Asgharian, Hossein; Sikström, Sverker

Working Paper

Predicting Stock Price Volatility by Analyzing Semantic Content in Media

Working Paper, No. 2014:38

Provided in Cooperation with:
Department of Economics, School of Economics and Management, Lund University

Suggested Citation: Asgharian, Hossein; Sikström, Sverker (2014) : Predicting Stock Price Volatility by Analyzing Semantic Content in Media, Working Paper, No. 2014:38, Lund University, School of Economics and Management, Department of Economics, Lund

This Version is available at:
https://hdl.handle.net/10419/260134

Working Paper 2014:38

Department of Economics
School of Economics and Management

Predicting Stock Price Volatility by Analyzing Semantic Content in Media

Hossein Asgharian
Sverker Sikström

November 2014

LUND
UNIVERSITY

Predicting Stock Price Volatility by Analyzing Semantic Content in Media

Hossein Asgharian[*]

Department of Economics, Lund University and *Knut Wicksell Centre for Financial Studies*

Sverker Sikström[**]

Department of Psychology, Lund University

Abstract

Current models for predicting volatility do not incorporate information flow and are solely based on historical volatilities. We suggest a method to quantify the semantic content of words in news articles about a company and use this as a predictor of its stock volatility. The results show that future stock volatility is better predicted by our method than the conventional models. We also analyze the functional role of text in media either as a *passive* documentation of past information flow or as an *active* source for new information influencing future volatility. Our data suggest that semantic content may take both roles.

Keywords: volatility, information flow, latent semantic analysis, GARCH

JEL classification: G19

[*] Corresponding author, professor of economics at the Department of Economics, Lund University, and Knut Wicksell Centre for Financial Studies. Department of Economics, Lund University Box 7082, S-22007 Lund, Sweden. Tel.: +46 46 222 8667; fax: +46 46 222 4118. Hossein.Asgharian@nek.lu.se. This research is supported by a grant from *Jan Wallanders och Tom Hedelius Stiftelse*.

[**] Professor of psychology at the Department of Psychology, Lund University, Box 7082, S-22007 Lund, Sweden. sverker.sikstrom@psychology.lu.se. This research is supported by a grant from the Swedish Research Council.

Predicting Stock Price Volatility by Analyzing Semantic Content in Media

Abstract

Current models for predicting volatility do not incorporate information flow and are solely based on historical volatilities. We suggest a method to quantify the semantic content of words in news articles about a company and use this as a predictor of its stock volatility. The results show that future stock volatility is better predicted by our method than the conventional models. We also analyze the functional role of text in media either as a *passive* documentation of past information flow or as an *active* source for new information influencing future volatility. Our data suggest that semantic content may take both roles.

Keywords: volatility, information flow, latent semantic analysis, GARCH

JEL classification: G19

1. Introduction

Volatility, defined as the variation of return around some expected value, is a commonly used estimate of risk in financial assets. The expected future volatility is therefore a key parameter for portfolio selection, risk management, and pricing-equity-related derivative instruments. Risk anticipation also has an important implication for policy makers such as central banks and financial regulators. Current models for predicting future volatility (e.g., GARCH, stochastic volatility) are based on information available on the historical price variations and try to fit statistical models on data to give a forecast of the future volatility. Thus, these models do not directly rely on the information per se, but on the market's interpretation of the available information.

Because of the cognitive biases in processing information about the market (e.g., Gärling et al., 2009), it is important to make a distinction between information available on the market and the interpretation of this information. Empirical psychological research has identified a number of such biases. Some examples are overconfidence (Glaser et al., 2004), where people believe that their knowledge is more accurate than it really is (Lichtenstein et al., 1982) or that their abilities are above average (Svenson, 1981), and optimism, where people have an overly optimistic belief about the future (Weinstein, 1980). In addition, evaluations of outcomes may systematically differ depending on whether the outcomes are framed as gains or losses (Kahneman and Tversky, 1979). The tendency of actors to imitate each other may also lead to information cascades, where investors ignore relevant information and focus on other actors' behaviors (Smith and Sørensen, 2000). These examples of cognitive biases suggest that investors do not always act rational on available information.

We suggest that a more direct measure of available information would be less sensitive to investors' biased processing of available information. The problem here is how to measure market information beyond traditional data on stock variability. We propose that an important source of information is the semantic content of the news. By semantic content, we mean the underlying meanings of words in articles rather than the specific words that are referenced. For example, the semantic content of the words *rise* and *increase* is very similar, whereas the reference to the specific words is often irrelevant. Here, we present a method for analyzing the underlying semantic content of stock-related media text by applying a computational method called semantic spaces, where the semantic representation of words can be computationally generated from information of their co-occurrence in large text corpora. The resulting semantic representation places a given word in the text as a point in a high-dimensional semantic space, where the meaning of a word is given by its distance from other words in this space.

The purpose of this paper is to use the semantic representation to analyze the information flow related to stock market volatility. We make two proposals: first, that the semantic content in media can be used to develop an automatic method for measuring and tracking the effect of company information in media on the company's stock price volatility. Second, that semantic information in media may be both a passive documentation of past information flow and an active source of new information influencing future stock volatility.

We compare our semantic method to a number of standard models of stock volatility, which rely on historical return data. Among the most commonly used models are those belonging to the General Autoregressive Conditional Heteroskedastic (GARCH) class of models (see Engle, 1982; Bollerslev, 1986), which aim to capture the volatility persistence or the

clustering pattern in volatility. A majority of previous research indicates a relatively better prediction power for different GARCH specifications compared to other available models (e.g., Akgiray, 1989; West and Cho, 1995; Pagan and Schwert, 1990; Franses and van Dijk, 1996; Brailsford and Faff, 1996). In order to assess the prediction ability of our semantic approach, we compare our method to two GARCH-related specifications (i.e., a simple GARCH and an Exponential GARCH [EGARCH]) and two simple predictions (i.e., the random walk in volatility and the moving average of past volatilities). We use a number of evaluation strategies to assess the prediction power of our model relative to the alternative volatility models.

Our findings strongly support the strength of our suggested volatility forecast model compared to the conventional volatility prediction methods, which rely solely on the historical return data to forecast volatility. Our results also indicate that the media both reflects previous events in the stock market and influences volatility in the future.

This paper contributes to the literature by presenting an automatic method which quantifies the information flow in media in order to improve prediction of future volatility. We also study whether the information flow in media acts as a passive summary of previous volatility or actively influences future volatility. We can perform this analysis since the predictions from the semantic method are made on content in media and not merely on historical volatilities.

The rest of the study is organized as follows: section 2 presents a review of volatility forecast models that are based on nonsemantic data, our proposed method of predicting volatility based on semantic content of information in media text, and our evaluation methods for comparing these models. Section 3 contains the empirical evaluations of the different prediction

methods. Section 4 studies the time dynamics of the predictions or whether the models are predictive of past or future data. Section 5 concludes the paper.

2. Semantic and Nonsemantic Models to Predict Stock Price Volatility

2.1. How to Quantify the Meanings of Words

The semantic content of language conveys a wealth of information that typically is immediately understood by people due to its meaningful nature. At the same time, this information is often ignored in scientific studies due to lack of methods to quantify the semantic content. However, more recently, methods that allow quantification of meanings of words have been emerging. These methods utilize the empirical fact that text tends to keep to a certain semantic theme so that words within a certain context (i.e., sentence, paragraph, or document) are more likely to have a more similar meaning than those within other contexts. To be able to quantify this semantic content, it is necessary to have access to huge collections of text data, typically in the order of 100 MB or larger, where appropriate statistical methods are required for identifying the semantic representation.

