LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

# MESTRADO
# ECONOMETRIA APLICADA E PREVISÃO

## TRABALHO FINAL DE MESTRADO
DISSERTAÇÃO

## SOCIAL MEDIA IMPACT ON STOCK PRICES

ELIANO PATRÍCIO MACEDO MARQUES

OUTUBRO 2015

LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

# MESTRADO EM
# ECONOMETRIA APLICADA E PREVISÃO

## TRABALHO FINAL DE MESTRADO
DISSERTAÇÃO

# SOCIAL MEDIA IMPACT ON STOCK PRICES

ELIANO PATRÍCIO MACEDO MARQUES

**ORIENTAÇÃO:**
PROF. DOUTOR JOÃO NICOLAU

OUTUBRO 2015

## *Abstract*

This dissertation explores the impact of social media on stock prices. In order to measure such an impact, an end-to-end online platform was created leveraging the Amazon Web Services cloud, open source R, web APIs from Twitter, Google News, Google Finance, Yahoo News, Yahoo Finance and Financial Times. The end goal was to create an intraday dataset that tracks all the news related to two companies Tesco PLC, SKY PLC and then perform several econometric models using a Multiplicative GARCH model to understand if the social media news from multiple sources are statistically relevant to measure the stock prices returns and if the forecast accuracy improves by including such covariates in the predictive equations. For all of the approaches it will be presented the respective theory background and their applications to high-frequency data. Overall it can be concluded that news, past market information and tweets didn't improve that much the stocks returns models and forecasting accuracy but, for models without Twitter, Tesco's 2$^{nd}$ best model and SKY's best model included social media news as covariates. Google Finance and Financial Times presented better results when compared to Yahoo Finance and/or Twitter, however, overall, the impact was very residual and closer to zero.

## *Resumo*

Este documento explora o impacto do *social media*[1] nos preços das ações. Para medir tal impacto, foi construído uma plataforma online na *Cloud* da Amazon (AWS) recorrendo a ferramentas como o R, web APIs do Twitter, Google News, Google Finance, Yahoo News, Yahoo Finance e Financial Times. O objetivo deste estudo passa por criar uma base de dados para modelação com uma granularidade de 15 minutos incluindo como possíveis variáveis explicativas todas as noticias publicadas das fontes acima referidas para o Tesco e SKY. Posteriormente, recorrendo a uma técnica de GARCH multiplicativos, diversos modelos econométricos foram desenvolvidos com o objetivo de efetuar previsões a um passo dos retornos dos preços das ações acima referidos. Se efetivamente a informação do social media for relevante, deverá ser observado uma melhoria nas previsões dos retornos das ações. No decorrer do documento, será apresentado a metodologia utilizada no estudo e a sua aplicação em dados com elevada frequência. No geral, podemos concluir que a informação do *social media* é residualmente relevante para modelar e prever os retornos dos preços das ações e, nos modelos sem informação do Twitter, o segundo melhor modelo para o Tesco e o melhor modelo para a SKY, inclui informação do social media. No que se refere às fontes utilizadas, pode-se concluir que o Google Finance e o Financial Times relevam maior importância nos movimentos dos retornos das ações do que o Yahoo Finance e/ou Twitter, apesar de tal relevância ser muito residual e próxima de zero.


[1] Social media refere-se a noticias publicadas online por diversas fontes, tais como, Google News ou Twitter

## *Acknowledgments*

First I would like to emphasize how much fun I had building this solution and writing this document. It was such a great opportunity to improve my skills in many different fields that go even beyond the econometric modelling lessons learned in the Master.

I would like to thank the authors of several R packages that I have used in order to create the end-to-end online platform. The solution heavily depends on tm.plugin.webming, twitter, rugarch, and dplyr so thank you for the great work, Mario Annau, Geoff Jentry, Alexis Ghalanos and rstudio, respectively. I would also like to thank Financial Times for providing a free user account for the online API to extract the Financial Times data used in this study.

Finally, I also like to thank the Professor and mentor João Nicolau for the guidance provided throughout this work, always providing fantastic inputs and suggesting new challenges.

**Keywords**: Social Media, Google, Yahoo, Financial Times, Twitter, Machine Learning, Amazon Web Services, R, Multiplicative GARCH, Tesco, SKY, Intraday, Volatility

# *Table of Contents*

# 1 Introduction

The purpose of this study was to measure the impact of social media news and tweets in the stocks returns of Tesco PLC and SKY PLC. In order to get the data for this study, several companies were approached but only Financial Times responded offering a free academic license for their online API. As a result of not having the data available (even with Financial Times an algorithm to extract the data from the API was required) it was build an online end-to-end solution (see Figure 1 – High-Level end-to-end online solution) to gather intraday stock prices, market returns information, news and tweets from several sources, including Financial Times, about Tesco and SKY. During this process of building and maintaining the online platform, several issues were found, e.g. the tweets from Twitter between end-May until mid-July were not extracted correctly, which cause issues during the estimation of the models. The initial intention of this study was to classify the news and then use that information to predict stock returns. Therefore, a machine learning technique suggested by Go, Bhayani, & Huang (2009) was used to predict the news and tweets as positive, negative or neutral. This technique was robust for news and tweets with small length, however Financial Times news are longer articles and to avoid biased classifications, only the total news was considered in the econometric models. Then, all the information collected was aggregated into a final "Modelling Dataset" with a 15 minutes' intervals frequency which was then used to perform the econometrics estimation and forecasting.

As described before, the main goal for this study was to use the information collected from the online platform and test if the short-term (up to 4 lags) past market returns, news and tweets were relevant to model and predict stock returns. In order to test that, and given the characteristics of the financial time series: high-volatility, negative asymmetry,

"volatility clustering" (which were also visible in two stock prices analyzed), it was adopted a recent methodology suggested by Engle & Sokalska (2012) that extends the standard GARCH model. For comparison purposes, the standard GARCH was also considered and in total, 7 econometric equations were estimated for models with and without Twitter as covariates. Then, a Weighted-Ljung-Box suggested by Fisher & Gallagher (2012) was used to diagnose the models quality. In order to measure the gain (if any) in the forecasting accuracy, four forecasting metrics were used to compared the results, two metrics for the returns prediction, Means Absolute Error (MAE) and Mean Direction Accuracy, and two loss function metrics for volatility forecasting suggested by Patton (2010). Overall it can be concluded that news, past market information and tweets didn't dramatically improve the stocks returns models and forecasting accuracy but at least with Tesco, the best model included social media news as covariates. Overall, Google Finance and Financial Times presented better results when compared to Yahoo Finance and/or Twitter, however, the impact was very residual and closer to zero.

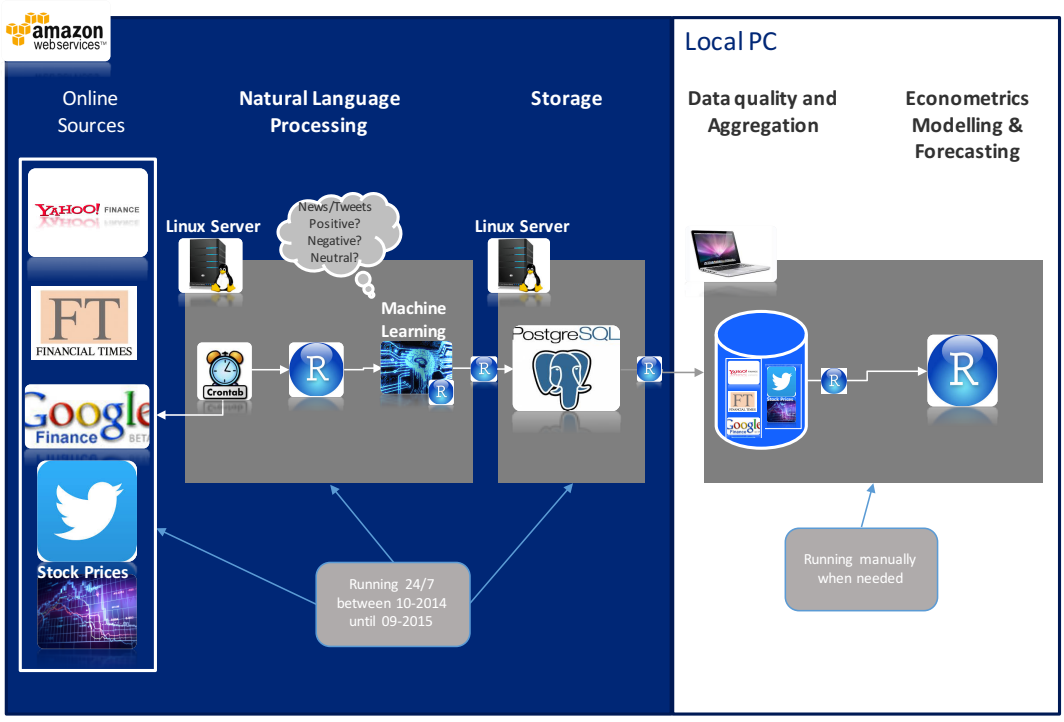



**Figure 1 – High-Level end-to-end online solution**

The document is organized in five sections, first an introduction, then a chapter for Data, where it is explained how the platform was created and data extracted, third Social Media, where it is presented the background around social media and sentiment analysis, including the algorithm for classification of news and tweets, fourth the econometrics analysis with respective methodologies adopted, models, tests and forecasts developed and finally the conclusions of the study, where it is also highlight limitations and potential next steps.

# 2 Data

In order to perform this research, it was necessary to create an online environment[2] that was continuously running to extract social media news and tweets as well as intraday stock prices and market information. The platform leverages the Amazon Web Services cloud[3], where 2 servers were rented, 1 Linux[4] server for the R[5] application and 1 Linux database server to store the data using Postgres[6]. Once the servers were launched, an R package[7] was developed with several functions to extract the data from online sources, predict the sentiment of news and tweets, send and extract the data from the database, extract stock prices and market information, further transform and join the news and tweets with the stock prices and market information and finally generate the 'Modelling Dataset' which then was used during the econometric analysis. The online platform was running from beginning of October 2014 until mid-September 2015.

## 2.1 Social Media News and Tweets

[2] Please see Figure 1 – High-Level end-to-end online solution
[3] Amazon Web Services - https://aws.amazon.com/
[4] Linux Operation System - https://en.wikipedia.org/wiki/Linux
[5] Open-source R - https://www.r-project.org/
[6] Postgres - http://www.postgresql.org/
[7] R Package is available in ElianoMarques github page

An online process was created in R to gather all the news and tweets about Tesco and SKY from several sources (Twitter, Yahoo Finance, Google Finance and Financial Times) during approximately 12 months (from 10 October 2014 until 13 of September 2015). Table 1 - Social Media APIs presents the source of each API used to gather news and tweets, the respective searching parameters and any additional text-mining filter applied. Please note that, with the exception of Financial Times API, the initial searching function was available in the R Packages "tm.webmining" and "twitteR". However, for each API it was required a function that saved the results of each search (i.e. news or tweets) to the database but only if the content of the news (tweets) was not saved previously. In the case of Financial Times, a function to search news for a given period was built, leveraging the date search parameters available in the API. In addition, if required, a function was created to filter only news or tweets related to financial markets (this was more important on twitter than the rest of the sources). In total, it was gathered about 30.000 news from Google and Yahoo, circa 2.9 million tweets from Twitter and circa 583.000 news from Financial Times.

| API | Source | R Package Support | Comments |
|---|---|---|---|
| Google Finance | http://www.google.com/finance | tm.webmining | The search parameter for Tesco was LON:TSCO and for SKY was LON:BSY |
| Yahoo Finance | http://developer.yahoo.com/rss/ | tm.webmining | The search parameter for Tesco was TSCO.L and for SKY was BSY.L |
| Financial Times | http://developer.ft.com/ | Custom built | The API returned news for all companies. Then using text mining it was filtered only news that matched the following words: "Tesco ", "TSCO.L", "LON:TSCO", "SKY ", "BSY.L" and "LON.BSY" |
| Twitter | https://twitter.com/apps | twitteR | The search parameters for Tesco usernames in Twitter: "tesco", "tescohelp","tescomedia","tescofood"; Tesco words search: "tesco", "tescohelp","tescomedia","tescofood"; "TSCO"; SKY usernames: "skyhd"; "skyhelpteam"; SKY words search: "skyhd"; "skyhd"; "sky ";"BSY". Furthermore all tweets collected were filtered if they had any of the following words: "TSCO","BSY","FTSE","stock prices","financial market","finance". |

**Table 1 - Social Media APIs**

Once the news were gathered from the respective APIs[8], additional transformations, such as, adding the timestamp retrieved or predicting the sentiment of the news as positive, negative or neutral using a machine learning algorithm introduced by Go, Bhayani, & Huang, (2009) were performed before the news were sent to the storage database. Then, using the Linux scheduler Crontab, a task was created to execute the above process every three hours.

## *2.2 Stock prices and market information*

Another online process was created in R to gather all the intraday (slots of 15 minutes) stock prices metrics (e.g. Open, Close) and market information from Google Finance. No transformation was required after extractions so the data was sent directly to the database. A scheduling task was added to Crontab to repeat the above process at 23:00 of each day, with the exceptions of weekends. In total, 58.000 intraday stock prices and market information was gathered.

| API | Source | R Package Support | Comments |
|---|---|---|---|
| Google finance | http://www.google.com/finance/getprices?q | Custom built | The list of stocks prices and markets information gatheres was TSCO, BSY, RWE, EDF, UKX, DAX, PX1, PSI20. |

**Table 2 – Stock and Market Information APIs**

## *2.3 Data quality and aggregation*

An R process was created to extract all the information from the database, access the quality of the information (e.g. removing duplicated news and tweets), generate a financial calendar based in the London Stock Exchange market and then join all the data together to create the final 'Modelling Dataset'. At this point several decisions had to be made, such as:

[8] API = "In computer programming, an application programming interface (API) is a set of routines, protocols, and tools for building software applications. An API expresses a software component in terms of it is operations, inputs, outputs, and underlying types." - Wikipedia

- Which market to use to generate the financial calendar
    - London Stock Exchange was defined
- What to do with the news and tweets that occur when the market is closed
    - All the news and tweets that occur outside of the market hours are linked with the opening time of the next day. So for example, if news comes out on Friday at 21, it will be considered in the first slot of 15 minutes in Monday (assuming Monday the market opens normally).
- What should be the granularity of the final dataset
    - It was defined a 15 min slot as granularity of the data.

Once the data has been through the above process a 'Modelling Dataset' is generated with the following fields:

| Fields Names | Description | Comments |
|---|---|---|
| TimeSeries | A timestamp of the data | By slots of 15 minutes |
| Tot_News (GN, GF, YN, YF, FT) | Total News | Also available for each source. GN = Google News; GF = Google Finance; YN = Yahoo News; YF = Yahoo Finance; FT = Financial Times |
| Pos_News (GN, GF, YN, YF, FT) | Total Positive News | |
| Neg_News (GN, GF, YN, YF, FT) | Total Negative News | |
| Neu_News (GN, GF, YN, YF, FT) | Total Neutral News | |
| Tot_Tweets (Stk) | Total Tweets | Also available vy source. GN = Google News; GF = Google Finance; YN = Yahoo News; YF = Yahoo Finance; FT = Financial Times |
| Pos_Tweets (Stk) | Total Positive Tweets | |
| Neg_Tweets (Stk) | Total Negative Tweets | |
| Neu_Tweets (Stk) | Total Neutral Tweets | |
| Stock | Stock name (Tesco PLC, SKY PLC, Npower, EDF Energy) | |
| Open (UKX, PSI20, DAX, PX1) | Open price for the 15 minutes slot | Also available for Market Indexes. UKX = London Stocks Index; PSI20 = Portugal Stock Index; DAX = DAX ; PX1 = PX1 |
| High (UKX, PSI20, DAX, PX1) | Highest price for the 15 minutes slot | |
| Low (UKX, PSI20, DAX, PX1) | Low price for the 15 minutes slot | |
| Volume (UKX, PSI20, DAX, PX1) | Volume price for the 15 minutes slot | |

**Table 3 - Modelling Dataset Fields**

After performing all the required transformations and aggregations, the final dataset had 7,854 for each of the 2 stocks gathered. Please note that during end of May until mid-July it occurred an error extracting the Twitter tweets and therefore this information cannot be used in the models during that period.

