Microsoft_BI_%Δ	0,00613	0,00242	2,53573	**0,01180**	0,02537
Coca-Cola_Price_%Δ	Coca-Cola_BI_%Δ	0,00091	0,00606	0,15029	0,88070	0,00009
Michael Kors_Price_%Δ	Michael Kors_BI_%Δ	0,00270	0,00884	0,30600	0,75990	0,00037
Facebook_Price_%Δ	Facebook_BI_%Δ	0,20053	0,02267	8,84606	**0,00000**	0,23839
General Motors_Price_%Δ	General Motors_BI_%Δ	0,00174	0,00361	0,48295	0,62960	0,00093
General Electricts_Price_%Δ	General Electricts_BI_%Δ	0,00088	0,00823	0,10715	0,91480	0,00005
Procter&Gamble_Price_%Δ	Procter&Gamble_BI_%Δ	-0,00092	0,00753	-0,12199	0,90300	0,00006
Starbucks_Price_%Δ	Starbucks_BI_%Δ	0,00125	0,00456	0,27503	0,78350	0,00030
SP_500 _Price_%Δ	SP_500 _BI_%Δ	0,02539	0,00837	3,03483	**0,00270**	0,03553
Delta Airlines_Price_%Δ	Delta Airlines_BI_%Δ	0,00922	0,01024	0,90042	0,36880	0,00323
Exxon_Price_%Δ	Exxon_BI_%Δ	0,01442	0,00725	1,98828	**0,04790**	0,01557
Goldman Sachs_Price_%Δ	Goldman Sachs_BI_%Δ	0,00283	0,00641	0,44149	0,65920	0,00078
Backberry_Price_%Δ	Backberry_BI_%Δ	0,28300	0,02928	9,66635	**0,00000**	0,27207
Yahoo_Price_%Δ	Yahoo_BI_%Δ	0,02098	0,01772	1,18411	0,23750	0,00558
Plug_Power_Price_%Δ	Plug_Power_BI_%Δ	0,45743	0,04200	10,89048	**0,00000**	0,32176
Twitter_Price_%Δ	Twitter_BI_%Δ	0,07665	0,01402	5,46671	**0,00000**	0,10678

Table 17: Results of Leas Squares for correlation between BI_%Δ and Price_%Δ.

Dependent Variable	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Google_TV_%Δ	Google_MV_%Δ	-0,04274	0,05838	-0,73220	0,46470	0,00214
Apple_TV_%Δ	Apple_MV_%Δ	-0,00352	0,00991	-0,35499	0,72290	0,00050
Tesla_TV_%Δ	Tesla_MV_%Δ	-0,01669	0,01031	-1,61973	0,10660	0,01039
Microsoft_TV_%Δ	Microsoft_MV_%Δ	-0,00664	0,01382	-0,48055	0,63130	0,00092
Coca-Cola_TV_%Δ	Coca-Cola_MV_%Δ	-0,01968	0,01813	-1,08533	0,27880	0,00469
Michael Kors_TV_%Δ	Michael Kors_MV_%Δ	-0,02547	0,02212	-1,15121	0,25070	0,00527
Facebook_TV_%Δ	Facebook_MV_%Δ	-0,00254	0,00470	-0,54024	0,58950	0,00117
General Motors_TV_%Δ	General Motors_MV_%Δ	0,04874	0,01194	4,08253	**0,00010**	0,06250
General Electricts_TV_%Δ	General Electricts_MV_%Δ	-0,00181	0,01237	-0,14621	0,88390	0,00009
Procter&Gamble_TV_%Δ	Procter&Gamble_MV_%Δ	0,02967	0,02155	1,37699	0,16970	0,00753
Starbucks_TV_%Δ	Starbucks_MV_%Δ	0,01069	0,01087	0,98323	0,32640	0,00385
SP_500 _TV_%Δ	SP_500 _MV_%Δ	0,00213	0,00406	0,52401	0,60070	0,00110
Delta Airlines_TV_%Δ	Delta Airlines_MV_%Δ	0,01229	0,01592	0,77207	0,44080	0,00238
Exxon_TV_%Δ	Exxon_MV_%Δ	0,01134	0,01709	0,66339	0,50770	0,00176
Goldman Sachs_TV_%Δ	Goldman Sachs_MV_%Δ	-0,00667	0,01639	-0,40677	0,68450	0,00066
Backberry_TV_%Δ	Backberry_MV_%Δ	-0,00112	0,01203	-0,09272	0,92620	0,00003
Yahoo_TV_%Δ	Yahoo_MV_%Δ	0,00689	0,00745	0,92563	0,35550	0,00342
Plug_Power_TV_%Δ	Plug_Power_MV_%Δ	-0,01484	0,01779	-0,83450	0,40480	0,00278
Twitter_TV_%Δ	Twitter_MV_%Δ	0,00786	0,01232	0,63789	0,52410	0,00163

