

# News Aware Volatility Forecasting: Is the Content of News Important?

**Calum S. Robertson**

Information Research Group  
Faculty of Information Technology  
Queensland University of Technology  
2 George Street, Brisbane, QLD,  
Australia 4000

cs.robertson@qut.edu.au

**Shlomo Geva**

Information Research Group  
Faculty of Information Technology  
Queensland University of Technology  
2 George Street, Brisbane, QLD,  
Australia 4000

s.geva@qut.edu.au

**Rodney C. Wolff**

School of Economics and Finance  
Faculty of Business  
Queensland University of Technology  
2 George Street, Brisbane, QLD,  
Australia 4000

r.wolff@qut.edu.au

## Abstract

The efficient market hypothesis states that the market incorporates all available information to provide an accurate valuation of the asset at any given time. However, most models for forecasting the return or volatility of assets completely disregard the arrival of asset specific news (i.e., news which is directly relevant to the asset). In this paper we propose a simple adaptation to the GARCH model to make the model aware of news. We propose that the content of news is important and therefore describe a methodology to classify asset specific news based on the content. We present evidence from the US, UK and Australian markets which show that this model improves high frequency volatility forecasts. This is most evident for news which has been classified based on the content. We conclude that it is not enough to know when news is released, it is necessary to interpret its content.

*Keywords.* Stock Market, News, Document Classification, Volatility, Forecast.

## 1. Introduction

The efficient market hypothesis states that the market incorporates all available information to provide an accurate valuation of the asset at any given time. There is large body of evidence that assets tend to react to public information, most often when the information contains a shock. This evidence includes the reaction to public information in the form of newspaper/magazine/real-time source (e.g. Cutler et al. 1989, Goodhart 1989, Goodhart et al. 1993, Melvin and Yin 2000, Mitchell and Mulherin 1994, Mittermayer 2004), macroeconomic announcements (e.g. Almeida et al. 1998, Ederington and Lee 1993, 1995, 2001, Graham et al. 2003, Kim et al. 2004, Nofsinger and Prucyk 2003), analyst recommendations Hong et al. 2000, Michaely and Womack 1999, (e.g. Womack 1996), and weather reports (e.g. Roll 1984).

Ederington and Lee (1993) found that volatility on Foreign Exchange and Interest Rate Futures markets increases within one minute of a macroeconomic news announcement, and the effect lasts for about 15 minutes. Ederington and Lee (1995) determined that the same markets begin to react within 10 seconds of macroeconomic news announcements, with weak evidence that they tend to overreact to news within the first 40 seconds after news, but settle within 3 minutes. Graham et al. (2003) established that the value of stocks on the S&P 500 index are influenced by scheduled macroeconomic news, however, they did not investigate any intraday effect. Nofsinger and Prucyk (2003) concluded that unexpected bad macroeconomic news is responsible for most abnormal intraday volume trading on the S&P 100 Index option.

Despite strong evidence that the stock market does react to macroeconomic news, there is far more asset specific news, i.e., news which is directly relevant to the asset, than macroeconomic news. Furthermore, unlike macroeconomic news, most asset specific news is not scheduled and therefore investors have not formed their own expectation, or adopted analysts' recommendations about the content of the news. Mittermayer (2004) investigated the effect of Press Announcements on the New York Stock Exchange and the NASDAQ and determined that the content of news can be used to predict, with reasonable accuracy, if the market will exhibit high return within 60 minutes of the announcement. Unfortunately press announcements are only a fraction of asset specific news, so further investigation is required to determine how the stock market reacts, if at all, to this type of news.

The Generalised Autoregressive Conditional Heteroskedasticity (GARCH) model introduced by Bollerslev (1986) has been shown to be a reliable model for forecasting the volatility of an asset. However, like virtually all volatility forecasting models, it completely disregards the impact of public information. Kaley et al. (2004) found that the forecast accuracy of GARCH(1,1) for 30 minute returns can be improved by factoring in the number of asset specific documents released to the market in the previous 30 minutes. Furthermore they found that the forecast accuracy could be further improved by restricting the news based on how the Australian Stock Exchange (ASX) categorised the news (e.g. Progress Report, Dividend Announcement, Mergers and Acquisitions). Whilst the ASX may classify news, it is not safe to assume that every asset specific news

Copyright © 2007, Australian Computer Society, Inc. This paper appeared at the Sixth Australasian Data Mining Conference (AusDM 2007), Gold Coast, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 70. Peter Christen, Paul Kennedy, Jiuyong Li, Inna Kolyskhina and Graham Williams, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

document for assets throughout the world will be classified in the same fashion at the time of their release. Therefore it is advisable to use an automated form of news classification, which can be applied to news from any source.

In this paper we propose a modification to the GARCH model proposed by Bollerslev (1986), to handle the arrival of asset specific news. Furthermore we describe an automated method to classify the news, which can be used to limit the number of documents which this model processes. Finally we demonstrate how this model can improve the volatility forecast accuracy using a large asset base and high frequency data.

## 2. Data

All data for this research were obtained using the Bloomberg Professional<sup>®</sup> service. The dataset consists of stocks which comprised the S&P 100, FTSE 100, and ASX 100 indices as at July 2005 and continued to trade through to November 2006, which is a total of 283 stocks. For each stock the Trading Data, and News were collected for the period beginning May 2005 through to and including the October 2006. There are over 500,000 documents (news articles) in this dataset, which we believe to be the largest used for the types of experiments we conduct.

### 2.1. Trading Data

The set defined in Eq. (1) consists of each distinct minute ( $z$ ) where trading occurred for the stock ( $s$ ), within all minutes for the period of data collection ( $T_A$ ). For each minute ( $d_{(s,z)}$ ) the average price ( $p_{(s,z)}$ ) for trades during that minute are stored.

$$I_{(s)} = \{I_1, I_2, \dots, I_m\} | I_{(s,z)} = (d_{(s,z)}, p_{(s,z)}) \wedge z \in T_A \quad (1)$$

However, only business time scale (minutes which occurred during business hours for the market on which the stock trades) is of interest. Furthermore it is necessary to have a homogenous time series (i.e., an entry for every business trading minute for the stock, regardless of whether any trading occurred). Therefore the date ( $D_{(s)}$ ) and price ( $P_{(s)}$ ) time series are produced for all minutes in the business time scale ( $T_B$ ) with the definitions in Eqs. (2) and (3). The price at time  $t$  is defined as the price of the last actual trade for the stock prior to or at the given time. Note that if the stock was suspended from trading for a whole day then the day is excluded from  $T_B$ .

$$D_{(s)} = \{D_1, \dots, D_n\} | D_{(s,t)} > D_{(s,t-1)} \wedge D_{(s,t)} \in T_B \wedge T_B \subseteq T_A \quad (2)$$

$$P_{(s)} = \{P_1, \dots, P_n\} | P_{(s,t)} = (p_{(s,z)} | z = \max(z | d_{(s,z)} \leq D_{(s,t)})) \quad (3)$$

### 2.2. News

The news search facility within the Bloomberg Professional<sup>®</sup> service was used to download all relevant documents for each stock within the dataset. These documents include Press Announcements, Annual Reports, Analyst Recommendations and general news which Bloomberg has sourced from over 200 different news providers.

