INTERNATIONAL HELLENIC UNIVERSITY

Twitter and Traditional News Media effect on Eurozone's Stock Market

MSc in Environmental Management and Sustainability

International Hellenic University

Author: Andreas Koutsouridis

Supervisor: Dr. Theologos Dergiades

Abstract

This study aims to shed light on the relationship between online activity in news media sources and stock market reaction. More precisely, I examine whether the use of the keyword "Grexit" on Twitter and Traditional News media has an effect on the Greek stock market. Afterwards, the potential contagion is investigated for the GIIPS countries. For the purpose of this analysis, Mixed Frequency VAR and Mixed Frequency Granger Causality Analysis is utilized to reveal the supremacy of mixed frequency analysis over common frequency sampling techniques implemented in the low frequency. I find that the Twitter "Grexit" Granger cause the Greek stock market in the short run, as well as, contagion effects are present in the other GIIPS stock markets.

Keywords: Behavioral Finance, Sentiment Analysis, Mixed frequency VAR, Mixed frequency Granger Causality, Grexit, Stock Market Indices, Eurozone, GIIPS

Table of contents

1. Introduction

From the beginning of the 21st century, the expansion of internet signifies the blooming era of social media and the emergence of big data analytics. The utilization of social media platforms and online news media have created vast amounts of data which could be used for a variety of scientific fields and business purposes (Liu, 2012). This condition has allowed web-based technologies to be embraced by firms that have exploited the promising environment for business opportunities through sentiment analysis and behavioral finance.

One of the leading social media platforms, Twitter, was founded in 2006 and since then has seen general acceptance from the public and especially from the financial community. As reported by Chung and Liu (2011), the registered users until 2011 reached 200 million and more than 1 billion tweets were posted every 5 days in this platform. Until the beginning of 2013 the total amount of users reached more than 500 million (Sul *et al*., 2014).

Based on the literature, multiple research about a wide variety of subjects has been conducted by utilizing the available data from Twitter, other social media platforms and online search engines, such as Google. Briefly, Asur and Huberman (2010) utilize data from Twitter by filtering the public sentiment from tweets in order to forecast box office collections. In their analysis the applied prediction outperforms the conventional market-based predictors. In another study, Aramaki *et al.* (2011) utilize Twitter to predict the transmission of diseases such as flu. O'Connor *et al.* (2010) examine the correlation of tweets with presidential polls by creating a sentiment index. Other analysts, such as Goel *et al*. (2010), have utilized online search engines to predict consumer behavior regarding films, music and games and Askitas and Zimmermann (2009) uses data from google searches about specific keywords in order to forecast the unemployment rates in Germany. In a similar manner, other researchers, as it is thoroughly presented in the section of literature review, examine the correlation of sentiment extracted from online activity with stock market performance.

Prior to the evolution of online network, the chain reaction from the price change of a stock to the interaction and the dissemination to the general public demanded a considerable amount of time compared to nowadays where it has a rapid effect through the use of internet and social media platforms as stated by Sharpley (2002). In a similar fashion, Acemoglu *et al.* (2010) indicates that markets required a significant amount of time to calm the public in cases of fraudulent rumors and false information.

With respect to the stock price prediction and the forecasting of stock market index, the topic is surrounded by great controversy. The two main pillars regarding stock price prediction are Efficient Market Hypothesis (EMH) and Random Walk (RW). Efficient Market Hypothesis assumes that rationality of investors is valid and that the current level of stock's price incorporates all available information. This information embodies news and events that have the ability to influence stock market. On the contrary, Random Walk theory supports that stock price prediction is random due to the randomness and the unpredictable nature of news, as stated by Pagolu *et al.* (2016).

The progress of social media and the abundance of relevant information incited the academic community to examine the potential correlation of social media and stock market. This condition allowed the analysts to question the principle of rationality and check the validity of behavioral finance. The notion of behavioral finance highlights the correlation of market performance and public mood. This theory, according to Makrehchi *et al*. (2013), rejects the rule of rationality among people and, hence, supports that public mood has an effect on market decision. In the case of social media, user's online activity that could reflect emotional condition, could be utilized as a predictor of a specific stock or market's index.

The effect of social media activity in Twitter could be traced in the fact that political leaders and owners of multinational companies use it as a news media instrument in order to spread information. One incident of Twitter effect on market, is the announcement from President Trump regarding Turkey and the doubling tariffs on Steel and Aluminum that caused the sink of the Turkish Lira and the activity of Elon Musk on Twitter that caused Tesla's stock to jump more than 6 percent. Similarly, fraudulent events could amplify the correlation of online activity with stock market. One clear example is the incident of Associated Press Hoax in 2013. In this case, on 23 April 2013, a fake tweet from Associated Press claimed the explosion of White House which resulted in the plunge of Dow Jones Industrial Average for more than 100 points. These events support the theory that online activity has the ability to affect stock market in the short run.

In this paper, within the scope of sentiment analysis, it is investigated whether the use of the keyword "Grexit" in social media and Traditional news media is associated with reactions in the stock market. The notion for this correlation is rooted in the fields of behavioral finance. Based on this emerging science in the sphere of economics, it is assumed that the use of this keyword reflects the concern of the public, as well as the stress of the financial community for periods of significant risks in the economy. Consequently, as the political and financial tension is rising, the augmenting use of keywords that reflect this risk is emerging. As presented in the work of Milas *et al*. (2018), it is pointed that several periods of high activity in Traditional and Twitter activity are related to economic and political crisis in Greece.

The main hypothesis of the analysis that is examined, is whether social media content and especially the use of the keyword “Grexit” influences the movement of the stock market indices. Within this broad hypothesis, several other assumptions are being tested. Firstly, the relationship of the Twitter data and the Traditional media sources as well as their potential effect on the Greek stock market is investigated. Apart from the effect of the keyword “Grexit” in the Greek stock market, it is examined whether contagious effects are present in other Eurozone member states. In order to examine the potential supremacy of Twitter over Traditional News media, the orthogonal Twitter effect that isolates the additional power of Twitter activity is taken also into consideration.

An additional hypothesis is whether GIIPS stock markets indices are affected in a higher degree from the Twitter and Traditional news activity in comparison with the more robust economies of the union. The majority of GIIPS states belong in the Southern Europe and they are selected due to the similarities in the economic, social and political turbulences after the burst of the global economic recession of 2008. For the purpose of a thorough investigation of potential contagion effects, apart from the major stock market indices of the Eurozone members, secondary indices are utilized. Through this addition the report has the capability to observe in a greater depth the effect of “Grexit” in firm’s stocks with different sizes in terms of capitalization. Finally, the last hypothesis tests whether the results of the analysis are in line with the findings of the Milas *et al*. (2018). More precisely, it is expected that the use of the keyword Grexit will similarly have a significant short-run relationship with the stock market as in the previous research regarding the relationship of the “Grexit” keyword with the EU spreads.

For the purpose of evaluating the results of the Mixed Frequency Granger Causality Analysis, the common frequency regression and common frequency VAR will be initially explored. Based on the advantages of Mixed Frequency Granger Causality test stated in the work of Ghysels *et al.* (2016), potential hidden correlations that are not distinguishable in low frequency formation, can be unveiled.

The findings of this paper agree with the remarks of Ghysels *et al.* (2016). The transformation of data from high to low frequency variables and the application of corresponding common frequency Granger Causality test, indicates no correlation between the “Grexit” data from Twitter and conventional media with the Greek stock market index. On the other hand, through the results of the Mixed Frequency Granger Causality Analysis, it is concluded that the Twitter activity regarding the use of the keyword “Grexit” explains the movement of the Greek stock market in the short run. Even though the traditional news media sources have insignificant effect in the Greek stock market, they seem to have the same insignificant effect in the stock market indices of GIIPS. Regarding the supremacy of Twitter data that is reflected through the orthogonal Twitter variable, it is resulted to be significant in most of the cases which underlines the value of Twitter data in the explanation of stock market and the significance of behavioral finance.

The GIIPS countries that have similarities with the Greek economy seem to react in a similar fashion with the use of the Twitter and Traditional news data. With respect to the analysis of several stock market indices that differentiate from the major index in terms of capitalization, it is observed that high capitalized firms that belong in GIIPS get mostly affected by the "Grexit" data and as the size of the capitalization size shrinks, the effect weakens. Last but not least, regarding the analysis of Milas *et al*. (2018) the findings of this research agree with the stated hypothesis. More precisely, the results of the Granger Causality Analysis in the work of Milas *et al*. (2018) highlight the direct and primary effect of Grexit data with the macroeconomic factors such as the spreads of the EU bonds which is in accordance with its secondary effect of Grexit data in the stock market. Hence, the findings of this analysis report similar results with the previous research and share the significant short-run effect of Twitter activity on stock and debt market.

In this section the main notion of behavioral finance, the evolution of news media, as well as the goal of this analysis is introduced. In section 2, the literature review demonstrates a wide variety of research in the fields of behavioral finance that examines the effect of social media, news media and online search engines on stock market. In section 3, the methodology of this empirical analysis is presented. The dataset along with the sources of data and their respective transformations are concentrated in section 4. The results and the concluding remarks of the analysis are displayed in sections 5 and 6 respectively.

2. Literature Review

In this section of the analysis, the retrospection of research related to behavioral finance is conducted. The following approaches mainly examine whether stock market activity could be reflected by sentiment analysis. Through literature review, several major stock market indices, sector indices, as well as, individual stocks are examined with a broad spectrum of techniques from linear models to machine learning techniques. Apart from the returns of the stocks, several analysts examine the presence of online activity on traded volume of stocks and the volatility of their returns. The sources and the transformation of online activity to sentiment data vary among literature, but principally the selected analysis encompasses research that utilizes data from Twitter and Google Trends.

In this early research, Fisher and Statman (2000) examine whether the sentiment extracted from individual investors, newsletter writers and Wall Street strategists could have an effect on stock market. The data regarding the Wall Street strategists which reflects the large investors is collected from Merrill Lynch and for the medium sized investors the Chartcraft Company provides data for 130 newsletter writers. The weekly data for individual investors is extracted from the American Association of Individual Investors (AAII) from September 1985 until July 1998. For the same period, the stock prices of S&P 500 as well as the CRSP 9-10 Index is collected to reflect the proxies of the large-cap and small-cap companies. The correlation of the sentiment of these groups identifies positive correlation between individual investors and newsletter writers but negative between Wall Street and the other two implying that large investors remain unaffected by small and medium investor's sentiment. Thus, the investor groups do not follow the same sentiment pattern. By forming scatterplots of the sentiment groups and the stock market future returns over 1 month it is resulted that sentiment of small and large investors are significantly and negatively related with the future returns of the S&P 500 Index. When the sentiment data is combined for all the three groups for the prediction of future returns it is highlighted that sentiment explains 8% of the stock market returns which could be useful approach for an asset allocation strategy. In contradiction to the small and medium investors that get affected instantly by stock market returns, it is observed that Wall Street strategists do not get significantly affected by short-term activity of the market.

Sehgal and Song (2007) investigate whether the sentiment data collected from message boards have predictive power on the stock market. By proposing an innovative tool called "TrustValue", the analysts overcome the problem of irrelevant messages that is common in message boards and attribute a score for each user of the message board. The score of the user increases when the sentiment of its messages is in accordance with the stock market, hence the authors are able to distinguish the relevant information in the data from the noisy messages. The online message data as well as stock prices are collected from Yahoo Finance for 52 stocks for a broad spectrum of companies from Oil to Technology sector, for a 6 months period. By utilizing the Weka toolkit that provides all the standard classifiers such as Naive Bayes, Decision Trees and Bagging, the messages are classified in the categories of: Strong Buy, Buy, Hold, Sell and Strong Sell. Regarding the results of the research, it is concluded that sentiment is correlated with the stock market and that data extracted from the "TrustValue" tool could increase the accuracy of the prediction of the Stock Market in the short run. When the authors include the "TrustValue" in the prediction model with the reduced noise from the message data, it yields 9% more accurate results.

Gilbert and Karahalios (2010) examine whether negative emotional condition from online community has an effect on stock market. The daily data for the sentiment analysis is extracted from the site LiveJournal which is a web-blog that contains a wide variety of topics not necessarily related to finance. The dataset is divided in 3 different periods in 2008: i) 25th of January – 13th of June, ii) 1st of August to 30th of September, iii) 3rd of November to 18th of December. For the classification of the posts into the emotional categories of anxiety, fear and worry and the construction of the Anxiety index, computationally identified linguistic features are utilized. As a proxy for the stock market, the closing prices, volume and volatility of S&P 500 index are used. The regression that is used is a linear-type and the equations are estimated by using OLS. By performing the Granger Causality Analysis, two models are compared where the first model contains only historical information up to 3 days about the S&P 500 Index and the second includes in addition lagged values of the Anxiety Index.

As it is resulted, the extended model performs better compared to the first model which implies that as the anxiety increase, the investors become more risk averse and the stock market is negatively affected. More precisely, if the Anxiety Index increases by one standard deviation, then the S&P 500 returns will decline 0,4% more than it is expected in actual returns. Furthermore, to overcome the issue of heteroscedasticity and non-normal residual distributions, the researchers perform Monte Carlo simulation in order to provide more robust results. The results of the simulation agree with the previous remarks. From the reversed Granger Causality, it is concluded that only Anxiety Index has an effect on S&P 500 and not vice versa. The limitations of this analysis refer to the special case of

2008 and the financial crisis, and more research should be done in order to investigate whether similar results appear when economy is more robust.

Zhang and Skiena (2010) examine potential relationships between data extracted from news sources such as newspapers and 3 blog sources, with the stock market. The data from news and blog sources are extracted with a high-speed text processing system called "Lydia" for 500 newspapers, Twitter, Spinn3r and LiveJournal blog and the content is classified into negative and positive. Due to the unavailability for several news sources, the time series vary in terms of observations. Based on the online activity, the polarity and subjectivity metrics are calculated. Regarding the stock market, the daily and monthly closing prices of 3238 stocks of the New York Stock Exchange (NYSE) is collected during the period of 2005 until 2009. By testing several correlations, it is observed that there is correlation up to 0.4 for logged normalized news data and logged stock market traded volume. In addition, as the market capitalization increases, the news feed increases accordingly for these companies. The polarity has a significant correlation with stock returns and subjectivity has a positive relationship with traded volume. It is observed that news from newspapers has no predictive power on stocks. On the other hand, sentiments from blogs are more persistent compared to news. Finally, correlations could be strengthened with the exclusion of neutral sentiment companies. Hence, in the last part of the analysis, the authors construct the market-neutral strategy which yields higher returns compared to worst sentiment and random selection strategies. This result underlines the importance of sentiment analysis for the prediction of stock market movement.