Table 18: Results of Leas Squares for correlation between MV_%Δ and TV_%Δ.

Dependent Variable	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Google_Price_%Δ	Google_MV_%Δ	0,00078	0,00177	0,44334	0,65790	0,00079
Apple_Price_%Δ	Apple_MV_%Δ	0,00094	0,00147	0,63956	0,52300	0,00163
Tesla_Price_%Δ	Tesla_MV_%Δ	0,00103	0,00119	0,86406	0,38840	0,00298
Microsoft_Price_%Δ	Microsoft_MV_%Δ	0,00082	0,00181	0,45317	0,65080	0,00082
Coca-Cola_Price_%Δ	Coca-Cola_MV_%Δ	0,00190	0,00254	0,74554	0,45660	0,00222
Michael Kors_Price_%Δ	Michael Kors_MV_%Δ	0,00148	0,00179	0,82457	0,41040	0,00271
Facebook_Price_%Δ	Facebook_MV_%Δ	0,00096	0,00085	1,12984	0,25960	0,00508
General Motors_Price_%Δ	General Motors_MV_%Δ	0,00088	0,00150	0,58486	0,55920	0,00137
General Electricts_Price_%Δ	General Electricts_MV_%Δ	-0,00459	0,00406	-1,13063	0,25930	0,00240
Procter&Gamble_Price_%Δ	Procter&Gamble_MV_%Δ	-0,00113	0,00437	-0,25717	0,79730	0,00026
Starbucks_Price_%Δ	Starbucks_MV_%Δ	0,00041	0,00164	0,25154	0,80160	0,00025
SP_500 _Price_%Δ	SP_500 _MV_%Δ	0,00119	0,00166	0,71541	0,47500	0,00204
Delta Airlines_Price_%Δ	Delta Airlines_MV_%Δ	0,00254	0,00272	0,93298	0,35170	0,00347
Exxon_Price_%Δ	Exxon_MV_%Δ	0,00419	0,00374	1,12179	0,26300	0,00501
Goldman Sachs_Price_%Δ	Goldman Sachs_MV_%Δ	0,00258	0,00292	0,88363	0,37770	0,00311
Backberry_Price_%Δ	Backberry_MV_%Δ	0,00148	0,00111	1,33738	0,18230	0,00710
Yahoo_Price_%Δ	Yahoo_MV_%Δ	0,00037	0,00104	0,35510	0,72280	0,00050
Plug_Power_Price_%Δ	Plug_Power_MV_%Δ	0,00018	0,00186	0,09457	0,92470	0,00004
Twitter_Price_%Δ	Twitter_MV_%Δ	0,00031	0,00099	0,30730	0,75890	0,00038

Table 19: Results of Leas Squares for correlation between MV_%Δ and Price_%Δ.