The set defined in Eq. (4) consists of each distinct news document ( $\lambda$ ) for the stock ( $s$ ) and contains the time ( $d_{(s,\lambda)}$ ) and content ( $C_{(s,\lambda)}$ ) of the document. Note that we allow the market time to react to news by ignoring any document which occurred within the last  $\Delta\tau$  minutes of a business day (i.e.,  $time(d_{(s,\lambda)}) < \max(time(T_B)) - \Delta\tau$ ). Furthermore we ignore the first  $\Delta\tau$  minutes of a business day as we expect investors are more focussed on opening their positions for the day rather than reading the latest news (i.e.,  $\min(time(T_B)) + \Delta\tau \leq time(d_{(s,\lambda)})$ ).

$$A_{(s,\Delta\tau)} = \{A_1, A_2, \dots, A_p\} | A_{(s,\Delta\tau,\lambda)} = (d_{(s,\lambda)}, C_{(s,\lambda)}) \wedge d_{(s,\lambda)} \in T_B \wedge \min(time(T_B)) + \Delta\tau \leq time(d_{(s,\lambda)}) < \max(time(T_B)) - \Delta\tau \quad (4)$$

All documents are pre-processed to remove numbers, URLs, email addresses, meaningless symbols, and formatting. Each term in the content  $C_{(s,\lambda)}$  of the document is stemmed using the Porter stemmer algorithm (Porter 1980). The Porter stemmer removes suffixes from words, using strict rules which apply to the English language, such that words with the same stem are considered to be the same word. For example the stems of “finance”, “finances”, “financed”, and “financing” are the same. Stemming is performed to reduce the number of terms which need to be investigated, and to help to find similar documents. The stemmed term index defined in Eq. (5) is created with the stemmed terms which appear in the document ( $S_{(s,\lambda,\omega)}$ ), and the number of times they appear within the document ( $SC_{(s,\lambda,\omega)}$ ), where  $\omega$  is the stemmed term identifier.

$$C_{(s,\lambda)} = \{T_1, T_2, \dots, T_q\} | T_{(s,\lambda,\omega)} = \{S_{(s,\lambda,\omega)}, SC_{(s,\lambda,\omega)}\} \wedge SC_{(s,\lambda,\omega)} = \#\{\forall S_{(s,\lambda,\omega)} \in C_{(s,\lambda)}\} \quad (5)$$

## 3. Methodology

The methodology section is divided into sections titled News Classification, News Aware GARCH, and Measuring Forecast Performance. In the first section we define a classifier which we use to predict whether a document will cause abnormal market behaviour based on its content. In the second section we describe how the classified documents are incorporated into a model to forecast volatility. In the final section we define how we measure the performance of the new model.

### 3.1. News Classification

In order to classify documents it is first necessary to categorise the documents and determine which documents are of more interest. Building a classifier which predicts whether a document is interesting requires the construction of training and test sets. To ascertain if these documents in the training set have anything in common it is then necessary to analyse the terms contained in the documents, and rank the terms which are most interesting. Subsequently the accuracy of the classifiers must be tested by comparing the predictions of the classifiers with the actual document category. Therefore this section is split into subsections covering document categorisation, training and test sets, term ranking, and classification, and testing.

### 3.1.1. Document Categorisation

In order to determine the accuracy of a classifier it is necessary to have specific measures of how the market reacts to news. To do so it is necessary to perform time series analysis on the trading data and categorise each document according to how the market behaved shortly after its arrival.

The return time series in Eq. (6) gives the log returns over the period  $\Delta t$  for the stock. The return time series is one of the most interesting to investors as it demonstrates the amount of money which can be made. However, at high frequencies it is impossible to predict returns as the market is far too noisy.

$$R_{(s,\Delta t)} = \{R_1, \dots, R_m\} | R_{(s,\Delta t,t)} = \log(P_{(s,t)}) - \log(P_{(s,t-\Delta t)}) \quad (6)$$

Realised volatility given by  $v_{(s,n,\rho,\Delta t)}$  in Eq. (7) is more commonly used within the finance community to estimate the risk of owning an asset. The variable  $n$  defines the number of previous minutes to sum and  $\rho$  is the exponent for the return.

$$v_{(s,n,\rho,\Delta t)} = \{v_1, v_2, \dots, v_u\} | v_{(s,n,\rho,\Delta t,t)} = \left[ \frac{1}{n} \sum_{j=0}^{n-1} R_{(s,\Delta t,t-j)}^\rho \right]^{\frac{1}{\rho}} \quad (7)$$

There are many methods used to forecast volatility, though the GARCH (Generalised Autoregressive Conditional Heteroskedasticity) model introduced by Bollerslev (1986) is one of the most common. The GARCH( $P,Q$ ) forecast volatility for the stock  $s$ , at time  $t$  is given by  $(\sigma_{(s,\Delta t,P,Q,t)})$  in Eq. (8). It combines autoregression in the variance with the lagged conditional variance. The variable  $P$  is used to define the number of autoregressive components, and  $Q$  is used to define the number of lagged conditional variances to include in the forecast. The variable  $\alpha_0$  is a constant, whilst the  $\alpha$ 's and  $\beta$ 's are used to scale the autoregressive and lagged conditional variances respectively.

$$\sigma_{(s,\Delta t,P,Q)} = \{\sigma_1, \sigma_2, \dots, \sigma_u\} \quad (8)$$

$$| \sigma_{(s,\Delta t,P,Q,t)} = \sqrt{\alpha_0 + \sum_{i=1}^P \alpha_i R_{(s,\Delta t,t-i)}^2 + \sum_{j=1}^Q \beta_j \sigma_{(s,\Delta t,P,Q,t-j)}^2}$$

The parameters for the model are optimised using the previous month's trading data, to ensure that they are not fitted to the given month's trading conditions. For the calendar month of January parameters which were optimised using all trading data for the stock for the calendar month of December are used. This achieved by maximising the log-likelihood function, given by Eq. (9) for the stock  $s$ , where there are  $n$  entries in the time series (Dacorogna et al. 2001). The parameters which produce the maximum likelihood function for the given data are chosen.

$$L(\theta) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^n \left( \ln(\sigma_{(s,\Delta t,P,Q,t)}^2) + \frac{R_{(s,\Delta t,t)}^2}{\sigma_{(s,\Delta t,P,Q,t)}^2} \right) \quad (9)$$

In our case we are trying to optimise the GARCH model after the arrival of news. Therefore we apply GARCH by calculating the forecast error given by Eq. (10), which is the difference between the forecast (Eq. (8)) and realised volatility (Eq. (7)). This highlights periods where the

GARCH model is poor at forecasting the volatility. We use  $P=Q=3$ , because we found in previous work that abnormal forecast errors with these parameters have a strong correlation with the arrival of asset specific news.