Zhang *et al.* (2011) investigate the correlation of twitter content and stock market indices. More precisely, they make an effort to forecast the following 3 market indices:

Dow Jones, NASDAQ and S&P 500, by utilizing daily tweets from twitter during the period from 30th of March 2009 to 7th of Sept 2009. The number of tweets per day, as well the number of followers and re-tweets are taken into account. After the filtering process and by attributing emotional content based on the usage of positive and negative words, they conclude that emotional tweets could be useful predictors of the stock market. The list of these words is: hope and happy for the positive ones and fear, worry, nervous, anxious, and upset for the negative ones. With respect to the results of the analysis, it is observed that emotional tweets are negatively correlated with the dependent variables but have a significant positive correlation to VIX. As it is underlined by the researchers, when the general public is overwhelmed by positive feelings, it is when people express a lot of emotions that are depicted through tweets containing words such as hope, fear, and worry. In this case the Dow Jones reacts by declining the following day. On the other hand, when people are more reserved and neutral, the Dow Jones is moving upwards.

Bollen *et al.* (2011) examine the correlation of Twitter feeds with Dow Jones Industrial Average (DJIA). The daily data concerning tweets is recorded from 28th of February 2008 until 19th of December 2008 including 9,853,498 tweets posted by approximately 2.7 million users. During the filtering process, Opinion Finder is utilized as a tool in order to attribute a positive, negative and neutral mood as well as the Google Profile of Mood States (GPOMS). By using the second process (GPOMS), 6-dimensional series are created that describe a calm, alert, sure, vital kind and happy mentality of the public. The report concentrates on two significant occasions of the period of the analysis which are the presidential election in US and the Thanksgiving Day. To perform the analysis, they conduct Granger Causality Analysis, as well as, Self Organizing Fuzzy

Neural Network which is primarily used to treat for non-linear relations among the variables and the non-linear time series. As it is concluded, the sentiments are predictive instruments of the DJIA. More precisely, changes in emotions could predict movements of the indicator after 3-4 days. The accuracy of the prediction of the stock movement reaches 87.6% when the model contains sentiment data. Calm and Happiness features provide better forecasts for the DJIA but this fact is contradictive with the positivity extracted from the Opinion Finder tool that do not yield similar results.

Chung and Liu (2011) examine the causal link between Twitter activity in the field of technology and the hourly stock price of the most and least successful companies in this specific sector. More precisely, they test the hypothesis that more tweets will occur for the successful firms as well as the opposite scenario. The data for this analysis is collected from Twitter and Google Intraday Stock Prices during the period of 29th of November 2011 until 2nd of December 2011. For the collection of the tweets, AND/OR and quotations style Boolean search method, as well as subject (hashtag) searches from users on Twitter is utilized. In order to attribute a sentiment in each tweet the Hu & Liu's list is utilized that classifies the data into positive and negative. The stock data is on 1-minute level but it is averaged to hourly level and finally it is standardized in order to be comparable for the 10 companies stocks. Based on the results of the analysis and the interpretation of the cross-correlation, it is observed that tweets are positively correlated with the high gainers of the technology sector and precede by 7 hours the stock reaction. The correlation is even higher with the prices change and the resulted Twitter activity. On the other hand, for the least successful companies, no significant correlation is observed which implies that in case of inefficient investments online users are more reserved and as a consequence, they do not publicly share online content related to their investment activity.

Mao *et al.* (2011) examine the relationship of sentiment extracted from several online sources with market indices, volatility measures and gold prices for both daily and weekly frequency. Regarding the public sentiment, the sources of online data that are taken into account are the Twitter, 8 news outlets and data from Google trend in order to form the 6 sentimental indicators in a daily (Twitter) and weekly frequency (Google). These indicators are: DSI bullish percentage, Investor Intelligence (II), Twitter Investor Sentiment (TIS), Tweet volumes of financial search terms (TV-FST), Negative News Sentiment (NNS) and Google search volumes of financial search terms (GIS). The dependent variables are consisted of DJIA, traded volumes, market volatility (VIX) and gold prices.

In the first part of the analysis where the search volume and the financial indicator correlation are held, the GIS and the dependent variables are examined during the period of January 2008 until September 2011 in a weekly frequency. Based on the Pearson correlation, it is observed that there is a strong negative relationship between GIS and DJIA and a positive one with GIS and VIX implying that when people search about a specific indicator, they feel unsecure. From the cross-correlation process and the performed Granger Causality Analysis, it is resulted that GIS has predictive power over market indices. After this part, a forecasting analysis is held with a 1-step ahead prediction over 20 weeks that concludes that data from Google trends about financial terms could provide better forecast based on MAPE in periods when the market is volatile and exhibits high trading volumes.

In the next section of the analysis, the researchers examine the daily correlation between the 4 mood indicators of Twitter and the dependent variables during the period from 1st of July 2010 until 29th September 2011. For this part, Granger Causality Analysis, multiple regression and forecasting analysis are conducted. It is concluded that the market returns as well as VIX are correlated with all the mood variables and that the inclusion of twitter data could forecast better the directional movement of the dependent variables. In addition, TIS and the TV-FTS lagged values could explain the current market returns while the other 2 indicators are less statistically significant.

Ruiz *et al.* (2012) examine the effect between the activity in Twitter and the price and traded volume of stocks. By using 150 randomly selected companies in the S&P 500 index, they collect daily data for their traded volume and closing price for the first half of 2010. The Twitter data is collected based on tweets related to these companies and the features of the analysis which create two categories of data. The first category is related to the total activity and the second is related to the induced interaction graph. The features of the second category are more correlated with the traded volume of stocks rather than with the price of the stocks. As it is concluded, the researchers perform stock trading simulations and highlight the fact that even the correlation of stock price with Twitter activity is weak, it improves investor's portfolio.

Mittal and Goel (2012) examine through machine learning techniques whether sentiment analysis is related to stock market. This research is conducted in a similar fashion with the previous work of Bollen *et al.* (2011). The dataset contains tweets and closing prices of DJIA from the 6-month period of June 2009 until December 2009. The Twitter feeds are classified into the categories of Calm, Happy, Alert and Kind by utilizing the

Profile of Mood Stated (POMS) questionnaire. From the Granger Causality Analysis, it is resulted that the lag values of 3 or 4 days in mood affected the present value of the DJIA. Calm and Happy lagged values of 3 and 4 days are both significant for 10%. After these results the authors examine potential non-linear relationships by utilizing Logistic Regression, SVM and Self Organizing Fuzzy Neural Networks. While SVMs and Logistic Regression perform poorly, the accuracy is approximately 75.56% when the Self Organizing Fuzzy Neural Networks and the SOFNN algorithm, using k-fold sequential cross validation are used. This fact confirm that the lagged values of 3-4 days of Calm and Happiness are causal factors of the DJIA. To conclude, the researchers apply the above strategy in a portfolio management and the result yield a profit for a total of 40 days. Compared to the work of Bollen et al. (2011), this paper shows higher correlation of Calm and Happy with the values of DJIA. In addition, due to the technique of k-fold sequential cross validation it is even more evident that the resulting remark is present not only in a sub-sample but behaves accordingly in the entire dataset.

In their research, Mao *et al*. (2012) instead of concentrating on sentiment, they examine whether S&P 500 stock indicators are related to the daily number of tweets that mention Standard & Poor 500 stocks and whether Twitter could be used to predict this specific stock indicator. Three approaches are used for the description of the market from the stock market level to the individual company stocks level (Apple Inc. stock). The daily data from Twitter is collected from 16th of February 2012 until 10th of May 2012 and linear regression model is utilized for the regression analysis. The results of the analysis demonstrate that the daily number of tweets are correlated with certain stock market indicators at each level and that Twitter could be used as an additional tool to forecast stock market.

More precisely, at the stock market level, S&P 500 movement could be forecasted with lower RMSE when Twitter data is included in the model. The daily number of tweets that mention S&P 500 stocks is significantly correlated with S&P 500 daily closing price and it is also correlated with S&P 500 daily price change and S&P 500 daily absolute price change. In respect of the sector level, 8 out of 10 sectors of the S&P 500 traded volume are highly correlated with the corresponding tweet for each sector. In the company level, the traded volume of Apple's stock as well as the absolute price change is strongly related with the tweet volume for this specific company. Regarding the forecasting in the last part of the analysis, for the stock market level and the sector level the forecasting accuracy of the indicator movement is 68% for the following 20 days, which is significantly larger than the random guess. On the contrary, for the company's stock level the prediction accuracy of the daily traded volume is only 52%.

Chen and Lazer (2013) examine the relationship of micro-blogging activity with the movement in stock market. This analysis is related with the paper of Bollen et al. (2010) and is aiming to test whether a more simplistic model could yield similar results. For the classification process of the Twitter data, the SentiWordNet tool is used instead of the more simplistic Alex Davies word list that attaches a "happy" and "sad" feature to each token of the tweet. The regression of the model is a linear regression and the results indicate that the lag of 3 days for Twitter data could yield up to 70% accuracy. In the next part of the analysis, 2 investment strategies that contain sentiment data are tested with the use of 2 simulations that have different training periods. The Classification strategy considers only the sing of the predicted value of the stock which implies that if it is positive the investor can purchase the maximum number of stocks with his capital. The Regression strategy is more complex and determines the amount of investment based on how certain the predictions are. Based on the results of the analysis and comparing them with the default strategy, it is concluded that both strategies increase the size of the invested fund.

Smailovic *et al.* (2013) investigate whether sentiment analysis extracted from Twitter, could forecast the movements of the closing prices of stocks. The daily tweets of 152.572 in total are collected during the period of 11th of April 2011 until 9th of December 2011. For the tweet classification process, the authors utilize the SVM (Support Vector Machine) tool in order to attribute a neutral emotion after the positive and negative emotion to each tweet and extract the ratio of positive sentiment probability. For the correlation between tweets and stock market indices, granger causality is performed for lags from 1 up to 3 days to check whether tweets affect stock market and vice versa. The corresponding p-values are significant for the cases of Netflix, Amazon, RIM and Microsoft but are not significant for Apple, CISCO and Google. The neutral zone has the most positive effect on the results of Baidu and the least positive effects on the results for RIM. This indicates that Twitter data could be an additional instrument for the prediction of stock prices, especially when the closing prices have a big decline or when the market is significantly volatile. Regarding the neutral zone, as it is stated by the authors, once there is no obvious declining trend of a stock the neutral zone could yield better forecast, but on the contrary, when the decline is apparent then its usage is not needed.

Si *et al.* (2013) examine whether extracted sentiment analysis from Twitter feeds, could be related with the stock market and could increase its forecasting accuracy. This research is related to previous work of various authors, such as Bollen et al. (2011) and Ruiz et al. (2012). The daily data from Twitter and the S&P 100 index which is used as a variable for the stock market are collected during the period of 2nd November 2012 until 7th of February 2013. Through the use of the Dirichlet Processes Mixture (DPM) model which is a non-parametric topic-based sentiment model and the use of lexicon O, the Twitter feed is filtered. With respect to the regression, the researchers use vector autoregression model (VAR) to regress the S&P100 index with the sentiment time series and also they apply different training window sizes ranging from 15 to 30 days. From the results of this analysis, it is concluded that the use of Twitter data that reflect emotional condition could benefit the forecasting accuracy of the S&P 100 Index.

Arias *et al.* (2013) examine whether data collected from Twitter could increase the prediction accuracy of commercial, economic and social indicators. The daily data is collected from Twitter during the period from 20nd of March 2011 until 20th of November 2011 for stock market and movie box office revenue based on specific use of words and ticker symbols. After the filtering and the processing of the public opinion data, the Sentiment Index is created which reflects the progress of the mood for a specific item. From the stock market data, the stocks of Apple, Google, Yahoo!, Microsoft as well as market indices such as S&P100, VXO, VIX and the historic volatility of each of the firms. The authors regress 3 different groups of models such as linear, SVM and neural networks. By utilizing a decision tree-based technique the researchers have the ability to pick the best forecasting models from a variety of modifications. The models are compared with their accuracy level, their Cohen's Kappa measure and their directional measure. As it is highlighted in this work, the non-linear models such as SVMs and Neural Networks yield higher accuracy by utilizing the data derived from Twitter in the form of tweets volume or public sentiment index. On the contrary, the accuracy of linear models lacks in the case of financial indicators. More precisely, SVMs with sigmoid kernel could predict volatility indices with either Twitter index (sentiment or volume) more accurately than machine models. Finally, neural networks provide with more precise forecast level when the order of lags is higher or equal than 3.

Makrehchi *et al.* (2013) investigate the potential predictive power of Twitter data on the movement of stock market indices. For the Twitter data, the tweets based on companies of the DJIA30 index are collected during 27th March 2012 until 13th 2012. For the detection of sentiment in each tweet the analysts apply the unsupervised lexicon-based mood and detection, as well as the approach which is based on supervised learning in order to explore peaks of emotions that signify good days and troughs which present the bad days. Regarding the dependent variable, daily prices for S&P 500 and computed daily returns as close/open− 1 are calculated. By computing the net sentiment, it is observed that there is a positive correlation between extreme sentiments and returns. When the public sentiment is extremely positive then the returns are significantly high. The opposite case is also valid. By applying a portfolio strategy that takes a long position if the previous day's sentiment is positive and the opposite, it is resulted that it yields higher returns compared to the unsupervised lexicon strategy and the holding of the index. Regarding the forecasting of the S&P 500 and the individual stocks, the aggregated sentiment is used from the analysts. From the forecasting of the index with the inclusion of sentiment, a training set of last ten days is calculated daily and regressed upon. With the use of supervised learning method, the returns are 15% for a period of 4 months. The results are in line with the prediction of individual stocks, where the long position is chosen when the sentiment is higher than the average and similarly the short position in the case where the sentiment is lower than the average.