Semantic spaces were early on proposed as a theory or model for how children acquire an understanding of the meanings of words (Landauer et al., 1998). Semantic spaces have been used in a number of fields—assessing the quality of essays (Miller, 2003); measuring context coherence (Foltz et al., 1998); measuring values of social groups (Gustafsson and Sikström, 2011); studying how object relations of mother, father, and self are influenced by long-term

psychotherapy (Arvidsson et al., 2011); studying semantic linguistic maturity in children and teenagers (Hansson et al., 2011); disambiguating different meanings of *holy* in blogs (Willander and Sikström, 2011); etc. Here, we show how semantic space can be applied for studying and predicting stock price volatilities.

Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) is probably the most well-known method for quantifying semantic representations; however, several other methods exist that produce similar representations—for example, probabilistic Latent Semantic Indexing (Hofmann, 1999), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), random indexing (Sahlgren, 2007), etc. Here, we focus on LSA as the original approach to create semantic representations because it provides a representation with sufficient good quality for our purpose.

LSA takes a text corpus as input and from this data builds a word-by-context frequency table. The rows in this table represent the words in the corpus, while the columns represent the contexts (i.e., document or paragraphs), and each cell in the table includes information on the number of occurrences for each word in a given context. A semantic representation of each word can be generated by a data compression algorithm called Singular Value Decomposition (SVD), which compresses the information on the large number of contexts (columns) into a smaller number of semantic dimensions.

2.2. Predicting Stock Price Volatility from Semantic Spaces

Changes in stock price are believed to be influenced by information flows through media. Here, we propose a method to predict stock price volatility based on the semantic content of news articles about companies. This semantic method consists of several steps: building a semantic representation of words in a language, collecting contextual news data related to the selected

companies, quantifying the contextual information in the semantic representation, and using machine learning methods to predict stock prices based on the contextual semantic information. For an illustrative example, see appendix A.

i. Collecting News Data

News media is perhaps the best source of daily information about stocks. This source of information also has the advantage of being highly accessible and having a specific time mark associated, namely the date of publication. The space described below is based on 100,000 news articles published in 100 different Swedish newspapers and magazines during the period 2000–2009. These data are made accessible through the courtesy of AffärsData, a company that provides access to Swedish news media.

ii. Creating a Semantic Representation

A semantic space is created by using the empirical fact that words that co-occur in the same context tend to have more comparable meanings than words in a different context. The space is created by LSA. A word-by-context frequency matrix, $\boldsymbol{x}$, is created, where each cell includes information on the number of occurrences a particular word has in a particular context (see Table A1 in appendix A). The context size is defined as 15 words preceding or following each target word. Approximately 100 stopwords, and extremely high-frequency words lacking a specific semantic content (*in*, *and*, etc.), are removed. The remaining 15,000 most frequent words are included in the rows of $\boldsymbol{x}$. The columns of $\boldsymbol{x}$ represent 20,000 contexts. The cell content is normalized by taking the log frequency plus 1. The information in this huge matrix is compressed while maintaining as much information as possible using singular value

decomposition (SVD). The SVD computes an approximation of $\boldsymbol{x}$ by factorizing matrix $\boldsymbol{x}$ in three components:

$$\boldsymbol{x} = \boldsymbol{u} \times \boldsymbol{s} \times \boldsymbol{v}', \quad (1)$$

where $\boldsymbol{x}$ is the $m \times n$ word-by-context frequency matrix. The element (i,j) gives the occurrence of word i in document j. Consequently, a row in matrix $\boldsymbol{x}$ will be a vector corresponding to a specific word, giving its relative occurrences in a different context, and a column of this matrix is a vector corresponding to a context, giving its relation to different words.

The matrix $\boldsymbol{u}$ is an $m \times r$ orthogonal matrix containing the eigenvectors of the matrix product $\boldsymbol{x}\boldsymbol{x}'$, where r is the rank of the matrix $\boldsymbol{x}$. The matrix $\boldsymbol{v}'$ is the transpose of $\boldsymbol{v}$, where $\boldsymbol{v}$ is an $n \times r$ orthogonal matrix containing the eigenvectors of the product matrix $\boldsymbol{x}'\boldsymbol{x}$. It should be noted that the product matrix $\boldsymbol{x}\boldsymbol{x}'$ gives the correlations among the word vectors over all the contexts, while the product matrix $\boldsymbol{x}'\boldsymbol{x}$ gives the correlations among the context vectors over all the words. The matrix $\boldsymbol{s}$ is an $r \times r$ diagonal matrix, where $s_{11} \geq s_{22} \ldots \geq s_{rr}$ and $s_{ij} = 0$ for all $i \neq j$. The diagonal elements of $\boldsymbol{s}$ are the square roots of the eigenvalues of the product matrix $\boldsymbol{x}'\boldsymbol{x}$.

The matrixes $\boldsymbol{u}$, $\boldsymbol{s}$, and $\boldsymbol{v}$ can be calculated from $\boldsymbol{x}$ by applying the known algorithm of SVD:

$$[\boldsymbol{u}\ \boldsymbol{s}\ \boldsymbol{v}] = \mathrm{SVD}(\boldsymbol{x}). \quad (2)$$

We call the resulting matrix, $\boldsymbol{u}$, the semantic space. The columns of $\boldsymbol{u}$ represent the dimensions in the space, and the rows represent the words. The columns of $\boldsymbol{u}$ which correspond to the higher eigenvalues (larger s_{ii} values) are supposed to be more informative. We select the 100

dimensions associated with the highest eigenvalues.[1] This reduction of the dimension is similar to the use of the principal component analysis where the first dimension corresponds to the factor that accounts for the highest variance, the second dimension is the second most important factor, and so on. The reduction of dimension, in addition to its computational efficiency, reduces the noise components from the original matrix (considered to be a better matrix). Each word is then normalized to a length of 1. This is done by calculating the length of the row vector representing each word and dividing the dimension (column) values of that word (row) with this length. The result is that each word becomes associated with an array of values representing semantic information of this word (for an illustrative example, see Table A2 in appendix A). Words with similar meanings, or synonyms, tend to have similar representation. The space is created by the public domain software Infomap (http://infomap-nlp.sourceforge.net/).

iii. Quantifying the Contextual News Data Related to Stocks

The semantic representations in matrix $\boldsymbol{u}$, described above, can now be used to summarize articles related to different companies. Contextual information related to a specific company in a specific article is generated by extracting 15 words preceding and 15 words following the name of the company in that article. We summarize the semantic information by summing the semantic representations for these words and then normalizing this vector to the length of 1 (see

[1] The number of semantic dimensions in the space can be chosen freely; however, a recommended procedure is to choose the number of dimensions that provides the highest quality of the space, which typically is in the order of a few hundred. Fewer dimensions tend not to carry sufficient information, whereas too many dimensions do not generalize well. The quality, or how well the semantic representation matches human associations, can be measured by, for example, a synonyms test, where semantic spaces have been found to pass TOEFL, a test used for entrance into American colleges (Landauer, Foltz, and Laham, 1998).