$$e_{(s,\Delta t,P,Q)} = \{e_1, e_2, \dots, e_u\} | e_{(s,\Delta t,P,Q,t)} = \sigma_{(s,\Delta t,P,Q,t)}^2 - v_{(s,1,2,\Delta t,t)}^2 \quad (10)$$

We want to categorise documents whose incidence correlates with abnormal forecast errors as interesting. To do so it is necessary to calculate the mean and standard deviation of the forecast error over a given period. The variable  $M$  in Eq. (11) defines the average number of trading minutes per month by using the average number of trading minutes per business day for the relevant country, and multiplying by the average number of trading days per month (20).

$$M = 20 \times m | \{m_{US} = 390, m_{UK} = 510, m_{AU} = 360\} \quad (11)$$

In Eq. (12) the mean  $(\mu_{(s,\Delta t,t)})$  for time  $t$  in the forecast error time series  $e_{(s,\Delta t,P,Q)}$  is defined by taking the mean value for the  $M$  trading minutes which preceded the start of the current trading day. In Eq. (13) the standard deviation  $(std_{(s,\Delta t,t)})$  for time  $t$  in the forecast error time series  $e_{(s,\Delta t,P,Q)}$  is defined by again using the  $M$  trading minutes which preceded the start of the current trading day. Note that if a stock was suspended from trading during the last 20 trading days for the stock exchange, only the last 20 days which the stock traded on are used.

$$\mu_{(s,\Delta t,P,Q,t)} = \frac{\sum_{j=t_0-M}^{t_0-1} e_{(s,\Delta t,P,Q,j)}}{M} \quad (12)$$

$$| t_0 = \min(\{\forall T_{(B,i)} | time(T_{(B,i)}) = \min(time(T_B)) \wedge T_{(B,i)} \leq t\})$$

$$std_{(s,\Delta t,P,Q,t)} = \sqrt{\frac{\sum_{j=t_0-M}^{t_0-1} (e_{(s,\Delta t,P,Q,j)} - \mu_{(s,\Delta t,t)})^2}{M}} \quad (13)$$

$$| t_0 = \min(\{\forall T_{(B,i)} | time(T_{(B,i)}) = \min(time(T_B)) \wedge T_{(B,i)} \leq t\})$$

The category  $\Psi$  of each document in Eq. (4) is calculated using the definition in Eq. (14). If the forecast error within  $\Delta \tau$  minutes equals or exceeds  $\delta$  standard deviations from the mean function value then the document is categorised as interesting (i.e., 1), for same  $\delta$ . Otherwise it is categorised as uninteresting (i.e., 0).

$$\Psi_{(s,\Delta t,\Delta \tau,P,Q,\delta)} = \{\Psi_1, \Psi_2, \dots, \Psi_p\} | \Psi_{(s,\Delta t,P,Q,\Delta \tau,\delta,\lambda)} = \left( \begin{array}{l} \exists t | d_{(s,\lambda)} < t \leq d_\lambda + \Delta \tau \\ \wedge \left( \begin{array}{l} e_{(s,\Delta t,P,Q,t)} \geq \mu_{(s,\Delta t,P,Q,t)} + \delta \times std_{(s,\Delta t,P,Q,t)} \\ \vee e_{(s,\Delta t,P,Q,t)} \leq \mu_{(s,\Delta t,P,Q,t)} - \delta \times std_{(s,\Delta t,P,Q,t)} \end{array} \right) ? 1 : 0 \end{array} \right) \quad (14)$$

### 3.1.2. Training and Test Sets

The stocks for each country  $c$  are grouped together using Eq. (15) to form a large dataset of related stocks. Each document for each stock within each country is then categorised using the forecast error time series with the chosen parameters. Training sets are created by taking  $N$  documents, of which  $R$  are categorised as interesting (i.e., those which correlated to abnormal behaviour), and the rest are not. The test set is a subset of the documents not included in the training set.

$$G_{(c)} = \{G_1, G_2, \dots, G_v\} \quad (15)$$

### 3.1.3. Term Ranking

A dictionary is created using Eq. (16) for each term which appears in at least one document for a stock in the training set. The term count ( $d_j$ ), document count ( $df_j$ ), and interesting document count ( $r$ ) are stored for each term. The term count  $d_j$  is the total number of times the given term appears in all documents in the training set. The document count  $df_j$  is the total number of documents which contain the given term. The interesting document count  $r$  is the total number of documents which are categorised as interesting in the training set which contain the given term. The subscript  $\eta$  refers to a distinct document within the training set.

A sub-dictionary is formed by taking the top  $\phi$  terms based on a given term ranking algorithm. For this research we chose three term ranking methods which we will subsequently define.

$$\begin{aligned} X_{(c,\Delta t,P,Q,\Delta\tau,\delta)} &= \{X_1, X_2, \dots, X_w\} \mid X_{(c,\Delta t,\delta,\eta)} = \\ &\{S_{(c,\Delta t,\delta,\eta)}, d_{j(\eta)}, df_{j(\eta)}, r_{j(\eta)}\} \\ \wedge d_{j(\eta)} &= \sum SC_{(s,\lambda,\eta)} \mid s \in G_{(c)} \\ \wedge df_{j(\eta)} &= \#\{\forall C_{(s,\lambda)} \mid SC_{(s,\lambda,\eta)} > 0 \wedge s \in G_{(c)}\} \\ \wedge r_{j(\eta)} &= \#\{\forall C_{(s,\lambda)} \mid SC_{(s,\lambda,\eta)} > 0 \wedge \Psi_{(s,\Delta t,P,Q,\Delta\tau,\delta,\lambda)} = 1 \wedge s \in G_{(c)}\} \end{aligned} \quad (16)$$

Firstly, we choose the term frequency inverse document frequency (TFIDF) method given by Eq. (17). Note the  $N$  in Eq. (17) is the number of documents in the training set. The inverse document frequency helps to bias against terms which occur in every document. The term frequency helps to favour terms which occur frequently. Note that, typically, TFIDF is used to measure the effect of a term within a single document, whilst here it is used to measure the effect of the term within the training set.

$$TFIDF = d_j \times \log_{10} \left( \frac{N}{df_j} \right) \quad (17)$$

Secondly, the binary version of the gain ratio introduced by Quinlan (1993), given by Eq. (19) was chosen. This method selects terms which provide the most information, i.e., splits the data between the classes most effectively. In Eq. (19)  $E(R, N)$  is the entropy value (Eq. (18)) for the ratio of interesting documents ( $R$ ) to documents ( $N$ ) in the training set. The next part calculates the entropy value for the ratio of interesting documents to documents which contain the term, scaled by the ratio of documents which contain the term. This helps to select terms which occur frequently in interesting documents. The last part of the equation calculates the entropy value for the ratio of uninteresting documents to documents which contain the term, scaled by the ratio of documents which do not contain the term. This helps to select terms which do not occur in interesting documents, i.e., documents which do not contain the term are interesting.