Preis *et al.* (2013) utilize data from Google searches related to finance in order to investigate potential correlation with stock market. More precisely, they collect the weekly volume of 98 search terms closely related to finance by utilizing Google Trends for the period of January 2004 until February 2011. As a dependent variable, the Dow Jones Industrial Average (DJIA) has been used. For several words from the list, it is concluded that increased search for these words which reflect the investors' concern, follow a fall in the value of the stock market. By utilizing the dataset from Google, both long position and short position strategies yield higher returns than random investment strategies. As an example, by running a simulation of 10000 iterations, the Google Trends strategy of the term "debt" increases the investor's portfolio by 326%.

Oliveira *et al.* (2013) examine the potential relationship between micro-blogging activity and the returns, volatility and traded volume of stocks and stock market indices. By aiming to create a more robust assessment of the analysis compared to previously related studies, they utilize data from the blog StockTwits which is specifically used by the financial community. In addition, they use a larger period of analysis and a fixed-sized rolling window that leads to 505 predictions. Regarding the market data, the daily adjusted closing prices of the 5 largest US companies and the index is collected during the period of 1st of June 2010 until 31st of October 2012. Afterwards, the differenced logarithms are calculated for all the dependent variables. The data from StockTwits is collected based on the cashtag "$" for Apple (AAPL), Amazon (AMZN), Goldman Sachs (GS), Google

(GOOG), IBM (IBM) as well as for the index Standard and Poor's 500 Index (SPX). From the online network source, the sentiment analysis and the posting volume is recorded. The regressions are based on multiple regression models for 5 different approaches which are compared based on their RMSE and MAPE. These approaches contain different indicators created from the StockTwits data such as: bearishness or bullishness, bullishness index, twitter investor sentiment (TIS) and TIS ratio. The results of this research highlight the fact that there is no relationship and predictive capability between sentiment and returns or volatility of the stock market. On the other hand, the volume of posts in the micro-blog could be utilized for the forecasting of the volume of traded stocks which could be used for the measurement of the stock liquidity.

Sprenger *et al.* (2014) investigate the association between sentiment from micro-blogging and stock market, as well as the mechanism of information diffusion in Twitter. In more detail they search for a link between: i) sentiment and stock returns, ii) volume of messages and trading volume and finally iii) disagreement in tweets and volatility in stock returns. The 250.000 daily tweets with a direct reference to the stock market with ticker $ from Twitter are selected during the period of 1st of January 2010 up to 30th of June 2010. This specific period is selected in order to deal with stable condition on the US financial markets and for netting out potential exogenous parameters that could affect the results. For the purpose of the filtering of the tweets, text classification methods are utilized in order to classify messages as either buy, hold or sell signals by implementing the multinomial Naïve Bayesian implementation of the Weka machine learning. Afterwards, by utilizing the Twitter data, the bullishness index is formed. As a dependent variable, the researchers use the S&P100 stocks to adequately describe the whole range of the largest and most established companies in the United States. The results conclude that bullishness is positively correlated with the market prices. Furthermore, message volume lags predict current day trading volume and positive correlation is observed between tweet volume and traded volume, but this is not the case for the relationship between message volume and stock returns. In addition, high volatility lead to increased message volume but the opposite relationship does not hold. Higher volatility in the stock market, contribute to disagreement between Twitter users and disagreement among traders that result to higher trading volumes. As it is stated, quality and content of tweets prevail over the quantity, since bullishness is related to returns more strongly than message volume. In the last part of the research where the mechanism of information weighting and diffusion is analyzed, it is found that users sharing high quality investment advice receive greater attention through higher levels of retweets and also gain a larger audience in the social media.

Rao and Srivastava (2014) investigate whether an efficient short-term hedging strategy can be constructed by utilizing public sentiment data. Based on previous work of Bollen *et al.* (2011), Gilbert *et al.* (2010), Zhang *et al.* (2011) and Sprenger *et al.* (2014), the authors extend the previously related work in order to seek potential hedging opportunities with the use of online data from social media. The dataset with daily frequency for public opinion is collected from Twitter during the period of approximately one year from June 2nd, 2010 until July 29th, 2011. During the filtering and classification of the tweets the features of positive, negative, bullishness, message volume and agreement are created through the "Twittersentiment" service and the Naïve Bayesian classification method. The variables that reflect the stock market are the DJIA and NASDAQ-100 indices, as well as 11 big cap companies in the field of technology. From the correlation matrix, a correlation of 0.88 occur between the returns of stocks and the sentiment data extracted from Twitter. By performing Granger Causality analysis, it is observed that in short term, public sentiment negative or positive with lags from 1 up to 3 weeks has an effect on the stock prices and indices. By applying the Expert Model Mining System (EMMS) the analysts provide accurate forecasting for the stock market variables and observe more efficient prediction with the inclusion of Twitter data. Based on the results, the MaxAPE of DJIA is 1.76% and the directional accuracy for the same index is 90.8% and 82.8 for NASDAQ-100. In the last part, the performed hedging strategy with the inclusion of public sentiment provides promising results for short term hedging strategies. The results of the directional accuracy in the hedging strategy provide up to 91% accuracy which exceeds the accuracy of 86.7% in the work of Bollen *et al.* (2011).

Curme *et al.* (2014) examine the hidden relationship of article search on specific topics, with the stock market movement. More precisely, they focus on the construction of a method to obtain the search categories that users select before extreme changes in the stock market. The weekly data regarding the online activity is extracted from Google, Wikipedia and Amazon during the period of January 2004 until December 2012. By extracting list of words from semantic topics on Wikipedia and engaging users in Amazon Mechanical Turk, as well as receiving search volume from Google Trends, the authors implement previous methodology from Preis *et al.* (2013) to find the search categories that precede the significant movement in stock market and to apply a corresponding trading strategy. As a dependent variable the S&P500 index is utilized. Based on the trading strategy from Preis et al. (2013), it is highlighted that when the trading option takes into consideration the augmented search for the topics of politics and business, it results higher yields for the S&P500 compared to random trading strategy. On the contrary, there is no significant relationship between a variety of search engine topics and the market movement. Finally, it is observed that the increase for the search of politics and business topics, lead to stock market decline which could be translated as a concern of the public about the economy.

Sul *et al.* (2014) investigate the correlation between cumulative emotional content extracted from tweets about specific companies on the S&P500 and the stock returns of these companies. The daily Twitter data is collected based on the ticker symbol "$" for an annual basis from March 2011 until February 2012. For the filtering and the classification process, by utilizing the Harvard-IV dictionary the tweets are divided into positive and negative ones. Regarding the financial data, the closing prices of the firms comprising the S&P 500 are collected. In order to test the main hypothesis that future returns in short and long terms are related with the current emotion of public, the Cumulated Abnormal Return (CAR) variables are created as control variables. Based on the results which are in accordance with the stated hypothesis, it is observed that cumulative tweets have the capability to forecast index return for 10 days in advance. As the number of followers for each Twitter user increases, the impact in stock price is higher in short-term. On the contrary, when the followers are fewer, the longer is the period for the stock price until its reaction.

Li *et al.* (2014) examine the predictive power of sentiment extracted by online news articles on stock market. More precisely, they compare 6 sentiment and non-sentiment models in a company, sector and stock market index level. The daily data from news articles is collected from FINET which is one of the dominant financial news sites during the period of January 2003 to March 2008. The Harvard dictionary and the Loughran McDonald financial dictionary are used for the classification process to the categories of positive, neutral and negative. Regarding the dependent variables, the time series of all the open-to-close price returns of stocks from the Hong Kong Exchange indicator are formed. The SVM method is utilized for the analysis of the data. Firstly, based on the high correlation between market capital of each stock and the news about this specific stock, it is observed that as the company grows in size, the more the media publish news about this company. Based on the results of the analysis, it is concluded that sentiment analysis could be a predictor of stock market prices which is implied by the fact that in all cases, from company to index level, they sentiment approaches yield higher accuracy in the forecasting of the financial index compared to the non-sentiment approaches. To conclude, it is stated that by utilizing only the categories of positive and negative emotions, it is not capable of providing useful knowledge for the forecast of stock returns.

Wu *et al.* (2014) examine the relationship between public opinion extracted from a popular financial website and the stock market. More precisely, they seek potential links from daily posts on "Sina Finance", which is a major finance portal in China with more than 100 million members, with 50 stocks that belong in the military sector. This specific sector is selected due to its significant activity during the period of the analysis. The data that is extracted from the online portal, as well as the stocks returns, and traded volume are collected during 15th of July 2012 until 15th November 2012. The military index is formed, and the volatility of stocks is calculated on a rolling window. The data from the online portal is classified in bearish, bullish and neutral categories. In order to take into account asymmetric links between the variables and the volatility prediction, SVM-GARCH is utilized for the analysis. In the first part of the research, it is observed that machine learning technique provide higher levels of accuracy compared to lexicon methods. Based on the results of the forecasting and after the normalization of the data, it is highlighted that online activity has a delayed effect of 4 to 7 days to stock market and that especially the negative posts have a greater effect in the military index. In the second part of this paper, the firms are separated based on age, firm size and price-to-book value. By running 10 equal-weighted portfolios, it is concluded that there are significant changes regarding the size of the stock, the price-to-book and the public sentiment. The results support that pessimistic posts have higher predictive accuracy on the index than optimistic labels and that the proposed method is more reliable for small stocks due to their volatile sentiment sensitivity.

Nguyen *et al.* (2015) investigate whether the sentiment analysis extracted from the online text message board of the company is able to give useful knowledge for the forecast of the stock market. Regarding the dataset, daily adjusted closing prices of 18 stocks are extracted, as well as data from the message boards concerning specific topics of the companies. This data is collected during the period 23rd of July 2012 until 19th of July 2013. For the filtering and classification process, the data is automatically classified in the following categories: strong buy, buy, hold, sell, strong sell. In order to construct the model that will more accurately forecast the movement of the stocks, the Support Vector Machine (SVM) is utilized for the comparison of 6 different regression models. These 6 models are: i) price only ii) human sentiment iii) sentiment classification iv) LDA based method v) JST based method and vi) Aspect Based Sentiment. The last 2 models are methods that contain algorithms proposed by the authors. Based on the results and comparing the forecasting results for each model, the accuracy of the prediction reaches 54.41% for the average of the 18 stocks when the analysts use the Aspect Based Sentiment method. When stocks are predicted separately, the accuracy reaches higher levels, for example 71.05% for AMZN stock and 64.47% for DELL stock.

Xiong *et al.* (2015) investigate whether data derived from Google searches could benefit the prediction accuracy of the stock market volatility. For this purpose, they examine a Long-Short Term Memory (LSTM) neural network that includes data from Google Trends in comparison with two linear regression models and a GARCH model. By using Google data as public sentiment for macroeconomic issues, 25 times series are created with relative searches to the term “bankruptcy”. As a proxy for the stock market, the S&P500 index is utilized and the necessary time series of index volatility are formed for the period of 1st of January 2004 until 24th of July 2015. Based on the results of the analysis, the more parsimonious LSTM model, yields lower MAPE and RMSE compared to the benchmark models. From this result, the hidden and complex importance of online data for the prediction of volatility is highlighted.

Yang *et al.* (2015) examine the relationship between Twitter and stock market by investigating whether influential Twitter users that belong in the financial community could bond the public sentiment with stock markets. The indices that are selected to reflect the stock market are Dow Jones Industrial Average, S&P 500, NASDAQ and Russel 3000 ETF. The data from Twitter is collected during the period 5th of October 2013 until 5th of February 2014. The use of “SentiWordNet” attributes a score from 0-1 that reflects the positive or negative sentiment respectively. By applying the sentiment analysis algorithm, they create a weighted sentiment measure. The critical nodes that represent the influential experts are consisted of 50 investment experts that are selected based on the frequency of their posts, the usefulness of their tweets and the number of their followers. Before the filtering, 2 more layers that represent their followers are created. From the results of the analysis it is marked that influential Twitter users have the ability to affect market returns as well as market volatility (VIX). The research reveals the positive correlation of public sentiment with market returns in the cases where the mood is positive. In addition, when the mood is positive the stock market is described by lower volatility. On the other hand, when the mood is becoming more negative, volatility peaks are present. Finally, by grouping the Twitter users, the authors achieve more accurate predictions for the financial indices. All these results reflect the importance of sentiment analysis for the stock market prediction.

Ranco *et al.* (2015) investigate the relationship between public sentiments extracted from Twitter about specific companies with the stock market. More precisely, based on the hypothesis that the decision of buying or selling could be derived from a peak in the tweet volume and the level of polarity in the corresponding direction, they examine whether aggregated sentiment collected from financial tweets with a ticker, could provide useful information for the pricing of stocks especially after the period where the Twitter activity is peaking. These peaks could reflect known events such as earning announcements but also unexpected events. The dataset of daily frequency is collected during the period from June 2013 until September 2014. After the manual labeling of a subsample by experts, the tweets are classified with the SVM classification method into the categories of negative, positive and neutral. In addition, the volume of tweets and the sentiment polarity are taken into account by the analysts. The stock market variables are consisted of the 30 stocks of the DJIA index and the differenced logarithms of the prices are calculated. The findings of this analysis are in accordance with the related work of Sprenger *et al.* (2014) where instead of utilizing the SVM, Naïve Bayesian classifier has been used. The Pearson correlation as well as Granger causality is applied in order to reveal the potential dynamics between Twitter data and stock market which is observed to be low in both cases. Furthermore, the polarity variable is observed to be insignificant and the tweets volume Granger causes the 1/3 of the companies' stocks. According to the cumulative abnormal returns (CAR) and the detected Twitter peaks, it is verified that Twitter activity affects stock returns.