# 3 Social Media Analysis

In the past years, several studies explored the impact of social media news and blogs and their respective sentiment (e.g. positive, negative) in order to find trends and patterns

in different fields such as stock prices, marketing and customer behavior. See for example Go, Bhayani & Huang (2009), Jiang, Yu, Zhou, Liu, & Zhao (2011), Bautin, Vijayarenu, & Skiena (2008), Oh & Sheng (2011), Asur & Huberman (2010), Chan (2003), Velikovich, McDonal, & Councill (2010), Skiena & Zhang (2010) and Garimella, Weber, & Cin (2014) for more details. The approach taken was similar in that regard (finding patterns and trends about social media news) but different in the way it was applied, i.e., the trends from social media news or tweets (e.g. number of positive news from Google Finance between 9:15am and 9:30am for a given day) are embedded into several econometric models[9] to understand and measure their impact on stock prices returns[10].

## *3.1 Machine learning classification algorithm*

As cited before, many previous studies tried to build classification algorithms to extract the content of blogs, news and tweets. The approach in this study was to leverage the work developed by Go, Bhayani & Huang (2009) for several reasons: 1) With the exception of the financial times articles, the news and tweets extracted had a small length, a characteristic of that research; 2) It was taken into account several features methods, such as the standard unigrams (a keyword e.g. “good”), bigrams (a combination of two keywords, e.g. “not bad”), unigrams and bigrams together and parts of speech; 3) The training data used in their study represented millions of tweets vs. several thousands from previous studies. This was possible through the Twitter API that enables easily extraction of tweets. 4) Standard sentiment classifications algorithms, including some of the ones available in R, are performed against a list of positive and negative words. The final polarity is then the

[9] Please see chapter Models specification for details

[10] Stock Prices Returns: method 1 = log (Stock $Price_t$)- log (Stock $Price_{t-1}$), where T= time; method 2 = $|r_t|^d$ , where T=time and d, typically, is 1, i.e. the absolute value of the returns

difference between the sum of positive and negative keywords. This tends to be very limited and potentially combining multiple words, emoticons and/ or parts of speech most likely will improve the accuracy of the forecast. This was also proved in Go, Bhayani & Huang (2009), where the machine learning methods were ~15-18 percentual points more accurate than the "keyword list method". During this study the *Naïve Bayes*[11], *Maximum Entropy*[12] methods had an average accuracy of 81.3%, 80.6%, respectively. 5) The most important one, the classification algorithm is available online via an API. As described by the author, this algorithm presents some limitations, given it is only trained in news and tweets with small length (in fact an average of 14 characters), don't consider emoticons ":)", ":(". This fact was important during this study as the characteristics of the news and tweets from Google, Yahoo and Twitter are similar to the ones used in this study but not the news from Financial Times, as they are big articles from journalists about a specific topic. After analyzing the results of the machine learning algorithm, the results for Google, Yahoo and Twitter were ~30% positive, ~30% negative and remaining neutral. However, for Financial Times, 90% of the news where negative and that didn't sound about right. Time pressures didn't enable the creation of a separated model for Financial Times so the econometrics models only included the total number of news as covariates.

# 4 Econometric Analysis

## 4.1 Background

Financial time series data have been an area of interest for several academics in the past, for example, example Andersen, Bollerslev & Cai (2000), Engle & Gallo (2006), Rossi &

[11] See Go, Bhayani, & Huang (2009) for details
[12] See Go, Bhayani, & Huang (2009) for details

Fantazzini (2012), Bollerslev, Cai & Song (2000), Singh, Allen, & Powell (2013), Engle & Sokalska (2012), Andersen & Bollerslev (1998). As described in many of those studies, but in particular in Engle & Sokalska (2012), daily volatility has been hugely study in the literature and conventional GARCH has proven to be satisfactory to model such data. However, for intraday data, and according to the authors at the Olsen conference on High Frequency Data Analysis in Zurich in March 1995, this type of approach wasn't proving to be robust enough. This was also proved in Anderson & Bollerslev (1997), where the estimation of a MA(1)[13]-GARCH(1,1) on top of multiple high-frequency intervals resulted in inconsistent parameters, because of the pronounced diurnal patterns of volatility and trading activity. Since then, several different approaches have been tested, Pierre (2015), Anderson & Bollerslev (1997) and in a later stage, Engle & Sokalska (2012), suggested an approach to model conditional variance expressed as a product of daily and diurnal components (see chapter Multiplicative GARCH (henceforth mcGARCH)*)*. During this study, Engle & Sokalska (2012), methodology was adopted to model the intraday volatility of Tesco and SKY using intervals of 15 minutes taking into account the FTSE100 market returns, news and tweets that are released into the market in the last four observations of time (i.e. in the last hour). Several equations will be performed to compare the impact of adding the news and tweets as repressors. In line with previous studies of intraday volatility, see Engle & Sokalska (2012) and Ghalanos (2015), the first return of the day (09:00-09:15) will be excluded from the models.

## *4.2 Standard GARCH (henceforth sGARCH)*

Before the Multiplicative GARCH (henceforth mcGARCH) it is presented, it will be provided the background of the ARCH and GARCH techniques. This type of technique,

[13] For details, see (Nicolau, Modelação de Séries Temporais Financeiras, 2012), chapter 6.

Autoregressive Conditional Heteroskedasticity (ARCH), was first introduce by Engle (1982) and then generalized to GARCH by Bollerslev (1986) and Taylor (1986) to model the conditional variance of the dependent variable. As mentioned previously, the volatility modelling has been an area of interest for several academics such as Andersen, Bollerslev & Cai (2000), Anderson & Bollerslev (1997), Engle & Sokalska (2012), Fuertes (2009), Skiena & Zhang (2010) and Oh & Sheng (2011) and GARCH was a technique of reference among those studies to model and forecast volatility among financial time series. There are also some other reasons to choose GARCH models, such as, understanding the risk of an asset (typically described by the standard deviation of the stock prices returns, the higher the metric the higher the risk), or, if proven that the returns have a time-dependent variance, robustly predict confidence intervals for the stock price.

Below it is presented the general GARCH model specification:

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right), P_t = Stock\ Price\ at\ period\ t \quad (1)$$

$$r_t = c + X_t{}'\beta + u_t \quad (2)$$

$$u_t = \varepsilon_t \sigma_t \quad (3)$$

$$\sigma^2{}_t = \omega + G_t{}'\gamma + \alpha_1 u^2{}_{t-1} + \cdots + \alpha_p u^2{}_{t-p} + b_1 \sigma^2{}_{t-1} + \cdots + b_q \sigma^2{}_{t-q} \quad (4)$$

Equation 2 describes the mean equation of a given stock price returns ($r_t$), explained by a set of covariates ($X_t$), which can include the market return of the respective stock, number of positive, negative or neutral news and tweets from any given source (e.g. Yahoo Finance and Financial Times), and an error term $u_t$, that has a time dependent variance $\sigma_t$ and a random variable $\varepsilon_t$ that, typically, is assumed to be independent and identically distributed (i.i.d.) with $\mathrm{E}(\varepsilon) = 0, Var(\varepsilon) = 1$, and independent of $u_{t-k}, k \in \mathbb{N}$. Equation 4 describes the conditional variance, which can be express as a set of external regressors ($G_t$), similar to $X_t$, past squared errors ($e.g.\ u^2{}_{t-p}$) and past conditional variances ($e.g.\ \sigma^2{}_{t-p}$).

### *4.3 Multiplicative GARCH (henceforth mcGARCH)*

As previously mentioned, Engle & Sokalska (2012) have suggested a multiplicative approach for modelling intraday financial time series that incorporates daily, diurnal and stochastic (intraday) components. Let us consider a continuous returns $r_{t,i}$, where *t* represents a specific day and *i* the intraday time interval at which the return was calculated. Hence, the conditional variance is a multiplicative product of daily, diurnal and stochastic (intraday) components, as described below:

$$r_{t,i} = c + X_{t,i}{}'\beta + u_{t,i} \tag{5}$$

$$u_{t,i} = \left(q_{t,i}\sigma_t s_i\right)z_{t,i} \tag{6}$$

$$s_i = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{u_{t,i}^2}{\sigma_t^2}\right) \tag{7}$$

$$\bar{u}_{t,i} = \frac{u_{t,i}}{\sigma_t\sqrt{s_i}} \tag{8}$$

$$q^2{}_{t,i} = \omega + G_t{}'\gamma + \alpha_1\bar{u}^2{}_{t-1} + \cdots + \alpha_p\bar{u}^2{}_{t-p} + b_1q^2{}_{t-1} + \cdots + b_qq^2{}_{t-q} \tag{9}$$

Equation 5 is very similar to Equation 1 with the only difference in the way the errors $u_{t,i}$ are specified, which is detailed Equation 6. In the latter, $q_{t,i}$ is the stochastic intraday volatility, $\sigma_t$is the daily exogenous forecasted volatility (we will discuss later how this can be generated), $s_i$ is the diurnal component evaluated at every time interval (in our case every 15 minutes) and finally $z_{t,i}$, a random variable (i.i.d.) which can take several distributions, e.g. normal-distribution or t-distribution. Equation 8 details how the standardized residuals are calculated and finally Equation 9 is the GARCH equation of the intraday volatility, which depends on a set of external covariates, similar to $X_t$ in Equation 1, past squared standardized residuals $\bar{u}^2{}_{t-p}$ and past intraday volatilities ($q^2{}_{t-q}$). Please note that this method requires an exogenous daily volatility forecast $\sigma_t$ which can be obtained by a

commercial company, as described in Engle & Sokalska (2012), or can be estimated using a sGARCH(1,1). The approach taken in this study was to gather the daily returns for each stock from January 2010 and then estimate and predict a daily volatility using a AR(1)[14]+GARCH(1,1) and then embed those results into the mcGARCH. This approach has also been suggested by Engle & Sokalska (2012) and (Ghalanos, 2015)[15].

### *4.4 Estimation*

There are several ways to estimate GARCH models, however, as discussed in Nicolau (2012), the method that typically is adopted is the Maximum Likelihood, which is detailed below. Please note that a 2-step GMM estimator can also be used. In fact, Engle & Sokalska (2012) describes the properties of this estimator in the context mcGARCH. However, in this study the mcGARCH is estimated using the Maximum Likelihood method, as explained in Ghalanos (2015). In addition, and as explained in Nicolau (2012), Bollerslev & Wooldridge (1992) and Ghalanos (2015), the distribution of $z_{t,i}$ from Equation 6 is typically unknown. However, it can still be assumed the same conditions for $z_{t,i}$, e.g. ~ N(0,1) and obtain consistent estimators for the GARCH mean and variance equations, even if the true distribution of $z_{t,i}$ is different. In this case, it is applied a Pseudo-Maximum-Likelihood estimator which uses a different but robust variance-covariance matrix, i.e. $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N(0, A(\theta_0)^{-1}B(\theta_0)A(\theta_0)^{-1})$ instead of the standard $I(\theta_0)^{-1}$; where $A(\theta_0)$ and $B(\theta_0)$ are the first and second order derivatives of the Log-Likelihood function, respectively, and $I(\theta_0)$ is the Fisher Information matrix, which equals $-I(\theta_0)$. More details can be found in Nicolau (2012).

[14] For details, see Nicolau (2012), chapter 6.

[15] See a tutorial here: http://unstarched.net/2013/03/20/high-frequency-garch-the-multiplicative-component-garch-mcsgarch-model/

## 4.5 Fitting

### 4.5.1 Pre-estimation

While in financial time series data it is expected that the conditional variance is correlated with the previous residuals and past conditional volatility (e.g. GARCH(1,1)), it is important to perform the statistical tests to validate such hypothesis. In this study two tests will be performed to prove the need for GARCH effects:

Arch Test:

Consider equations 2 and 3 and $$u^2{}_t = \omega + \alpha_1 u^2{}_{t-1} + \cdots + \alpha_p u^2{}_{t-p} + e_t \quad (10)$$

Under H0: $\alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$, $$nR^2 \xrightarrow{d} \mathcal{X}^2{}_{(p)} \quad (11)$$

Where $R^2$ is determination coefficient of the regression of $u^2{}_t$ in $(1, u^2{}_{t-1}, \ldots, u^2{}_{t-p})$.

Ljung-Box[16] test of ACF[17] of $u^2{}_t$:

This test is an equivalent test as the ARCH Test and under the null hypothesis, it is tested if $\rho_1(u^2{}_t) = \cdots = \rho_m(u^2{}_t) = 0$, where $\rho_i$ is the autocorrelation coefficient between $u^2{}_t$ and $u^2{}_{t-i}$. More specifically, under H0:

$$Q = n(n+2)\sum_{i=1}^{m} \frac{1}{n-1} \hat{\rho}_i{}^2(u^2{}_t) \xrightarrow{d} \mathcal{X}^2{}_{(m-k)} \quad (12)$$

Where k is the number of parameters estimated minus the constant.

Fisher & Gallagher (2012) have suggested an improvement for both of the above tests. In their paper, they introduce a weighted matrix which better account for the distribution of the statistics of the parameters of the tests described above. During this study it is used Fisher & Gallagher (2012) methodology for diagnostics purposes.

[16] This test will also be use in pos-estimation diagnostics.
[17] Autocorrelation function can be seen in Nicolau (2012)

In both tests, evidence against the null hypothesis suggests a GARCH effect.

### 4.5.2 Pos-estimation

#### 4.5.2.1 Diagnostic Test:

Following the previous example, If the model has been correctly specified, $\varepsilon_t$ in (3) won't be correlated and should be conditionally homoscedastic. A way to confirm this will be to perform a weight Ljung-Box test described by Fisher & Gallagher (2012) on both $\varepsilon_t$ and ${\varepsilon_t}^2$ and expect that the null hypothesis is not rejected on both cases.

## *4.6 Forecasting*

The main purpose of this study is to perform n-ahead forecasts of the stocks returns and volatilities and understand if there is a gain in accuracy by adding short-term news and tweets as covariates of the mean and/or variance equation of the GARCH model. Therefore, and after taking into account what other authors have accomplished, e.g. Engle & Sokalska (2012), performing one-step ahead forecast for intraday data seems the most appropriate, given the context of intraday trading. With this in mind, it was developed a rolling-forecasting methodology with the following characteristics:

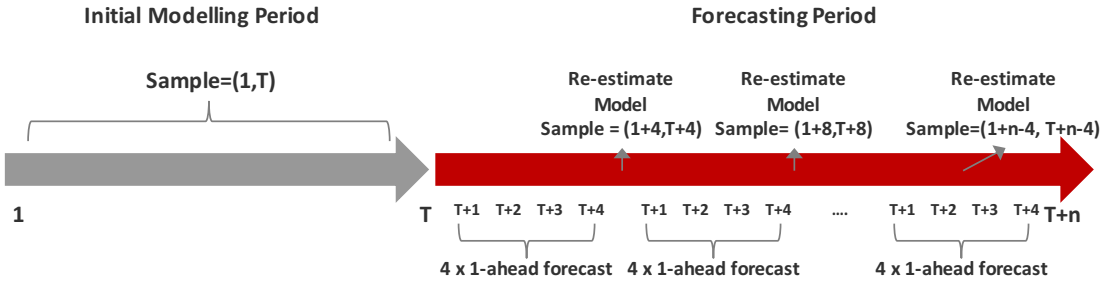



**Figure 2 – Rolling Forecasting High-level methodology**

As can be seen by Figure 2, the rolling forecasting methodology applied implies that after 4 times 1-ahead forecasts the initial model be re-estimated using a new sample (1+4,T+4). This concept is known as rolling-window forecasting and can be found in the

literature in several references, e.g. in Ghalanos (2015). It is important to highlight two assumptions about this methodology: First, it is assumed that the properties of the first model will remain valid in the future re-estimated models, i.e. if the model diagnostics for the first model are consider to be valid, it is assumed that for all of the following models those results remain valid. Secondly, it is assumed that during 3-ahead steps, the model coefficients remain valid to predict the 1-ahead forecast. In a context of a trading agency, this is saying that the models will only be updated every hour and within the hour the 1-ahead predictions will use the same model. Below, it will be explained how those forecasts are calculated and evaluated.