$$E(n, m) = - \left( \frac{n}{m} \log_2 \left( \frac{n}{m} \right) + \left( 1 - \frac{n}{m} \right) \log_2 \left( 1 - \frac{n}{m} \right) \right) \mid n \leq m \quad (18)$$

$$GAIN = E(R, N) - \frac{df_j}{N} \times E(r, df_j) - \frac{N - df_j}{N} \times E(df_j - r, df_j) \quad (19)$$

Finally, the BM25 algorithm (Best Match) introduced by Robertson and Spärck Jones (2006) was adapted to get the Average Document BM25 value (ADB25). This is given by Eq. (20), where  $k_1$  and  $b$  are constants,  $dl_{(i)}$  is the length of the document  $i$ , and  $avdl$  is the average document length for documents in the training set. The ADB25 algorithm is the same as the BM25 algorithm if  $N$  were equal to 1, or in other words if there was only one document. The first part of the equation normalises the term frequency by taking into account the length of the document which contains the term and the average document length. This ensures that, if a term occurs frequently in a very long document, it is not given unwarranted significance. The log part of the equation normalises results by factoring in the number of interesting documents which contain the term ( $r$ ), the number of documents which contain the term ( $df_j$ ) and the total number of interesting documents ( $R$ ) and documents ( $N$ ). This favours terms which provide more information, i.e., splits the two classes most efficiently.

$$ADB25 = \frac{1}{N} \sum_{i=1}^N \frac{(k_1 + 1) \times d_j}{\left( k_1 \times \left( (1 - b) + b \times \frac{dl_{(i)}}{avdl} \right) \right) + d_j} \times \log \left( \frac{(r + 0.5)(N - df_j - R + r + 0.5)}{(df_j + 0.5)(R - r + 0.5)} \right) \quad (20)$$

### 3.1.4. Classification

A binary vector is created for each document in the training and test sets where each entry specifies whether the given term (which is a member of the sub-dictionary) occurred in the document. These vectors are used to train and test the C4.5 decision tree introduced by Quinlan (1993), and the support vector machine (SVM) introduced by Vapnik (1999) using the SVM Light Classifier released by Joachims (2007).

The C4.5 decision tree introduced by Quinlan (1993) classifies documents by building a tree where the root node is the term which produces the highest Gain value (Eq. (18)). The root node contains two leaf nodes, the first is for all documents which contain the term and the second is for all documents which exclude the term. The tree is grown by recursively repeating the process at each node on the documents which contain/exclude each term contained in the path directly from the root node to the current node. However, only terms which are contained in the remaining documents are included in the search for the next term.

The support vector machine (SVM) introduced by Vapnik (1999) projects the terms and their values into higher dimensional space (e.g. one dimension per term). It produces a classifier by identifying the hyperplane which most effectively separates the two classes.

### 3.1.5. Testing

To compare the performance of different classifiers there are several statistical measures which are commonly used. The most important of these is the classification accuracy, given by Eq. (21), which is the ratio between the number of vectors correctly classified ( $\#TP$  is the

number of true positives, and  $\#TN$  is the number of true negatives, and  $N$  is the number of documents).

$$Accuracy = \frac{\#TP + \#TN}{N} \quad (21)$$

The True Positive Rate also known as Sensitivity, given by Eq. (22) is the percentage of documents whose incidence correlated with abnormal behaviour which were correctly classified.

$$True\ Positive\ Rate = Sensitivity = \frac{\#TP}{\#TP + \#FN} \quad (22)$$

The False Positive Rate which is equivalent to 1 subtract the Specificity, given by Eq. (23), is the percentage of documents whose incidence did not correlate with abnormal behaviour which were incorrectly classified.

$$False\ Positive\ Rate = 1 - Specificity = \frac{\#FP}{\#TN + \#FP} \quad (23)$$

It is common practice when demonstrating the performance of a classifier to plot a Receiver Operating Characteristic (ROC) Curve. This has the True Positive Rate on the Y axis and the False Positive Rate on the X axis.

### 3.2. News Aware GARCH

In this section we define a variation of the GARCH model which is aware of the arrival of news. In Eq. (4) we defined the set of each distinct news document for the stock. For the purpose of forecasting the reaction to news we are more concerned whether news occurred at the given time for the stock. Therefore we produce the news time series defined in Eq. (24) such that each trading minute for the stock contains the count of the documents forecast to cause a shock. Note that  $\Gamma(A_{(s,\Delta t,\lambda)}, \delta)$  denotes the outcome of the classifiers defined in 3.1 where  $\delta$  is the given threshold. Note also that when we refer to  $\delta=0$ , we simply mean that no classification was used so every document at the given time is included.

$$N_{(s,\Delta t,\delta)} = \{N_1, N_2, \dots, N_q\} | N_{(s,\Delta t,\delta,t)} = \#\{\forall A_{(s,\Delta t,\lambda)} | D_{(t-1)} < d_{(s,\lambda)} \leq D_{(t)} \wedge \Gamma(A_{(s,\Delta t,\lambda)}, \delta) = 1\} \quad (24)$$

#### 3.2.1. NAGARCH-S

Let us assume that the GARCH model is effective at forecasting future volatility when news has not been released to the market. Furthermore let us assume that when news is released to the market investors process this information and their behaviour makes it difficult for GARCH to forecast volatility. Therefore the state of the GARCH model must change in order to take advantage of the knowledge that news has been released.

The Baseline GARCH model for predicting  $\Delta t$  minutes into the future for the stock  $s$ , at time  $t$  is given by  $(\sigma_{B(s,\Delta t,P,Q,t)})^2$  in Eq. (25) where  $\alpha_{B0}$ ,  $\alpha_{Bi}$ , and  $\beta_{Bj}$  are constants.

$$\sigma_{B(s,\Delta t,P,Q,t)} = \sqrt{\alpha_{B0} + \sum_{i=1}^P \alpha_{Bi} R_{(s,\Delta t,t-i\Delta t)}^2 + \sum_{j=1}^Q \beta_{Bj} \sigma_{B(s,\Delta t,P,Q,t-j\Delta t)}^2} \quad (25)$$

We define the News Aware GARCH Switching model (NAGARCH-S) for predicting  $\Delta t$  minutes into the future for the stock  $s$ , at time  $t$  is given by  $(\sigma_{S(s,\Delta t,P,Q,\delta,t)})^2$  in Eq. (26), where  $\alpha_{S0}$ ,  $\alpha_{Si}$ , and  $\beta_{Sj}$  are constants. Furthermore  $N_{(s,\Delta t,\delta,t-k)}$  is the number of articles at time  $t-k$  classified using the threshold  $\delta$  to correlate with abnormal market behaviour. Note that the conditional variance (i.e., the forecast volatility) of the Baseline GARCH model is used within NAGARCH-S. This is to ensure that forecasts are unaffected by a period when news occurs frequently, which is a concern for parameter optimisation.

$$\sigma_{S(s,\Delta t,P,Q,\delta,t)} = \left( \left( \sum_{k=1}^M N_{(s,\Delta t,\delta,t-k)} \right) = 0 ? \sigma_{B(s,\Delta t,P,Q,t)} : \sqrt{\alpha_{S0} + \sum_{i=1}^P \alpha_{Si} R_{(s,\Delta t,t-i\Delta t)}^2 + \sum_{j=1}^Q \beta_{Sj} \sigma_{B(s,\Delta t,P,Q,t-j\Delta t)}^2} \right) \quad (26)$$

Parameters which are evaluated in a given test month are optimised to maximise the Log Likelihood function defined in Eq. (9) for the models during a training set for each stock. The training set comprises of a limited number of months which occurred prior to the test month. The classifier used to classify documents during the training period is trained during a period which excludes both the training and test months. This is to ensure that the classifier does not use prior knowledge to determine how the market will react after the news is released.