Maragoudakis and Serpanos (2016) investigate whether textual information could provide more accurate forecast for the stock market. In order to obtain daily textual data, the Greek financial newspapers "Naftemporiki" and "Capital" are utilized to analyze the articles, as well as, Twitter for the social media analysis. From this data, based on a manual classification from a financial expert, the observations are classified in the following categories: clearly negative, relatively negative, neutral, relatively positive, clearly positive. Regarding the dependent variables, 4 Greek companies' stocks from 3 different fields are collected during the period from January 2013 until January 2014. These companies are: Eurobank, NBG, O.T.E. and Aegean Airlines. In addition, major indices from other main markets and commodities are selected. In order to overcome the instability of data mining occurring when the number of features increase, a Markov Chain Monte Carlo Bayesian Inference approach (MCMC-TAN) is utilized. The primary classifier MCMC-TAN is used in comparison with the traditional Bayesian classifier (NB), an ensemble algorithm called Random Forest (RF), Radial Basis Functions neural network (RBF) and the Support Vector Machine (SVM). For the experiment phase, 2 experiments are executed. The first experiment is a 10-fold cross validation of stock prices with datasets containing technical analysis up to technical analysis plus indices, news articles and public opinion from users. By evaluating the results based on the F-measures and the MAE and RMSE, it is concluded that MCMC-TAN is the most accurate methodology. In the second experiment, the authors perform a portfolio simulation by using the MCMC-TAN model. From the results of the simulation, the MCMC-TAN method outperforms the simple buy-and-hold strategy. The limitation of this method is the high requirements of time for the completion of the algorithms compared to the traditional machine learning algorithms.

Pagolu *et al.* (2016) examine the potential correlation of a tweet about a specific company and its products and services with the stock price of this company. Regarding the dataset, the researchers extract daily stock prices and tweets about Microsoft from the period of 31st of August 2016 until 25th of August 2016. By applying two different classifiers of tweets: Word2vec and N-gram, the tweets are grouped into sentimental categories such as positive, negative and neutral. Through the examination of the sentiment analysis and the utilization of machine learning, it is concluded that positive tweets about a specific company will result on the increased stock price. The results indicate that there is a significant correlation between the sentiment analysis and the stock price as well as that an increasing dataset would provide more accurate results.

Perlin *et al.* (2017) investigate whether finance-related data derived from Google Trend could have an effect in the return, traded volume and volatility of stock market capitalization for 4 English speaking countries. The countries that are included in the analysis are: USA, UK, Australia and Canada. The corresponding market indices that are utilized for these countries are the S&P 500, FTSE, S&P/ASX 200 and S&P/TSX respectively. The weekly data is collected during the period 2005 until 2014. For the filtering of the online data, a web-based finance dictionary in combination with 4 different textbooks create the list of the 15 words that occur most, are more suitable and finally are going to be used in this research. Regarding the regression analysis, vector autoregression (VAR) and Granger Causality are utilized to shed light on the relationship of online data with financial markets and vice versa. As it is concluded, Google Trend data inclusion yields higher accuracy for the prediction of financial markets but not vice versa.

It is highlighted that an increased frequency of the words “stock”, “finance” and ”market” has a positive correlation with the volatility for at least 3 out of 4 market indices.

In the case of market indices returns, it is observed that an increased search for the words “stock”, “finance” and “debt” decreased the market index price and are used usually before a sell strategy from the investors. In this case, the opposite Granger Causality holds. Finally, concerning the last dependent variable, the words “finance”, “value” “journal” and “risk” have a negative effect with market trading volume for at least 3 out of 4 countries. The word “stock” is statistically significant in all countries and models which provide the intuition that an increase in this search could forecast an increase in volatility and decrease in stock prices. Due to the robustness of the word “stock”, it is used in a portfolio strategy. This strategy that utilizes the frequency of this specific word has higher returns compared with 2 other strategies: a naive buy-and-hold strategy and a strategy based on an ARMA–GARCH model. It is worth mentioning that higher predictability of the model during the crash of 2009 leads to the conclusion that these data could perform ideally as warning calls for the prediction of financial crashes.

Weng *et al.* (2018) investigate whether data derived from online data sources could become predictors for short-term stock prices. The daily data are collected from Yahoo, Wikipedia, Quandl and Google trend during the period of 1st of January 2016 until 31st of December 2016 and the stock prices of City Group are used for forecasting. This is the first research to combine these specific data sources. With the use of artificial intelligence (AI) platform that builds the knowledge base from the data sources, 4 ensemble methods of individual classifiers over multiple periods are trained : i) Neural Networks, ii) SVM, iii) Boosted Regression Tree (BRT), iv) Random Forest Ensemble (RFR). Based on the results of this paper, BRT and RFR methods are more accurate for the prediction of the stock prices in short-term. For the evaluation and comparison of the models, the RMSE, MAPE, MSE are utilized. For 1-day and 1-10-day ahead prediction, these methods have the lowest deviation from the real values. In the last part of the analysis, 19 stocks of companies from different fields are used in order to be forecasted with the method used in the City Group case. Again, the results underline the significance of online data for forecasting. It is concluded that information derived from online sources, could benefit the prediction accuracy in the short run as an addition to the traditional models and that as the forecasting period enhanced, the forecasting ability could become more inaccurate.

Broadstock and Zhang (2019) examine the relationship between public sentiment collected from Twitter about a specific stock or an index, and the stock market. By utilizing intraday tick-level stock returns collected during the month of August 2018, the data is aggregated up to 1-,5-, and 30-minute intervals. The twitter data is extracted based on hashtags and cashtags about each firm and it is classified in emotion categories and in the SMOG index. The companies' stocks that are collected for the analysis are from: Exxon Mobil, General Electric, Chesapeake Energy, Ford Motor Company, Disney and Walmart. For the purpose of the regression analysis, the Capital Asset Pricing Model (CAPM) extended with sentiment features. It is observed that prices got affected by sentiments from online users. Based on the results of this paper, the stock returns have clear relationship with sentiment and lagged sentiment terms are significant in each stock.

By focusing on the Greek debt crisis, Milas *et al.* (2018) examine the relationship between traditional news sources and Twitter regarding the use of word “Grexit” and its potential effect on sovereign bond markets. The reaction of the bond markets is not limited only in the Greek spreads, but also possible contagion effects are examined for EU bond markets and especially the GIIPS and France. All the variables of the dataset are collected during the period of 5th of March 2012 until 24th of June 2016 in daily frequency. Due to the widely use of Twitter in the financial community as a source of news media, data is extracted from this specific source based on the presence of the keyword “Grexit” or #Grexit. In addition, data is extracted from traditional sources for comparison purposes. The analysts utilize the Dufour et. Al (2006) causality test which is based on the multiple-horizon Vector Autoregressive (VAR) specification and afterwards they construct the impulse response function (IRF). By dividing the sample into 2 subsamples which reflect 2 periods of significant financial frustration in the Greek economy, the researchers perform the stated methodology which leads to important conclusions. The Twitter data has a more persistent effect than traditional media sources in periods of financial tension which are reflected in high activity of the term “Grexit”. The Greek spreads are affected positively and in greater amount from the Twitter data and regarding the contagion effects, during the full sample period and the first subsample, the effect is weak for the cases of Portugal and Ireland whereas in the second subsample this effect is not present. This paper provides useful knowledge about the relationship of public sentiment with bond markets as well as portrays the use of a specific term in financial events that occurred in Greece during the period of the analysis.

3. Data

The dataset of this analysis contains observations in mixed frequencies which have been collected during the period of 5th of March 2012 until 24th of June 2016. This set of information includes data from social media and traditional media news sources in daily frequency, as well as financial indices of Eurozone members that accurately reflect the stock market in a weekly frequency. For the stacking of data in a Mixed Frequency VAR formation with daily (7 days per week) "Grexit" observations from Sunday to Saturday and weekly data of stock market indices, the final dataset is restricted to the period of 11th of March 2012 until 19th of June 2016.

3.1. Dependent variables

The financial data is extracted from the online database of Investing.com in a weekly frequency. Each financial index that is included in this work is published each Sunday and in terms of occurrence, follows the daily "Grexit" data from Twitter and conventional news media. These times series are collected during the period of 11th of March 2012 until 19th of June 2016 and represent the stock market index for the total amount of 223 weeks. These variables reflect the major stock market indices of the member states of the Eurozone for 11 out of 19 countries of the union and mostly focuses on the GIIPS countries (Greece, Ireland, Italy, Portugal, Spain) plus France. The GIIPS countries comprise the group of European states which were primarily affected by the European debt crisis and during the period of the analysis, they were characterized by their dwindling gross domestic product, their financially unstable condition and their potential economic default. Regarding the transformation of data, each stock market index is transformed into returns by using the first logarithmic differences of the prices. The list of the Eurozone member-states, and their corresponding major and secondary stock market indices that are utilized in this research are displayed in Table 1.

Table 1. Eurozone's primary and secondary stock market indices

Eurozone Members	Group	Major Index	Secondary Indices
Greece	GIIPS	AGC	-
Italy	GIIPS	FTSE MIB	FTSE Small Cap, FTSE Mid Cap
Ireland	GIIPS	ISEQ	ISEQ Small Cap, ISEQ General
Portugal	GIIPS	PSI 20	PSI All-Share
Spain	GIIPS	IBEX 35	IBEX Small Cap, IBEX Mid Cap, General Madrid
France	GIIPS	CAC 40	CAC Small Cap, CAC Mid 60, CAC Large 60
Belgium	Non-GIIPS	BEL 20	-
Finland	Non-GIIPS	OMX Helsinki 25	-
Germany	Non-GIIPS	DAX	-
Netherlands	Non-GIIPS	AEX	-
Slovenia	Non-GIIPS	Blue-Chip SBITOP	-

3.2. Independent Variables

The time series of social and traditional media reflect the daily use of the term "Grexit" or "#Grexit" which is utilized as a secondary data from the analysis of Milas *et al.* (2018). By utilizing the premium Twitter historical database of Followthehashtag,15, Milas *et al.* (2018) collect 936,837 unique tweets that contain this specific keyword. Regarding the traditional news media, the authors utilize the Lexis Academic database in order to gain access to a total amount of 66.246 mentions of "Grexit" through a wide variety of media sources such as: newspapers, magazines, broadcast transcripts from television, radio news as well as wire Services. In order to provide a valid comparison of the Twitter and conventional time series, both of them act in accordance in terms of topic homogeneity, geographical and linguistic coverage. The total amount of daily "Grexit" mentions in Traditional media and Twitter are collected during the period of 5^{th} of March 2012 until 24^{th} of June 2016, which are then transformed into logarithms. Afterwards both series are converted to 7-days matrices for the total amount of weeks in order to perform the mixed frequency vector autoregressive model (MF-VAR) and the mixed frequency Granger Causality Analysis (MF-GCA). After the transformation of data in 7-day matrices, from Sunday to Saturday, for the total amount of the weeks in the period of the analysis, the observations that are utilized are restricted to the period of 11^{th} of March 2012 until 19^{th} of June 2016 for the total 1562 days of the report.