Table A3 for an illustrative example). Words that are missing in the semantic space are ignored. We denote the resulting matrix by $\boldsymbol{u}^*$.

iv. Predicting Stock Price Volatility Based on Contextual Semantic Information

In this step, we match the semantic content related to each company in a specific article to the company's volatility a period after the article is published. More specifically, the volatilities of the selected companies are estimated for each week as the standard deviation of the daily returns on the closing prices. The contextual information of each stock is matched with the stock's volatility with a delay of one week relative to the publication date of the news context. These data sets are "trained" to find the best-fitting weights for the 100 dimensional contextual semantic representations of the volatility measure. This training is made by multiple linear regression, where we find the coefficient, $\boldsymbol{C}$, that best describes the linear relation between the semantic space ($\boldsymbol{u}^*$) and the empirical values of volatility ($\boldsymbol{V}$):

$$\boldsymbol{V} = \boldsymbol{u}^* \boldsymbol{C} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\boldsymbol{V}$ is a $T \times 1$ vector of the estimated volatility for a specific company, $\boldsymbol{u}^*$ is a $T \times K$ matrix representing the semantic space, $\boldsymbol{C}$ is a $K \times 1$ coefficient vector, K is the dimension of the semantic space (100), and T is the number of observations on volatility. The ordinary least squares estimation of this regression model gives

$$\hat{\boldsymbol{C}} = \left(\boldsymbol{u}^{*\prime} \boldsymbol{u}^* \right)^{-1} \boldsymbol{u}^{*\prime} \boldsymbol{V} . \tag{4}$$

The predicted volatility for period t can then be calculated by the following formula:

$$\hat{V}_t = \boldsymbol{u}^*_{t-1} \hat{\boldsymbol{C}} , \tag{5}$$

where $\boldsymbol{u}_{t-1}^{*}$ is K, the dimensional row vector of the semantic space based on the articles published at time $t - 1$. This training is made separately for each stock. Furthermore, training is based on articles published prior to a certain week. The weights generated on this training are then used to predict the stock price associated with articles published the week following the time period of the articles used in the training set. Thus, there is no overlap between the training set (that is always published prior to the test set) and the test set. This procedure is repeated for all the weeks in the data set. To avoid overfitting of the prediction, the number of dimensions in the semantic space that were used for prediction are limited to one-third of the number of volatility data points. For example, if the training set included 30 data points of volatility, then only the first 10 dimensions in the semantic spaces are used.

2.3. Models Based on Historical Volatilities

In addition to our suggested prediction model based on the semantic content of information in media text, we use several other commonly used models for volatility prediction. According to these models, historical volatilities can be used to predict future volatility. Our first model is a simple moving average of all the past volatilities[2] (e.g., Brooks and Persand, 2003):

$$E_t(V_{t+1}) = \frac{1}{t}\sum_{j=0}^{t-1} V_{t-j}\,, \qquad t = 1, \ldots, T, \tag{6}$$

where V_t is the standard deviation of the log returns on week t and $E_t[.]$ is a forecast formed at time t.

[2] We use several alternative lengths to compute the moving average. Since our inferences are robust to the choices of the length, for the sake of space, we only report the result when we use all the past observations to compute the moving average.

The other models, described below, follow the idea that large price fluctuations usually trigger large price movements over subsequent dates. A basic model is the random walk in volatility, according to which the best prediction of the future volatility is the current volatility (e.g., Pagan and Schwert, 1990):

$$E_t\left(V_{t+1}\right) = V_t . \tag{7}$$

This is equivalent to a unit root in volatility implying that a shock in volatility has a permanent impact on the volatility process.

We also select two specifications from the GARCH family of the volatility models, a simple GARCH(1,1) model (i.e., a GARCH with one moving average and one autoregressive component), which is by far the most popular specification for modeling financial time series, and an EGARCH(1,1), which is an exponential form of the GARCH(1,1) model. A GARCH(1,1) is defined as

$$r_t = \mu + \eta_t , \qquad \eta_{t+1}|\Omega_t \sim N\left(0, h_{t+1}^2\right), \tag{8}$$

$$h_{t+1}^2 = w + a\eta_t^2 + bh_t^2 ,$$

while the variance equation in an EGARCH(1,1) is given by

$$\log h_{t+1}^2 = w + a\left(\frac{|\eta_t|}{h_t} - \sqrt{\frac{2}{\pi}}\right) + c\left(\frac{\eta_t}{h_t}\right) + b\log h_t^2 . \tag{9}$$

The advantage of the EGARCH model to the GARCH model is that it allows for the so-called leverage effect, that negative shocks may increase volatility more than positive shocks of the same magnitude.

The forecast for volatility in both models is

$$E_t\left(V_{t+1}\right) = h_{t+1} . \tag{10}$$

2.4. Evaluation Methods

We use a number of measures to evaluate the predictive power of our suggested approach relative to the rival models. In all these measures, we compare the volatility prediction of a specific model with the realized volatility. Realized volatility for week t is defined as the standard deviation of the daily return in that week.

In addition to analyzing the correlations between predicted and realized volatility, we run the following regression of the realized volatility on the predicted volatility (e.g., Andersen and Bollerslev, 1998; Hansen, 2001):

$$V_{t+1} = \alpha + \beta E_t(V_{t+1}) + \varepsilon_t . \tag{11}$$

If the predicted volatility has some information about the future realized volatility, then the parameter β should be significantly different from 0. Furthermore, for an unbiased prediction, we expect the parameter α to be 0 and the parameter β to be equal to 1. Therefore, we test for three hypotheses in this regression: $\beta = 0$, $\alpha = 0$, and $\beta = 1$.

We use two loss functions to compare the ability of the different approaches to forecast the realized volatility, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), defined as

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\sigma_{t+1} - E_t(\sigma_{t+1}))^2} , \tag{12}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|\sigma_{t+1} - E_t(\sigma_{t+1})\right| .$$

RMSE is a quadratic loss function and gives a larger weight to large prediction errors compared to the MAE measure and is therefore proper when large errors are more serious than small errors (see Brooks and Persand, 2003).

The loss functions above do not provide any statistical inferences. We therefore use the test suggested by Diebold and Mariano (1995), the DM test, to compare the prediction accuracy of two competing models:

$$DM = \frac{E(d_t)}{\sqrt{\text{var}(d_t)}} \sim N(0,1), \tag{13}$$

$$d_t = e_{A,t}^2 - e_{B,t}^2,$$

where $e_{A,t}$ and $e_{B,t}$ are prediction errors of two rival models A and B, respectively, and E(d_t) and var(d_t) are the mean and the variance of the time series of d_t, respectively. The estimated test value has a standard normal distribution.

3. Empirical Evaluations of Volatility Predictions

We selected five stocks with the highest market values in the Swedish stock market. The stock Ericsson is omitted because of its ambiguous meaning of either a family name or a company name. The selected stocks are *Nordea*, *SCA*, *SHB*, *TeliaSonera*, and *Volvo*. Initial data cover the period from June 2000 until June 2009. Weekly volatilities are estimated as the standard deviation of the daily observations within each week. These volatilities are used to obtain our suggested semantic forecast approach. GARCH and EGARCH are estimated on the weekly returns, while a sample of 100 observations is used as the first estimation window. Therefore, the out-of-sample period starts in May 2002 and ends in June 2009. We then extend the estimation window by one week each time to get a new volatility prediction.

We start the analysis by looking at the number of articles that are used in our volatility prediction. Table 1 reports the mean and the standard deviation of the number of articles within each week for all the five firms. The statistics are reported for the entire period as well as for two subperiods, 2002–2005 and 2006–2009, respectively. Volvo has the largest number of articles, while the smallest number belongs to SHB. Telia has quite a small number of articles in the first period. The positive and highly significant time trend in the number of articles for all the companies shows that the number of articles has increased over time.

[Please insert Table 1 about here.]

3.1 Correlation Analysis

Table 2 gives an overview of the relationship between the weekly volatility of different firms. Panel A of the table gives the correlations between the realized volatilities, while the correlations between the predicted volatilities based on contextual semantic information (article-based predicted volatility) are reported under panel B. All the values are statistically significant from 0 except the correlation of predicted volatilities for Telia and SHB. In most of the cases, the correlations between the predicted volatilities are lower than those between the realized volatilities, particularly for Telia. This shows that the predicted volatilities fail to entirely capture the common trend present in the realized volatilities.

[Please insert Table 2 about here.]

To assess the prediction power of the different volatility models, we first look at the correlations between the forecasted volatilities and the realized volatilities. Table 3 reports the average correlation over all the five firms, while the correlation matrixes of the individual firms are given in Table B1 in appendix B. The results show that our suggested semantic forecast

approach has, in general, a considerably higher correlation with the realized volatility than the other prediction methods except the lagged volatility. The average correlation between different prediction methods is mostly around 0.40, while the average correlation between the two GARCH-related specifications is 0.83. We can almost draw the same conclusions by looking at the correlations at the firm level in Table B1. All the correlations are significantly different from 0 at the 5% level except in a few cases for the moving average approach.