Parameters for the Baseline GARCH model are optimised over the entire time series in the training set for the stock. Parameters for the NAGARCH-S model are optimised during the  $\Delta t$  minutes after the release of a document classified to correlate with abnormal market behaviour using the threshold  $\delta$  in the training set. This is because it is the only time when the model produces a different forecast from the Baseline GARCH model.

In the event that parameters could not be found to improve the NAGARCH-S model over the Baseline GARCH model, parameters from a previous month for the stock are chosen.

### 3.3. Measuring Forecast Performance

In order to evaluate whether the NAGARCH-S model is any better than the GARCH model it is necessary to measure the difference in forecast accuracy. In this section we define several measures which we use to compare the models.

The benchmark signal ( $b$ ), given by Eq. (27), is the error between the forecast and realised volatility for the stock  $s$  at time  $t$  using the GARCH model. We define the realised volatility at time  $t$  using the volatility definition in Eq. (7) using  $n=1$  and  $\rho=2$ .

$$b_{(s,\Delta t,P,Q,t)} = \sigma_{B(s,\Delta t,P,Q,t)}^2 - v_{(s,1,2,\Delta t,t)}^2 \quad (27)$$

The forecast signal ( $f$ ) is the error between the NAGARCH-S forecast and the realised volatility for the stock  $s$  at time  $t$ . We define the realised volatility at time  $t$  using the volatility definition in Eq. (7) using  $n=1$  and  $\rho=2$ .

$$f_{(s,\Delta t,P,Q,\delta,t)} = \sigma_{S(s,\Delta t,P,Q,\delta,t)}^2 - v_{(s,1,2,\Delta t,t)}^2 \quad (28)$$

We want to evaluate the performance of the model across multiple stocks from the same country so we group the stocks as defined in Eq. (29). This is useful for calculating the average performance improvement for the model for each country.

$$S = \{S_1, S_2, \dots, S_h\} | h \geq 1 \quad (29)$$

### 3.3.1. Unscaled Forecast Quality ( $Q_u$ )

The unscaled forecast quality ( $Q_u$ ), given by Eq. (30), measures the improved performance of the model over the benchmark by comparing the sum of the absolute errors for all stocks in the set. Note that term ‘‘unscaled’’ is used as typically the forecast quality factors in the change in the realised volatility (Dacorogna et al. 2001). Note also that  $t \in T_{B(s)}$  means that the minute  $t$  is a member of business time  $T_B$  for the stock  $s$ . In other words the minute occurred during a business day when the stock was not suspended from trading.

$$Q_u(s, \Delta t, P, Q, \delta) = 1 - \frac{\sum_{s \in S} \sum_{t \in T_{B(s)}} \left\{ \left| f_{(s, \Delta t, P, Q, \delta, t)} \right| \mid t \in T_{B(s)} \right\}}{\sum_{s \in S} \sum_{t \in T_{B(s)}} \left\{ \left| b_{(s, \Delta t, P, Q, t)} \right| \mid t \in T_{B(s)} \right\}} \quad (30)$$

### 3.3.2. Superior Quality ( $Q_s$ )

The superior quality ( $Q_s$ ), given by Eq. (31), finds the percentage of times that the forecast signal is better than the benchmark signal. If the value is 0 then it is not worth using the model as the forecast is never better than the benchmark. Note that  $t \in T_{B(s)}$  means that the minute  $t$  is a member of business time  $T_B$  for the stock  $s$ . In other words the minute occurred during a business day when the stock was not suspended from trading.

$$Q_s(s, \Delta t, P, Q, \delta) = \frac{\sum_{s \in S} \# \left\{ \forall t \mid \left| f_{(s, \Delta t, P, Q, \delta, t)} \right| < \left| b_{(s, \Delta t, P, Q, t)} \right| \mid t \in T_{B(s)} \right\}}{\sum_{s \in S} \# \left\{ \forall t \mid t \in T_{B(s)} \right\}} \quad (31)$$

## 4. Results

We have separated the results in sections titled News Classification and Model Performance. The first section describes how the best classifiers were chosen for each country. The second section describes how the NAGARCH-S model performed using the classified news.

### 4.1. News Classification

We have divided this section into two subsections. The first addresses the issue of the size of the time window for finding abnormal market behaviour. The second addresses the problem of how much historical knowledge is necessary to produce the best classifiers.

#### 4.1.1. Choice of Time Window

In order to choose an effective news classifier it is first necessary to determine an effective time window for measuring abnormal behaviour. For this purpose the documents are categorised using various time window sizes ( $\Delta t = \Delta \tau$ ) and  $\delta = 6$  standard deviations for the forecast

error time series with  $P=Q=3$  (approximately the 99.7<sup>th</sup> percentile for 30 minute returns) (Robertson et al. 2007).

There were 10 training sets created by selecting  $N=1,000$  documents and  $R=500$  documents at random which correlated to abnormal market behaviour from the entire collection of news documents for the country. The test set for the respective training sets contained all the documents for the country which were not included in the training set. An equal allocation of documents which correlated to abnormal behaviour and those that did not was chosen so as not to bias the classifier. Tests were run using both the SVM and the C4.5 classifiers and each term ranking algorithm with varying  $\phi$  values (100, 200, 500, 1,000, 2,000, and 5000 terms).

In Fig. 1 the effect of increasing the time window size ( $\Delta t = \Delta \tau$ ) is investigated on the mean accuracy of the classifiers using every variation of  $\phi$  terms (Note that the mean is calculated over the 10 test sets). The most accurate term ranking algorithm and classifier combination are displayed.

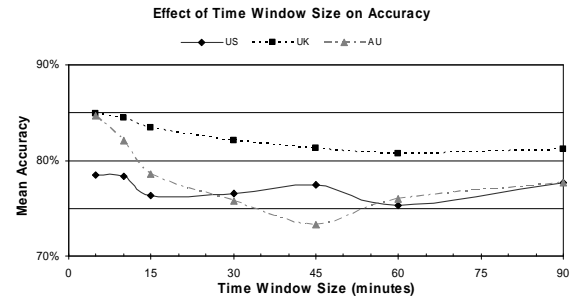


Fig. 1. Effect of Time Window Size on Accuracy.