### 4.6.1 One-step ahead forecast

One-step ahead forecast of the returns and conditional volatility is relatively straight-forward to be calculated, given the fact that, in the majority of the GARCH models, both the returns and conditional variance depend on previous values. Let us consider the following AR(1)-GARCH(1,1) model:

$$r_t = c + \phi_1 r_{t-1} + u_t, |\phi_1| < 1 \tag{13}$$

$$u_t = \varepsilon_t \sigma_t \tag{14}$$

$$\sigma^2{}_t = \omega + \alpha_1 u^2{}_{t-1} + b_1 \sigma^2{}_{t-1} \tag{15}$$

The one-step ahead forecasts for $r_{t+1}$ and $\sigma^2{}_{t+1}$ are given by:

$$\widehat{\sigma^2}_{t+1,t} = \hat{E}(\omega + \alpha_1 u^2{}_{t-1} + b_1 \sigma^2{}_{t-1} | \, \mathfrak{T}_t) = \hat{\omega} + \hat{\alpha}_1 u^2{}_t + \hat{b}_1 \sigma^2{}_t \tag{16}$$

$$\hat{r}_{t+1,t} = \hat{E}(c + \phi_1 r_t + u_t | \, \mathfrak{T}_t) = \hat{c} + \hat{\phi}_1 r_t \tag{17}$$

Where the parameters $\hat{\omega}, \hat{\alpha}_1, \hat{b}_1, \hat{c}$ and $\hat{\phi}_1$ are the estimated values of the models. In the context of the mcGARCH there are some additional parameters that are calculated, more specifically:

$$\hat{s}_i = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{u}_{t,i}^2}{\hat{\sigma}_t^2}\right) \quad (18)$$

$$\hat{z}^2{}_{t,i} = \frac{r_{t,i}^2}{\hat{\sigma}_t \hat{s}_i} \quad (19)$$

Where $\hat{z}^2{}_{t,i}$, a squared return metric, is a random variable drawn from a distribution with a variance that we are trying to measure with the mcGARCH model. More details about this measure can be found in Engle & Sokalska (2012) and Patton (2010). These additional parameters will be important in the forecasting evaluation analytics, described below.

### 4.6.2 Evaluation of forecasts

Given the fact that both returns and volatility are predicted, the evaluation of forecasts will be made by comparing four metrics, two for each forecasting type, the Mean-Absolute-Error metric (MAE), Mean Directional Accuracy (MDA) for the returns predictions and the mean of two loss functions to measure the accuracy of the volatility predictions, introduced by Patton (2010) and also explored by Engle & Sokalska (2012). Please note that the selection of this two loss functions are related with recent results that prove that this two metrics are consistent as long as volatility proxies are unbiased, see Hansen and Lunde (2006) and Patton (2010) for details.

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|r_t - \hat{r}_t| \quad (20)$$

$$MDA = \frac{1}{T}\sum_{t=1}^{T} DA\ Flag \quad (21)$$

$$Loss\ function\ 1\ (L1)_{t,i} = \log \hat{q}_{t,i} + \frac{z^2{}_{t,i}}{\hat{q}_{t,i}};\ \text{Mean-L1} = \frac{1}{T}\sum_{t=1}^{T} L1 \quad (22)$$

$$Loss\ function\ 2\ (L2)_{t,i} = (z^2{}_{t,i} - \hat{q}_{t,i})^2;\ \text{Mean-L2} = \frac{1}{T}\sum_{t=1}^{T} L2 \quad (23)$$

Where $\hat{r}_t\hat{}$, is the forecasted returns given by a specific model, "DA Flag" is a binary variable that equals 1 if the $sign(\hat{r}_t) = sign(r_t)$, 0 otherwise, $\hat{q}_{t,i}$ is the intraday volatility

given in Equation 9 and $z^2$ was explained in equation 19. With the exception of the MDA, the smaller the metric the better the forecast is.

## *4.7 Use Case Introduction*

Potentially, using the news and tweets about Tesco and SKY in the short-term (e.g. in the last hour) can be used as a proxy for what the market believes to be the future value of the stock. Engle & Sokalska (2012), in the context of Anderson & Bollerslev (1997) suggestion of adding dummies for the macroeconomic news that arrived in the market, said that the traders may react to the macroeconomic news that are coming to the market only if they are unexpected, i.e. they couldn't be forecasted in advance. Also, this type of news typically are released when the US Market is still closed and therefore not make an impact in the intraday volatility (more on the open volatility, which by default, is excluded in the intraday volatility studies). In this situation, the equity of interest is traded in the European Market (FTSE100) but, the remaining points from Engle & Sokalska (2012) could remain valid, i.e. only news that are not expected may change their trading behavior. One interesting aspect to keep in mind is how fast online companies publish the news and tweets that are considered relevant for impacting the trading behavior, i.e. if Google Finance, Yahoo Finance, Financial Times[18] publish the same news but in different timings, which one is causing more effect in the volatility and stock returns?

It in fact, we should expect to see some type of correlation between the stock price returns and the news and tweets that come out in the market. Indeed, if this effect proves to be

[18] This study doesn't include other online news providers such as Bloomberg or Reuters. This is a limitation of the current study given the fact that there wasn't any free APIs at the point where the online platform was created.

valid, we should expect that predicting volatility (even for intraday returns) including the most recent news and tweets about Tesco will improve the forecast accuracy.

## *4.8 Use Case 1 – Tesco*

As can be seen in Figure 3 – Tesco Stock Price vs. FTSE100, Tesco stocks prices have a negative trend in the recent months. Part of this trend can be explained by the negative trend in the market that Tesco is traded, FTSE100, however there is a remaining part of this trend that have to be explained by Tesco's performance, expectation of growth, innovation, among many other financial and economical reasons.

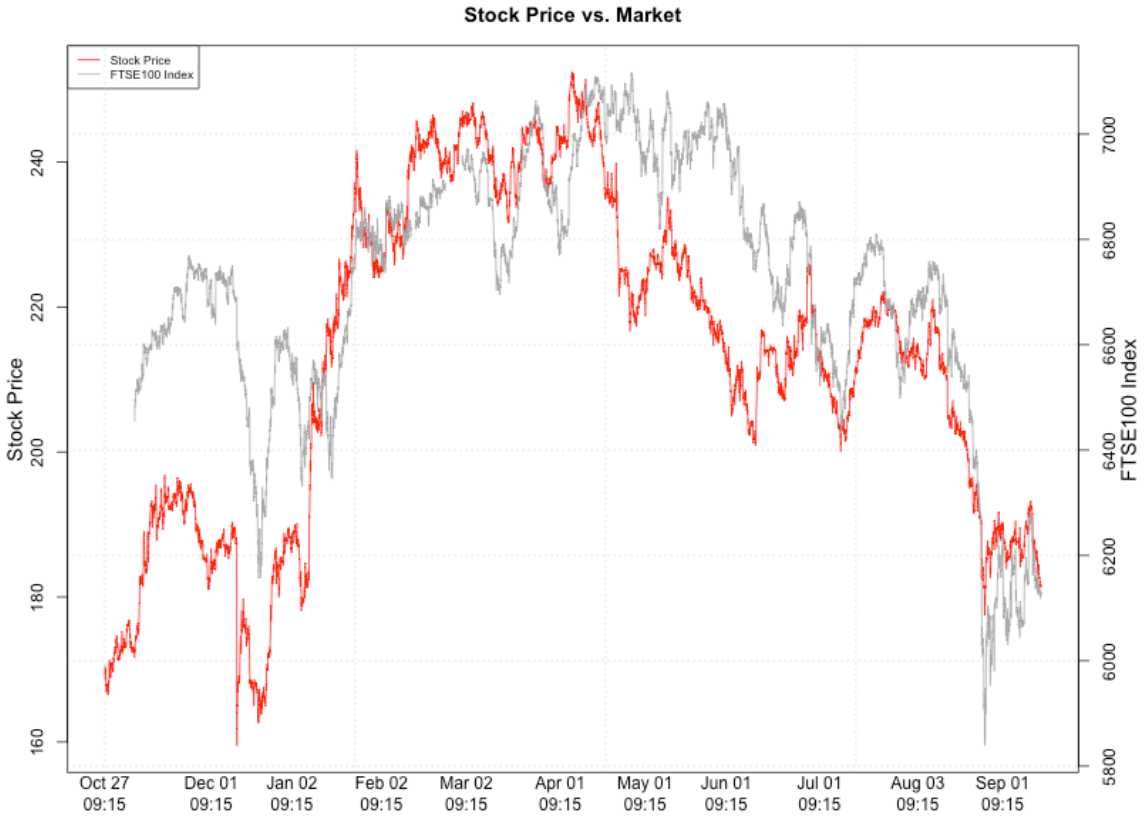



**Figure 3 – Tesco Stock Price vs. FTSE100**

In the next section, several econometric models will be created to measure the impact of news and tweets in the stock returns. Then, a forecasting analysis will be performed to understand the forecast accuracy of those models and for the best model.

### 4.8.1 Time series analysis

Before the definition of the models, an analysis of the multiple time series of interest will be performed to understand the key patterns within the data.

From Figure 4 – Tesco Returns vs. FTSE100 Returns vs. News and Tweets is not that clear the relationship between news and tweets and the returns. It is possible to see a period of higher volatility in November till December (with some persistence throughout the way), and some picks in volatility in May, April and most recently in August. The biggest spike in Tesco's volatility in December is explained by the unexpected announcements of Tesco's financials, which ended with the entire executive team being replaced. The period of higher volatility in Tesco returns is slightly visible in the FTSE returns volatility (Nov-Dec), however, the market biggest volatility occurs between August till September and this pattern is not totally visible in Tesco.

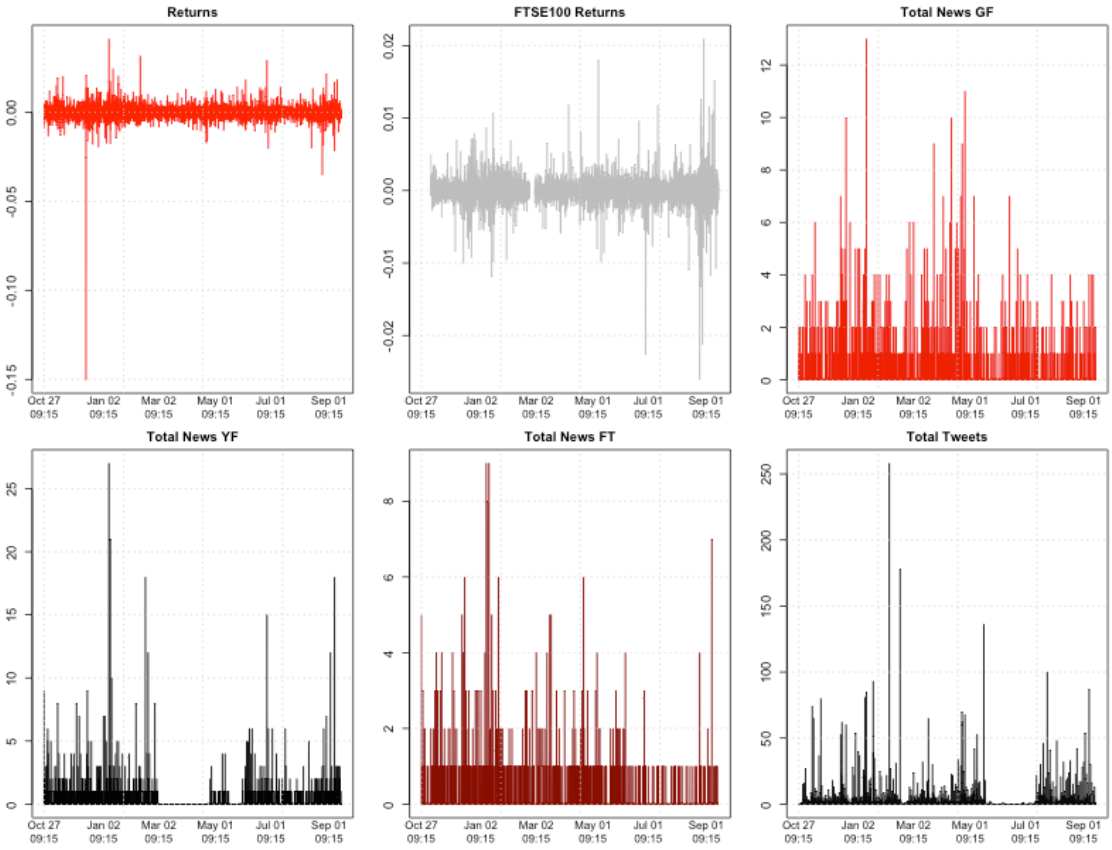



**Figure 4 – Tesco Returns vs. FTSE100 Returns vs. News and Tweets**

It is also possible to see some issues with the data extractions from Twitter in the period between end of May until 20th of July. For this reason, the models that included Tesco will only use a smaller sample, i.e. the models will start in 20/07/2015.

The correlation matrix below describes the key cross-correlation between the above variables lagged up-to one hour (t=-1,-2,-3,-4):

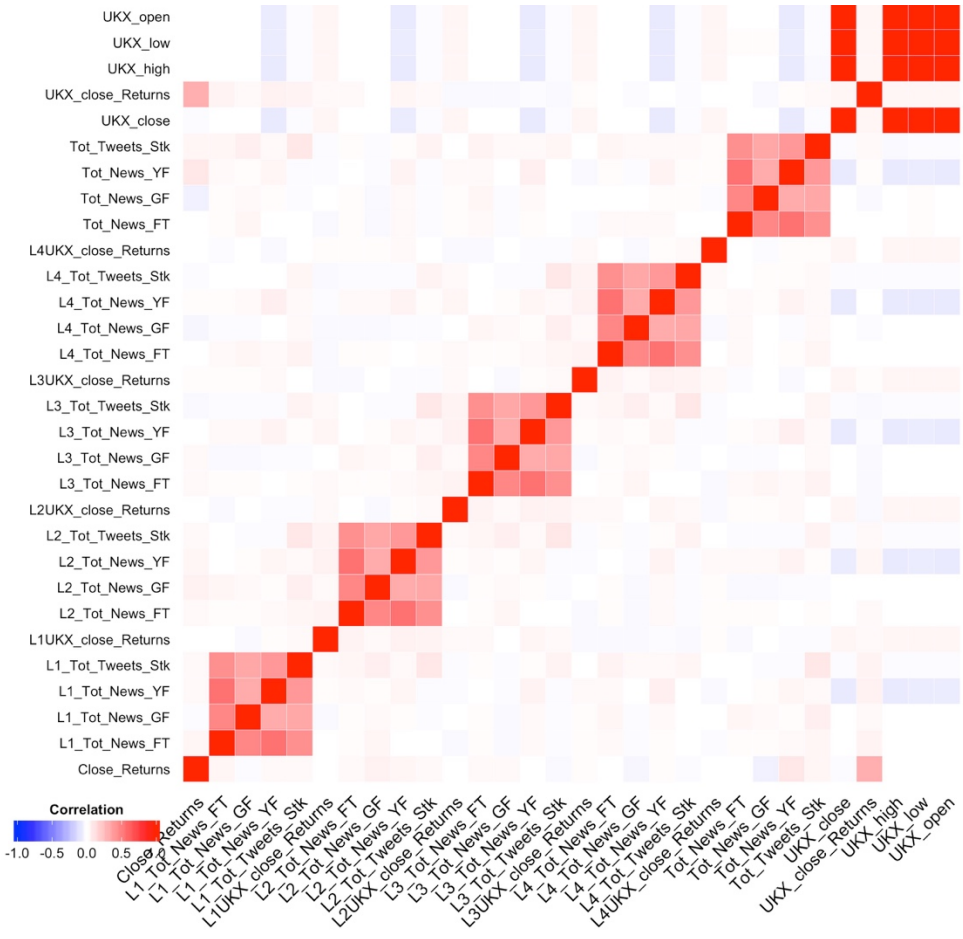


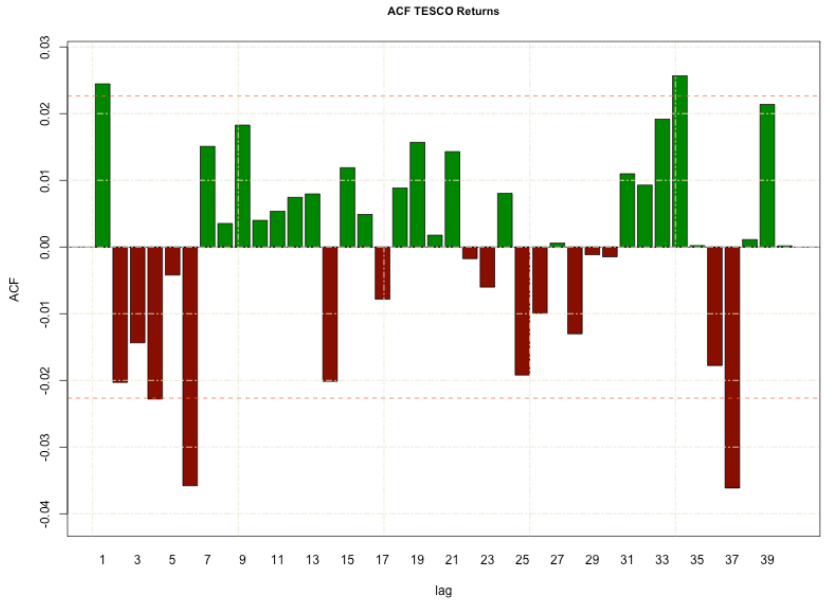

**Figure 5 – Cross-Correlation between Tesco Returns vs. FTSE100 Returns vs. News and Tweets up-to 4 Lags**

From the figure above, a couple of insights can be taken, 1) Very low correlations between news and tweets and returns, 2) Positive, but still not strong correlation between the contemporaneous Market return (UKX_close_Returns in the figure) and Tesco Return (Close_Returns), 3) Positive correlation among the contemporaneous news and tweets. This analysis was repeated up-to 80 lags to understand the main differences. The results don't really differ apart from a slightly higher correlation (but still small) between the news from financial times and the returns in the lag 34 (one day before). One potential explanation is

that Financial Times news are journalist articles, bigger in length and typically with more analysis when compared with the news and tweets from Google Finance, Yahoo Finance and Twitter. As the first return of the day is excluded by the analysis (similar to many other studies of intraday volatility) this makes no difference from an inference perspective.