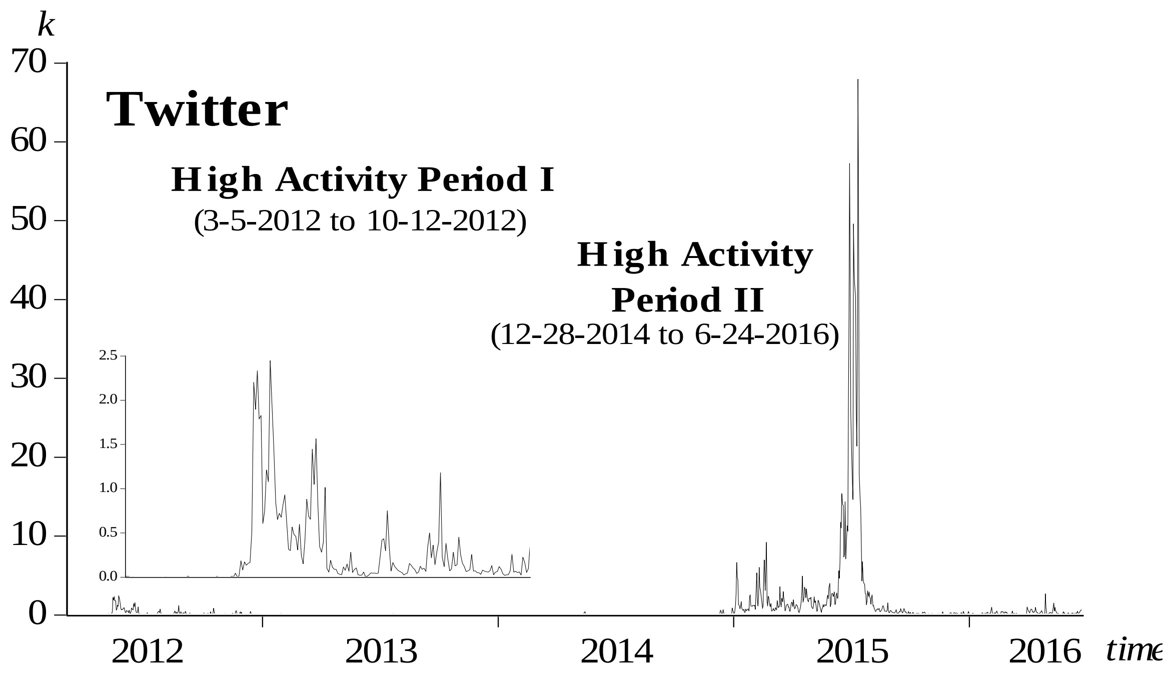

Figure 1. Grexit mentions in Twitter

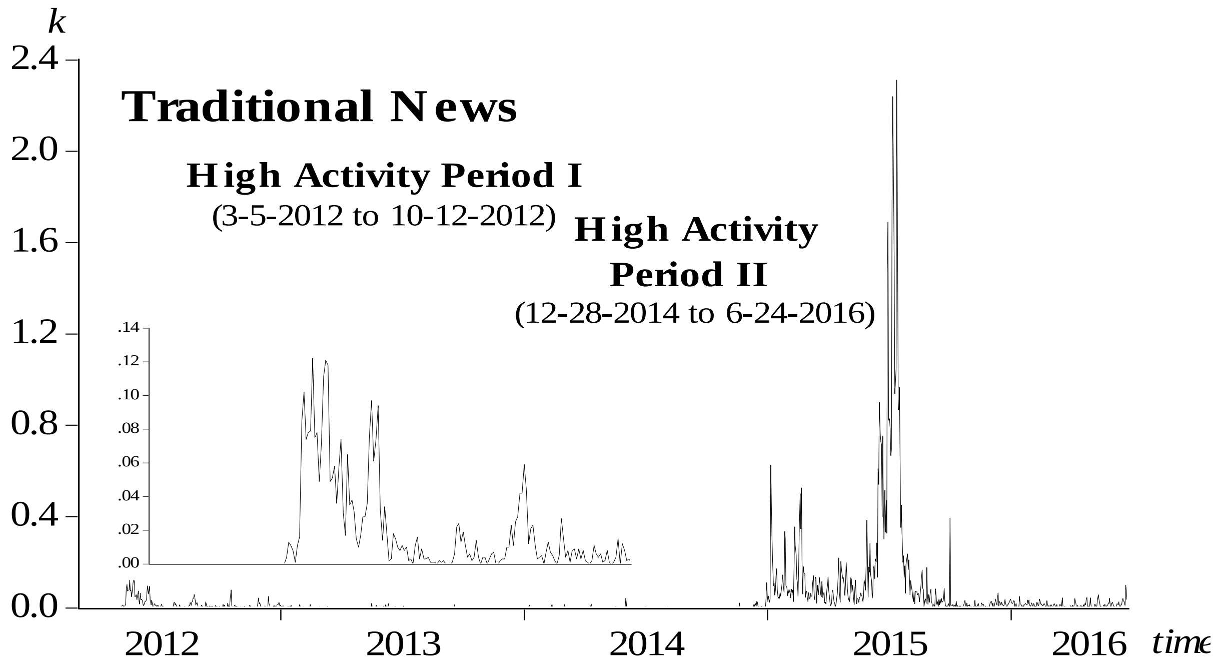

Figure 2. Grexit mentions in Traditional News

As it is represented in Figure 1. and Figure 2. both the Twitter data and Traditional news data behave in a similar manner during the observed high activity periods that took place during the year of 2012 and the year of 2015-2016. The explosive increase of the use of the keyword "Grexit" is related to several economic and political events that jeopardized the membership of Greece in the Eurozone. Both the high correlation of the Twitter and Traditional means variables and the events occurred during the Greek crisis are thoroughly analyzed in the work of Milas *et al.* (2018). The descriptive statistics in Table 1. exhibit the usage of "Grexit" in the social media and the conventional news media. The amplified appearance of "Grexit" observations in Twitter compared to the significantly fewer observations of this keyword in conventional news media, is in accordance with the claim that Twitter reflects directly the mentality of the financial community and probably represents the behavior of the investors in a refined fashion.

Table 2. Descriptive Statistic for the constructed time-series

Statistic	Variable	
	Twitter mentions	Traditional news mentions
	S	S
mean	850.6586	42.11443
median	49	2
minimum	67.948	2312
maximum	0	0
st. deviation	4263.594	166.4308
skewness	9.777915	8.23827
kurtosis	113.5344	85.68952

4. Methodology

For the analysis of the relationship between news sources (Twitter and Traditional news media) and stock market, Granger Causality Analysis is applied. With the utilization of this analysis introduced by Granger (1969), a statistical hypothesis test is performed in order to determine whether a variable $X(t)$ has predictive power on another variable $Y(t)$ by attempting to reject the null hypothesis that $X(t)$ does not Granger-cause $Y(t)$. On the other hand, the alternative hypothesis rejects that $X(t)$ does not Granger Cause $Y(t)$.

As an evolution of the common frequency Granger Causality Analysis, Ghysels *et al*. (2016) feature the unrevealed value of Mixed Frequency Granger Causality Analysis (MF-GCA). By evolving the conventional approach of this test, through the use of Mixed Frequency VAR (MF-VAR) and Mixed Frequency Granger Causality Analysis (MF-GCA), the authors conclude that MF-GCA is indicating the causality relationships more accurately compared to the traditional low frequency method. Another disadvantage of common frequency analysis is that the VAR models that are performed in common frequency variables may also suffer from spurious relationships or generated causality. In addition, as it is stated in their work, the tests performed in Mixed Frequency, have higher asymptotic power over local alternatives which is also proved by a simulation in finite samples. As it is concluded, this approach is efficient when the ratio of high to low frequency (m) frequency is relatively small. In the case of this analysis, the high frequency variable which refers to the “Grexit” data is in daily observations and the low frequency variable which reflects stock market indices is in weekly observations, hence, the ratio of high to low frequency (m) is equal to 7.

The principal notion of the Mixed Frequency VAR models is to arrange the observations in a K dimensional vector. This arrangement is based on their frequency, in the form of $K = K_L + mK_H$ where $K_L < K$. K_L represents the vector of low frequency and K_H represents the high frequency vector that is multiplied by the ratio *m*. More precisely the stacking of the data forms a matrix:

$$X(\tau_L) = \left[x_H(\tau_L, 1)', \ldots, x_H(\tau_L, 7)', x_L(\tau_L)' \right]'$$

As in the part of the analysis, two sampling frequencies are considered where the $x_L(\tau_L)$ represents the low frequency LF (multivariate) process that is stacked in the last block and is observed after the high frequency HF, and the $x_H(\tau_L, k_H)$ is the high frequency process. The high frequency series are represented by a vector process $x_H(\tau_L, k_H)$ which is observed at (high) frequency periods $k_H = 1, \ldots, m$ during period τ_L. Thus, in the case of this analysis, the daily observations for each day of the week that are concentrated as well as the total amount of weekly stock market indices for a total amount of 223 weeks, forms a matrix with dimensions equal to 8×223 and the corresponding Mixed Frequency Vector $X(\tau_L)$ is a 8×1 vector.

Several assumptions have been made in order to proceed to the estimation of the VAR, such as that the process $X(\tau_L)$ is governed by a $VAR(p)$:

$$X(\tau_L) = \sum_{k=1}^{P} A_k X(\tau_L - k) + \varepsilon(\tau_L)$$

all the roots of the $(I_K - \sum_{k=1}^{p} A_k z^k) = 0$ lie outside the unit circle and that ε(τL) has an absolutely continuous distribution with a bounded joint density.

Regarding the Granger causality analysis, the multiple-horizon Granger Causality is performed in this research. This methodology allows the researcher to investigate potential correlations non only in a single horizon but in an expanded period. Based on Dufour and Renault (1998) who expand the single horizon causality to long run or multiple-horizon causality, the authors Ghysels *et al.* (2016) combine this methodology with the work of Dufour *et al.* (2006) who apply Wald tests of multiple-horizon non-causality.

The MF-VAR causality tests utilize the OLS estimation of the (*p*, h) autoregression parameter set:

$$B(h) = \left[A_1^{(h)}, ..., A_p^{(h)} \right]' \in \mathbb{R}^{pK \times K}$$

Where A_K are $K \times K$ matrices for $k = 1, ..., p$ that embody the multi-horizon causal patterns. Based on the assumptions made for the MF-VAR process, consistency is proved, as well as asymptotic normality of ordinary least squares estimator of A_K. The linear parametric restrictions of A_K can be tested by utilizing Wald test and the Wald statistic that is defined in this approach is represented in the following equation:

$$W[H_0(h)] \equiv T_L^*(Rvec\left[\hat{B}\right] - r)' \times (R\hat{\Sigma}_p(h)R')^{-1} \times (Rvec\left[\hat{B}(h)\right] - r)$$

Regarding the Mixed Frequency Granger Causality Analysis and the individual variables in the mixed sampling frequency, there are four main categories that indicate the direction of the effect: *(low to low)* effect, *(high to low)* effect, *(low to high)* effect, *(high to high)* effect.

For the purpose of observing potential effect of “Grexit” data on stock market indices, the second category of high to low effect is utilized. After the estimation of the optimal lag length, the MF-GCA is performed for multiple horizons h, with $h \in \{1,...,10\}$. For the purpose of avoiding the effects of heteroscedasticity and size distortions, bootstrapped replications are used with $N = 999$.

5. Results

5.1 Common frequency Analysis

For the purpose of investigating whether the Mixed Frequency Granger Causality Analysis unveil potential correlation between "Grexit" data and stock market indices, it is beneficial to use as benchmark variables in common frequency and observe the results of the single regression and Granger Causality Analysis. Through this initial research, the validity of main advantage of Mixed Frequency variables will be explored, as it is stated in the work of Ghysels *et al.* (2016).

With respect to the transformation of data to common frequency, the daily tweets of "Grexit" from Twitter and conventional news media are averaged per week. Hence, both the variables of stock market indices and "Grexit" data have the same frequency and the same amount of observations. Regarding the stock market index in the common frequency regression, the Greek stock market index AGC is selected, due to the expected direct relationship with Grexit.

The results of common frequency regression between AGC and "Grexit" data from Twitter are displayed in Table 1a and the software Eviews10 is utilized for these estimations. As it is observed, the corresponding p-value of Twitter data is 0.5085 which fails to reject the null hypothesis. Consequently, Twitter data is observed to have an insignificant effect in AGC when variables are transformed in low frequency.

Table 1a. Results of Common Frequency Regression

Independent Variable	Coefficient	Std. Error	t-Statistic	Prob.
Twitter	-0.001165	0.001758	-0.66229	0.5085
C	0.003185	0.008111	0.392668	0.6949

In a similar fashion, in Table 1b, the results of the regression between "Grexit" data from Traditional News media and Greek stock index are displayed. The p-value of Traditional News variable has an insignificant effect in the movement of Greek stock index. The corresponding p-value is 0.9808 which fails to reject the null hypothesis. In comparison with Table 1a, both regressions in common frequency appear to have no effect on Greek stock market index.

Table 1b. Results of Common Frequency Regression

Independent Variable	Coefficient	Std. Error	t-Statistic	Prob.
Traditional	-6.10E-05	0.002527	-0.02413	0.9808
C	-0.001479	0.006361	-0.23253	0.8163

A further approach in order to investigate the potential correlation of the two variables, is to form a Bivariate Vector Autoregressive model and observe the corresponding Granger Causality test. The information criterion signifies that the optimal lag length is equal to 1 based on Swartz and Hannan and Quinn criterion. From Table 2a, where the results of the common frequency Granger Causality Analysis are presented, it is concluded that Twitter data does not Granger Cause AGC and vice versa. The corresponding p-value of Twitter data Granger Causing AGC is equal to 0.6143 which fails to reject the null hypothesis of non-causation.

Table 2a. Results of Granger Causality Analysis for common frequency variables

Null Hypothesis:	Obs.	F-Statistic	Prob.
Twitter does not Granger Cause AGC	222	0.25464	0.6143
AGC does not Granger Cause Twitter		0.08542	0.7704

Similarly, the Vector Autoregressive model is formed between Greek stock market index and the "Grexit" variable from Traditional News media. The corresponding results from the Granger Causality Analysis are displayed in Table 2b. As it is observed, the use of the keyword "Grexit" in Traditional News media does not Granger Cause the Greek stock market index.

Table 2b. Results of Granger Causality Analysis for common frequency variables

Null Hypothesis:	Obs	F-Statistic	Prob.
Traditional does not Granger Cause AGC	222	0.26436	0.6077
AGC does not Granger Cause Traditional		5.87508	0.0162

Thus, from the above analysis with low frequency variables, it is concluded that there is no apparent effect between "Grexit" data and the Greek stock market. For the purpose of examining the potential hidden correlation between these variables, the Mixed Frequency Granger Causality is going to be explored.

5.2 Mixed Frequency Analysis

5.2.1 Grexit data and GIIPS stock market indices

In this part of the analysis, the results of the Mixed Frequency Granger Causality from high to low frequency will be displayed with the utilization of mapping tools, graphical depictions and tables. With respect to the methodological approach, the code of Ghysels et al. (2016) is utilized for the construction of the Mixed frequency VAR and the Mixed Frequency Granger Causality test, as well as MATLAB software is used for the purpose of the analysis. Through this section the accordance of the results with the stated hypotheses will be investigated.

Through the manual examination of the optimal lag length and the automatic lag length from the information criterion, the set lag that is used in this MF-VAR model is equal to 5, with varying horizon from 1 to 10 weeks. In contradiction to the previous section of the common frequency analysis, the findings of the Mixed Frequency Granger Causality Analysis between "Grexit" data and Greek stock market index represent significant correlation in the short run. These results are presented in Table 3., where the findings of the Mixed Frequency Granger Causality Analysis between "Grexit" data from Twitter and GIIPS stock market are grouped. As it is observed the bootstrapped p-values indicate 0.002 significance in the first horizon of AGC index and 0.012 for the following horizon. These findings reject the null hypothesis of the Granger Causality, hence, reject the non-cause of Twitter data on Greek stock market in the short run for 1% significance in the first horizon and 5% in the second horizon. After the second horizon and up to the tenth horizon, it is clear that the correlation vanishes, and the corresponding p-values are insignificant for 10% significance.

In addition, regarding the main indices of the GIIPS stock markets, it is observed that Greek stock market index (AGC) appears to lead significantly in terms of significance the other countries of this specific group. With respect to the other major stock market indices of the GIIPS, the results indicate a varying significance for 1% and 5% in the first horizon, a slightly insignificant p-values in the second horizon and mostly insignificant results in 10% for the third horizon. More precisely, for the first horizon the Twitter data causes the stock market of Spain in 1% significance (0.006) as well as the stock market of Italy and France for 1% significance.

Table 3. Mixed Frequency Granger Causality Analysis for Twitter data

h	Greece		Italy		Ireland		Portugal		Spain		France	
	W	p-value	W	p-value	W	p-value	W	p-value	W	p-value	W	p-value
1	136	0.002	103.2	0.01	171	0.052	155	0.048	114.1	0.006	250	0.01
2	172	0.012	330.4	0.048	111	0.198	182	0.178	237.7	0.04	364.9	0.1976
3	119	0.218	454.4	0.054	110	0.343	929	0.541	172	0.05	288.6	0.6826
4	87.8	0.577	118.4	0.272	106	0.585	68.4	0.86	157.2	0.21	106.7	0.5609
5	453	0.343	170.5	0.142	124	0.401	130	0.609	228.8	0.02	348.1	0.1437
6	131	0.531	357.2	0.351	118	0.571	145	0.375	163.1	0.287	122	0.5868
7	190	0.26	185.3	0.309	473	0.611	175	0.327	182.1	0.319	131.6	0.7046
8	201	0.331	262.9	0.148	175	0.437	193	0.351	370.2	0.254	208.8	0.3453
9	738	0.254	131	0.85	198	0.794	177	0.587	204.2	0.667	210.3	0.4212
10	139	0.91	403.8	0.493	264	0.319	171	0.661	142.7	0.886	192.1	0.6786

Notes: h represents horizons in weeks

After the examination of correlation between Twitter data and stock market, the same analysis is performed for the conventional news data and the stock market. As it is depicted in Table 4, the resulting p-values from the Mixed Frequency Granger Causality Analysis between "Grexit" use in conventional news media and each stock market index of the GIIPS, fail to reject the null hypothesis. Especially in the case of short run effect, Conventional News media effect on stock market deviates from the corresponding Twitter activity effect. The findings of this test reveal the weak or insignificant impact of the use of the keyword "Grexit" on Traditional News media in the stock market of GIIPS.