The most accurate results (i.e., those with the highest mean accuracy) for every country are achieved within 5 minutes. As the time window size is increased there is a slight reduction in the accuracy in the UK, though a substantial reduction in Australia. The US is a little more stable than Australia but not as efficient as the UK. Therefore it appears that investors in all countries react quickly and decisively to news. This indicates that investors in all countries are rational. Increasing the time window size reduces the accuracy as  $\Delta \tau$  increases the number of documents which spuriously correlate to abnormal market behaviour. Increasing the value of  $\Delta t$  however could yield better results as there is too much noise in the market at extremely high frequencies.

#### 4.1.2. Choice of Historical Time Window

The tests in the previous section were useful for highlighting the time window size ( $\Delta t = \Delta \tau$ ) to use for classifying news. However, it is not practical to use a classifier trained on a sample of all documents. This is because it is possible that priori information is used to classify the document. Therefore it is necessary to produce classifiers for each month which have no knowledge of the immediate future.

The training sets were created using the past  $\Omega$  months. Documents are categorised using the forecast error time series with  $P=Q=3$ ,  $\Delta t = \Delta \tau = 5$  minutes, and  $\delta = 6$  standard deviations. Each document categorised as interesting during this period ( $R$ ) is included in the training set. To

avoid biasing the classifier we use  $N=2R$  and therefore chose  $R$  uninteresting documents at random in the same period. In the event that there are not  $\Omega$  months prior to the given month then extra months from the end of the dataset are used. It is unlikely that an event which caused a major shock will be referred to in a document released a long time afterwards.

A training set is created for each month and each  $\Omega$  value (3, 6, 9 and 12 months) using both the SVM and the C4.5 classifiers and each term ranking algorithm with varying  $\phi$  values (100, 200, 500, 1,000, 2,000, and 5000 terms).

The results in Fig. 2 show the mean accuracy of the best classifier for each  $\Omega$  value (Note that the mean is calculated over the test set for each month). It is clear that more historical knowledge is advantageous as it provides the classifier with a wide selection of different types of documents which correlated to shocks. If the classifier were only trained on documents which were released during annual reporting season, it is likely that there would be a strong bias towards words such as “earnings”, “profit”, and “loss”. These words are less likely to cause a shock throughout the rest of the year, unless the document reports an unexpected large profit or loss. Note that the mean and standard deviation of classifiers which are trained on only immediate history are very similar to those which also use months from the end of the dataset.

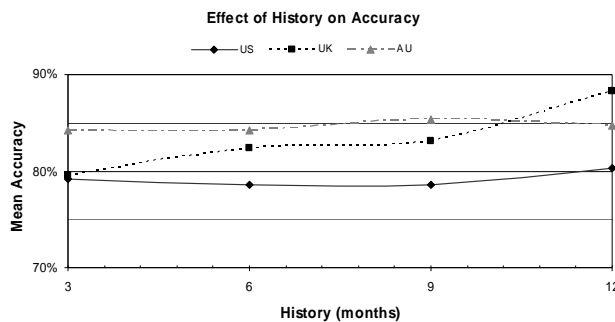


Fig. 2. Effect of History on Accuracy.

The results in Table 1 and Table 2 show the classification details for the best classifier for each country. The mean true and false positive rates are provided in the TPR and FPR columns respectively of Table 2. The  $\Omega=12$  value yielded the best results for the US and UK, whilst the  $\Omega=9$  value produced the best results for Australia.

Table 1. Characteristics of Best Classifiers.

$\delta$	Country	Classifier	Term Ranking	Terms ( $\phi$ )	Documents	
					Total	Positive
4	US	SVM	GAIN	1,000	133,019	28,742
	UK	C4.5	GAIN	100	81,522	6,907
	AU	SVM	ADBM25	5,000	33,098	5,187
6	US	SVM	ADBM25	100	133,019	25,944
	UK	C4.5	GAIN	100	81,522	8,995
	AU	SVM	ADBM25	2,000	33,098	4,753

The true positive rate (TPR) value in Table 2 shows that despite the UK having a high accuracy there is a low

percentage of documents which actually correlated to a shock which were correctly classified. However, the accuracy rates for all tests are promising as Mittermayer (2004) only achieved 58%. It should be noted though the Mittermayer was attempting predict the direction of return, which is harder to do.

Table 2. Accuracy of Best Classifiers.

$\delta$	Country	Accuracy	TPR	FPR
4	US	77.73%	34.36%	78.67%
	UK	90.19%	19.44%	91.77%
	AU	83.77%	39.13%	84.94%
6	US	80.31%	42.26%	80.77%
	UK	88.25%	25.60%	89.18%
	AU	85.19%	37.07%	86.04%

## 4.2. Model Performance

In this section we evaluate the performance of the NAGARCH-S model using several thresholds for the news time series. Initially we investigate the unscaled forecast quality for each country to determine if the NAGARCH-S model improves on the Baseline GARCH model. This includes tests to determine whether the results are statistically significant. Then we evaluate the superior quality for each country to determine how frequently the NAGARCH-S model provides a better forecast than the Baseline GARCH model.

For all tests we used the previous 3 months of trading data for each stock to optimise parameters. Furthermore the news time series were assembled using the classified documents for the same period for the stock. Specifically this means that 3 separate classifiers were used for each test as the classifiers were each produced to predict one month ahead. We did so because there are not enough samples for the  $\delta=6$  news time series with only one month of data.

We forecast the volatility of returns for every minute in the time series for every stock in each country, using  $P=Q=1$  to limit the cost of parameter optimisation. We make no attempt to predict the delay between news arrival and market reaction, but simply use regression to optimise the parameters for the  $\Delta t$  minutes after news. Therefore if the market tends to take 3 minutes to react to news then the forecast volatility of the NAGARCH-S for the first 3 after news will probably be worse than the Baseline GARCH model. Note that as we classified documents with  $\Delta\tau=5$  minutes, articles which occurred within the first or last 5 minutes of the trading day were excluded.

### 4.2.1. Unscaled Forecast Quality

In Fig. 3 - Fig. 5 the unscaled forecast quality ( $Q_u$ ) for the US, UK, and Australia respectively is evaluated for all time windows ( $\Delta t$ ). Note that  $Q_u$  is calculated for the  $\Delta t$  minutes after news as the models are the same without news. In each figure the legends STD0, STD4, and STD6 correspond to the model using news classifiers with the  $\delta=0$ ,  $\delta=4$ , and  $\delta=6$  thresholds respectively. Note that the  $\delta=0$  threshold means that all news is processed by the NAGARCH-S model.