Correlation was proven only between Bullish Index and price as for other relations between bullish index trade volume, messages volume and trade volume, and message volume and price there was no statistically significant effect. As for the correlation between changes in bullish index and changes in stock price, it applies mostly for firms operating in IT and modern technogy industries. Namely; Apple, Tesla, Facebook, Plug Power, Yahoo, Exxon and also index S&P 500. For those I will continue testing with ARIMA models and then I will try to determine whether Bullish index expressing sentiment between users of StockTwits is a leading or lagging indicator of stock prices movements.

2.3.2. ARIMA method

A mentioned before I will again use ARIMA (2,2) specification as one model that fits all. So the results will not be maximally accurate but they should asses the information value hidden inside the sentiment on StockTwits.

$$\text{PRICE\_\%}\Delta_t = \beta_0 + \beta\text{BI\_\%}\Delta_t + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_t + \alpha_1\varepsilon_{t-1} + \alpha_2\varepsilon_{t-2}$$

Dependent Variable:	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Apple_PRICE_%Δ	Apple_BI_%Δ	0,19668	0,02257	8,71251	**0,00000**	0,24965
Blackberry_PRICE_%Δ	Blackberry_BI_%Δ	0,31811	0,03187	9,98215	**0,00000**	0,29118
Exxon_PRICE_%Δ	Exxon_BI_%Δ	0,01749	0,00534	3,27290	**0,00120**	0,05456
Facebook_PRICE_%Δ	Facebook_BI_%Δ	0,20109	0,02166	9,28432	**0,00000**	0,26278
Microsoft_PRICE_%Δ	Microsoft_BI_%Δ	X	X	X	X	X
Plug Power_PRICE_%Δ	Plug Power_BI_%Δ	0,43773	0,03969	11,02946	**0,00000**	0,35882
S&P 500_PRICE_%Δ	S&P 500_BI_%Δ	0,01896	0,00601	3,15624	**0,00180**	0,05738
Tesla Motors_PRICE_%Δ	Tesla Motors_BI_%Δ	0,07222	0,00821	8,80106	**0,00000**	0,24125
Twitter_PRICE_%Δ	Twitter_BI_%Δ	0,08611	0,01349	6,38209	**0,00000**	0,13183

Table 20: Results of ARIMA (2, 2) between BI_%Δ and PRICE_%Δ.

The effect of change in bullish index on StockTwits remained significant in all cases and R squared rose when treated for autocorrelation in price development. Only problem was in case of Microsoft for which there was not continuous data sample because for some days there were no post with sentiment regarding Microsoft stocks which prevents from running ARIMA model on Microsoft Stocks.

2.3.3. Granger Causality

Again as in the part for Google Trends the last test will be focused on determining whether bullish index is leading or lagging indicator of stock price movement. I will focus only on those companies for which the change in bullish index proved to have a statistically significant effect on change in stock price. Testing whether bullish index is a leading indicator will be based on ARIMA models.

$$\begin{aligned}\text{PRICE\_\%}\Delta_i = {} & \beta_0 + \beta_1\text{BI\_\%}\Delta_{t-1} + \beta_2\text{BI\_\%}\Delta_{t-2} + \beta_3\text{BI\_\%}\Delta_{t-3} + \beta_4\text{BI\_\%}\Delta_{t-4} \\ & + \beta_5\text{BI\_\%}\Delta_{t-5} + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_t + \alpha_1\varepsilon_{t-1} + \alpha_2\varepsilon_{t-2}\end{aligned}$$

Whereas the reverse testing if the bullish index is a lagging indicator of stock price movement will be based on Least Squares method as the autocorrelation in bullish index development it not so strong.

$$\text{BI\_\%}\Delta_i = \beta_0 + \beta_1 \text{PRICE\_\%}\Delta_{i-1} + \beta_2 \text{PRICE\_\%}\Delta_{i-2} + \beta_3 \text{PRICE\_\%}\Delta_{i-3} + \beta_4 \text{PRICE\_\%}\Delta_{i-4} + \beta_5 \text{PRICE\_\%}\Delta_{i-5} + \varepsilon_t$$

When testing I will go as far as five days back as I believe that examine sentiment more than five days back is, with respect of how volatile and fast changing the mood between StockTwits users is, useless..