Table 4. Mixed Frequency Granger Causality Analysis for Traditional News data

h	Greece		Italy		Ireland		Portugal		Spain		France	
	W	p-value	W	p-value	W	p-value	W	p-value	W	p-value	W	p-value
1	46.4	0.623	96.8	0.303	222	0.058	91.9	0.313	549.3	0.186	81.24	0.319
2	61.2	0.709	71.38	0.836	197	0.441	60.6	0.579	88.67	0.529	63.4	0.583
3	274	0.731	354.2	0.224	73.7	0.705	121	0.134	155.3	0.068	192.1	0.182
4	91.8	0.505	95.01	0.539	99.6	0.525	184	0.423	494.8	0.248	121.1	0.315
5	64.6	0.932	393.9	0.433	191	0.842	684	0.513	136.2	0.455	189.4	0.148
6	107	0.741	258.4	0.028	127	0.523	1309	0.721	280.4	0.044	388.8	0.02
7	158	0.499	327.2	0.03	122	0.771	472	0.607	292.7	0.052	293.9	0.064
8	210	0.331	347.6	0.08	150	0.649	153	0.695	210.7	0.363	174.4	0.609
9	188	0.669	97.99	0.996	110	0.934	96.5	0.97	98.42	0.984	111.9	0.954
10	230	0.573	164.9	0.838	111	0.96	314	0.826	310.3	0.287	625.5	0.78

Notes: h represents horizons in weeks

After the examination of correlation of Twitter data and Traditional data with the stock market, it is worth to investigate the potential supremacy of Twitter over conventional means. For the purpose of this examination and the creation of the orthogonal Twitter series, the twitter data are regressed over traditional series. From this regression the residuals are collected which reflect the unexplained part of the Twitter that Traditional News cannot explain. Thus, the variable of Orthogonal Twitter nets out any effect of Traditional News on Twitter. In table 5, in a similar fashion with the previous tests in tables 3 and 4, the results of the Granger Causality Analysis highlight the supremacy of Twitter over conventional news media. In addition, the lowest p-value, which is occurred in the AGC horizon 1, indicates that Granger Cause exists for 5% significance in the case of Greek stock market index for the first horizon, in contradiction to the other GIIPS member where the p-values are slightly significant in 10% or insignificant. As in the case of Twitter and Traditional data in tables 3 and 4, the correlation disappears after the first 3 horizons.

Table 5. Mixed Frequency Granger Causality Analysis for Orthogonal data

h	**Greece**		**Italy**		**Ireland**		**Portugal**		**Spain**		**France**	
	W	p-value	W	p-value	W	p-value	W	p-value	W	p-value	W	p-value
1	129	0.004	117.8	0.09	76.4	0.613	72.8	0.082	234.5	0.118	60.71	0.18
2	145	0.042	89.44	0.198	84.1	0.561	90.2	0.064	96.74	0.028	191	0.379
3	272	0.05	71.88	0.771	60.1	0.794	111	0.19	125	0.09	71.37	0.437
4	172	0.355	87.02	0.485	55.1	0.984	115	0.333	131	0.074	103.9	0.19
5	108	0.321	126.5	0.214	73.3	0.792	135	0.174	269.7	0.014	164.4	0.028
6	110	0.507	114.9	0.481	89	0.743	116	0.513	193.7	0.056	124	0.359
7	160	0.262	121.6	0.593	111	0.697	123	0.886	175.3	0.2	163.3	0.309
8	194	0.238	133.7	0.611	177	0.335	192	0.262	188.4	0.313	147.6	0.517
9	139	0.719	191.8	0.385	217	0.275	141	0.906	186.1	0.451	923.3	0.301
10	289	0.094	282.7	0.128	170	0.647	179	0.585	296.9	0.144	273.8	0.142

Notes: h represents horizons in weeks

The resulted p-values of the MF-GCA for the GIIPS for both Twitter, Traditional News media, as well as Orthogonal Twitter are graphically depicted in the figure 3. The graphs clearly reflect the effectiveness of Twitter activity on stock market and its supremacy over the insignificant Traditional News media.

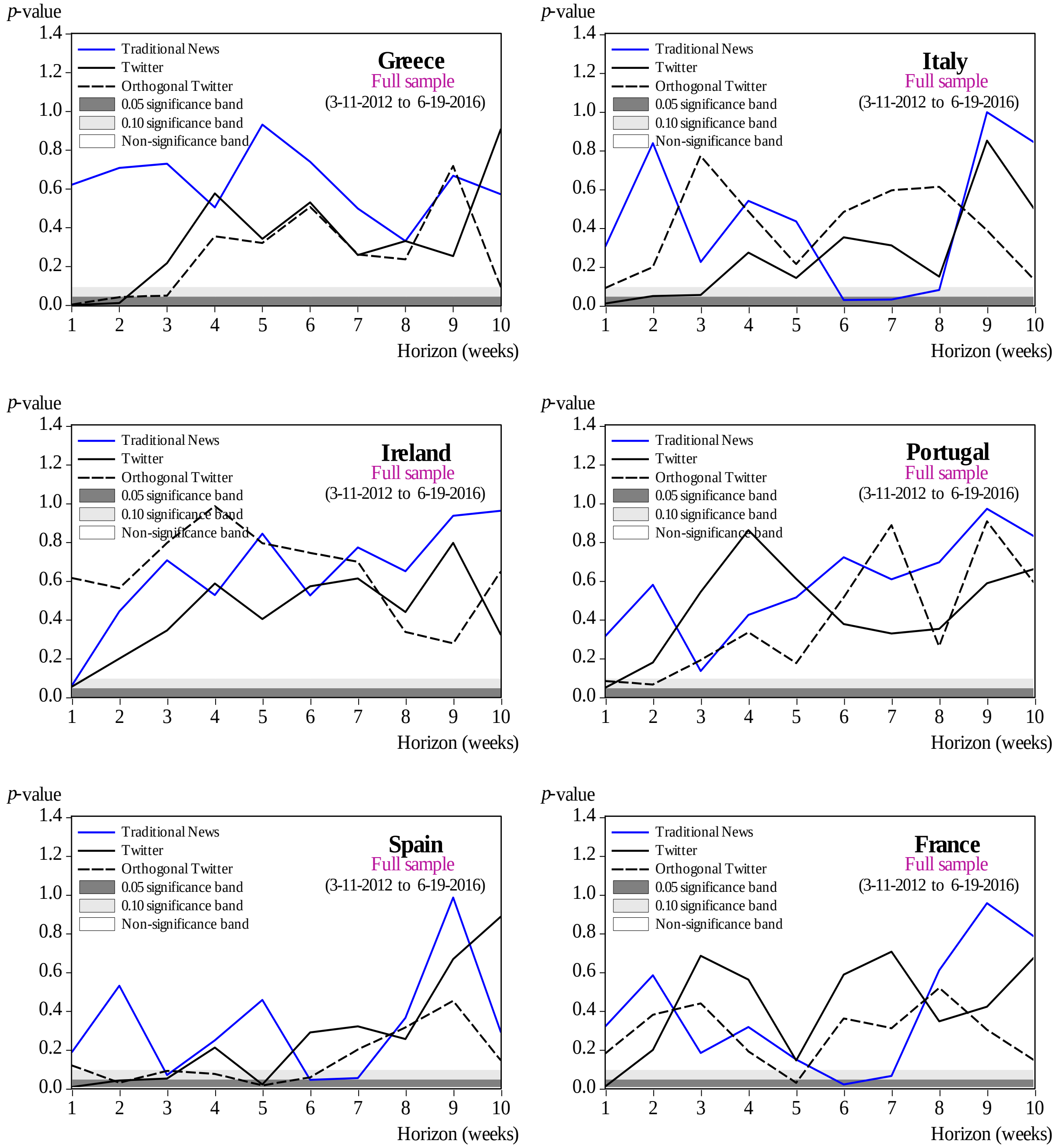

Figure 3. MF-GCA results for GIIPS stock market indices

5.2.2 GIIPS and Non-GIIPS comparison

After the comparison of the relationship between "Grexit" observations and the major indices of the GIIPS stock market, the investigation of the potential deviation of significance between other members of the Eurozone is a subject that has an undeniable merit. For the purpose of this analysis, the group of non-GIIPS is formed with the inclusion of Belgium, Finland, Germany, Netherlands and Slovenia. For these members of the Eurozone, their corresponding major stock market indices are collected and transformed as presented in Table 2.

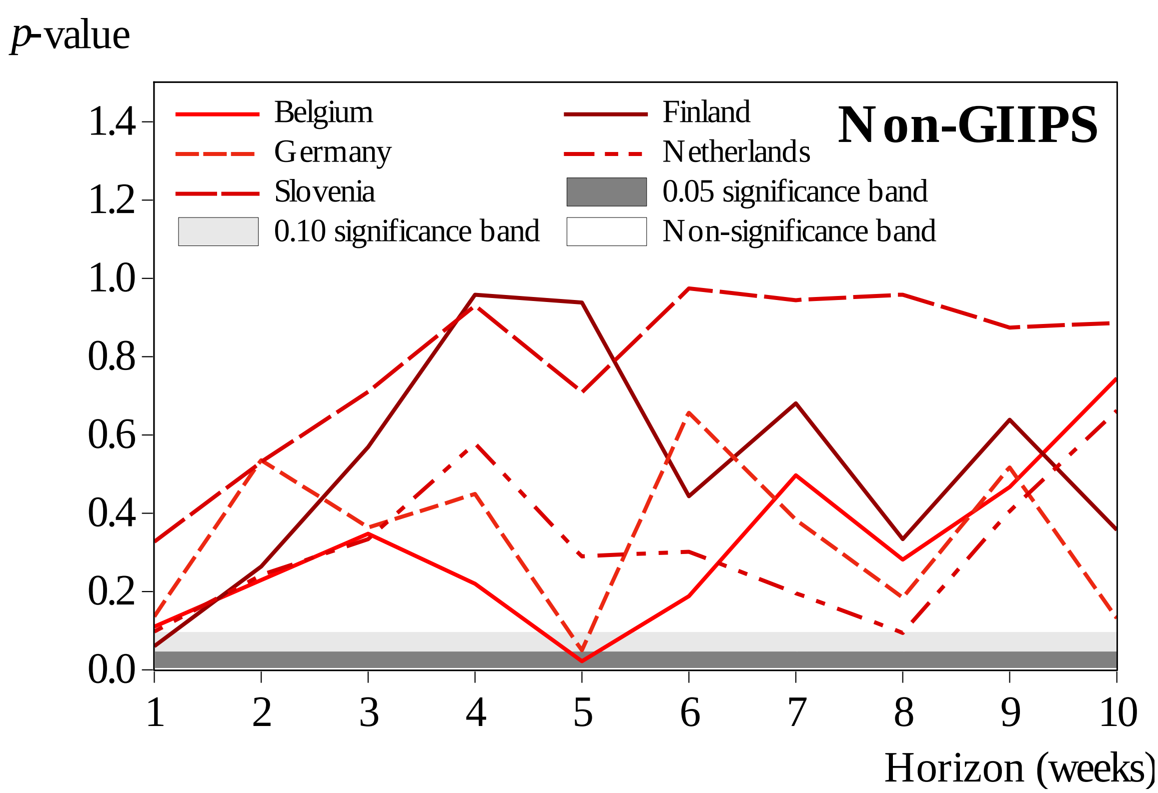

Figure 4. Granger Causality Analysis of Twitter data on Eurozone stock market indices (non-GIIPS)

The results of the MF-GCA for the non-GIIPS members of the Eurozone highlight the significant deviation of p-values in comparison with the GIIPS members. As it is graphically displayed in Figure 4., where the results are grouped, the correlation between Twitter data and stock market indices for this specific group of countries, is insignificant in all cases for 5% significance and insignificant in most cases for 10% significance.

For the sake of illustrative comparison among the two groups of Eurozone members, the comparison of the corresponding results is conducted. As it is displayed in Figure 5, it is clearly observable that the pattern of GIIPS p-values exhibit a more significant arrangement compared to the non-GIIPS states in Figure 4.