[Please insert Table 3 about here.]

The relatively higher correlation between the lagged volatility and the realized volatility indicates that this method gives a better prediction of the direction of the future volatility compared to the other methods. This may be due to the presence of volatility clustering over time. However, to assess the forecast ability of the models, in addition to the prediction of the direction, we need to investigate the precision of the volatility forecasts. The next two sections apply a number of evaluation methods to evaluate the accuracy of the different models to forecast future volatilities.

3.2 Forecast Evaluation Using Regression Analysis

Table 4 gives a summary of the results of the regressions of the realized volatilities on the predicted volatilities. The detailed results at the firm level are reported in Table B2 (see appendix B). For a perfect fit, we expect an intercept equal to 1 and a slope equal to 0. Further, a positive slope shows that the realized volatility moves in the same direction as our predicted volatility, and the sign of the estimated intercept gives an indication of an average over/underestimation of the volatility.

[Please insert Table 4 about here.]

Given that the semantic method works well, we would expect that it would have a relative advantage over the nonsemantic models during time periods where there is a large number of articles compared to time periods where there is a low number of articles. To verify this, in addition to the entire sample from 2002 to 2009, we report the results for two subsets: the subsample 2006–2009, which according to Table 1 contains more articles than the earlier period, and the sample of weeks with a larger-than-median number of articles.

The semantic forecast approach in general outperforms other volatility models. In most cases, this model shows an intercept equal to 0 (particularly when using the second part of the sample or the weeks with a relatively large number of articles) and gives a significantly positive slope coefficient. In addition, in three out of five cases, the beta is not significantly different from 1. However, based on the regression results, we cannot see a considerable improvement in the second part of the sample or when we use the weeks with a larger number of articles. The GARCH and EGARCH models both perform poorly; although all the estimated slopes (beta coefficient) are significantly different from 0 for the regression over the entire period, the estimated slopes are significantly lower than 1. In addition, three out of five intercept terms are significant. This scenario is roughly the same even when we use the lagged volatility as a predictor of the future volatility. Moving average is probably the worst model, both over the entire sample (only three out of five slope coefficients are significant) and for the second part of the sample (all intercepts are significant, and the slope coefficients are significantly different from 1).

3.3 Forecast Evaluation Using Loss Functions

In this section, we evaluate the forecast ability of the alternative volatility models by using two loss functions, MAE and RMSE. Table 5 reports the average results of the loss functions, and Table 6 gives a summary of the DM test, which contrasts the semantic predictions with other models (the result of the individual firms is illustrated in Tables B3 and B4 in the appendix). The semantic forecast clearly outperforms the GARCH and the EGARCH predictions for all the firms and samples; the average values of both loss functions for the semantic forecast is about half of the values for these models. Moreover, based on the DM test, the semantic method has significantly lower prediction errors than the GARCH-related models for all the five firms (Table 6). It is also significantly better than the moving average for almost all of the cases. The result is slightly weaker when we compare our model with the lagged realized volatility but still supports the relatively better performance of our suggested model, particularly regarding the number of significant DM tests; the semantic model significantly outperforms the lagged volatility in two out of five cases, while both cases with positive DM tests are insignificant.

[Please insert Table 5 about here.]

[Please insert Table 6 about here.]

We expect that the semantic model should have a relative advantage when there is a lot of semantic material available. Data support this conjecture. Using the MAE and RMSE measures, the MA and lagged volatility predictions perform relatively worse in the two subsamples where there is a large number of articles compared to the whole data set, whereas the semantic model shows approximately the same performance over the three data sets. Similarly, the DM test shows relatively better performance (i.e., more negative DM values) for the semantic compared

to lagged volatility and MA measures in these two subsamples. Since the last part of the sample coincides with the start of the financial crisis in the middle of 2008, it is possible that the predictions given by all the models become less accurate for the subsamples containing this period. This might explain why the semantic approach has almost the same performance over the three data sets, while its relative performance is better than the other models for the two subsamples with a high number of articles.

4. Does Information Flow Influence Future Volatility or Summarize Past Volatility?

The information flow in media may serve as an input that influences stock volatility in the future; however, it may also summarize information relevant for the stock volatility in the past. The question is fundamental because it is connected to the causality between volatility and information in media (i.e., whether information influences stock volatility or whether it is the other way around). In our view, this is not an all-or-none question. Information flow in the media serves several purposes, including summarizing previous information of volatility as well as providing new information that influences future volatility.

Here, we study the strength of the correlation between predicted and realized volatility changes depending on the time lags between these measures. Our predictions regarding the information flow are as follows: (a) there is a positive correlation to past realized volatilities, indicating that media information covers information related to past volatilities; (b) there is a positive correlation to future realized volatilities, indicating that current information flow influences future volatilities; and (c) for a given time lag, the correlation is stronger to realized

volatility in the past compared to the future as new and unexpected information is continuously introduced over time.

To investigate these hypotheses, we calculate the correlation between the estimated-based predicted volatility at zero lag and the realized volatility at different time leads and lags between these two measures. The average correlations over five firms are plotted in Figure 1.

The results show that the maximum correlation is reached at lag zero (i.e., when both the predicted and the realized volatilities belong to the same week). As hypothesized, the predicted volatilities correlate higher with the realized volatilities in the past compared to the future. The correlation almost monotonically increases before lag zero and decreases sharply after this lag. This may indicate that the text in media to a large extent reflects information about the very near future (lag zero) and most recent history. The figure also plots the correlations between the GARCH prediction and the realized volatility. In comparison with the semantic prediction, the GARCH prediction has a lower average correlation from approximately 20 weeks in the past to 8 weeks in the future. Thus, the semantic content in the media within this time span seems to carry important information regarding stock volatility, where the correlation with the past may indicate that media contains a passive summary of previous historical events affecting volatility, whereas the predictability toward the future may reflect the possibility that media influences future volatility.

[Please insert Figure 1 about here.]

The finding above is strongly supported by the results of the regression of the predicted volatilities by the semantic and GARCH approaches on the realized volatilities at different lags and leads. The regressions are estimated separately for each firm. Figure 2 illustrates the average

intercept (alpha) and the average slope (beta) estimated over all the firms as well as the difference between the parameters given by the two alternative prediction methods and the corresponding 95% confidence interval.

The result of the slope coefficient (beta), analogous to that from the correlation analysis, gives a clear indication of the better performance of the semantic approach compared to the GARCH model; almost all the differences are large and statistically significant. Moreover, the slope diminishes faster in the regression on the future realized volatilities than for those on the past values.[3] For both models, the estimated intercept converges to 0 at lag zero (i.e., when the prediction for time *t* is regressed on the realized volatility at time *t*). As expected, the intercept increases more for the leads than for the lags, indicating that both models incorporate more information from the past than about the future. The difference between two intercept terms is not statistically significant in any case.

[Please insert Figure 2 about here.]

5. Conclusion

The prices in the financial market reflect the inflow of information via media. However, this information is difficult to quantify and statistically measure. We develop an automatic method for measuring and tracking the effect of company information in media on the company's stock price volatility. The suggested method for volatility forecasting is based on the semantic content of information in media text. The method utilizes the co-occurrence of words in large text

[3] We perform a similar comparison between the semantic method and the EGARCH model. The results are almost the same as those from the comparison with GARCH. For the sake of space, the results are not reported but are available upon request.

corpora to quantify the semantic content of words. Based on this representation, we quantify the semantic content of news articles describing a stock and use this as a predictor of the associated volatility.