Finally, it will be analyzed the ACF for the $r_t$ and ${r_t}^2$, the kurtosis and skewness to validate some of the properties of the Tesco returns.

Staring by the kurtosis, and as expected, the value is higher than 3 (the normal-distributed value), ~6.55. Also as expected, the skewness is negative and ~-0.073.



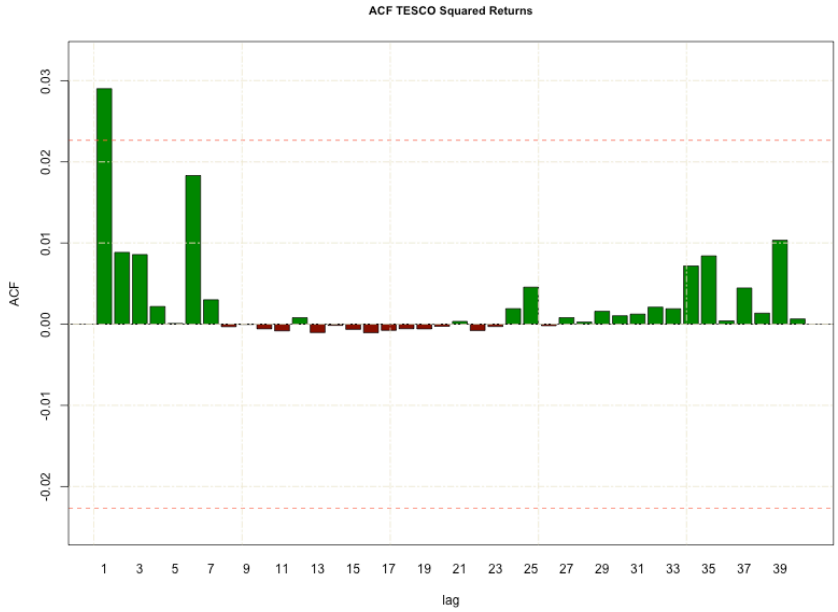



**Figure 6 – ACF for Tesco Returns and Tesco Squared Returns**

From the above figures it can be seen that Tesco Returns don't seem to be auto-correlated with past values, however, Tesco Squared Returns seem to be (an indication of GARCH effects), specially in the first lags. The Ljung-Box test shows evidence of that.

| Lag | Tesco Returns | | | Tesco Squared Returns | | |
|---|---|---|---|---|---|---|
| | ACF | P_Value | Above 5% | ACF | P_Value | Above 5% |
| 1 | -0.02 | 0.17 | | 0.08 | 0.00 | |
| 2 | -0.01 | 0.25 | | 0.13 | 0.00 | |
| 3 | 0.00 | 0.41 | | 0.12 | 0.00 | |
| 4 | 0.00 | 0.57 | | 0.05 | 0.00 | |
| 5 | 0.00 | 0.71 | | 0.02 | 0.00 | |
| 6 | 0.02 | 0.59 | FALSE | 0.04 | 0.00 | TRUE |
| 7 | 0.02 | 0.50 | | 0.10 | 0.00 | |
| 8 | 0.01 | 0.54 | | 0.06 | 0.00 | |
| 9 | 0.01 | 0.62 | | 0.02 | 0.00 | |
| 10 | 0.00 | 0.71 | | 0.07 | 0.00 | |
| 11 | -0.02 | 0.63 | | 0.00 | 0.00 | |
| 12 | 0.00 | 0.71 | | 0.03 | 0.00 | |

**Table 4 – Ljung-Box test for Tesco Returns and Squared Returns**

### 4.8.2 Models specification

Below it is detailed the model equations developed for Tesco. Each model was run twice, one with Twitter tweets and one without, for the reason mentioned above.

| Model Type | Modelling Sample | Sample size | Rolling Forecasting Sample | Sample size | Twitter |
|---|---|---|---|---|---|
| 1 | 01/04/2015 09:30-20/08/2015 17:30 | 5,280 | 21/08/2015 9:30-11/09/2015 17:30 | 495 | No |
| 2 | 20/07/2015 09:30:00-28/08/2015 17:30 | 990 | 01/09/2015 09:30-11/09/2015 17:30 | 297 | Yes |

**Table 5 – Model and Forecasting sample sizes**

Please note that for simplicity, when the index of a variable only includes t it refers to daily returns, when it includes t,i it is referring to an intraday period of a given day. To avoid repeating the same equation for the daily variance, please consider that all equations below that use mcGARCH have a daily variance forecasted with $r_t \rightarrow \mathrm{AR}(1)$; $\sigma_t \rightarrow s\mathrm{GARCH}(1,1)$. In addition, the equations below reflect the models with Twitter tweets, but it is easy to make the parallels with the models without twitter as is just a change in the index *j* from 4 to 3.

| | | |
|---|---|---|
| Model 1: | $r_{t,i} = c + \phi_1 r_{t-1,i} + u_{t,i}$<br>$u_{t,i} \rightarrow s\text{GARCH}(1,1)$ | (24) |
| Model 2: | $r_{t,i} = c + \phi_1 r_{t-1,i} \sum_{k=1}^{4} \gamma_k {r^m}_{t-k,i} + \sum_{j=1}^{4}\sum_{k=1}^{4} \beta_{jk} X_{t-k,j,i} + u_{t,i}$<br>$u_{t,i} \rightarrow \text{sGARCH}(1,1)$ | (25) |
| Model 3: | $r_{t,i} = c + \phi_1 r_{t-1,i} + u_{t,i}$<br>$u_{t,i} \rightarrow \text{mcGARCH}(1,1);$ | (26) |
| Model 4: | $r_{t,i} = c + \phi_1 r_{t-1,i} + \sum_{k=1}^{4} \gamma_k {r^m}_{t-k,i} + \sum_{j=1}^{4}\sum_{k=1}^{4} \beta_{jk} X_{t-k,j,i} + u_{t,i};$<br>$u_{t,i} \rightarrow \text{mcGARCH}(1,1);$ | (27) |
| Model 5: | $r_{t,i} = c + \phi_1 r_{t-1,i} + \sum_{k=1}^{4} \gamma_k {r^m}_{t-k,i} + \sum_{j=1}^{4}\sum_{k=1}^{4} \beta_{jk} X_{t-k,j,i} + u_{t,i};$<br>$u_{t,i} \rightarrow \text{mcGARCH}(1,1)$ with<br>${q^2}_{t,i} = \omega + \sum_{j=1}^{4}\sum_{k=1}^{4} \gamma_{jk} G_{t-k,j,i} + \alpha_1 {\bar{u}^2}_{t-1} + b_1 {q^2}_{t-1}$ ; | (28) |
| Model 6: | $r_{t,i} = c + \phi_1 r_{t-1,i} + \sum_{j=1}^{4}\sum_{k=1}^{4} \beta_{jk} X_{t-k,j,i} + u_{t,i};$<br>$u_{t,i} \rightarrow \text{mcGARCH}(1,1);$ | (29) |
| Model 7: | $r_{t,i} = c + \phi_1 r_{t-1,i} + \sum_{k=1}^{4} \gamma_k {r^m}_{t-k,i} + u_{t,i};$<br>$u_{t,i} \rightarrow \text{mcGARCH}(1,1);$ | (30) |

Where *i* is the intraday interval, *j* is the index for the news and tweets source (1 = Google Finance, 2 = Yahoo Finance, 3 = Financial Times and 4 = Twitter[19]), *k* = the number of lags (1 till 4 included), $r^m$ = FTSE100 returns, $X_{t-k,j,i}$ = $G_{t-k,j,i}$ is the sum of the news (tweets) for the period *t-k*, from the source *j* for the intraday interval *I*. sGARCH(1,1) is defined in (4) with p=q=1 and without any external repressors *G*. $\text{mcGARCH}(1,1)$ is defined in (9) with p=q=1 and without external covariates G. Equation (24), estimates the AR(1)+sGARCH(1,1). This model should confirm what the literature says about sGARCH not presenting satisfactory results to model intraday volatility. Model 2 expands Model 1 to include the short-term

[19] Twitter was only included in the models with a smaller sample size.

market returns, news and tweets. Again, this model should not present satisfactory results, hence, should be use as a comparison. Equation (24), estimates an AR(1) + mcGARCH(1,1) with daily volatility being computed by a AR(1)+GARCH(1,1). This model can be seen as a baseline for comparison purposes, as is very similar to the model presented by Engle & Sokalska (2012). Equation (28), is almost an extension of the previous model, where it is added the lagged market returns, news from multiple sources and tweets. (28) is an extension of Model 4, where the stochastic volatility metric $q^2_{t,i}$ incorporates the short-term news and tweets $G_{t-k,j,i}$. The idea behind this model is to understand if the short-term news and tweets are also important to predict the intraday volatility. Equations (29) and (30) are reductions from (28), where in the first excludes market returns and in the second news and tweets are excluded.

### 4.8.3 Models results

From Table 6 – Tesco model results it can be analyzed the coefficients of the models with and without Twitter, robust standard errors (the values inside parenthesis) and the statistical relevance of the parameter at three confidence levels, "***", "**", "*" - 99%, 95% and 90%, respectively. Additional information is available at the end of the table, for example the Log-Likelihood, kurtosis and skewness of the standardized residuals and the p-value of the normality test Jarque Bera.

From the outputs of the models without Twitter it can be seen that the AR(1) parameters, $\phi_1$, have a very low negative value across all models, varying from -0.016 to -0.051 and is only relevant on model 2 at 10%. These results are expected, as the returns shouldn't have a higher dependence on previous returns (ACF for Tesco also showed this). In fact, some authors don't even include any AR or MA component in the mean equation.