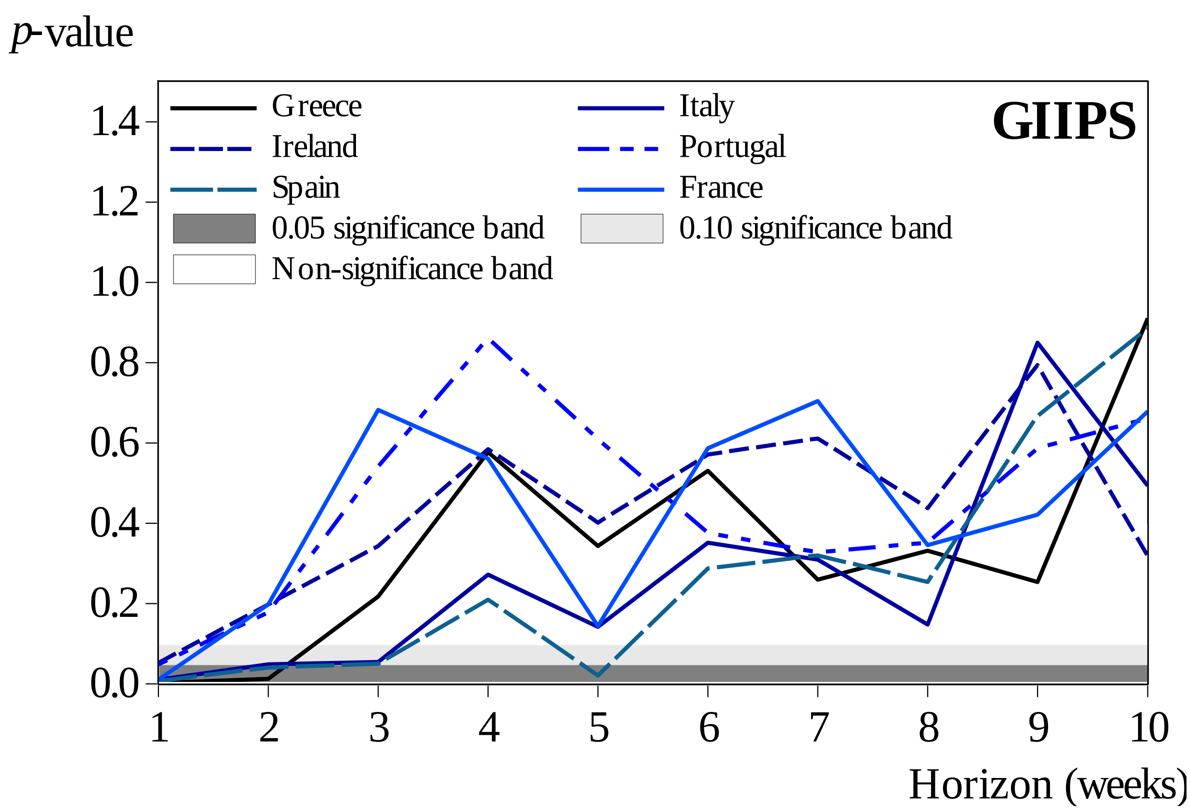

Figure 5. Granger Causality Analysis of Twitter data on Eurozone stock market indices (GIIPS)

The contradictive result of Twitter activity on Eurozone's stock market indices and the corresponding segregation in the Northern and Southern Europe is represented in Figure 6. As it is observed in Figure 3, countries that belong in the geographic area of Northern and Central Europe seem to be unaffected by the effect of "Grexit" use in Twitter. On the contrary, Southern economies exhibit significant effect in the short run. The intense effect of Twitter activity is apparent in horizon 1 and gradually vanishes in horizon 4.

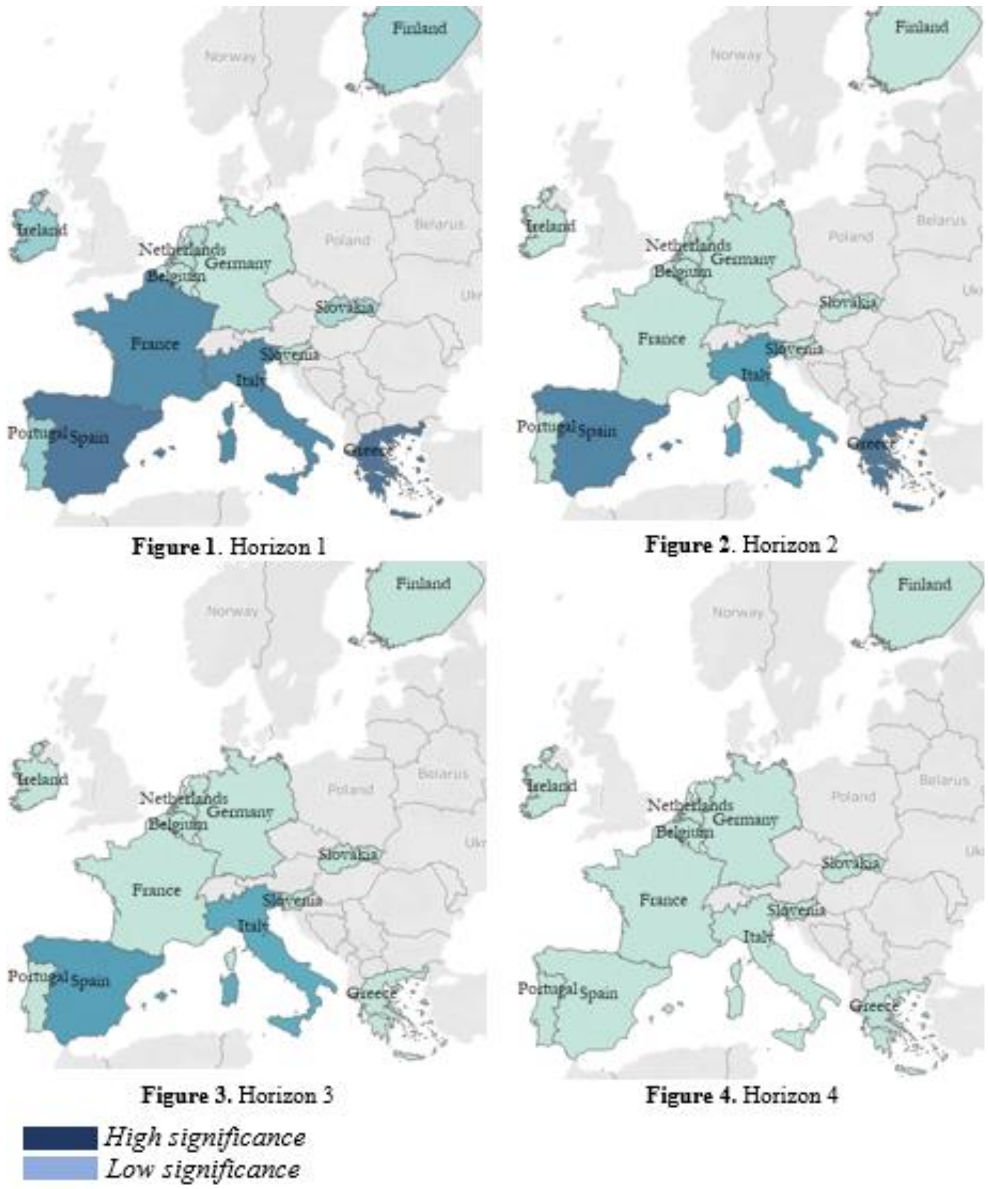

Figure 6. Mixed Fequency GCA for Eurozone's stock markets

5.2.3 Secondary stock market indices - Capitalization size

After the analysis of "Grexit" data with the major stock market indices of the Eurozone members, it is worth to delve more deeply into each major index of the GIIPS members. With respect to the analysis of each main index, several secondary indices are collected that, more or less, reflect the stock market index based on the capitalization size of the firms which deviates them from the components of the main index.

From this perspective of the analysis, it is going to be investigated whether the "Grexit" data has a greater impact in the returns of the higher or lower capitalized firms of the GIIPS states. The secondary indices that are utilized in this report are displayed in Table 2. The analytical results of the MF-GCA for each index, including major and secondary indices, are grouped in Tables 6-10 and are graphically displayed in Figure 7.

Regarding the case of Italy, it is observed that the higher correlation of Twitter data is exhibited in the major index FTSE MIB that reflects the 40 most traded firms of the Italian stock market and less on the FTSE Mid cap and FTSE Small cap. With respect to Ireland the ISEQ General has lower p-values in the first horizon compared to the broader ISEQ index that reflects the overall stock market, as well as, the ISEQ Small Cap that is significant only in the 10% significance. In a similar fashion the Portuguese PSI 20 that comprises the 20 largest firm in terms of capitalization is significant for 5% significance in contradiction with the broader PSI All-Share index that is insignificant even for the 10% significance. Regarding the French stock market indices, it is observed that the major index CAC 40 is more correlated to the Twitter data and as the index focuses more in smaller capitalization sizes, the correlation weakens. Finally, in the case of Spain, the IBEX Mid

Cap is more correlated with Twitter data compared to the major index IBEX 35 and the other secondary indices.

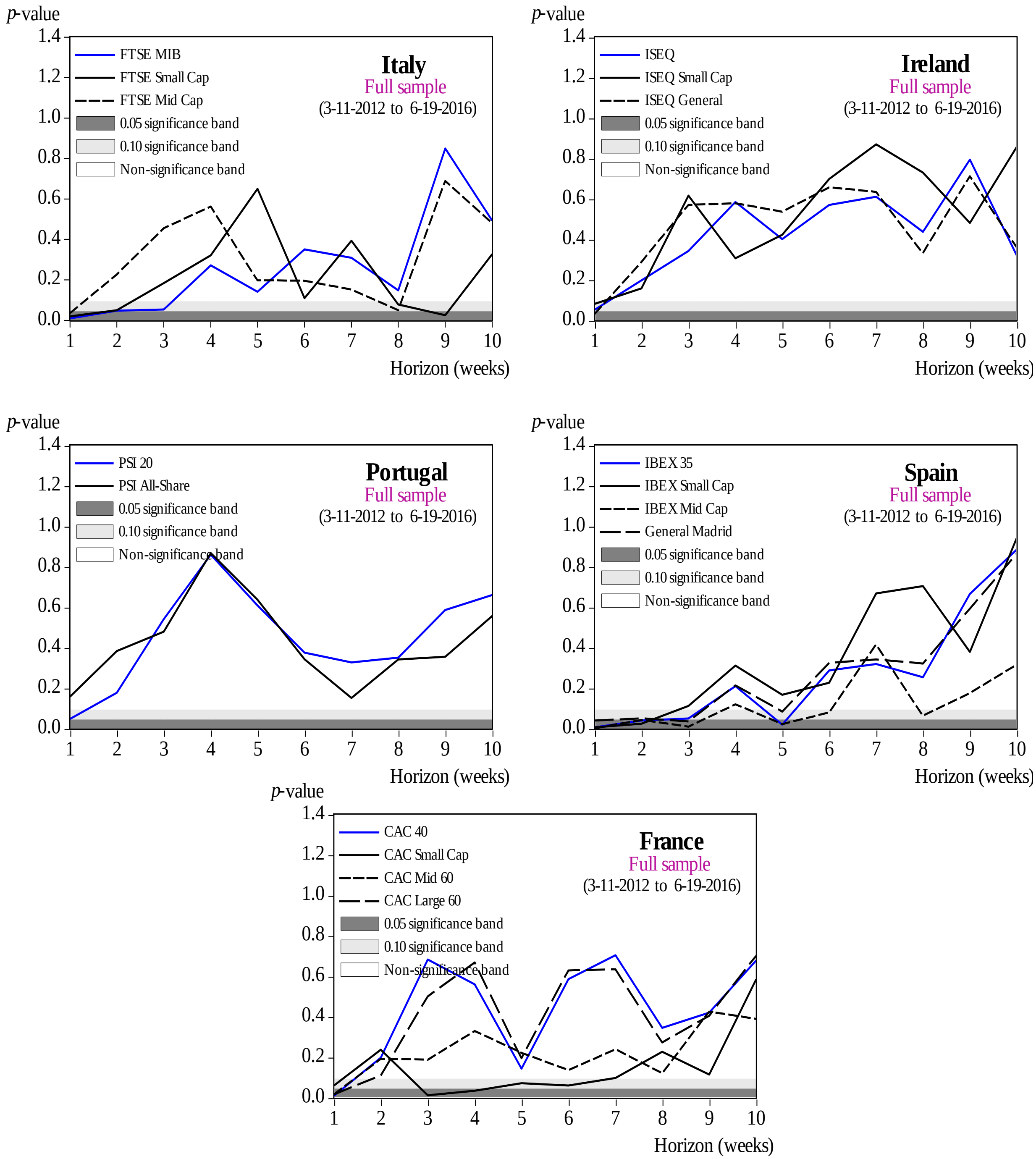

Figure 7. MF-GCA for Twitter data and GIIPS stock market indices (major and secondary indices)

Table 6. Mixed Frequency Granger Causality Analysis for Twitter activity on Italy's Stock Market

h	FTSE MIB		FTSE Small Cap		FTSE Mid Cap	
	W	p-value	W	p-value	W	p-value
1	103.1702	0.01	216.6513	0.02	83.5732	0.0359
2	330.3527	0.0479	106.5665	0.0499	143.0091	0.2275
3	454.4446	0.0539	226.8709	0.1836	231.4474	0.4551
4	118.4464	0.2715	109.5336	0.3214	247.9189	0.5629
5	170.5143	0.1417	263.2797	0.6507	206.521	0.1976
6	357.2468	0.3513	195.1991	0.1098	481.1522	0.1956
7	185.307	0.3094	199.663	0.3932	483.1876	0.1517
8	262.8669	0.1477	273.515	0.0778	314.5507	0.0499
9	131.0231	0.8503	421.155	0.0259	161.419	0.6886
10	403.7955	0.493	247.5705	0.3273	223.0412	0.479

Notes: h represents horizons in weeks

Table 7. Mixed Frequency Granger Causality Analysis for Twitter activity on Irish Stock Market

h	ISEQ		ISEQ Small Cap		ISEQ General	
	W	p-value	W	p-value	W	p-value
1	171.2708	0.0519	211.2398	0.0818	112.8999	0.0319
2	111.1409	0.1976	122.3108	0.1577	91.9594	0.2894
3	109.8261	0.3433	95.0177	0.6168	115.1684	0.5709
4	106.2015	0.5848	116.2305	0.3074	152.7943	0.5788
5	124.1476	0.4012	118.7831	0.4232	108.3115	0.5369
6	117.9294	0.5709	108.1964	0.6986	105.5152	0.6567
7	473.2758	0.6108	100.7858	0.8703	126.7374	0.6347
8	175.1354	0.4371	136.985	0.7305	203.1692	0.3333
9	197.9857	0.7944	186.875	0.481	159.1488	0.7126
10	264.3607	0.3194	143.0442	0.8583	245.1877	0.3593

Notes: h represents horizons in weeks

Table 8. Mixed Frequency Granger Causality Analysis for Twitter activity on Portuguese Stock Market

h	PSI 20		PSI All Share	
	W	p-value	W	p-value
1	154.7745	0.0479	371.96	0.1577
2	182.3266	0.1776	159.81	0.3832
3	928.6771	0.5409	597.37	0.4790
4	68.3889	0.8603	69.70	0.8683
5	129.851	0.6088	732.52	0.6367
6	145.1645	0.3752	145.67	0.3433
7	175.317	0.3273	332.58	0.1517
8	192.9012	0.3513	249.00	0.3413
9	177.042	0.5868	214.40	0.3553
10	170.7547	0.6607	1427.50	0.5569