As alternative models for comparison, we use the most commonly used specifications in the literature: the random walk model and the moving average of past volatilities as well as the GARCH and EGARCH models. A number of evaluation strategies are used to assess the relative prediction power of our suggested model relative to the alternative volatility models. The analysis is performed on five listed Swedish firms over a period from June 2000 until June 2009.

A simple correlation analysis shows that our suggested semantic forecast approach has a higher correlation with the realized volatility than the alternative models, except the lag volatility. A regression analysis of the realized volatility on the semantic prediction shows a strong ability to give an unbiased forecast of the future volatility and in general outperforms all other models. The strong forecast ability of the semantic forecast is supported by findings from two alternative loss functions, MAE and RMSE, as well as the DM test. The semantic forecast outperforms the GARCH and the EGARCH specifications for all the firms and samples. It is also significantly better than the moving average for almost all of the cases. Furthermore, the semantic model seems to provide a relatively better prediction compared to the nonsemantic models during time periods when there is a lot of semantic data available. This provides additional validity to the finding that the semantic model is actually basing the prediction on semantic information rather than some other unknown artifact. It also shows that the predictability of the semantic models depends on the availability of semantic data.

An analysis of the relation between the semantic forecasted volatility and the realized volatility at different leads and lags gives a maximum correlation at lag zero, indicating that the text in media reflects mostly information about the very near future. However, this relation is stronger than that given by alternative models both in the past and in the future, suggesting that the media both reflects previous events on the stock market and influences volatility in the future.

All in all, our findings show that quantifying the information flow in media can considerably enhance the prediction of future volatility in comparison with the volatility models relying solely on past return data. This might indicate that the market participants' cognitive biases in interpreting the available information lead to an inefficient pricing of financial assets in the sense that the market prices do not incorporate all available information.

References

Akgiray, V., 1989, Conditional Heteroskedasticity in Time Series of Stock Returns: Evidence and Forecasts, *Journal of Business* 62, 55–80.

Andersen, T. G., and T. Bollerslev, 1998, Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts, *International Economic Review* 39, 885–905.

Arvidsson, D., A. Werbart, and S. Sikström, 2011, Changes in Self- and Object Representations following Psychotherapy Measured by a Theory-Free, Computational, Semantic Space Method, *Psychotherapy Research*, Forthcoming.

Blei, D. M., A. Y. Ng, and M. Jordan, 2003, Latent Dirichlet Allocation, *Journal of Machine Learning Research* 3, 993–1022.

Bollerslev, T., 1986, Generalized Autoregressive Conditional Heteroskedasticity, *Journal of Econometrics* 31, 307–328.

Brailsford, T. J., and R. W. Faff, 1996, An Evaluation of Volatility Forecasting Techniques, *Journal of Banking and Finance* 20, 419–438.

Brooks, C., and G. Persand, 2003, Volatility Forecasting for Risk Management, *Journal of Forecasting* 22, 1–22.

Diebold, F. X., and R. Mariano, 1995, Comparing Predictive Accuracy, *Journal of Business and Economic Statistics* 13, 253–263.

Engle, R. F., 1982, Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation, *Econometrica* 50, 987–1007.

Foltz, P. W., W. Kintsch, and T. K. Landauer, 1998, The Measurement of Textual Coherence with Latent Semantic Analysis, *Discourse Processes* 25, 285–307.

Franses, P. H., and D. van Dijk, 1996, Forecasting Stock Market Volatility Using Non-Linear GARCH Models, *Journal of Forecasting* 15, 229–235.

Gärling, T., E. Kirchler, A. Lewis, and F. van Raaij, 2009, Psychology, Financial Decision Making, and Financial Crises, *Psychological Science* 10, 1–47.

Glaser, M., M. Nöth, and M. Weber, 2004, Behavioral Finance. In D. J. Koehler, and N. Harvey (Eds.), *Blackwell Handbook of Judgment & Decision Making* (pp. 527–546). Oxford, UK: Blackwell.

Gustafsson, M., and S. Siktröm, 2011, Redistributing Resource through Language to Improve Fitness, Working Paper.

Hannson, K., R. Bååth, S. Löhndorf, B. Sahlen, and S. Sikström, 2011, Semantic Language Maturity (SELMA), Working Paper.

Hansen, C. S., 2001, The Relation between Implied and Realised Volatility in the Danish Option and Equity Markets, *Accounting and Finance* 41, 197–228.

Hofmann, T., 1999, Probabilistic Latent Semantic Indexing, Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval, SIGIR-99.

Kahneman, D., and A. Tversky, 1979, Prospect Theory: An Analysis of Decision under Risk, *Econometrica* 47, 263–291.

Landauer, T., P. W. Foltz, and D. Laham, 1998, An Introduction to Latent Semantic Analysis, *Discourse Processes* 25, 259–284.

Landauer, T. K., and S. T. Dumais, 1997, A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge, *Psychological Review* 104, 211–240.

Lichtenstein, S., B. Fischhoff, and L. D. Phillips, 1982, Calibration of Probabilities: The State of the Art to 1980. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases* (pp. 306–334). New York: Cambridge University Press.

Miller, T., 2003, Essay Assessment with Latent Semantic Analysis, *Journal of Educational Computing Research* 29, 495–512.

Pagan, A. R., and G. W. Schwert, 1990, Alternative Models for Conditional Stock Volatilities, *Journal of Econometrics* 45, 267–290.

Sahlgren, M., 2007, An Introduction to Random Indexing, Working Paper, Stockholm University, Stockholm.

Smith, L., and P. Soerensen, 2000, Pathological Outcomes of Observational Learning, *Econometrica* 68, 371–398.

Svenson, O., 1981, Are We All Less Risky and More Skillful than Our Fellow Drivers? *Acta Psychologica* 47, 143–148.

Weinstein, N. D., 1980, Unrealistic Optimism about Future Life Events, *Journal of Personality and Social Psychology* 39, 806–820.

West, K. D., and D. Cho, 1995, The Predictive Ability of Several Models of Exchange Rate Volatility, *Journal of Econometrics* 69, 367–391.

Willander, E., and S. Sikström, 2011, Categorizing Cultural Meanings of the Word *Sacred*, *Innovating Methods in the Study of Religion*, Ed. Linda Woodhead, in Press.

Table 1. Statistics on the number of articles within a week

		Nordea	*SCA*	*SHB*	*Telia*	*Volvo*
2002–2009	*Mean*	58.22	22.41	9.23	26.75	86.56
	std. dev.	68.40	34.08	12.36	42.61	78.71
2002–2005	*Mean*	24.17	9.96	4.23	4.98	47.89
	std. dev.	16.33	10.17	8.63	6.83	29.86
2006–2009	*Mean*	94.03	35.45	14.48	49.58	127.11
	std. dev.	16.33	10.17	8.63	6.83	29.86
Trend in no. of articles	*Coef.*	0.39	0.07	0.15	0.25	0.42
	t-value	15.17	13.56	10.29	15.68	13.78

Note: The table shows the means and standard deviations of the number of articles within each week for the period April 2002 to June 2009 as well as for two subperiods. The last two rows show the estimated time trend expressed as coefficient values and the related *t*-values.

Table 2. Correlation of realized volatility/predicted volatility between different firms

Panel A. Realized volatility

	Nordea	*SCA*	*SHB*	*Telia*	*Volvo*
Nordea	1.00				
SCA	0.63	1.00			
SHB	0.80	0.69	1.00		
Telia	0.49	0.39	0.42	1.00	
Volvo	0.65	0.65	0.72	0.41	1.00

Panel B. Semantic volatility predictions

	Nordea	*SCA*	*SHB*	*Telia*	*Volvo*
Nordea	1.00				
SCA	0.75	1.00			
SHB	0.24	0.48	1.00		
Telia	0.37	0.29	0.12	1.00	
Volvo	0.85	0.85	0.49	0.28	1.00

Note: Panel A (panel B) of the table shows the correlations of the realized (predicted) volatility between different firms. The predicted volatility is estimated by the semantic content of the text in media, while the realized volatility is the standard deviation of the returns within each week.