| Coefficient | Model 1 - AR(1)+GARCH(1,1) | | Model 2 - AR(1)+GARCH(1,1) | | Model 3 - AR(1)+msGARCH(1,1) | | Model 4 - AR(1)+msGARCH(1,1) | | Model 5 - AR(1)+msGARCH(1,1) | | | | Model 6 - AR(1)+msGARCH(1,1) | | Model 7 - AR(1)+msGARCH(1,1) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Standard Coefficients | without Twitter | Twitter | without Twitter | Twitter | without Twitter | Twitter | without Twitter | Twitter | without Twitter | | Twitter | | without Twitter | Twitter | without Twitter | Twitter |
| c | -0.0001 (0) | -0.0001 (0.0007) | -0.0001 (0) | 0 (0.0001) | -0.0001 (0) ** | -0.0001 (0.0001) ** | -0.0001 (0) * | -0.0001 (0.0001) | -0.0001 (0) * | | -0.0001 (0.0001) | | -0.0001 (0) ** | 0 (0.0001) | -0.0001 (0) ** | -0.0001 (0.0001) ** |
| $\phi_1$ | -0.0331 (0.0238) | -0.0168 (0.0534) | -0.0511 (0.0261) * | -0.0549 (0.0777) | -0.0197 (0.0202) | 0.001 (0.0356) | -0.0347 (0.0222) | -0.0197 (0.0498) | -0.0347 (0.0225) | | -0.0197 (0.0742) | | -0.0165 (0.0199) | 0.0226 (0.0417) | -0.0343 (0.0226) | -0.038 (0.0481) |
| w | 0 (0) | 0 (0.0002) | 0 (0) | 0 (0) | 0.0676 (0.0375) * | 0.2109 (0.1435) | 0.0731 (0.0403) * | 0.1906 (0.1471) | 0.0731 (0.049) | | 0.1906 (0.3707) | | 0.0624 (0.0372) * | 0.1401 (0.1953) | 0.0677 (0.038) * | 0.2204 (0.1205) * |
| $a_1$ | 0.1591 (0.0592) *** | 0.1611 (2.5783) | 0.1607 (0.074) ** | 0.2284 (0.1478) | 0.1203 (0.0371) *** | 0.1432 (0.0413) *** | 0.1276 (0.039) *** | 0.153 (0.0557) *** | 0.1276 (0.0501) ** | | 0.153 (0.1648) | | 0.0943 (0.0261) *** | 0.115 (0.0626) * | 0.1182 (0.0368) *** | 0.157 (0.0427) *** |
| $b_1$ | 0.701 (0.0151) *** | 0.7591 (2.6208) | 0.7096 (0.0158) *** | 0.6549 (0.1232) *** | 0.8172 (0.0668) *** | 0.6471 (0.1764) *** | 0.8044 (0.0709) *** | 0.6582 (0.1911) *** | 0.8044 (0.0874) *** | | 0.6582 (0.5107) | | 0.8405 (0.062) *** | 0.7426 (0.2608) *** | 0.8188 (0.0674) *** | 0.6239 (0.1486) *** |
| Market Returns | | | | | | | | | | | | | | | | |
| $\Upsilon_1$ | - | - | 0.0657 (0.0455) | 0.0896 (0.1173) | - | - | 0.0586 (0.0369) | 0.0793 (0.0664) | 0.0586 (0.0371) | | 0.0793 (0.0957) | | - | - | 0.0589 (0.0368) | 0.1048 (0.0628) * |
| $\Upsilon_2$ | - | - | 0.0038 (0.0322) | -0.0637 (0.0705) | - | - | -0.0137 (0.03) | -0.0661 (0.0398) * | -0.0137 (0.0295) | | -0.0661 (0.0518) | | - | - | -0.0187 (0.0297) | -0.0689 (0.0378) * |
| $\Upsilon_3$ | - | - | -0.0043 (0.0307) | 0.0396 (0.0599) | - | - | 0.0004 (0.0283) | 0.0332 (0.0404) | 0.0004 (0.0315) | | 0.0332 (0.0442) | | - | - | -0.0031 (0.0291) | 0.0303 (0.0354) |
| $\Upsilon_4$ | - | - | -0.0205 (0.0349) | -0.0811 (0.1072) | - | - | -0.0047 (0.0282) | -0.02 (0.0485) | -0.0047 (0.028) | | -0.02 (0.0574) | | - | - | -0.0044 (0.0284) | -0.0207 (0.049) |
| Google Finance | | | | | | | | | Mean | Variance | Mean | Variance | | | | |
| $\beta_{11}$ | - | - | 0.0001 (0.0002) | 0 (0.0004) | - | - | 0 (0.0001) | 0.0002 (0.0002) | 0 (0.0001) | 0 (0.0403) | 0.0002 (0.0005) | 0 (0.3968) | 0.0001 (0.0001) | 0.0002 (0.0002) | - | - |
| $\beta_{12}$ | - | - | 0.0002 (0.0001) * | -0.0002 (0.0003) | - | - | 0.0002 (0.0001) ** | -0.0003 (0.0002) | 0.0002 (0.0001) * | 0 (0.1364) | -0.0003 (0.0004) | 0 (0.325) | 0.0002 (0.0001) ** | -0.0004 (0.0002) * | - | - |
| $\beta_{13}$ | - | - | 0 (0.0001) | -0.0001 (0.0002) | - | - | 0 (0.0001) | 0 (0.0002) | 0 (0.0001) | 0 (0.1083) | 0 (0.0004) | 0 (0.4421) | 0 (0.0001) | 0 (0.0002) | - | - |
| $\beta_{14}$ | - | - | -0.0002 (0.0001) | -0.0005 (0.0002) ** | - | - | -0.0002 (0.0001) * | -0.0004 (0.0002) ** | -0.0002 (0.0001) | 0 (0.0549) | -0.0004 (0.0003) | 0 (0.5615) | -0.0002 (0.0001) * | -0.0004 (0.0002) ** | - | - |
| Yahoo Finance | | | | | | | | | Mean | Variance | Mean | Variance | | | | |
| $\beta_{21}$ | - | - | 0 (0.0002) | -0.0005 (0.0003) * | - | - | 0 (0.0001) | -0.0003 (0.0003) | 0 (0.0002) | 0 (0.0967) | -0.0003 (0.0004) | 0 (0.411) | 0 (0.0001) | -0.0004 (0.0003) | - | - |
| $\beta_{22}$ | - | - | 0.0001 (0.0002) | 0.0001 (0.0002) | - | - | 0.0002 (0.0001) | 0.0002 (0.0002) | 0.0002 (0.0001) | 0 (0.092) | 0.0002 (0.0012) | 0 (1.1398) | 0.0002 (0.0001) | 0.0001 (0.0002) | - | - |
| $\beta_{23}$ | - | - | 0.0001 (0.0001) | -0.0001 (0.0002) | - | - | 0.0001 (0.0001) | 0 (0.0002) | 0.0001 (0.0001) | 0 (0.0836) | 0 (0.0003) | 0 (0.9395) | 0.0001 (0.0001) | -0.0001 (0.0002) | - | - |
| $\beta_{24}$ | - | - | 0.0001 (0.0001) | -0.0002 (0.0002) | - | - | 0.0001 (0.0001) | -0.0001 (0.0002) | 0.0001 (0.0001) | 0 (0.0739) | -0.0001 (0.0003) | 0 (0.3728) | 0.0001 (0.0001) | -0.0001 (0.0002) | - | - |
| Financial Times | | | | | | | | | Mean | Variance | Mean | Variance | | | | |
| $\beta_{31}$ | - | - | -0.0001 (0.0003) | -0.0001 (0.001) | - | - | -0.0004 (0.0002) ** | -0.0005 (0.0005) | -0.0004 (0.0003) | 0 (0.0684) | -0.0005 (0.0015) | 0 (0.3361) | -0.0004 (0.0002) * | -0.0006 (0.0005) | - | - |
| $\beta_{32}$ | - | - | -0.0001 (0.0002) | -0.0002 (0.0011) | - | - | -0.0001 (0.0002) | 0.0004 (0.0007) | -0.0001 (0.0002) | 0 (0.1704) | 0.0004 (0.0015) | 0 (0.6114) | -0.0001 (0.0002) | 0.0004 (0.0007) | - | - |
| $\beta_{33}$ | - | - | -0.0002 (0.0002) | -0.0004 (0.0006) | - | - | -0.0002 (0.0002) | -0.0003 (0.0004) | -0.0002 (0.0002) | 0 (0.2465) | -0.0003 (0.0016) | 0 (4.696) | -0.0003 (0.0002) | -0.0004 (0.0004) | - | - |
| $\beta_{34}$ | - | - | -0.0002 (0.0002) | -0.0002 (0.0005) | - | - | -0.0001 (0.0002) | -0.0004 (0.0005) | -0.0001 (0.0002) | 0 (0.1473) | -0.0004 (0.0007) | 0 (3.0165) | -0.0001 (0.0002) | -0.0005 (0.0005) | - | - |
| Twitter | | | | | | | | | Mean | Variance | Mean | Variance | | | | |
| $\beta_{41}$ | - | - | - | 0 (0) | - | - | - | 0 (0) | - | - | 0 (0.0001) | 0 (0.034) | - | 0 (0) | - | - |
| $\beta_{42}$ | - | - | - | 0 (0) | - | - | - | 0 (0) | - | - | 0 (0) | 0 (0.0543) | - | 0 (0) | - | - |
| $\beta_{43}$ | - | - | - | 0 (0) | - | - | - | 0 (0) | - | - | 0 (0.0001) | 0 (0.0997) | - | 0 (0) | - | - |
| $\beta_{44}$ | - | - | - | 0 (0) * | - | - | - | 0 (0) | - | - | 0 (0) | 0 (0.0764) | - | 0 (0) | - | - |
| Diagnostic and Stats | | | | | | | | | | | | | | | | |
| Log-Likelihood | 14990.663 | 4598.178 | 15004.812 | 4614.502 | 15223.270 | 4683.922 | 15238.876 | 4696.468 | 15238.876 | | 4696.468 | | 15234.352 | 4692.165 | 15225.058 | 4687.235 |
| kurtosis (Res. Stand) | 4.875 | 4.985 | 4.640 | 4.568 | 2.401 | 1.873 | 2.253 | 1.720 | 2.253 | | 1.720 | | 2.236 | 1.673 | 2.407 | 1.791 |
| Skewness (Res. Stand) | 0.384 | -0.653 | 0.350 | -0.547 | 0.151 | -0.112 | 0.144 | -0.092 | 0.144 | | -0.092 | | 0.149 | -0.039 | 0.145 | -0.122 |
| Jarque Bera P.Value (Res. Stand) | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 |

**Table 6 – Tesco model results**

In this study however, it was assumed an AR(1) for all the equations. The constant term, c, has a value close to zero for all models and is only significant for model 6. The results from the model with Twitter for the above parameters were very similar and the only difference that is important to callout is the fact that the constant term is statistically relevant in different models, but the value is still close to zero. In what regards the volatility equations and their respective constant ($w$), ARCH ($a_1$) and GARCH ($b_1$) parameters, it can be observed the following:

- In the model without Twitter, the constant term assumes a value of 0 in the sGARCH models. For the mcGARCH, the value varies from 0.05-0.07 and is always statistical relevant with the exception of model 6. Within mcGARCH's and excluding model 6, the constant term value is ~0.05. This value is lower than the values presented by Engle & Sokalska (2012) but this maybe linked with the granularity of this study (15 min intervals) being higher than in Engle & Sokalska (2012) (1 min intervals) and, as expected, lower granularity typically has higher volatility. The same pattern can be seen in the models with Twitter, however, with a slightly different order of magnitude for the coefficients in the mcGARCH models, ~.14-.21 vs. 0.05-0.07 in the mcGARCH models without Twitter. The differences in the coefficients magnitude might not be just explained by the fact that Twitter tweets were added as covariates but maybe because the modelling sample reduced from 5,280 (Model without Twitter) to 990 (Model with Twitter).
- The ARCH parameter $a_1$ is statistically relevant in all models without Twitter. From a value perspective, the estimated coefficient value ranges between 0.09 and 0.161, however, the sGARCH models present higher values, 0.159 and 0.161, respectively. In the models with Twitter, the same parameter presented higher values across all models 0.11-0.15 and 0.16-0.22 in the mcGARCH and sGARCH,

respectively. One main difference between both types (with and without Twitter) is that in the model with Twitter, $a_1$ is only statistically relevant in the mcGARCH models (with exception of model 5) and in the models without Twitter, the parameter is always relevant.

- The GARCH parameter $b_1$ is statistically significant in all the models without Twitter, an expected results given the "volatility" clustering phenomenon. It can also be observed that the sGARCH models present a smaller value for the parameter (~0.701) when compared with the mcGARCH models (~0.82). In the models with Twitter, this variable was not statistically relevant in models 1 and 6 and their values were much smaller when compared with the models without Twitter. These latter results are less in line with some of the papers mentioned previously, where the coefficient $b_1$, typically presents a higher value, similar to the ones presented in the mcGARCH models without Twitter.

The market returns parameters didn't show any relevant relationship in any of the models without Twitter. These results are expected as only the contemporaneous market return is expected to be relevant to explain the part of the stock returns movement. However, including this variable in the model would exclude the ability to forecast the returns and therefore this contemporaneous variable was not included in the models. In contrast, the models with Twitter exhibit some significant correlation between the returns and past values of the market returns (model 4 and 7). These results are less easy to explain, as you would expect if there is a dependency between previous values of the market and the next returns, the trading agencies would have taken advantage of it.

Finally, the social media impact on the Tesco's stock returns and volatility, i.e. the impact of Google Finance, Yahoo Finance, Financial Times news and Twitter tweets:

- Google Finance: In the models without Twitter, the news at *i-2* (i.e. 30 min) was statistically relevant. Notwithstanding, the positive coefficient is closer to zero in all models (~0.0002). All the other periods seemed not relevant in any model. In the models with Twitter, Google Finance remain statistically important but the lags changed from *t-2* to *t-4* (apart from model 6 which *t-2* was also relevant). The values of the coefficients remained relatively low (~0.0004).
- Yahoo Finance: All the values were close to zero and not statistically significant in the models without Twitter. The models with Twitter, the first lag was relevant in the model 2, but the values were ~0.
- Financial Times: In the models without Twitter, the sGARCH models didn't show any important relationship parameters. In the mcGARCH models 4 and 6, the parameter at *t-1* was statistically relevant but with a negative value closer to zero (-0.0004). In the models with Twitter, none of the models show any significant relationship with the returns.
- Twitter: The value was zero for all the parameters and only relevant in the sGARCH model 2 at *t-4*.

### 4.8.4 Diagnostics Results

Table 7 – Tesco Weighted-Ljung-Box P-values at several lags, describe the results of the Weighted-Ljung-Box test for the Standardized Residuals ($\bar{u}_{t,i}$ presented in Equation 8) and Squared Standardized Residuals at different lags.

| Variable | Standardized Residuals | | | | | Squared Standardized Residuals | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Models \ Lags | Lag 1 | Lag 4 | Lag 8 | Lag 15 | Lag 33 | Lag 1 | Lag 4 | Lag 8 | Lag 15 | Lag 33 |
| Model 1 | 0.180 | 0.476 | 0.767 | 0.914 | 0.909 | 0.727 | 0.814 | 0.438 | 0.312 | 0.392 |
| Model 2 | 0.164 | 0.455 | 0.755 | 0.921 | 0.916 | 0.664 | 0.811 | 0.464 | 0.347 | 0.455 |
| Model 3 | 0.440 | 0.648 | 0.804 | 0.902 | 0.925 | 0.395 | 0.791 | 0.803 | 0.921 | 0.922 |
| Model 4 | 0.415 | 0.603 | 0.786 | 0.895 | 0.919 | 0.427 | 0.776 | 0.743 | 0.852 | 0.867 |
| Model 5 | 0.415 | 0.603 | 0.786 | 0.895 | 0.919 | 0.427 | 0.776 | 0.743 | 0.852 | 0.867 |
| Model 6 | 0.599 | 0.789 | 0.898 | 0.951 | 0.948 | 0.077 | 0.268 | 0.414 | 0.669 | 0.812 |
| Model 7 | 0.433 | 0.556 | 0.733 | 0.866 | 0.896 | 0.373 | 0.775 | 0.796 | 0.922 | 0.927 |

**Table 7 – Tesco Weighted-Ljung-Box P-values at several lags – Model without Twitter**

| Variable | Standardized Residuals | | | | | Squared Standardized Residuals | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Models \ Lags | Lag 1 | Lag 4 | Lag 8 | Lag 15 | Lag 33 | Lag 1 | Lag 4 | Lag 8 | Lag 15 | Lag 33 |
| Model 1 | 0.958 | 0.899 | 0.879 | 0.888 | 0.932 | 0.682 | 0.944 | 0.978 | 0.994 | 0.955 |
| Model 2 | 0.935 | 0.814 | 0.746 | 0.779 | 0.776 | 0.800 | 0.935 | 0.977 | 0.998 | 0.919 |
| Model 3 | 0.589 | 0.631 | 0.727 | 0.820 | 0.938 | 0.761 | 0.974 | 0.981 | 0.994 | 0.851 |
| Model 4 | 0.489 | 0.374 | 0.482 | 0.619 | 0.715 | 0.695 | 0.969 | 0.979 | 0.996 | 0.941 |
| Model 5 | 0.489 | 0.374 | 0.482 | 0.619 | 0.715 | 0.695 | 0.969 | 0.979 | 0.996 | 0.941 |
| Model 6 | 0.252 | 0.466 | 0.581 | 0.662 | 0.751 | 0.345 | 0.677 | 0.784 | 0.937 | 0.869 |
| Model 7 | 0.620 | 0.349 | 0.490 | 0.671 | 0.861 | 0.772 | 0.987 | 0.994 | 0.999 | 0.866 |

**Table 8 – Tesco Weighted-Ljung-Box P-values at several lags – Model with Twitter**

As can be seen from the above tables, the null hypothesis that the $\bar{u}_{t,i}$ and $\bar{u}^2{}_{t,i}$ are not correlated with previous values are never rejected in any of the of the lags analyzed (40 lags were analyzed but just a summary is presented). As mentioned before in the document, the Weighted-Ljung-Box test suggested by Pascual, Romo & Ruiz (2006) present more robust results when compared with the standard Ljung-Box test. Notwithstanding, the standard Ljung-Box test was performed and the results were no different from the above. It is important to highlight that even taking into account that several authors concluded that sGARCH are not a robust method to model intraday volatility, the diagnostics performed don't show that. However, the magnitude of the coefficients and as will be seen later the sign are not always aligned with the results from the mcGARCH.

### 4.8.5 Forecasting results

As described previously, in order to compare the forecasting results four metrics will be analyzed, MAE and MDA for the returns and Mean-L1 and Mean-L2 for the volatility. As can be seen from Table 9 – Tesco Forecasting results, in the models without Twitter, model 7 presents the best results in the volatility metrics analyzed and in the MAE and Model 3 is the best in the MDA. Both models don't include any news or tweets but model 7 include past market information. Model 6, which includes news, is the second best in the MDA and Mean-L2 metrics. Model 1, a sGARCH model actually is the second best in the MAE metric. In the models with Twitter, model 3 is the best in volatility (Mean-L1 and Mean-L2) and returns metric MDA and Model 7 remains the best in MAE. Model 6, still comes 3rd from Mean-L2 and MDA. Overall, the accuracy of the models without Twitter are slightly higher when compared to the models with Twitter. This is not say that Twitter should not be included in the predictive models but instead, the sample size might not be enough to capture all the relevant patterns and correlations in the data.

| | without Twitter | | | | | | | | Twitter | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable | Returns | | | | Volatility | | | | Returns | | | | Volatility | | | |
| Models \ Metric | MAE | Rank | MDA | Rank | Mean-L1 | Rank | Mean-L2 | Rank | MAE | Rank | MDA | Rank | Mean-L1 | Rank | Mean-L2 | Rank |
| Model 1 | 0.002126 | 2 | 0.529293 | 3 | - | | - | | 0.001883 | 2 | 0.511785 | 2 | - | | - | |
| Model 2 | 0.002160 | 7 | 0.503030 | 6 | - | | - | | 0.001928 | 7 | 0.459459 | 7 | - | | - | |
| Model 3 | 0.002128 | 3 | 0.547475 | 1 | 1.029191 | 2 | 0.091791 | 3 | 0.001885 | 3 | 0.525253 | 1 | 1.011324 | 1 | 0.025599 | 1 |
| Model 4 | 0.002154 | 5 | 0.523232 | 4 | 1.029917 | 3 | 0.095140 | 4 | 0.001913 | 5 | 0.500000 | 4 | 1.013707 | 4 | 0.032475 | 5 |
| Model 5 | 0.002154 | 6 | 0.523232 | 4 | 1.029918 | 4 | 0.095141 | 5 | 0.001913 | 6 | 0.500000 | 4 | 1.013691 | 3 | 0.032443 | 4 |
| Model 6 | 0.002150 | 4 | 0.545455 | 2 | 1.030160 | 5 | 0.091447 | 2 | 0.001910 | 4 | 0.510135 | 3 | 1.013962 | 5 | 0.032222 | 3 |
| Model 7 | 0.001338 | 1 | 0.458333 | 7 | 1.011273 | 1 | 0.019000 | 1 | 0.001881 | 1 | 0.476351 | 6 | 1.011444 | 2 | 0.026107 | 2 |

**Table 9 – Tesco Forecasting results**

### 4.8.6 Conclusions

Seven models with and without Twitter were estimated to understand the impact of the news and tweets in the stock returns. The results shows high dependency in previous values of volatility $(b_1, a_1)$. In the models without Twitter (which have a bigger sample size), the market returns were not relevant in any model, while Google Finance news in *t-2* and Financial Times news in *t-1* were statistically relevant, however with values very small and closer to zero. The results from mcGARCH models are more aligned with the literature, i.e.