According to result (see the table 22 on next page) bullish index does not seem to be a good leading indicator of future stock price movement. Users of StockTwits seem to be best in predicting future price development of Tesla Motors, for which the first and third lags are significant with positive coefficient. This accordingly with my expectations suggests that higher bullish index predicts increase in stock price in next days. But the results for other companies are not so promising and clear, for example for Twitter the second and fourth lags are significant with negative coefficient. That means that, at least as Twitter is concerned, higher bullishness between users of StockTwits predicts a decrease in price during next days.

Dependent Variable:	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Apple_PRICE_%Δ	Apple_I,_LAG_BI_%Δ	-0,04802	0,03004	-1,59888	0,11120	0,07051
	Apple_II,_LAG_BI_%Δ	-0,03197	0,03172	-1,00761	0,31470	
	Apple_III,_LAG_BI_%Δ	-0,04907	0,03478	-1,41080	0,15960	
	Apple_IV,_LAG_BI_%Δ	-0,05694	0,03095	-1,83973	**0,06710**	
	Apple_V,_LAG_BI_%Δ	-0,01798	0,03053	-0,58897	0,55640	
Blackberry_PRICE_%Δ	Blackberry_I,_LAG_BI_%Δ	0,03295	0,04649	0,70869	0,47920	0,03083
	Blackberry_II,_LAG_BI_%Δ	0,03518	0,06020	0,58437	0,55950	
	Blackberry_III,_LAG_BI_%Δ	0,06547	0,06057	1,08089	0,28090	
	Blackberry_IV,_LAG_BI_%Δ	0,01722	0,05993	0,28733	0,77410	
	Blackberry_V,_LAG_BI_%Δ	0,02298	0,04620	0,49745	0,61930	
Exxon_PRICE_%Δ	Exxon_I,_LAG_BI_%Δ	-0,02756	0,00780	-3,53284	**0,00050**	0,06167
	Exxon_II,_LAG_BI_%Δ	-0,00372	0,01488	-0,24968	0,80310	
	Exxon_III,_LAG_BI_%Δ	-0,00656	0,01626	-0,40308	0,68730	
	Exxon_IV,_LAG_BI_%Δ	-0,00196	0,01403	-0,13944	0,88920	
	Exxon_V,_LAG_BI_%Δ	-0,00263	0,00917	-0,28663	0,77460	
Facebook_PRICE_%Δ	Facebook_I,_LAG_BI_%Δ	-0,04460	0,03581	-1,24572	0,21410	0,01318
	Facebook_II,_LAG_BI_%Δ	-0,03714	0,04436	-0,83732	0,40330	
	Facebook_III,_LAG_BI_%Δ	-0,00804	0,04752	-0,16926	0,86570	
	Facebook_IV,_LAG_BI_%Δ	0,00571	0,04277	0,13360	0,89380	
	Facebook_V,_LAG_BI_%Δ	-0,02002	0,03343	-0,59898	0,54980	
Plug Power_PRICE_%Δ	Plug Power_I,_LAG_BI_%Δ	0,03036	0,06933	0,43799	0,66180	0,03467
	Plug Power_II,_LAG_BI_%Δ	0,05282	0,06908	0,76463	0,44530	
	Plug Power_III,_LAG_BI_%Δ	-0,02050	0,06355	-0,32260	0,74730	
	Plug Power_IV,_LAG_BI_%Δ	0,11915	0,06538	1,82234	**0,06970**	
	Plug Power_V,_LAG_BI_%Δ	-0,03979	0,06237	-0,63788	0,52420	
S&P 500_PRICE_%Δ	S&P 500_I,_LAG_BI_%Δ	-0,01232	0,00895	-1,37682	0,16990	0,04607
	S&P 500_II,_LAG_BI_%Δ	-0,00760	0,01227	-0,61897	0,53650	
	S&P 500_III,_LAG_BI_%Δ	-0,00452	0,01434	-0,31533	0,75280	
	S&P 500_IV,_LAG_BI_%Δ	0,00925	0,01194	0,77472	0,43930	
	S&P 500_V,_LAG_BI_%Δ	0,00848	0,00892	0,94964	0,34330	
Tesla Motors_PRICE_%Δ	Tesla Motors_I,_LAG_BI_%Δ	0,16540	0,02629	6,29165	**0,00000**	0,16854
	Tesla Motors_II,_LAG_BI_%Δ	0,00105	0,03533	0,02980	0,97620	
	Tesla Motors_III,_LAG_BI_%Δ	0,06250	0,03238	1,93010	**0,05480**	
	Tesla Motors_IV,_LAG_BI_%Δ	-0,00892	0,02493	-0,35774	0,72090	
	Tesla Motors_V,_LAG_BI_%Δ	0,03915	0,02378	1,64626	0,10100	
Twitter_PRICE_%Δ	Twitter_I,_LAG_BI_%Δ	-0,01953	0,01605	-1,21632	0,22510	0,11003
	Twitter_II,_LAG_BI_%Δ	-0,05921	0,01745	-3,39341	**0,00080**	
	Twitter_III,_LAG_BI_%Δ	0,00718	0,01835	0,39151	0,69580	
	Twitter_IV,_LAG_BI_%Δ	-0,03209	0,01785	-1,79776	**0,07350**	
	Twitter_V,_LAG_BI_%Δ	0,00131	0,01678	0,07821	0,93770	