In Fig. 3 it is shown that the  $Q_u$  of the models using the  $\delta=4$  and  $\delta=6$  thresholds in the US are consistently better than for the  $\delta=0$  threshold. This suggests that the content of the news is important, and the classifiers are performing well. The  $\delta=4$  and  $\delta=6$  thresholds provide very similar values until after the 15 minute time window. This implies that news which is classified using the  $\delta=6$  is not significantly different from that classified using the  $\delta=4$  threshold. However, for time windows larger than 15 minutes the  $\delta=4$  threshold yields higher  $Q_u$  values. This is most likely because there are less documents are classified to correlate with abnormal volatility forecast errors using the  $\delta=6$  threshold. Therefore it is difficult to optimise parameters for this threshold as there are fewer periods around news, and therefore regression tends to overfit parameters.

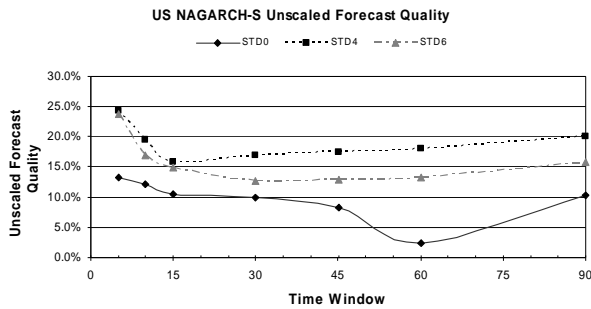


Fig. 3. Unscaled Forecast Quality in the US.

The results in Fig. 4 show that until the 15 minute window the  $Q_u$  of the model using the  $\delta=4$ , and  $\delta=6$  thresholds in the UK are higher than for the  $\delta=0$  threshold. It is also clear that the  $\delta=6$  threshold yields better results than  $\delta=4$  during this period. However, for larger time windows the  $Q_u$  values tend to be negative. This is because the forecast accuracy of the Baseline GARCH model with these time windows is substantially lower than for smaller time windows. Therefore as NAGARCH-S attempts to improve on the Baseline GARCH model it overfits parameters to the training set which leads to significantly worse performance in the test set.

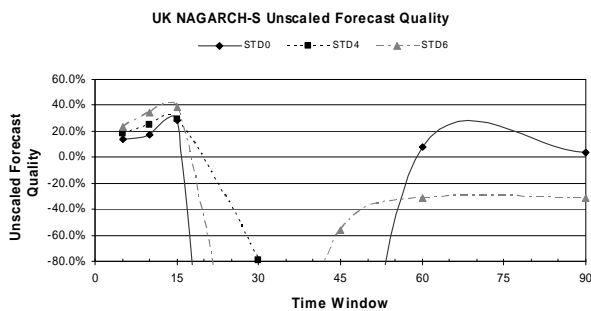


Fig. 4. Unscaled Forecast Quality in the UK.

For the 60 and 90 minute time windows in the UK, as shown in Fig. 4, the  $\delta=0$  threshold provides positive  $Q_u$  values. This indicates that the Baseline GARCH performance begins to improve and the large number of documents aids parameter optimisation. Therefore it appears that the content of news is important in the UK

and the classifiers are performing well. However, it is difficult to forecast volatility a long time into the future. This suggests that the volatility does not persist for long after the release of news.

The 5 minute time window in Fig. 5 reveals that the  $\delta=0$  threshold provides higher  $Q_u$  values than the  $\delta=4$ , and  $\delta=6$  thresholds in Australia. This is possibly because there is the potential for a large improvement over the Baseline GARCH model during this period, and the other thresholds do not have sufficient documents to optimise parameters effectively. However, for all other time windows the  $\delta=4$  and  $\delta=6$  thresholds yield higher  $Q_u$  values. This suggests that the content of news is important in Australia and that the classifiers are performing well.

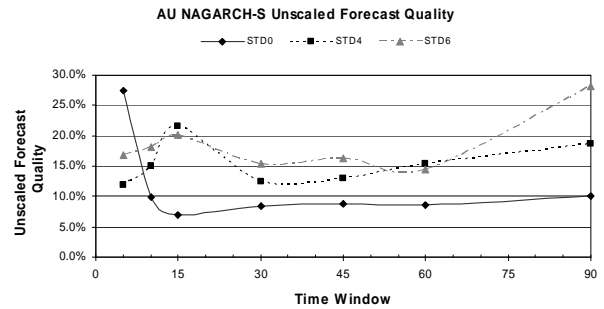


Fig. 5. Unscaled Forecast Quality in Australia.

We test the null hypothesis that the NAGARCH-S model produces the same forecasts as the Baseline GARCH model using an F-Test. This compares the average forecast error for each model for each month and each stock in the given country for the 5 minute time window. The p values of these tests are shown in Table 3. They reveal that, apart for the  $\delta=0$  threshold in the US, the NAGARCH-S model produces statistically significant different forecasts than the Baseline GARCH model.

Table 3. Significance of Forecasts for the 5 minute window.

Country	Threshold ( $\delta$ )		
	0	4	6
US	69.79%	0.00%	6.63%
UK	0.00%	0.00%	0.00%
AU	0.00%	0.42%	0.00%

The results in this section indicate that documents classified to correlate with abnormal volatility forecast errors improve the NAGARCH-S model more than all documents. This implies that the content of the news is important and investors do not tend to react to all news.

#### 4.2.2. Superior Quality

In Fig. 6 - Fig. 8 the superior quality ( $Q_s$ ) for the US, UK, and Australia respectively is evaluated for all time windows ( $\Delta t$ ). Note that  $Q_s$  is calculated for the  $\Delta t$  minutes after news as the models are the same without news. In each figure the legends STD0, STD4, and STD6 correspond to the model using news classifiers with the  $\delta=0$ ,  $\delta=4$ , and  $\delta=6$  thresholds respectively. Note that the  $\delta=0$  threshold means that all news is processed by the NAGARCH-S model.

The results in Fig. 6 - Fig. 8 reveal that the  $\delta=4$  and  $\delta=6$  thresholds provide better forecasts than the  $\delta=0$  threshold for all time windows. Note that it is difficult to tell for the 5 minute time window in Australia, though it is the case.

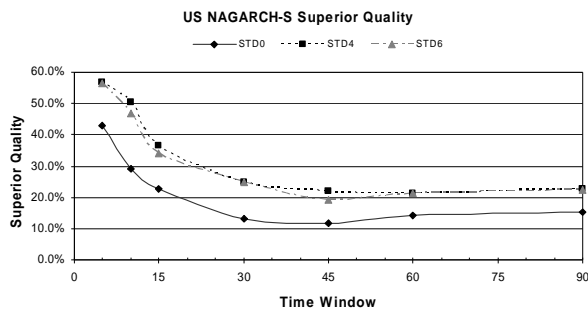


Fig. 6. Superior Quality in the US.

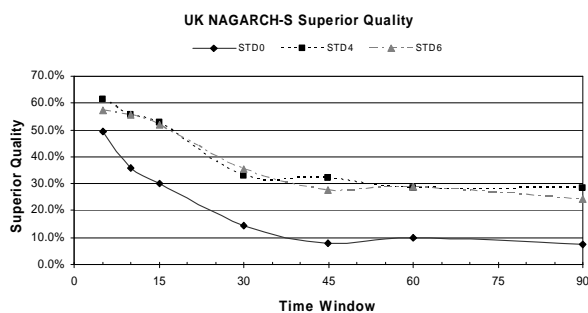


Fig. 7. Superior Quality in the UK.