$b_1$ closer to 0.9 than the ones from the sGARCH. It is difficult to comment if the sGARCH results are not robust (as highlighted by many authors), however, at least the parameters w and $a_1, b_1$seems to agree with the fact that the results are inconsistent (i.e. it is expected that all parameters are statistically relevant, and they are not). In addition, the parameters $a_1, b_1$in the sGARCH models are quite different in their magnitude when compared with the mcGARCH models. In the models with Twitter (which have a smaller sample), the results seemed less consistent with the theory, i.e. $a_1, b_1$ values were smaller than expected and, in contrast, the parameter *w* higher than expected. Google Finance was again statistically relevant but with a different lag (*t-4*) but Financial Times didn't show any significant relationship. The models without Twitter presented higher dependency on the past volatility parameter $b_1$ when compared with the values from the models with Twitter. In the latter, the constant term of the variance equation shows a bigger value when compared with models without Twitter.

According to the 4 metrics analyzed, the MAE and MDA for returns and Mean-L1 and Mean-L2 for the volatility, in the models without Twitter, the model 7 (which only include past information of the market) had the best accuracy in 3 of the 4 metrics analysed MAE, Mean-L1, Mean-L2. Model 3 was the second best model, scoring first in MDA. In the models with Twitter, model 3 was the best model in 3 of the 4 metrics, MDA, Mean-L1 and Mean-L2 and model 7 the best in MAE. Overall, the models without Twitter performed better, probably not because of Twitter it itself but because the sample size was not bigger enough to capture the trends and patterns between the covariates and the returns and volatility.

## *4.9 Use Case 2 – SKY*

As can be seen in Figure 7 – SKY Stock Price vs. FTSE100, SKY stock prices had an upward trend between October 2014 until end of July 2015 and then a negative trend in the recent months. There was a big spike by the end of July, specially in the morning of the 29th of July given the positive results presented by the company in that morning. It is also possible to see a positive correlation between FTSE100 and SKY's stock price, specially in the later months of the year analyzed. However, similar to what was mentioned about Tesco there is a part of stock movement that has to be explained by SKY's performance, expectation of growth, innovation, among many other financial and economical reasons.


**Figure 7 – SKY Stock Price vs. FTSE100**

In the same line as before, it will be estimated 7 models that incorporate the short-term news and past market returns to understand if these variables improve or not the forecasting accuracy. Again, the market expectation of news varies, i.e. if the market has

expectation that certain news will be release it will automatically adjust to them. However, if the market is not expecting specific news then the stock prices movement can present high jumps such as the one seen in the 29th of July. It is interesting to see that in that day, at 9:15, Google Finance had 4 news about SKY and Yahoo Finance 5, but only one from Financial Times. Clearly, the market was not expecting such results for the company and that had definitely an impact in the traders' strategy, but the question remains about the impact of the all the other news that don't add any change to the traders' strategy.

Notwithstanding, if some news proved to be correlated with SKY's returns, it would be expected a gain in the forecasting accuracy of SKYs returns and volatility predictions.

In the next section, several econometric models will be created to measure the impact of news in the stock returns. Then, a forecasting analysis will be performed to understand the forecast accuracy of those models.

### 4.9.1 Time series analysis

From Figure 8 – SKY Returns vs. FTSE100 Returns vs. News and Tweets, the correlation with news is difficult to extract, however looks like periods with high-volatility are followed by an increase in the total news of Google Finance and Yahoo Finance. As previously, it can be seen the issue with the Twitter extractions. For that reason, Twitter tweets will only be included in the models with a smaller sample.

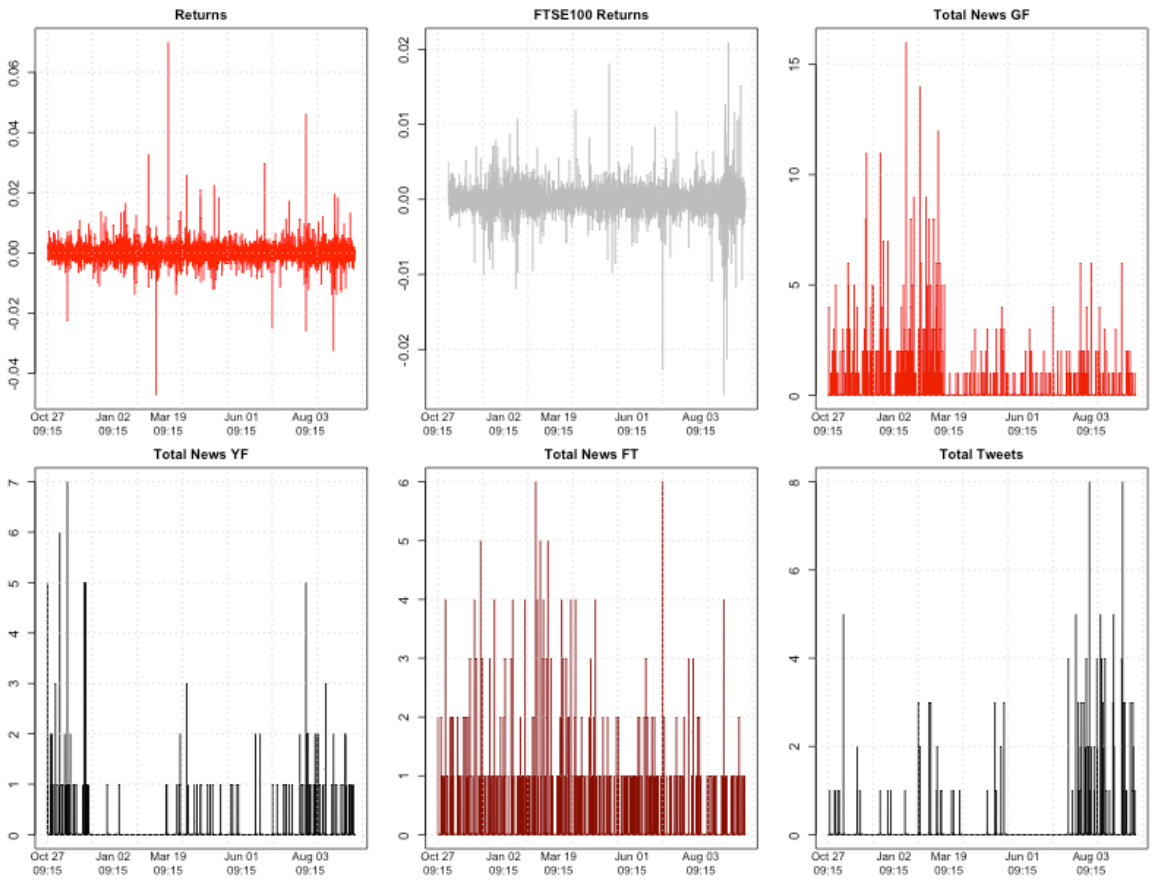



**Figure 8 – SKY Returns vs. FTSE100 Returns vs. News and Tweets**

The correlation matrix below describes the key cross-correlation between the above variables lagged up-to one hour (t=-1,-2,-3,-4):

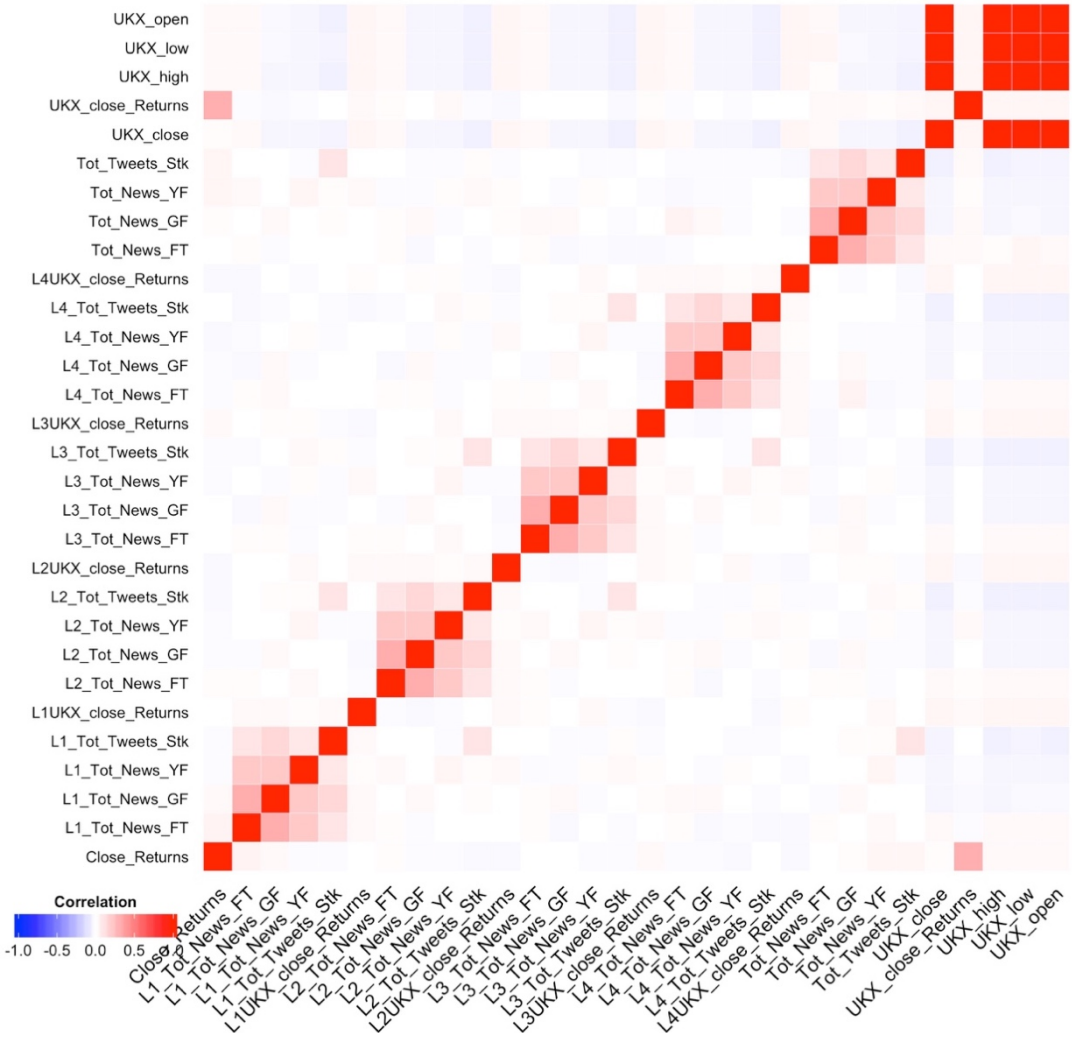



**Figure 9 – Cross-Correlation between SKY Returns vs. FTSE100 Returns vs. News and Tweets up-to 4 Lags**

From the figure above, some insights can be taken, 1) Very low correlations between news and tweets and returns, 2) Positive, but still not strong correlation between the contemporaneous Market return (UKX_close_Returns in the figure) and SKY Returns (Close_Returns), 3) Positive correlation among the contemporaneous news and tweets.

Finally, it will be analyzed the ACF for the $r_t$ and ${r_t}^2$, the kurtosis and skewness to validate some of the properties of the SKY returns.

Starting by the kurtosis, and as expected, the value is higher than 3 (the normal-distributed value), ~11.9. However, the skewness was not negative, ~0.073.

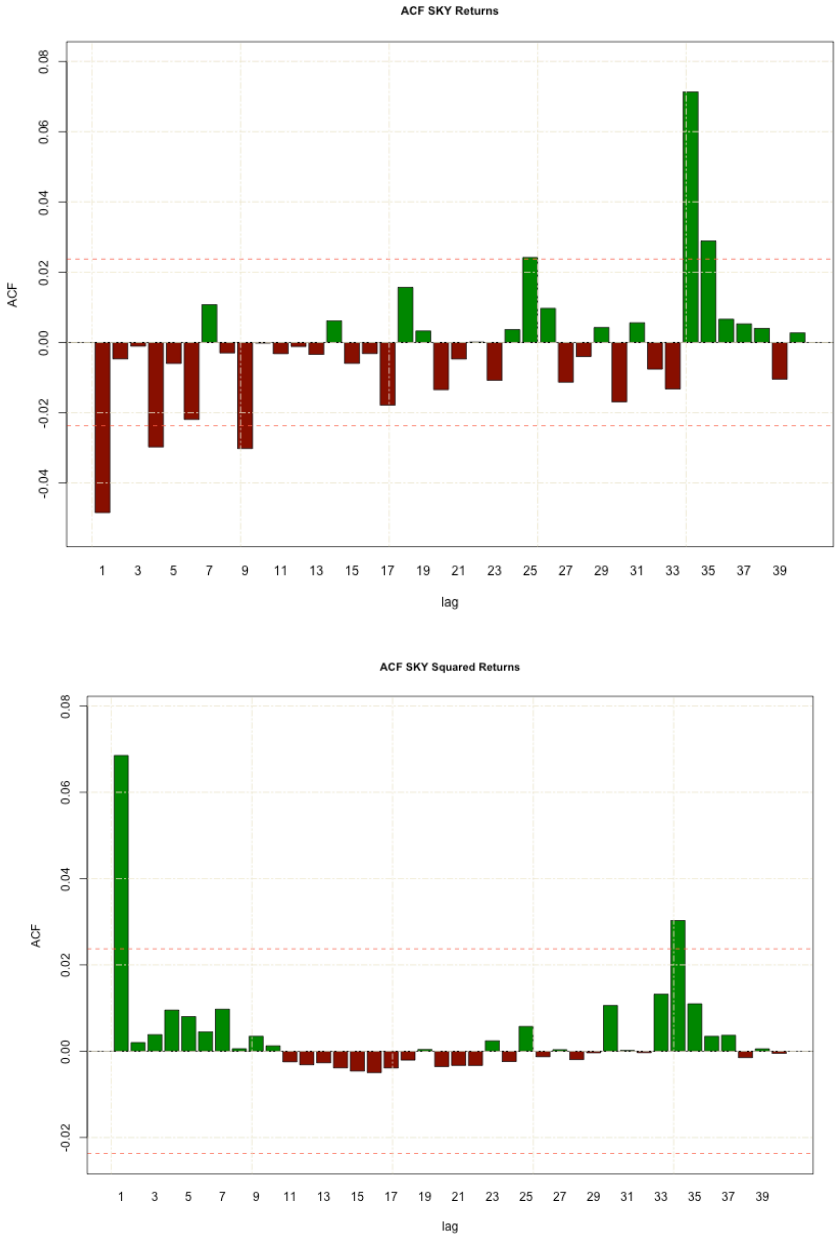



**Figure 10 – ACF for SKY Returns and SKY Squared Returns**

From the above figures it can be seen that both SKY's Returns and Squared Returns seems to be auto-correlated with past values, specially in the first lags (the Ljung-Box test proves exactly that). These results suggest a combination of auto-regressive models with msGARCH models.

| Lag | Sky Returns | | | Sky Squared Returns | | |
|---|---|---|---|---|---|---|
| | ACF | P_Value | Above 5% | ACF | P_Value | Above 5% |
| 1 | -0.07 | 0.00 | TRUE | 0.07 | 0.00 | TRUE |
| 2 | -0.02 | 0.00 | | 0.03 | 0.00 | |
| 3 | -0.01 | 0.00 | | 0.09 | 0.00 | |
| 4 | -0.01 | 0.00 | | 0.06 | 0.00 | |
| 5 | 0.00 | 0.00 | | 0.10 | 0.00 | |
| 6 | -0.01 | 0.00 | | 0.04 | 0.00 | |
| 7 | 0.02 | 0.00 | | 0.02 | 0.00 | |
| 8 | 0.02 | 0.00 | | 0.04 | 0.00 | |
| 9 | -0.06 | 0.00 | | 0.06 | 0.00 | |
| 10 | 0.01 | 0.00 | | 0.00 | 0.00 | |
| 11 | 0.00 | 0.00 | | 0.03 | 0.00 | |
| 12 | -0.01 | 0.00 | | 0.02 | 0.00 | |

**Table 10 - Ljung-Box test for SKY Returns and Squared Returns**

### 4.9.2 Models specification

The same models developed for Tesco (see equations 24 till 30) were also consider for SKY including the sample size (see Table 5). After analyzing the results and respective diagnostics, the models proved to have a higher dependency in the volatility as the null hypothesis of Weighted-Ljung-Box tests were rejected at a 5% level confidence, specially in the first 4 lags. Therefore, the variance equation of SKY models without Twitter are estimated using a GARCH (1,2) and the variance equation of SKY models with Twitter are estimated using a GARCH(1,3) instead of a GARCH(1,1) used for all models in Tesco.