Table 21: Results of ARIMA (2, 2) between lagged BI_%Δ and PRICE_%Δ.

StockTwits bullish index as a lagging indicator of stock price development.

Dependent Variable:	Variable	Coefficient	Std, Error	t-Statistic	Prob,	R-squared
Apple_BI_%Δ	Apple_I,_LAG_PRICE_%Δ	-1,37541	0,73493	-1,87148	**0,06250**	0,02644
	Apple_II,_LAG_PRICE_%Δ	-0,90923	0,73778	-1,23238	0,21900	
	Apple_III,_LAG_PRICE_%Δ	-0,22667	0,74118	-0,30583	0,76000	
	Apple_IV,_LAG_PRICE_%Δ	-0,52044	0,73856	-0,70467	0,48170	
	Apple_V,_LAG_PRICE_%Δ	-0,48205	0,73694	-0,65412	0,51370	
Blackberry_BI_%Δ	Blackberry_I,_LAG_PRICE_%Δ	-0,53133	0,27484	-1,93324	**0,05440**	0,01976
	Blackberry_II,_LAG_PRICE_%Δ	-0,03291	0,27571	-0,11938	0,90510	
	Blackberry_III,_LAG_PRICE_%Δ	-0,25073	0,27382	-0,91566	0,36080	
	Blackberry_IV,_LAG_PRICE_%Δ	0,11133	0,27269	0,40827	0,68340	
	Blackberry_V,_LAG_PRICE_%Δ	0,01780	0,27171	0,06552	0,94780	
Exxon_BI_%Δ	Exxon_I,_LAG_PRICE_%Δ	0,59814	3,45457	0,17314	0,86270	0,01102
	Exxon_II,_LAG_PRICE_%Δ	0,96406	3,43713	0,28048	0,77930	
	Exxon_III,_LAG_PRICE_%Δ	-5,46555	3,46304	-1,57826	0,11580	
	Exxon_IV,_LAG_PRICE_%Δ	-0,12952	3,42404	-0,03783	0,96990	
	Exxon_V,_LAG_PRICE_%Δ	1,66427	3,44413	0,48322	0,62940	
Facebook_BI_%Δ	Facebook_I,_LAG_PRICE_%Δ	-1,87838	0,45719	-4,10854	**0,00010**	0,08039
	Facebook_II,_LAG_PRICE_%Δ	-0,93616	0,45344	-2,06455	**0,04000**	
	Facebook_III,_LAG_PRICE_%Δ	-0,52198	0,45114	-1,15703	0,24840	
	Facebook_IV,_LAG_PRICE_%Δ	-0,45481	0,44961	-1,01156	0,31280	
	Facebook_V,_LAG_PRICE_%Δ	-0,35437	0,45393	-0,78068	0,43580	
Plug Power_BI_%Δ	Plug Power_I,_LAG_PRICE_%Δ	-0,48957	0,10899	-4,49188	**0,00000**	0,08686
	Plug Power_II,_LAG_PRICE_%Δ	-0,23206	0,10787	-2,15131	**0,03240**	
	Plug Power_III,_LAG_PRICE_%Δ	-0,05308	0,10051	-0,52815	0,59790	
	Plug Power_IV,_LAG_PRICE_%Δ	-0,00717	0,10042	-0,07138	0,94320	