Table 3. Correlations among different volatility measures

	Realized vol.	*Lagged vol.*	*Semantic*	*GARCH*	*EGARCH*	*MA*
Realized vol.	1.00					
Lagged vol.	0.50	1.00				
Semantic	0.41	0.42	1.00			
GARCH	0.33	0.40	0.45	1.00		
EGARCH	0.30	0.34	0.38	0.83	1.00	
MA	0.13	0.15	0.47	0.49	0.47	1.00

Note: The table illustrates the correlations among different volatility measures (i.e., the realized volatility at time *t*, the one-period lagged realized volatility, the predicted volatility using the semantic content of the text in media, predicted volatilities given by GARCH and EGARCH models, and a moving average of the past volatilities [MA]). The values are the average correlations over the individual firms.

Table 4. Regression of the realized weekly volatility on the predicted volatilities

.		*2002 – 2009*		*2006 – 2009*		*Large no. of artic.*	
		α	β	α	β	α	β
Semantic approach	Average	−0.001	0.954	−0.003	1.214	−0.004	1.213
	No. of sig. pos.	1	5	0	5	0	5
	No. of sig. neg.	1	0	1	0	1	0
	No. of sig. diff. from 1		2		2		2
GARCH	Average	0.001	0.517	−0.006	0.829	−0.002	0.657
	No. of sig. pos.	2	5	1	4	1	4
	No. of sig. neg.	1	0	2	0	2	0
	No. of sig. diff. from 1		4		5		4
EGARCH	Average	0.002	0.471	0.001	0.575	0.001	0.564
	No. of sig. pos.	3	5	3	3	3	3
	No. of sig. neg.	1	0	1	0	2	0
	No. of sig. diff. from 1		4		4		4
Lagged volatility	Average	0.008	0.502	0.010	0.474	0.010	0.494
	No. of sig. pos.	5	5	5	5	5	5
	No. of sig. neg.	0	0	0	0	0	0
	No. of sig. diff. from 1		5		5		5
Moving average	Average	0.000	1.016	−0.082	6.684	−0.024	2.705
	No. of sig. pos.	1	3	1	4	1	3
	No. of sig. neg.	0	0	4	0	2	0
	No. of sig. diff. from 1		2		5		3

Note: A summary of the results of the regression of the realized volatility on the alternative volatility forecasts. The results are reported for three different samples (i.e., the entire sample from 2002 to 2009, the period from 2006 to 2009, and finally, the sample from weeks with a larger-than-median number of articles). We report the average value of the coefficients over the individual firms and the number of coefficients which are significantly different from 0 at the 5% level as well as the number of betas which are significantly different from 1 at the 5% level.

Table 5. Evaluation using loss functions

		2002–2009	2006–2009	Large no. of articles
MAE	Semantic	83.2	82.2	83.2
	GARCH	185.0	164.3	168.9
	EGARCH	174.4	156.6	160.7
	Lagged volatility	80.0	91.8	88.2
	MA	89.7	93.2	94.7
RMSE	Semantic	112.5	118.5	117.7
	GARCH	206.6	185.3	190.1
	EGARCH	197.2	180.2	182.8
	Lagged volatility	114.8	131.0	127.7
	MA	122.8	137.5	136.4

Note: Illustration of the evaluation results for the alternative volatility forecasts using two loss functions, mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The results are reported for three different samples (i.e., the entire sample from 2002 to 2009, the period from 2006 to 2009, and finally, the sample from weeks with a larger-than-median number of articles). The values are averages over the individual firms and are enlarged by a factor of 10,000 for illustrative purposes.

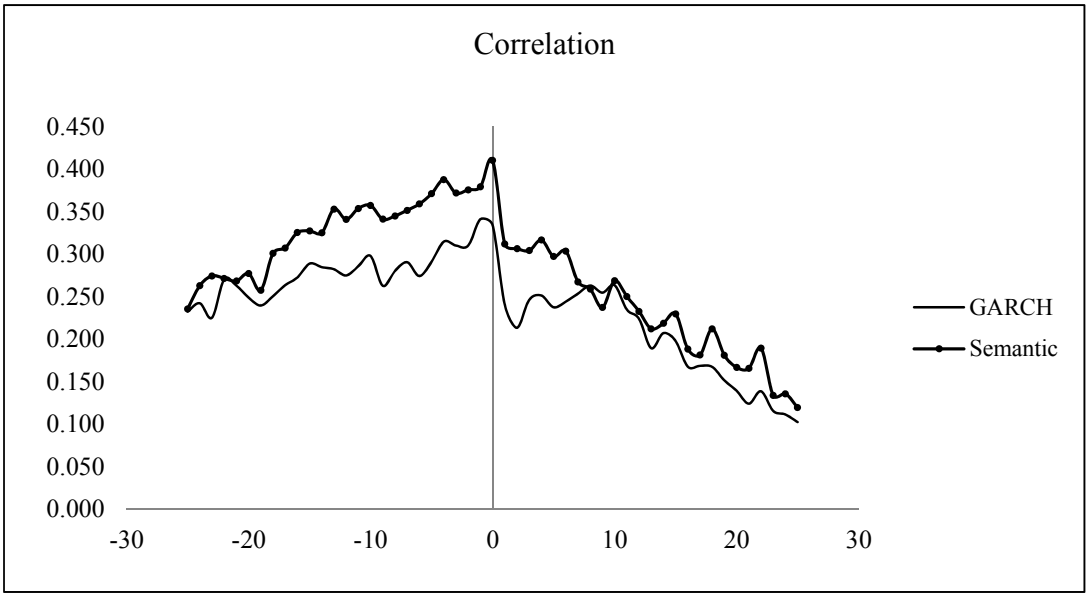
Table 6. Comparing the semantic model with other models using the DM test

.

		DM test			
		GARCH	*EGARCH* .	*Lagged volatility*	*Moving average*
2002–2009	*Average*	−12.51	−11.27	−0.57	−2.73
	No. of pos.	0	0	2	0
	No. of neg.	5	5	3	5
	No. of sig. pos.	0	0	0	0
	No. of sig. neg.	5	5	2	4
2006–2009	*Average*	−6.32	−5.99	−1.43	−3.52
	No. of pos.	0	0	1	0
	No. of neg.	5	5	4	5
	No. of sig. pos.	0	0	0	0
	No. of sig. neg.	5	5	2	5
Large no. of articles	*Average*	−7.145	−6.751	−1.302	−3.939
	No. of pos.	0	0	1	0
	No. of neg.	5	5	4	5
	No. of sig. pos.	0	0	0	0
	No. of sig. neg.	5	5	2	5