Notes: h represents horizons in weeks

Table 9. Mixed Frequency Granger Causality Analysis for Twitter data on Spanish Stock Market

h	IBEX 35		IBEX Small Cap		IBEX Mid Cap		General Madrid	
	W	p-value	W	p-value	W	p-value	W	p-value
1	114.1154	0.006	143.40	0.0060	232.2685	0.002	209.7879	0.0399
2	237.72	0.0399	120.94	0.0240	206.3342	0.0419	143.9	0.0499
3	172.0414	0.0499	114.00	0.1118	161.4007	0.01	149.1327	0.0359
4	157.2382	0.2096	131.35	0.3114	190.1259	0.1198	173.7376	0.2136
5	228.8425	0.02	261.03	0.1677	212.417	0.022	192.8484	0.0838
6	163.0802	0.2874	164.53	0.2275	215.5106	0.0798	155.5248	0.3253
7	182.1234	0.3194	124.76	0.6687	155.5033	0.4152	179.1688	0.3413
8	370.1689	0.2535	138.74	0.7046	308.6776	0.0639	278.9278	0.3214
9	204.1517	0.6667	209.05	0.3792	282.7017	0.1756	183.7201	0.5928
10	142.7202	0.8862	118.31	0.9441	287.1611	0.3174	146.6708	0.8643

Notes: h represents horizons in weeks

Table 10. Mixed Frequency Granger Causality Analysis for Twitter data on France's Stock Market

h	CAC 40		CAC Small Cap		CAC Mid 60		CAC Large 60	
	W	p-value	W	p-value	W	p-value	W	p-value
1	250.0051	0.01	97.72	0.0599	139.8379	0.016	113.9068	0.016
2	364.877	0.1976	236.04	0.2375	188.5969	0.1916	179.1229	0.1098
3	288.5789	0.6826	215.53	0.0120	143.4442	0.1876	93.9517	0.501
4	106.726	0.5609	132.56	0.0339	132.9308	0.3293	97.2773	0.6687
5	348.1432	0.1437	171.59	0.0719	151.0672	0.2216	241.683	0.1956
6	121.9697	0.5868	250.52	0.0599	948.7629	0.1357	258.6102	0.6287
7	131.5844	0.7046	216.68	0.0978	191.3174	0.2395	158.0089	0.6347
8	208.8493	0.3453	215.78	0.2275	268.1129	0.1198	219.6209	0.2735
9	210.3228	0.4212	316.07	0.1138	199.454	0.4251	212.3589	0.4052
10	192.0582	0.6786	514.59	0.5848	252.396	0.3892	186.6333	0.7006

Notes: h represents horizons in weeks

6. Conclusion

As it is stated in the section of the introduction, the purpose of this analysis is to investigate the potential correlation between "Grexit" data from Twitter and conventional news media with the stock market indices of Eurozone members. Through this scope, several hypotheses have been investigated through the methodological approach of the mixed frequency VAR and Mixed Frequency Granger Causality Analysis. The resulting findings in the previous section provide useful knowledge about the validity and the affirmation of these assumptions, as well as, highlight the importance of sentiment analysis in the fields of stock market prediction.

The first hypothesis of the analysis examines whether social media, conventional media content and especially the use of the keyword "Grexit" influences the movement of the Greek stock market. From the results of the MF-GCA it is observed that the "Grexit" data from Twitter does Granger Cause the stock market of Greece in the short run. On the contrary, the Traditional news media content does not have a significant effect on the Greek stock market index.

The second hypothesis that is tested, is whether the contagion effect exists in other Eurozone members and specifically in a group of the Eurozone members that includes Greece and several other states which represent similarities in their economic and political conditions, known as GIIPS (Portugal, Italy, Ireland, Greece, Spain), as well as, France. Regarding this assumption, it is concluded that the highest correlation of "Grexit" data is present in the Greek stock market and this effect is spread in a weakened fashion in the stock market of the GIIPS states. Consequently, it is observed that Twitter data has a primary effect on the Greek stock market which is significantly transmitted in the GIIPS stock market with a slightly debilitated effect. On the other hand, as in the case of Greek stock market index, it is concluded that Traditional news media data have no effect for each one of the GIIPS stock market indices.

From the formation of the Orthogonal Twitter series and the repetition of the MF-GCA it is concluded that Twitter data has a prevailing effect over the stock market, which clearly indicates its direct relationship with the financial community and gives merit to the notion of sentiment analysis within the fields of behavioral finance.

Since the clear supremacy of Twitter over the conventional means concerning the effect of Grexit data on the stock market in several Eurozone members is evident, the fifth hypothesis tests whether GIIPS stock markets get affected in a higher degree from the Twitter activity in comparison with the more robust economies of the union such as Germany, Finland, etc. After the interpretation of the Mixed Frequency Granger Causality results, it is clear that GIIPS stock market get affected significantly in a higher level from the Twitter data in contradiction with the non-GIIPS members that in most cases have insignificant relationship with the "Grexit" data. This fact supports the claim that the similarities of the GIIPS countries regarding their dwindling economies and their political instability could be leading factors for the similar behavior towards the "Grexit" concerns. These concerns have a greater impact in less robust economies, that are mostly located in the Southern Europe, than the Northern Europe economies.

The sixth hypothesis assesses whether the contagion effect in the GIIPS countries from the Twitter data is related to the capitalization size of the firms or whether the capitalization size is not part of the causal chain. From the analysis of multiple indices of the GIIPS stock markets that differentiate in terms of capitalization size, it is observed that highly capitalized firms get mostly affected by the "Grexit" data in contradiction to the small capitalization firms. This result is probably rooted to the reliance of highly capitalized firms on the economic and political stability in the Eurozone. These prosperous conditions allow the continuous export of goods and services and the global perspective of the firms. Hence, events that trigger agitation and forecast an ominous future in the Eurozone, such as the case of Grexit, seem to have a significant effect in these firms. On the other hand, the firms that are less capitalized and probably have higher local activity than exporting activity and less direct dependence from the overall condition of the Eurozone, seem to be affected in a less significant fashion by the "Grexit" tweets.

The last hypothesis of the analysis investigates whether the results of the analysis are in line with the results of the work of Milas *et al*. (2018) that examine the impact of "Grexit" data from Twitter and conventional media on the debt market and precisely on the level of the spreads. As in the analysis of the Milas *et al*. (2018), the effect of Grexit data is present in the case of Greek stock market. In a similar manner, the contagious effect transmitted in the GIIPS states is more feeble compared to the impact of Grexit tweets in the Greek stock market. In both cases, the effect of online activity has a short run effect. The main difference in these two reports concern the significance of the Traditional news media content. The Traditional news media have a weakened impact in the GIIPS stock market compared to the Twitter data, but in the case of this analysis the effect is insignificant in every stock market index of the GIIPS.

Through this empirical analysis in the fields of behavioral finance, the effect of "Grexit" tweets is apparent in the Southern economies of the Eurozone. The interwoven behavior of social media activity into the fabric of society is highlighted through the significance of Twitter data in contradiction to the feeble effect of conventional media. Online activity has the ability to reflect public behavior and more precisely, in the case of Grexit accurately represents an augmenting unrest among the financial community. Further research in this field could shed light on the direction of the Twitter effect in the returns of the stock market indices, as well as, the examination of the corresponding volatility. Another variation of this analysis could examine the effect of “Grexit” data on different sectors of the economy and their corresponding indices. Last but not least, the robustness of the results could be examined by using different frequencies and probably a smaller ratio of high to low frequency (*m*).

References

Acemoglu, D., A. Ozdaglar and A. Parandeh Gheibi (2010). Spread of (mis) information in social networks. *Games and Economic Behavior* 70(2), 194-227.

Aramaki, E., S. Maskawa and M. Morita (2011). Twitter catches the flu: detecting influenza epidemics using Twitter. *Proceedings of the conference on empirical methods in natural language processing*, 1568-1576.

Arias, M., A. Arratia and R. Xuriguera (2013). Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology* 5(1), 8.

Askitas, N. and K. F. Zimmermann (2009). Google econometrics and unemployment forecasting.

Asur, S. and B. A. Huberman (2010). Predicting the future with social media. *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* 1, 492-499.

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

Broadstock, D. C. and D. Zhang (2019). Social-media and intraday stock returns: The pricing power of sentiment. *Finance Research Letters* 30, 116-123.

Chen, R. and M. Lazer (2013). Sentiment analysis of twitter feeds for the prediction of stock market movement. *stanford edu Retrieved* January 25, 2013.

Chung, S. and S. Liu (2011). Predicting stock market fluctuations from twitter. *Berkeley, California.*

Chung, S. and S. Liu (2011). Predicting stock market fluctuations from twitter. Berkeley, California.

Curme, C., T. Preis, H. E. Stanley and H. S. Moat (2014). Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences* 111(32), 11600-11605.

Dufour, J. M., D. Pelletier and E. Renault (2006). Short run and long run causality in time series: inference. *Journal of Econometrics*, *132*(2), 337-362.

Dufour, J. M. and E. Renault (1998). Short Run and Long Run Causality in Time Series Theory," *Econometrica*, 66, 1099{1125).

Fisher, K. L. and M. Statman (2000). Investor sentiment and stock returns. *Financial Analysts Journal* 56(2), 16-23.

Ghysels, E., J.B. Hill and K. Motegi, (2016). Testing for Granger Causality with Mixed Frequency Data. *Journal of Econometrics* 192, 207-230.

Gilbert, E. and K. Karahalios (2010). Widespread worry and the stock market. *Fourth International AAAI Conference on Weblogs and Social Media*.

Goel, S., J. M. Hofman, S. Lahaie, D. M. Pennock and D. J. Watts (2010). Predicting consumer behavior with Web search. *Proceedings of the National academy of sciences* 107(41), 17486-17490.

Granger, C. (1969) Investigating Causal Relations by Econometric Models and Cross-Spectral Methods, *Econometrica*, 3, 424–438.

Li, X., H. Xie, L. Chen, J. Wang and X. Deng (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems* 69, 14-23.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5(1), 1-167.

Makrehchi, M., S. Shah and W. Liao (2013). Stock prediction using event-based sentiment analysis. *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* 1, 337-342.

Mao, H., S. Counts and J. Bollen (2011). Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv*:1112.1051.

Mao, Y., W. Wei, B. Wang and B. Liu (2012). Correlating S&P 500 stocks with Twitter data. *Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research*, 69-72.

Maragoudakis, M. and D. Serpanos (2016). Exploiting financial news and social media opinions for stock market analysis using MCMC Bayesian inference. *Computational Economics* 47(4), 589-622.

Milas, C., T. Panagiotidis and T. Dergiades (2018). *Twitter versus Traditional News Media: Evidence for the Sovereign Bond Markets*.

Mittal, A., and A. Goel (2012). Stock prediction using twitter sentiment analysis. *Stanford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 15.

Nguyen, T. H., K. Shirai and J. Velcin (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications* 42(24), 9603-9611.

O'Connor, B., R. Balasubramanyan, B. R. Routledge and N. A. Smith (2010). From tweets to polls: Linking text sentiment to public opinion time series. *Fourth International AAAI Conference on Weblogs and Social Media.*

Oliveira, N., P. Cortez and N. Areal (2013). On the predictability of stock market behavior using stock tweets sentiment and posting volume. *Portuguese Conference on Artificial Intelligence*, 355-365.

Pagolu, V. S., K. N. Reddy, G. Panda and B. Majhi (2016). Sentiment analysis of Twitter data for predicting stock market movements. *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*, 1345-1350.

Perlin, M. S., J. F. Caldeira, A. A. Santos and M. Pontuschka (2017). Can we predict the financial markets based on Google's search queries ? *Journal of Forecasting* 36(4), 454-467.

Preis, T., H. S. Moat and H. E. Stanley (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific reports* 3, 1684.

Ranco, G., D. Aleksovski, G. Caldarelli, M. Grčar and I. Mozetič (2015). The effects of Twitter sentiment on stock price returns. *PloS one* 10(9), e0138441.

Rao, T. and S. Srivastava (2012). Analyzing stock market movements using twitter sentiment analysis. *Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012),* 119-123.

Rao, T. and S. Srivastava (2014). Twitter sentiment analysis: How to hedge your bets in the stock markets. *State of the art applications of social network analysis,* 227-247.

Ruiz, E. J., V. Hristidis, C. Castillo, A. Gionis and A. Jaimes (2012). Correlating financial time series with micro-blogging activity. *Proceedings of the fifth ACM international conference on Web search and data mining*, 513-522.

Sehgal, V. and C. Song (2007). Sops: stock prediction using web sentiment. *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, 21-26.

Shapley, D. (2002). John Seely Brown and Paul Duguid The Social Life of Information. *INFORMATION COMMUNICATION AND SOCIETY* 5, 450-452.

Si, J., A. Mukherjee, B. Liu, Q. Li, H. Li and X. Deng (2013). Exploiting topic based twitter sentiment for stock prediction. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* 2, 24-29.

Smailović, J., M. Grčar, N. Lavrač and M. Žnidaršič (2013, July). Predictive sentiment analysis of tweets: A stock market application. *International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, 77-88.

Sprenger, T. O., A. Tumasjan, P. G. Sandner and I. M. Welpe (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management* 20(5), 926-957.

Sul, H., A. R. Dennis and L. I. Yuan (2014). Trading on twitter: The financial information content of emotion in social media. *2014 47th Hawaii International Conference on System Sciences*, 806-815.

Weng, B., L. Lu, X. Wang, F. M. Megahed and W. Martinez (2018). Predicting short-term stock prices using ensemble methods and online data sources. *Expert Systems with Applications* 112, 258-273.

Wu, D. D., L. Zheng and D. L. Olson (2014). A decision support approach for online stock forum sentiment analysis. *IEEE transactions on systems, man, and cybernetics: systems* 44(8), 1077-1087.

Xiong, R., E. P. Nichols and Y. Shen (2015). Deep learning stock volatility with google domestic trends. *arXiv preprint arXiv*:1512.04916.

Yang, S. Y., S. Y. K. Mo and A. Liu (2015). Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance* 15(10), 1637-1656.

Zhang, W. and S. Skeena (2010). Trading strategies to exploit blog and news sentiment. *Fourth international AAAI conference on weblogs and social media*.

Zhang, X., H. Fuehres and P. A. Gloor (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences* 26, 55-62.