	Plug Power_V,_LAG_PRICE_%Δ	-0,00090	0,10047	-0,00895	0,99290	
S&P 500_BI_%Δ	S&P 500_I,_LAG_PRICE_%Δ	-20,70298	3,99810	-5,17820	**0,00000**	0,11576
	S&P 500_II,_LAG_PRICE_%Δ	-3,50950	4,01470	-0,87416	0,38290	
	S&P 500_III,_LAG_PRICE_%Δ	1,37022	4,01519	0,34126	0,73320	
	S&P 500_IV,_LAG_PRICE_%Δ	-6,79214	4,01275	-1,69264	**0,09180**	
	S&P 500_V,_LAG_PRICE_%Δ	0,16276	4,01704	0,04052	0,96770	
Tesla Motors_BI_%Δ	Tesla Motors_I,_LAG_PRICE_%Δ	-1,63936	0,41417	-3,95816	**0,00010**	0,08942
	Tesla Motors_II,_LAG_PRICE_%Δ	-0,54540	0,41409	-1,31709	0,18910	
	Tesla Motors_III,_LAG_PRICE_%Δ	-0,66980	0,41418	-1,61717	0,10720	
	Tesla Motors_IV,_LAG_PRICE_%Δ	-0,34712	0,41417	-0,83810	0,40280	
	Tesla Motors_V,_LAG_PRICE_%Δ	0,62060	0,41441	1,49753	0,13560	
Twitter_BI_%Δ	Twitter_I,_LAG_PRICE_%Δ	-2,73086	0,51059	-5,34841	**0,00000**	0,11539
	Twitter_II,_LAG_PRICE_%Δ	-0,89501	0,51103	-1,75139	**0,08120**	
	Twitter_III,_LAG_PRICE_%Δ	-0,66683	0,50692	-1,31546	0,18960	
	Twitter_IV,_LAG_PRICE_%Δ	-0,29147	0,50617	-0,57583	0,56530	
	Twitter_V,_LAG_PRICE_%Δ	-0,02351	0,50480	-0,04657	0,96290	

Table 22: Results of Least Squares method for correlation between lagged PRICE_%Δ and BI_%Δ.

As for bullish index as lagging indicator of stock price development the results are quite uniform. The first lag is significant for all of the tested companies with exception of

Exxon. For Twitter, Plug Power, Facebook also the second lag is statistically significant. Also the coefficients are uniformly negative for all companies. Meaning that increase in price is connected with decrease of bullishness between users of StockTwits or that decrease in price causes a more bullish sentiment next day. This is truly interesting as it suggest that smaller investors, at least those actively posting on StockTwits are operating against the market.