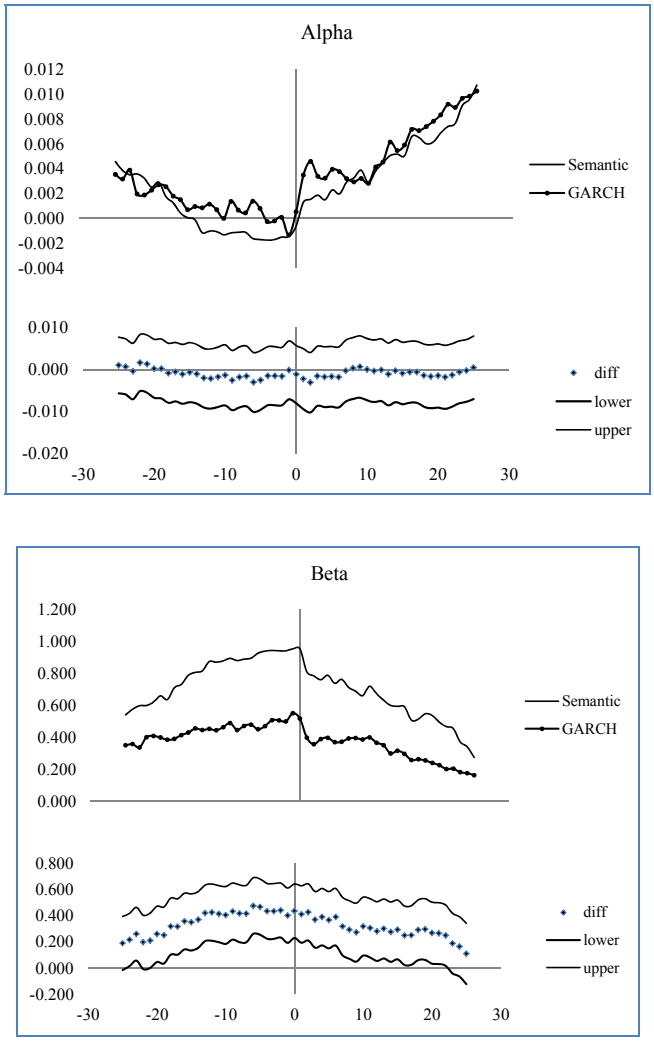
Note: The table illustrates the summary of the statistics of the DM test for relative performance of the semantic forecast approach against the alternative volatility forecasts. The results are reported for three different samples (i.e., the entire sample from 2002 to 2009, the period from 2006 to 2009, and finally, the sample from weeks with a larger-than-median number of articles). A negative value corresponds to a relatively better prediction of the semantic forecast approach. The estimated statistics have a standard normal distribution.

Figure 1. Correlations between predicted and realized volatility at different lags and leads



Note: The figure illustrates the correlations between the predicted volatilities and the realized volatilities at different lags and leads. We compare two different forecast approaches, the semantic approach and the GARCH model. We plot the average correlation over the five firms included in the analysis.

Figure 2. Regression of predicted volatility on realized volatility at different lags and leads



Note: The figure illustrates the intercept (alpha) and the slope (beta) estimated from the regression of predicted volatilities on the realized volatilities at different lags and leads. We compare two different forecast approaches, the semantic approach and the GARCH model. We also plot the difference between the parameters and the corresponding 95% confidence interval.

Appendix A. An example for generating semantic representation

Here, we use an example to illustrate how we generate a semantic representation and how this can be used to predict volatility. To generate a semantic representation, a huge corpus of text is required. Here, we use a corpus consisting of four sentences as an illustrative example:

1. It was a *calm* trading day, but the IBM stock did much better than expected in the extremely *slow market*.
2. Microsoft's CFO was *calm* despite the depressingly *slow market* conditions.
3. Chrysler shares had large trading volumes in a high *volatility bear market*.
4. The *volatility* of the Volvo followed the general trend of the *bear market*.

The first step is to construct a *word-by-context table*. We select all words occurring at least twice in the corpus (set in italics above). The words with lower frequencies are omitted due to insufficient statistics for constructing a semantic representation with good quality. Very high-frequency words lacking semantic content (stopwords such as *a* and *on*) are also omitted because they do not add any useful semantic information. These words are manually selected. We place the selected words in the rows and count the number of occurrences of these words in each sentence (context), represented in the columns (1–4). See Table A1.

Table A1. Word-by-context frequency table

Word/context	Art. 1	Art. 2	Art. 3	Art. 4
Slow	1	1	0	0
Calm	1	1	0	0
Volatility	0	0	1	1
Bear	0	0	1	1
Market	1	1	1	1

We normalize this table by adding 1 to each cell and then taking the logarithm of this value. Then we apply a data compression algorithm (SVD specified in section 2.2), which is similar to the principal component analysis. This semantic representation is presented in Table A2. Here, we have chosen to reduce the dimensions from 4 to 2, where the length of each row vector is normalized to 1.

Table A2. The semantic representation of words

Word	Dim. 1	Dim. 2
Slow	−0.57	−0.81
Calm	−0.57	−0.81
Volatility	−0.57	0.81
Bear	−0.57	0.81
Market	−1.00	0.00

We call the resulting representation a semantic space, where each word in the corpora is represented by an array of semantic features. Words that have similar meanings, or synonyms, tend to be located near each other in the space. Analogous to the principal component analysis, the first dimension contains the most important information (i.e., that factor accounts for the highest variance, the second dimension is the second most important factor, and so on).[4] In this paper, we use 100 semantic dimensions. Normally, the first dimension is related to the frequency of the words in the corpus.

These semantic representations can now be used to summarize articles related to different companies. These articles may come from another corpus rather than the one used to create the

4. The space includes semantic information; however, it does not discriminate between other nonsemantic information such as word classes, phonological representations, or lexical representations, where other methods are required for quantifying these aspects.

semantic representation, and size/number of articles can be much smaller. Assume that articles 1 and 2, including IBM's and Microsoft's, were written at a time with low volatility (denoted by regime 1 in Table A3) and articles 3 and 4, including Chrysler's and Volvo's, were written at a time of high volatility (regime 2). We summarize the articles by summing the semantic representations for the words included in the article (excluding the company names that were used for selecting the articles) and then normalizing this vector to the length of 1 (see Table A3). For example, in article 1, we sum up the semantic representations for the words *slow*, *calm*, and *market* and then normalize the values.

Table A3. The semantic representation of articles

	Dim. 1	Dim. 2	Regime
Art. 1	−0.57	−0.81	1
Art. 2	−0.57	−0.81	1
Art. 3	−0.57	0.81	2
Art. 4	−0.57	0.81	2

In the example illustrated in Table A3, it is easy to see that the second dimension differentiates well between the articles related to low (negative values) or high volatility (positive values).

With real data, the relation between semantic representation and volatility is more complex and typically involves the weighing of several dimensions. To identify this relation, we need a statistical algorithm to predict volatility. In this paper, we base the prediction on a multiple linear regression model in which the stock volatility is the dependent variable and the semantic dimensions are used as explanatory variables.

Appendix B. Results of the individual firms

Table B1. Correlations among different volatility measures

	Realized vol.	*Lagged vol.*	*Semantic*	*GARCH*	*EGARCH*	*MA*
		Nordea				
Realized vol.	1.00					
Lagged vol.	0.63	1.00				
Semantic	0.42	0.45	1.00			
GARCH	0.39	0.42	0.70	1.00		
EGARCH	0.41	0.45	0.57	0.83	1.00	
MA	0.10	0.12	0.72	0.70	0.52	1.00
		SCA				
Realized vol.	1.00					
Lagged vol.	0.42	1.00				
Semantic	0.42	0.42	1.00			
GARCH	0.32	0.40	0.32	1.00		
EGARCH	0.14	0.17	0.15	0.68	1.00	
MA	0.04	0.06	0.42	0.28	0.59	1.00
		SHB				
Realized vol.	1.00					
Lagged vol.	0.67	1.00				
Semantic	0.49	0.49	1.00			
GARCH	0.51	0.57	0.42	1.00		
EGARCH	0.50	0.57	0.42	0.92	1.00	
MA	0.15	0.14	0.01	0.35	0.33	1.00

Note: The table illustrates the correlations among different volatility measures (i.e., the realized volatility at time *t*, the one-period lagged realized volatility, the predicted volatility using the semantic content of the text in media, predicted volatilities given by GARCH and EGARCH models, and a moving average of the past volatilities [MA]).The correlation matrix is given separately for each firm.