### 4.9.3 Models results

From Table 11 – SKY model results it can be analyzed the coefficients of the models with and without Twitter, robust standard errors (the values inside parenthesis) and the statistical relevance of the parameter at three confidence levels, “***”, “**”, “*” - 99%, 95% and 90%, respectively.

| Coefficient | Model 1 | | Model 2 | | Model 3 | | Model 4 | | Model 5 | | | | Model 6 | | Model 7 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Formula | AR(1)+GARCH(1,2) | AR(1)+GARCH(1,3) | AR(1)+GARCH(1,2) | AR(1)+GARCH(1,3) | AR(1)+GARCH(1,2) | AR(1)+GARCH(1,3) | AR(1)+GARCH(1,2) | AR(1)+GARCH(1,3) | | | | | AR(1)+GARCH(1,2) | AR(1)+GARCH(1,3) | AR(1)+GARCH(1,2) | AR(1)+GARCH(1,3) |
| Standard Coefficients | without Twitter | Twitter | without Twitter | Twitter | without Twitter | Twitter | without Twitter | Twitter | AR(1)+msGARCH(1,1) | | AR(1)+msGARCH(1,1) | | without Twitter | Twitter | without Twitter | Twitter |
| c | 0 (0.0001) | -0.0001 (0.0001) | 0 (0.0001) | 0 (0.0001) | 0 (0) | 0 (0.0004) | 0 (0) | 0 (0.0001) | 0 (0) | | 0 (0.0006) | | 0 (0) | 0 (0.0018) | 0 (0) | -0.0001 (0.0001) |
| $\phi_1$ | -0.0653 (0.0797) | -0.0395 (0.1281) | -0.0942 (0.1406) | -0.0953 (0.2066) | -0.0992 (0.0182) *** | -0.0447 (0.0926) | -0.1189 (0.0215) *** | -0.0441 (0.0466) | -0.1189 (0.0245) *** | | -0.0441 (0.2893) | | -0.0262 (0.0168) | -0.0575 (0.2636) | -0.1153 (0.0219) *** | -0.0285 (0.0409) |
| w | 0 (0) | 0 (0) *** | 0 (0) | 0 (0) | 0.0188 (0.0124) | 0.0321 (0.0259) | 0.0182 (0.0118) | 0.0279 (0.0303) | 0.0182 (0.0344) | | 0.0279 (0.0303) | | 0.0178 (0.0479) | 0.0263 (0.1214) | 0.0184 (0.0122) | 0 |
| $a_1$ | 0.0333 (0.3047) | 0.1585 (0.1443) | 0.0343 (0.3088) | 0.1107 (0.8) | 0.0631 (0.0139) *** | 0.1381 (0.0852) | 0.0633 (0.0138) *** | 0.1494 (0.057) *** | 0.0633 (0.0293) ** | | 0.1494 (0.6966) | | 0.0755 (0.1197) | 0.1424 (1.2437) | 0.0625 (0.0141) *** | 0.1437 (0.0322) *** |
| $b_1$ | 0.3343 (0.1478) ** | 0.6702 (0.2427) *** | 0.3792 (0.1412) *** | 0.0003 (3.6497) | 0.6457 (0.0118) *** | 0.1903 (1.021) | 0.6666 (0.0114) *** | 0.1822 (0.126) | 0.6666 (0.0292) *** | | 0.1822 (2.2012) | | 0.5799 (0.0709) *** | 0.2175 (6.6871) | 0.6713 (0.0118) *** | 0.1573 (0.1026) |
| $b_2$ | 0.6231 (0.1477) *** | 0 (0.939) | 0.5807 (0.1416) *** | 0.5817 (0.324) * | 0.2736 (0.0119) *** | 0 (3.8541) | 0.253 (0.0115) *** | 0 (1.4006) | 0.253 (0.0256) *** | | 0 (5.6593) | | 0.3385 (0.0537) *** | 0 (34.1528) | 0.249 (0.0119) *** | 0 (0.7478) |
| $b_3$ | | 0 (1.0893) | | 0.2678 (3.0815) | | 0.6443 (4.8199) | | 0.6466 (1.4544) | | | 0.6466 (3.3959) | | | 0.6195 (39.9395) | | 0.6707 (0.8166) |
| Market Returns | | | | | | | | | | | | | | | | |
| $\Upsilon_1$ | - | - | 0.0003 (0.0005) | -0.0319 (0.3791) | - | - | 0.0549 (0.0315) * | -0.0435 (0.1199) | 0.0549 (0.0337) | | -0.0435 (0.129) | | - | - | 0.0493 (0.0325) | -0.0532 (0.0661) |
| $\Upsilon_2$ | - | - | 0 (0.0004) | -0.0329 (0.0923) | - | - | -0.0435 (0.028) | -0.0241 (0.0924) | -0.0435 (0.0294) | | -0.0241 (0.4252) | | - | - | -0.0483 (0.0276) * | -0.0228 (0.0478) |
| $\Upsilon_3$ | - | - | 0.0001 (0.0002) | -0.0759 (0.2236) | - | - | -0.0053 (0.0241) | -0.0662 (0.0541) | -0.0053 (0.0255) | | -0.0662 (0.0541) | | - | - | -0.0101 (0.0242) | -0.0708 (0.0517) |
| $\Upsilon_4$ | - | - | 0.0001 (0.0003) | -0.0674 (0.0474) | - | - | -0.0058 (0.0264) | 0.0002 (0.052) | -0.0058 (0.0285) | | 0.0002 (0.0621) | | - | - | -0.0044 (0.0263) | 0.0202 (0.0649) |
| Google Finance | | | | | | | | | Mean | Variance | Mean | Variance | | | | |
| $\beta_{11}$ | - | - | 0 (0) | -0.0001 (0.0007) | - | - | 0 (0.0001) | -0.0001 (0.0007) | 0 (0.0001) | 0 (0.3022) | -0.0001 (0.0018) | 0 (6.6009) | 0 (0.0001) | -0.0001 (0.0154) | - | - |
| $\beta_{12}$ | - | - | 0.0232 (0.0837) | 0.0002 (0.0005) | - | - | 0.0002 (0.0002) | 0.0002 (0.0005) | 0.0002 (0.0002) | 0 (0.2445) | 0.0002 (0.0005) | 0 (3.71) | 0.0002 (0.0002) | 0.0002 (0.0056) | - | - |
| $\beta_{13}$ | - | - | -0.0384 (0.0879) | -0.0004 (0.0005) | - | - | -0.0002 (0.0001) | -0.0002 (0.0002) | -0.0002 (0.0001) | 0 (0.4656) | -0.0002 (0.001) | 0 (2.8252) | -0.0001 (0.0001) | -0.0002 (0.0002) | - | - |
| $\beta_{14}$ | - | - | 0.0119 (0.028) | 0.0002 (0.0002) | - | - | 0 (0.0002) | -0.0001 (0.0003) | 0 (0.0002) | 0 (0.409) | -0.0001 (0.0022) | 0 (7.0028) | 0 (0.0002) | -0.0001 (0.0034) | - | - |
| Yahoo Finance | | | | | | | | | Mean | Variance | Mean | Variance | | | | |
| $\beta_{21}$ | - | - | -0.0325 (0.0466) | -0.0022 (0.0011) ** | - | - | -0.0003 (0.0005) | -0.0007 (0.001) | -0.0003 (0.0005) | 0 (0.2877) | -0.0007 (0.0093) | 0 (5.6451) | -0.0002 (0.0005) | -0.0007 (0.0049) | - | - |
| $\beta_{22}$ | - | - | 0.0001 (0.0002) | -0.0006 (0.0005) | - | - | -0.0002 (0.0004) | -0.0004 (0.0007) | -0.0002 (0.0004) | 0 (0.4287) | -0.0004 (0.009) | 0 (4.1853) | -0.0002 (0.0004) | -0.0005 (0.0198) | - | - |
| $\beta_{23}$ | - | - | -0.0001 (0.0003) | 0.0001 (0.0007) | - | - | -0.0004 (0.0003) | -0.0003 (0.0009) | -0.0004 (0.0003) | 0 (0.5569) | -0.0003 (0.0024) | 0 (3.7477) | -0.0005 (0.0003) * | -0.0003 (0.0167) | - | - |
| $\beta_{24}$ | - | - | 0.0001 (0.0003) | -0.0007 (0.0005) | - | - | -0.0002 (0.0003) | -0.0003 (0.0006) | -0.0002 (0.0003) | 0 (0.7416) | -0.0003 (0.0008) | 0 (4.023) | -0.0003 (0.0003) | -0.0003 (0.0124) | - | - |
| Financial Times | | | | | | | | | Mean | Variance | Mean | Variance | | | | |
| $\beta_{31}$ | - | - | -0.0009 (0.001) | 0.0006 (0.0017) | - | - | 0 (0.0001) | 0.0009 (0.0007) | 0 (0.0001) | 0 (0.0939) | 0.0009 (0.0019) | 0 (2.3325) | 0 (0.0001) | 0.0009 (0.0126) | - | - |
| $\beta_{32}$ | - | - | -0.0004 (0.0005) | 0.0001 (0.0007) | - | - | 0 (0.0001) | 0 (0.0005) | 0 (0.0002) | 0 (0.0887) | 0 (0.0077) | 0 (1.3705) | 0.0001 (0.0001) | 0.0001 (0.0029) | - | - |
| $\beta_{33}$ | - | - | -0.0003 (0.0004) | -0.0001 (0.0002) | - | - | 0.0001 (0.0001) | 0 (0.0004) | 0.0001 (0.0002) | 0 (0.2755) | 0 (0.0048) | 0 (14.7767) | 0.0001 (0.0001) | 0 (0.0004) | - | - |
| $\beta_{34}$ | - | - | -0.0005 (0.0004) | 0.0003 (0.0009) | - | - | 0.0002 (0.0001) ** | 0.0003 (0.0007) | 0.0002 (0.0001) * | 0 (0.272) | 0.0003 (0.0018) | 0 (6.121) | 0.0002 (0.0001) * | 0.0004 (0.0184) | - | - |
| Twitter | | | | | | | | | Mean | Variance | Mean | Variance | | | | |
| $\beta_{41}$ | - | - | - | 0.0004 (0.0003) | - | - | - | 0.0001 (0.0001) | - | - | 0.0001 (0.0011) | 0 (2.7143) | - | 0.0001 (0.0003) | - | - |
| $\beta_{42}$ | - | - | - | -0.0001 (0.0004) | - | - | - | 0 (0.0001) | - | - | 0 (0.0012) | 0 (4.1062) | - | 0 (0.0012) | - | - |
| $\beta_{43}$ | - | - | - | -0.0002 (0.0005) | - | - | - | -0.0002 (0.0002) | - | - | -0.0002 (0.0014) | 0 (4.2993) | - | -0.0002 (0.0026) | - | - |
| $\beta_{44}$ | - | - | - | 0.0001 (0.0001) | - | - | - | 0.0002 (0.0002) | - | - | 0.0002 (0.0038) | 0 (0.4828) | - | 0.0002 (0.0003) | - | - |
| Diagnostic and Stats | | | | | | | | | | | | | | | | |
| Log-Likelihood | 15504.641 | 4552.432 | 15526.254 | 4590.155 | 15843.308 | 4692.024 | 15853.187 | 4705.837 | 15853.187 | | 4705.837 | | 15830.772 | 4703.871 | 15846.603 | 4694.358 |
| kurtosis (Res. Stand) | 13.622 | 23.023 | 9.786 | 5.743 | 1.650 | 1.447 | 1.594 | 1.386 | 1.594 | | 1.386 | | 1.597 | 1.330 | 1.652 | 1.514 |
| Skewness (Res. Stand) | -0.378 | -1.730 | -0.064 | -0.229 | 0.029 | -0.057 | 0.032 | -0.108 | 0.032 | | -0.108 | | 0.044 | -0.101 | 0.030 | -0.070 |
| Jarque Bera P.Value (Res. Stand) | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 |

**Table 11 – SKY model results**

From the outputs of the model without Twitter it can be seen that the AR(1) parameter, $\phi_1$, have a much bigger value when compared to the values obtained by Tesco's models, varying from -0.02 to -0.12. This parameter is statistically relevant in 3 mcGARCH models (3,4 and 7) but not relevant in the two sGARCH models. Taking into account the ACF charts (Figure 10) it was expected that this parameter was relevant, however is difficult to explain this behavior economically, as you would expect the market would adjust to this automatically. In the models with Twitter, this parameter had smaller values, between -0.04 and -0.09 and was not relevant in any model. The constant parameter *c*, was not relevant in any model (both with and without Twitter) and had a value of 0 in all models. Regarding the volatility equations and their respective constant ($w$), ARCH ($a_1$) and GARCH ($b_1$) parameters, it can be observed the following:

- In the model without Twitter, the constant term assumes a value of 0 in the sGARCH models, similar to Tesco's. For the mcGARCH, the value was ~0.02. This pattern was also true in the Tesco's models (0 in sGARCH and >0 in mcGARCH), however the latter had bigger values and in some models the parameter was statistically relevant. Twitter models pretty much followed the pattern as the models without Twitter but with bigger coefficient value (~0.03).
- The ARCH parameter $a_1$ is statistically relevant in all mcGARCH models without Twitter (with the exception of model 6 – later it will be seen that the estimation for this model is not valid) but not in the sGARCH's. From a value perspective, the estimated coefficient value ranges between 0.03 and 0.06, however, the sGARCH models present lower values, ~0.03. In the models with Twitter, the same parameter presented higher values across all models ~0.14 and 0.11-015 in the mcGARCH and sGARCH, respectively.

- In the models without Twitter, the GARCH parameter $b_1$ is statistically significant in models. As similar to Tesco's, the sGARCH models present a smaller value for the parameter (0.334-0.379) when compared with the mcGARCH models (0.60-0.671). The models with Twitter cannot be directly compared given it is estimated with a GARCH(1,3) vs. a GARCH(1,2) in the models without Twitter. Notwithstanding, $b_1$ was only relevant in model 1. The value of this coefficient varied significantly in the sGARCH models (0.670 and 0.003) and were more stable in the mcGARCH's (0.16-0.19).
- The GARCH parameter $b_2$ was relevant in all models without Twitter. It can be observed the opposite behavior as for the parameter $b_1$, i.e. the mcGARCH have lower values than the sGARCH models, 0.250-0.338 vs. 0.580-0.623, respectively. For the Twitter models, this parameter is only relevant in model 2 with a value 0.581 and is zero in any other model.
- Finally, the parameter $b_3$, only available in the Twitter models, was not relevant in any model. The value of this parameter varied, in the mcGARCH models had an average value of ~0.646 and in the sGARCH models of 0 and 0.267, respectively.

The market returns parameters, strangely, were statistically relevant in the models without Twitter 4 and 7. In the model 4, the lag *t-1* was relevant with a value of 0.054 and in the model 7, the lag *t-2* was relevant with a value of -0.048. In the models with Twitter, as expected, none of this parameters were relevant.