Conclusion

Big data and information contained in actions of users on internet is lately becoming very frequently discussed topic. This revolution touches almost all industries from marketing to even most conservative areas as insurance business. Never before in history was the tracking of behavioral patents of individuals, groups or even nations and races so easy as it is now when thanks to our activity in online world, which effectively transmits our lives, characters and moods into zeros and ones in a way that our character and our moods (fear, happiness, calmness or distress) are becoming quantifiable. This also has enourmous implications for economic theory, as it allows more precise testing of economic theories. Logically the possibilities of usage of these data for predictions of business development started being exploited soon. And stocks and stock market development were not left out. Through multiple works on this topic have empirically proven the fact that data created by actions of internet user do contain valuable information for business and stock market development and it is also the conclusion of my work. As I focused on eighteen different stocks from across all industries and on whole market index and two sources of data I conclude that both Google Trends and StockTwits data do contain information which are to some extent viable for stock markets.

As for Google Trends, here is the correlation strongest in case of trade volume, where the correlation between percentages changes in Google Trends and trade volume were statistically significant for fourteen companies out of eighteen and also for whole market index S&P500. When testing for Granger Causality Google I have found statistically significant first lag only in cases of Google, Apple, Starbucks and Yahoo, where for Yahoo the coefficient was positive whereas for the others negative. That suggest us that in some cases we can predict changes of next week trade volume but not very reliably. Results for correlation between percentages changes Google Trends and stock price were significant for Tesla Motors, Facebook, Delta Air Lines and Exxon with positive coefficients. Out of these five Google Trends have not proved a leading indicator for none of them. When tested for absolute values of percentages changes Google Trends and stock price the results shows that Google Trends are to some extent viable for predicting next week volatility mainly for stock of firms operating in IT business, namely the first lags were significant for Tesla Motors, Facebook, Apple and BlackBerry.

Coefficients are positive in all cases suggesting so that higher volumes of searches in one week signal higher volatility during the next week.

The problem with Google Trends data is that I was only able to get weekly data because my request for daily data was denied. From that came the need to transpose all other data also into weekly format. That have caused a lot of distortion in results. And it also clearly suggests that one of the possible future areas worth looking into are the correlations between daily data.

For StockTwits, which is one nowadays probably the biggest social portal for stock traders, I was more successful and I managed to get and process data for the whole year 2014. I ran tests in similar structure as for Google Trends. From the four examined categories the results were significant only for one as the effect of change in bullish index proved to be significant for changes in stock price, but not for changes in trade volume. Whereas the effect of change in messages volume haven't proved significant for neither trade volume nor stock price movements.

With Least Squares Method I have proven correlation between percentage changes in bullish index and price in case of companies Apple, Blackberry, Exxon, Facebook, Microsoft, Plug Power, Tesla Motors, Twitter and also for the S&P500 index. This again supports expectations that the correlation will be strongest for companies operating in IT business.

When testing for Granger Causality result were mixed up. Lagged changes in bullish index proved to be significant in cases of Apple; fourth lag, negative coefficient, Exxon; first lag, negative coefficient, Plug Power; fourth lag, positive coefficient, Tesla Motors; first and third lags, positive coefficients and Twitter; second and fourth lags, negative coefficients. From these only the case of Tesla Motors shows the type of results I have expected. That is positive correlation between lagged bullish index and next day stock returns. As the result are not uniform in any way I can only conclude that sentiment on StockTwits does contain valuable information. But more specific conclusions are not possible and as I assed all companies with one model with the same ARIMA specification, this offers another opportunity for future work, to analyze specific companies individually and try to create trading strategy based on portfolio of these companies with incorporating these information into trading decisions.

On the other hand for the reversed test, for determining if the bullish index is more lagging indicator, the results were much clearer. For Apple, Blackberry, Facebook, Plug Power, Tesla Motors, Twitter and S&P500 index the first lag proved to be statistically significant and in all cases with negative coefficient. Which is really interesting as it signals that investors active on StockTwits are clearly betting against market as when the price decrease on one day marks next day increase in bullish sentiment on next day.