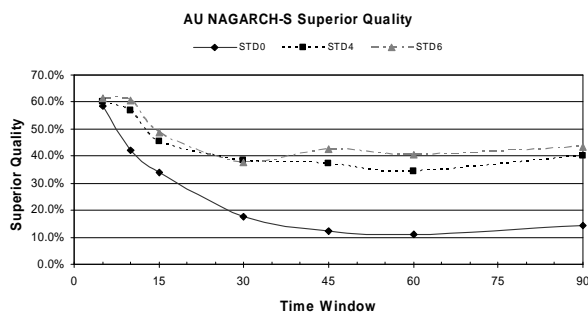


Fig. 8. Superior Quality in Australia.

The results in Fig. 6 - Fig. 8 also demonstrate that as the models attempt to forecast volatility further into the future there is less chance of producing better forecasts than the Baseline GARCH model. However, this does not mean that the models tend to be worse than the Baseline GARCH model. They actually have forecast accuracies greater than or equal to the Baseline GARCH model over 70% of the time for all time windows. Despite forecasts being worse for up to 30% of the time the results in the previous section reveal that the models tend to be better than the Baseline GARCH model. This suggests that when the models provide worse forecasts they are not large compared to the periods of better forecasts.

### 4.2.3. Summary

These results demonstrate that substantially greater forecasts can be achieved when considering news. However, the unscaled forecast quality results in the UK demonstrate that this model is not universally effective. Therefore it is necessary to perform comprehensive tests on a large dataset before determining what conditions are best for applying this model.

## 5. Conclusions

We have introduced a variation of the GARCH model which is aware of the arrival of news. We have shown that it is very effective at improving the forecast accuracy around news for the US and Australia for forecasts up to 90 minutes into the future. However, in the UK it is best not to forecast more than 15 minutes into the future as the model tends to be worse than the Baseline GARCH model.

We have demonstrated that classifying news based on the content improves the performance of this model more than by using all news. To our knowledge we have achieved higher classification accuracy rates for forecasting the market reaction to news than any previously reported by other authors.

Furthermore we have provided evidence that these models are statistically better than GARCH except in the US when using all news. Therefore it is clear that knowledge of news arrival is not enough, and it is very important to interpret the content of the news before forecasting how the market will react.

In future research we plan to investigate ways to improve the forecasts.

## 6. References

- Almeida, A., Goodhart, C. A. E. and Payne, R. (1998): The Effects of Macroeconomic News on High Frequency Exchange Rate Behavior. *Journal of Financial & Quantitative Analysis*, 33(3):383-408.
- Bollerslev, T. (1986): Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, 31(3):307-27.
- Cutler, D. M., Poterba, J. M. and Summers, L. H. (1989): What Moves Stock Prices? *Journal of Portfolio Management*, 15(3):4-12.
- Dacorogna, M. M., Gencay, R., Müller, U., Olsen, R. B. and Pictet, O. V. (2001) *An Introduction to High-Frequency Finance*, Academic Press, London.
- Ederington, L. H. and Lee, J. H. (1993): How markets process information: News releases and volatility. *Journal of Finance*, 48(4):1161-1191.
- Ederington, L. H. and Lee, J. H. (1995): The short-run dynamics of the price adjustment to new information. *Journal of Financial & Quantitative Analysis*, 30(1):117-134.
- Ederington, L. H. and Lee, J. H. (2001): Intraday Volatility in Interest-Rate and Foreign-Exchange Markets: ARCH, Announcement, and Seasonality Effects. *Journal of Futures Markets*, 21(6):517-552.
- Goodhart, C. A. E. (1989): News and the foreign exchange market. *Proc. Manchester Statistical Society*, 1-79.

- Goodhart, C. A. E., Hall, S. G., Henry, S. G. B. and Pesaran, B. (1993): News Effects in a High-Frequency Model of the Sterling-Dollar Exchange Rate. *Journal of Applied Econometrics*, 8:1-13.
- Graham, M., Nikkinen, J. and Sahlstrom, P. (2003): Relative Importance of Scheduled Macroeconomic News for Stock Market Investors. *Journal of Economics and Finance*, 27(2):153-165.
- Hong, H., Lim, T. and Stein, J. C. (2000): Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *Journal of Finance*, 55(1):265-95.
- Joachims, T., *SVM Light Classifier*(2007). Available: <http://svmlight.joachims.org/>.
- Kalev, P. S., Liu, W.-M., Pham, P. K. and Jarnecic, E. (2004): Public Information Arrival and Volatility of Intraday Stock Returns. *Journal of Banking and Finance*, 28(6):1441-1467.
- Kim, S.-J., McKenzie, M. D. and Faff, R. W. (2004): Macroeconomic News Announcements and the Role of Expectations: Evidence for US Bond, Stock and Foreign Exchange Markets. *Journal of Multinational Financial Management*, 14(3):217-232.
- Melvin, M. and Yin, X. (2000): Public Information Arrival, Exchange Rate Volatility, and Quote Frequency. *Economic Journal*, 110(465):644-661.
- Michaely, R. and Womack, K. L. (1999): Conflict of Interest and the Credibility of Underwriter Analyst Recommendations. *Review of Financial Studies*, 12(4):653-86.
- Mitchell, M. L. and Mulherin, J. H. (1994): The Impact of Public Information on the Stock Market. *Journal of Finance*, 49(3):923-50.
- Mittermayer, M.-A. (2004): Forecasting Intraday Stock Price Trends with Text Mining Techniques. Proc. 37th Annual Hawaii International Conference on System Sciences (HICSS'04), Big Island, Hawaii, 30064b.
- Nofsinger, J. R. and Prucyk, B. (2003): Option volume and volatility response to scheduled economic news releases. *Journal of Futures Markets*, 23(4):315-345.
- Porter, M. F. (1980): An Algorithm for Suffix Striping. *Automated Library and Information Systems*, 14(3):130-137.
- Quinlan, J. R. (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann.
- Robertson, C. S., Geva, S. and Wolff, R. C. (2007): The Intraday Effect of Public Information: Empirical Evidence of Market Reaction to Asset Specific News from the US, UK, and Australia. SSRN Working Paper Series: <http://ssrn.com/abstract=970884>.
- Robertson, S. and Spärck Jones, K. (2006): Simple, Proven Approaches to Text Retrieval. University of Cambridge Computer Laboratory Technical Report no. 356.
- Roll, R. (1984): Orange Juice and Weather. *American Economic Review*, 74(5):861-80.
- Vapnik, V. (1999) *The Nature of Statistical Learning Theory*, Springer-Verlag.
- Womack, K. L. (1996): Do Brokerage Analysts' Recommendations Have Investment Value? *Journal of Finance*, 51(3):137-67.