Finally, the social media impact analysis:

- Google Finance: Was not relevant in any models with and without Twitter and, with the exception of model 2 without Twitter (where the coefficients were in average 0.02), the value of all the parameters were close to zero ~0.002.
- Yahoo Finance: The lag *t-3* in the model 7 without Twitter was relevant but with a value closer to zero (-0.0005). All the other models the parameters were close to zero and not relevant. In the models with Twitter, the lag *t-1* in model 1 was relevant with a value of -0.002. All the other models were not relevant and had a value close to zero.
- Financial Times: In the models without Twitter, the sGARCH model didn't show any important relationship parameters. In the mcGARCH's, the parameter at *t-4* was statistically relevant but with a value closer to zero (0.0002). In the models with Twitter, none of the models show any significant relationship with the returns.
- Twitter: The lag *t*-4 in the mcGARCH models was statistically relevant with a value closer to zero (0.0002). The remaining parameters was not relevant.

### 4.9.4 Diagnostic Results

Table 12 – SKY Weighted-Ljung-Box P-values at several lags and Table 13 – SKY Weighted-Ljung-Box P-values at several lags – Models with Twitter, describe the results of the Weighted-Ljung-Box test for the Standardized Residuals ($\bar{u}_{t,i}$ presented in Equation 8) and Squared Standardized Residuals at different lags for the models with and without Twitter.

| Variable | Standardized Residuals | | | | | Squared Standardized Residuals | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Models \ Lags | Lag 1 | Lag 4 | Lag 8 | Lag 15 | Lag 33 | Lag 1 | Lag 4 | Lag 8 | Lag 15 | Lag 33 |
| Model 1 | 0.542 | 0.374 | 0.514 | 0.243 | 0.222 | 0.492 | 0.770 | 0.880 | 0.949 | 0.923 |
| Model 2 | 0.199 | 0.130 | 0.209 | 0.088 | 0.100 | 0.285 | 0.439 | 0.590 | 0.687 | 0.519 |
| Model 3 | 0.483 | 0.277 | 0.324 | 0.239 | 0.255 | 0.071 | 0.095 | 0.203 | 0.319 | 0.361 |
| Model 4 | 0.494 | 0.292 | 0.346 | 0.238 | 0.249 | 0.073 | 0.091 | 0.216 | 0.355 | 0.360 |
| Model 5 | 0.494 | 0.292 | 0.346 | 0.238 | 0.249 | 0.073 | 0.091 | 0.216 | 0.355 | 0.360 |
| Model 6 | 0.001 | 0.001 | 0.002 | 0.002 | 0.004 | 0.122 | 0.146 | 0.315 | 0.465 | 0.469 |
| Model 7 | 0.467 | 0.371 | 0.429 | 0.306 | 0.293 | 0.073 | 0.101 | 0.227 | 0.341 | 0.358 |

**Table 12 – SKY Weighted-Ljung-Box P-values at several lags – Models without Twitter**

| Variable | Standardized Residuals | | | | | Squared Standardized Residuals | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Models \ Lags | Lag 1 | Lag 4 | Lag 8 | Lag 15 | Lag 33 | Lag 1 | Lag 4 | Lag 8 | Lag 15 | Lag 33 |
| Model 1 | 0.805 | 0.785 | 0.802 | 0.503 | 0.622 | 0.899 | 0.994 | 0.999 | 1.000 | 1.000 |
| Model 2 | 0.485 | 0.674 | 0.718 | 0.497 | 0.690 | 0.488 | 0.878 | 0.877 | 0.950 | 0.750 |
| Model 3 | 0.949 | 0.989 | 0.975 | 0.924 | 0.886 | 0.104 | 0.290 | 0.555 | 0.702 | 0.959 |
| Model 4 | 0.827 | 0.995 | 0.968 | 0.841 | 0.767 | 0.056 | 0.173 | 0.383 | 0.548 | 0.880 |
| Model 5 | 0.827 | 0.995 | 0.968 | 0.841 | 0.767 | 0.056 | 0.173 | 0.383 | 0.548 | 0.880 |
| Model 6 | 0.861 | 0.941 | 0.934 | 0.809 | 0.785 | 0.080 | 0.240 | 0.478 | 0.603 | 0.911 |
| Model 7 | 0.914 | 0.990 | 0.978 | 0.923 | 0.848 | 0.075 | 0.233 | 0.478 | 0.664 | 0.944 |

**Table 13 – SKY Weighted-Ljung-Box P-values at several lags – Models with Twitter**

As can be seen, only the model 6 without Twitter didn't pass the test at different lags. Therefore, any interpretation from the models coefficients are not valid.

### 4.9.5 Forecasting results

As can be seen from Table 14, model 6 without Twitter presents the best results in the two volatility metrics. In the returns metrics, model 1 and model 3 presents the best results MAE and MDA, respectively. In the models with Twitter, model 3 presents the best results for the 4 metrics. The models with and without Twitter have similar accuracy in the returns metrics but the model with Twitter presents better results in the volatility metrics.

| | without Twitter | | | | | | | | Twitter | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable | Returns | | | | Volatility | | | | Returns | | | | Volatility | | | |
| Models \ Metric | MAE | Rank | MDA | Rank | Mean-L1 | Rank | Mean-L2 | Rank | MAE | Rank | MDA | Rank | Mean-L1 | Rank | Mean-L2 | Rank |
| Model 1 | 0.001848 | 1 | 0.460606 | 2 | - | | - | | 0.001632 | 2 | 0.457912 | 5 | - | | - | |
| Model 2 | 0.001871 | 5 | 0.432323 | 7 | - | | - | | 0.001694 | 7 | 0.466216 | 2 | - | | - | |
| Model 3 | 0.001853 | 2 | 0.468687 | 1 | 1.034037 | 3 | 0.087370 | 3 | 0.001629 | 1 | 0.468013 | 1 | 1.013232 | 1 | 0.020256 | 1 |
| Model 4 | 0.001876 | 6 | 0.452525 | 3 | 1.034185 | 5 | 0.088360 | 5 | 0.001685 | 6 | 0.456081 | 6 | 1.015829 | 3 | 0.024490 | 3 |
| Model 5 | 0.002163 | 7 | 0.440413 | 5 | 1.032352 | 2 | 0.084713 | 2 | 0.001682 | 5 | 0.466216 | 2 | 1.019979 | 5 | 0.029478 | 5 |
| Model 6 | 0.001860 | 3 | 0.440404 | 6 | 1.031643 | 1 | 0.084459 | 1 | 0.001681 | 4 | 0.462838 | 4 | 1.016192 | 4 | 0.024826 | 4 |
| Model 7 | 0.001866 | 4 | 0.446465 | 4 | 1.034150 | 4 | 0.087558 | 4 | 0.001635 | 3 | 0.456081 | 6 | 1.014132 | 2 | 0.021469 | 2 |

**Table 14 – SKY Forecasting Results**

### 4.9.6 Conclusions

As expected, the results from the models manifested high dependency in previous values of volatility, with $b_1 + b_2 \approx 0.9$. Different from Tesco's results, and only in the models without Twitter, past values of market returns were statistically relevant in two models (lags *t-1* and *t-2*). Google Finance in contrast to what was observed in Tesco's results were not relevant. However, Financial Times news remained statistically relevant in the models without Twitter but in a different lag *t-4*. In line with Tesco's results, the social media parameters were all close to zero. Similar to Tesco's, the models without Twitter presented higher values of $b_1 + b_2$ when compared with the models with Twitter.

A major difference from Tesco's results is the order of the GARCH models of the variance equation. For Tesco, a "simple" mcGARCH(1,1) was used to estimate the variance equation and that was enough, however, after analyzing the Weighed-Ljung-Box results, it was required to increase the GARCH order to (1,2) and (1,3) in the models without and with Twitter, respectively. By applying this strategy, only the results from model 6 without Twitter cannot be used, given the diagnostics performed.

From a forecasting perspective, model 6 without Twitter presented the best results for the volatility metrics. For the models with Twitter model 3 presented the best results in all 4 metrics. Comparing these results with Tesco's, it can be concluded that Tesco's models

performed better overall and their parameters had results more aligned with the theory, specially in the models that did not include Twitter as covariates.

# 5 Conclusions

## 5.1 Study key results

The main goal of this study was to understand and measure the impact of short-term news and tweets from several sources on Tesco and SKY stock returns. Overall it can be concluded that news, past market information and tweets didn't improve that much the stocks returns models and forecasting accuracy but, for models without Twitter, Tesco's 2nd best model and SKY's best model included social media news as covariates. Google Finance and Financial Times presented better results when compared to Yahoo Finance and/or Twitter, however, overall, the impact was very residual and closer to zero.

The methodology suggested by Engle & Sokalska (2012) proved to present more robust results when compared with the sGARCH models, a conclusion already made in several papers described throughout the document. In addition, it seems that models with smaller intraday sample sizes (e.g. the models with Twitter) have less dependency in past values of volatility, when compared with models with a bigger intraday sample size (models with Twitter).

Tesco models and forecasting key conclusions:

- The results from mcGARCH are well aligned with the literature and show a high dependency in past values of the volatility.
- From the results of the sGARCH models it can be concluded that this approach presents some values that are more questionable (e.g. $a_1$ not being statistically relevant and $b_1$ presenting lower value than mcGARCH) however the forecasting results didn't differ much from the mcGARCH, but where never better

- It is possible to see a correlation between news from Google Finance and Financial Times with Tesco's stock returns, but with a very low impact
- Model 7 (which includes past market information) was the best model from a forecasting perspective, followed by the model 6 (which include news) and model 3 (which don't include news or market information).
- Models with Twitter where less accurate than the models without, from a forecasting perspective.

SKY models and forecasting key conclusions:

- Modelling results were less aligned with the literature and a higher order GARCH was required to model the variance equations in models without and with Twitter, mcGARCH(1,2) and mcGARCH(1,3), respectively.
- Notwithstanding, the results from the high order GARCH's still show high dependency in the previous values of volatility ($b_1, b_2, b_3$)
- Financial Times was also relevant but with a different lag from Tesco's results
- Strangely, some models had a statistically relevant relationship between past value of the market return, past returns with the returns. Economically speaking this results are difficult to explain as you would expect the market to automatically adjust to this kind of behavior
- In the same line as Tesco's the models without Twitter had more dependency in the past parameter of volatility than the models with Twitter
- The model 6 (which include news) was the most accurate model in the volatility metrics. The models with Twitter were more accurate than the models without Twitter.

## 5.2 Key limitations of the study

- The social media news and tweets gathered for this study only included the Stock in cause and excluded any information for their competitors. Embedding this additional information would potentially impact the model estimations and forecasting accuracy but this would have required identification of the key competitors and additional work to collect, transform and store the information.
- In addition to the above, social media news and tweets about the market were not considered and clearly this might improve the accuracy of the models and forecasts
- Currently it was only considered as sources Google, Yahoo, Financial Times and Twitter but other sources maybe relevant as well, e.g. Bloomberg or Reuters
- Selecting which news to consider for the analysis from Twitter can be a difficult and never ending exercise given the amount of news related with a company that comes out in Twitter. In this document, only tweets with tags related to Financial Markets were included.

## 5.3 Potential next steps

- As explained above, gather more news and tweets related with the competitors of a given stock of interest. For the example of Tesco, news and tweets about Sainsbury and ASDA for example.
- Re-perform the analysis with datasets with lower and higher intraday granularity to understand the differences in contribution from the news and tweets
- The machine learning classification method used in the study worked well with news and tweets with smaller length but not with articles from Financial Times.

By improving this algorithm in the future, a better of way of filtering the news and the noise around then may help to improve the models and forecasting accuracy.

- Expand the techniques used in the econometrics models. For example, in parallel to mcGARCH model, try a recent approach suggested by Damasio & Nicolau (2014) using multivariate Markov-Chains.

# 6 Bibliography


Andersen, T., & Bollerslev, T. (1998). DM-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies. *Journal of Finance, 53*, 219-265.

Andersen, T., Bollerslev, T., & Cai, J. (2000). Intraday and interday volatility in Japanese stock market. *Journal of International Financial Markets, Institutions and Money, 10*, 107-130.

Anderson, T., & Bollerslev, T. (1997). Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance, 4*, 115-158.

Asur, S., & Huberman, B. (2010). Predicting the future with social media. *International Conference on Web Intelligence and Intelligent Agent Technology*, (pp. 492-499 ).

Bautin, M., Vijayarenu, L., & Skiena, S. (2008). *International Sentiment Analysis for News and Blogs.* Association for the Advancement of Artificial Intiligence (www.aaai.org).

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics, 31*, 307-327.

Bollerslev, T., & Wooldridge, J. (1992). Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time Varying Covariances. *Econometric Reviews, 11*, 143-172.

Bollerslev, T., Cai, J., & Song, F. (2000). Intraday periodicity, long memory volatility, and macroeconomic announcement effects in the US Treasury bond market. *Journal of Empirical Finance, 7*, 37-55.

Bollerslev, T., Chou, R., & Kroner, K. (1992). ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence. *Jornal of Econometrics*(52), 5-59.

Bollerslev, T., Engle, R., & Nelson, D. (1994). *Handbook of Econometrics, Volume 4 - Chapter 49.* (E. S. B.V., Ed.)

Chan, W. (2003, 11). Stock price reaction to news and no-news: drift and reversal after headlines . *Journal of Financial Economics, 70*(2), 223-260.

Damasio, B., & Nicolau, J. (2014). Combining a regression model with a multivariate Markov chain in a forecasting problem. *Statistics and Probability Letters*.

Engle, R. (1982). Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation. *Econometrica, 50*, 987-1008.

Engle, R., & Gallo, G. (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics, 131*, 3-27.

Engle, R., & Sokalska, M. (2012). Forecasting intraday volatility in the US equity market. Multiplicative component GARCH. *Journal of Financial Econometrics, 2012, 10*(1), 54-83.

Engle, R., & Sokalska, M. E. (2012). Forecasting intraday volatility in the US equity market. Multiplicative component GARCH. *Journal of Financial Econometrics, 10*, 54-83.

Fisher, T., & Gallagher, C. (2012, 5). New Weighted Portmanteau Statistics for Time Series Goodness of Fit Testing. *Journal of the American Statistical Association, 107*(498), 777-787.

Fuertes, A. M. (2009). On forecasting daily sock volatility: The role of intraday information and market conditions. *International Journal of Forecasting, 25*, 259-281.

Garimella, K., Weber, I., & Cin, S. (2014, 11). *From "I love you babe" to "leave me alone" - Romantic Relationship Breakups on Twitter.* Retrieved from Cornell University: http://arxiv.org/abs/1409.5980

Ghalanos, A. (2015). *Introduction to rugarch package.*

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter Sentiment Classification using Distant Supervision*. Retrieved from Stanford University Papers: http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

Hansen, P., & Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*(131), 97-121.

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter Sentiment Classification. *49th Annual Meeting of Association for Computational Linguistiscs* (pp. 151-160). Portland: Association for Computational Linguistiscs.

Nicolau, J. (2012). *Modelação de Séries Temporais Financeiras.* Almedina.

Nicolau, J. (2013). *Slides Séries Temporais Financeiras, Mestrado de Econometria Aplicada e Previsão 2011-2013.*

Oh, C., & Sheng, O. (2011). Investigating Predictive Power of Stock Micro Blog Sentiment Forecasting Future Stock Price Directional Movement. *Thirty Second International Conference on Information Systems.*

Pascual, L., Romo, J., & Ruiz, E. (2006, 5). Bootstrap prediction for returns and volatilities in GARCH models. *Computational Statistics and Data Analysis, 50*(9), 2293–2312.

Patton, A. (2010). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics, 10*.

Pierre, G. (2015, 8). Market Risk Models for Intraday Data. *The European Journal of Finance, 11*(4), 309-324.

Rossi, E., & Fantazzini, D. (2012, 11). Long memory and Periodicity in Intraday Volatility. *DEM Working Paper Series - Universitia di Pavie*.

Singh, A., Allen, D., & Powell, R. (2013). Intraday Volatility Forecast in Australian Equity Market. *20th International Congress on Modelling and Simulation.*

Skiena, S., & Zhang, W. (2010). Trading Strategies to exploit blog and news sentiment. *Fourth Int. Conf. on Weblogs and Social Media (ICWSM 2010).* Washington DC.

Taylor, S. (1986). *Modeling Financial Time Series.* John Wiley & Sons.

Velikovich, L., McDonal, R., & Councill, I. (2010). What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. *'10 Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, (pp. 51-59).

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society, 50*(1), 1-25.

# 7 List of Tables



# 8 List of Figures