Problem with analysis of StockTwits data lies in a fact that StockTwits data are not universal. Meant in a way that volume of messages expressing sentiment is for each company significantly different, when for some companies I have tested for, the average daily volume of messages expressing sentiment can be measured in order of hundreds of messages per day where for other companies the message volume does not even reach order of tens. Not surprisingly in the case of companies with low message volume the correlation between price and sentiment was denied. And even in the case when the effect of change of sentiment proved to be significant in the case of Microsoft the volume of messages containing sentiment was not sufficient enough for running ARIMA models. This represents another notable fact that the cross industry approach to this topic is contra productive. More efficient would be to focus from the beginning on stock for which users of StockTwits are generally most active.

Finally I can conclude that sentiment of StockTwits user and sentiment expressed by volume of searches for specific terms connected with investing in to firm do contain viable information. Another conclusion is that the viability of the information rises with popularity of the company and popularity of its stock in investors' community. Great example of this is the case of Tesla Motors, lately a hot deal on financial markets, for which also the results are the most promising. Analyses of the effects of user's sentiment connected with these glamour stocks and appropriate incorporating of these relations into day to day trading strategies represents an area yet to be exploited. And it can be only expected that the attractiveness of this area is going to increase during next years as the also the numbers of active users are constantly rising.

LIST OF ACRONYMS

GT – Google Trends

TREND – value of Google Trends

ATV – Average trade volume for a week as given by Google Trends. Total sum of week trade volume divided by number of trading days.

AWP – Average price for a week as given by Google Trends. Average of closing prices in a given week.

ABS – Absolute value

BI – Bullish Index

YF – Yahoo Finance

EMH – Efficient Market Hypothesis

AMH – Adaptive Market Hypothesis

LSM – Least Squares Method

ARIMA model – Autoregressive Integrated Moving Average model

%Δ – Percentage change

MV – Message volume (for StockTwits)

LIST OF FIGURES

LIST OF TABLES

REFERENCES

Fama, E. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), p.383.

Malkiel, B. (2005). Reflections on the Efficient Market Hypothesis: 30 Years Later. Financial Review, 40(1), pp.1-9.

Lo, A. (2004). The Adaptive Markets Hypothesis. The Journal of Portfolio Management, 30(5), pp.15-29.

Wysocki, P. (1998). Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards. SSRN Journal.

Tumarkin, R. and Whitelaw, R. (2001). News or Noise? Internet Postings and Stock Prices. *Financial Analysts Journal*, 57(3), pp.41-51.

Antweiler, W. and Frank, M. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *J Finance*, 59(3), pp.1259-1294.

Das, S. and Chen, M. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, 53(9), pp.1375-1388.

Bollen, J., Mao, H. and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), pp.1-8.

Sprenger, T., Tumasjan, A., Sandner, P. and Welpe, I. (2013). Tweets and Trades: the Information Content of Stock Microblogs. European Financial Management, 20(5), pp.926-957.

Loughlin, C. and Harnish, E. (2013). The Viability of StockTwits and Google Trends to Predict the Stock Market. [online] stocktwits.com. Available at: http://stocktwits.com/research/Viability-of-StockTwits-and-Google-Trends-Loughlin_Harnisch.pdf [Accessed 27 Apr. 2015].

Barber, B. and Odean, T. (2007). All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. *Review of Financial Studies*, 21(2), pp.785-818.

Stephens-Davidowitz, S. (2014). The cost of racial animus on a black candidate: Evidence using Google search data. Journal of Public Economics, 118, pp.26-40.

StockTwits, (2015). *About StockTwits*. [online] Available at: http://stocktwits.com/about [Accessed 27 Apr. 2015].

Stephens-Davidowitz, S. (2014). The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics*,*118*, 26-40

Box, G. and Jenkins, G. (1976). Time series analysis. San Francisco: Holden-Day.