Table B1. Correlations among different volatility measures (continued)

	Realized vol.	*Lagged vol.*	*Semantic*	*GARCH*	*EGARCH*	*MA*
		Telia				
Realized vol.	1.00					
Lagged vol.	0.31	1.00				
Semantic	0.28	0.27	1.00			
GARCH	0.24	0.33	0.47	1.00		
EGARCH	0.16	0.19	0.36	0.81	1.00	
MA	0.19	0.20	0.46	0.63	0.54	1.00
		Volvo				
Realized vol.	1.00					
Lagged vol.	0.47	1.00				
Semantic	0.45	0.48	1.00			
GARCH	0.21	0.28	0.36	1.00		
EGARCH	0.28	0.33	0.40	0.90	1.00	
MA	0.18	0.22	0.73	0.49	0.40	1.00

Table B2. Regression of the realized weekly volatility on the predicted volatilities

.		*2002 – 2009*		*2006 – 2009*		*Large No. of Artic.*	
		α	β	α	β	α	β
Semantic approach	*Nordea*	−0.005	1.136*	−0.011*	1.715^{*+}	−0.012*	1.656^{*+}
	SCA	−0.001	0.944*	0.002	0.924*	−0.001	1.016*
	SHB	0.002	0.890*	0.001	1.015*	−0.002	1.156*
	Telia	0.010*	0.336^{*+}	0.000	0.964*	0.003	0.780*
	Volvo	−0.009*	1.462^{*+}	−0.006	1.452^{*+}	−0.006	1.454^{*+}
GARCH	*Nordea*	0.000	0.553^{*+}	−0.033*	1.757^{*+}	−0.015	1.137*
	SCA	0.001	0.445^{*+}	0.006	0.364^{*+}	0.003	0.424^{*+}
	SHB	−0.013*	1.056*	−0.025*	1.535^{*+}	−0.019	1.322^{*+}
	Telia	0.008*	0.227^{*+}	0.015*	0.048^{+}	0.017	0.000^{+}
	Volvo	0.007*	0.304^{*+}	0.006	0.443^{*+}	0.006	0.400^{*+}
EGARCH	*Nordea*	−0.003	0.670^{*+}	−0.011	1.024*	−0.012	1.045*
	SCA	0.009*	0.156^{*+}	0.013	0.131^{+}	0.012	0.126^{+}
	SHB	−0.012*	1.039*	−0.021	1.386^{*+}	−0.019	1.320^{*+}
	Telia	0.012*	0.135^{*+}	0.017	−0.023^{+}	0.018	−0.048^{+}
	Volvo	0.006*	0.353^{*+}	0.009	0.358^{*+}	0.008	0.377^{*+}
Lag vol.	*Nordea*	0.007*	0.631^{*+}	0.008*	0.617^{*+}	0.008*	0.625^{*+}
	SCA	0.008*	0.418^{*+}	0.011*	0.350^{*+}	0.009*	0.413^{*+}
	SHB	0.006*	0.697^{*+}	0.007*	0.657^{*+}	0.007*	0.680^{*+}
	Telia	0.011*	0.295^{*+}	0.012*	0.254^{*+}	0.013*	0.247^{*+}
	Volvo	0.010*	0.469^{*+}	0.011*	0.489^{*+}	0.011*	0.506^{*+}
Moving average	*Nordea*	0.008	0.521	−0.148*	9.758^{*+}	−0.011	1.780*
	SCA	0.009	0.304	−0.082*	7.555^{*+}	0.007	0.636
	SHB	−0.006	1.566*	−0.124*	10.338^{*+}	−0.070*	6.320^{*+}
	Telia	0.009*	0.330^{*+}	0.063*	−2.573^{*+}	0.020*	−0.186^{+}
	Volvo	−0.022	2.360^{*+}	−0.119*	8.343^{*+}	−0.064*	4.976^{*+}

Note: The table illustrates the results of the regression of the realized volatility on the alternative volatility forecasts. The results are reported for three different samples (i.e., the entire sample from 2002 to 2009, the period from 2006 to 2009, and finally, the sample from weeks with a larger-than-median number of articles). The * is for values significantly different from 0 at the 5% level, and + is for values significantly different from 1 at the 5% level (only tested for the beta coefficient). The results are reported for each firm.

Table B3. Evaluation using loss functions

		MAE					***RMSE***				
		Semantic	*GARCH*	*EGARCH.*	*Lag vol.*	*Mov. aver.*	*Semantic*	*GARCH*	*EGARCH.*	*Lag vol.*	*Mov. aver.*
2002–2009	*Nordea*	0.011	0.019	0.017	0.009	0.011	0.014	0.021	0.020	0.013	0.015
	SCA	0.006	0.016	0.018	0.007	0.007	0.008	0.018	0.020	0.010	0.009
	SHB	0.009	0.015	0.015	0.008	0.009	0.012	0.017	0.016	0.011	0.014
	Telia	0.008	0.021	0.017	0.008	0.009	0.011	0.024	0.020	0.011	0.011
	Volvo	0.008	0.021	0.020	0.009	0.008	0.011	0.024	0.022	0.012	0.012
2006–2009	*Nordea*	0.009	0.015	0.016	0.010	0.011	0.014	0.017	0.019	0.015	0.017
	SCA	0.006	0.015	0.014	0.008	0.007	0.009	0.017	0.016	0.012	0.011
	SHB	0.010	0.015	0.015	0.009	0.011	0.014	0.017	0.017	0.013	0.017
	Telia	0.007	0.019	0.015	0.008	0.008	0.009	0.021	0.018	0.011	0.010
	Volvo	0.009	0.018	0.018	0.010	0.010	0.013	0.021	0.021	0.014	0.015
Large no. of articles	*Nordea*	0.010	0.016	0.016	0.010	0.012	0.015	0.018	0.019	0.015	0.017
	SCA	0.006	0.016	0.016	0.008	0.007	0.009	0.017	0.017	0.011	0.010
	SHB	0.010	0.015	0.015	0.009	0.011	0.014	0.017	0.016	0.013	0.016
	Telia	0.007	0.019	0.016	0.008	0.008	0.009	0.021	0.018	0.011	0.010
	Volvo	0.008	0.019	0.018	0.010	0.010	0.012	0.021	0.021	0.014	0.014

Note: The table illustrates the evaluation results for the alternative volatility forecasts using two loss functions, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The results are reported for three different samples (i.e., the entire sample from 2002 to 2009, the period from 2006 to 2009, and finally, the sample from weeks with a larger-than-median number of articles). The results are reported for each firm.

Table B4. Results of the DM test

		DM test			
		GARCH	*EGARCH .*	*Lag vol.*	*Mov. aver.*
2002–2009	*Nordea*	−9.67**	−8.87**	1.02	−3.35**
	SCA	−14.89**	−16.15**	−2.47*	−2.51*
	SHB	−6.41**	−6.26**	1.33	−3.10**
	Telia	−14.94**	−10.26**	−0.58	−0.75
	Volvo	−16.64**	−14.79**	−2.17*	−3.93**
2006–2009	*Nordea*	−3.73**	−5.69**	−0.34	−4.28**
	SCA	−7.18**	−6.99**	−2.67**	−2.84**
	SHB	−2.94**	−2.98**	0.64	−3.52**
	Telia	−11.13**	−7.68**	−3.44**	−2.77**
	Volvo	−6.59**	−6.60**	−1.36	−4.20**
Weeks with a large no. of articles	*Nordea*	−3.58**	−4.18**	−0.12	−4.30**
	SCA	−9.01**	−9.69**	−2.32*	−2.86**
	SHB	−3.72**	−3.55**	0.59	−3.94**
	Telia	−11.37**	−8.95**	−3.53**	−4.61**
	Volvo	−8.05**	−7.39**	−1.13	−3.98**

Note: The table illustrates the statistics of the DM test for relative performance of the semantic forecast approach against the alternative volatility forecasts. The results are reported for three different samples (i.e., the entire sample from 2002 to 2009, the period from 2006 to 2009, and finally, the sample from weeks with a larger-than-median number of articles). A negative value corresponds to a relatively better prediction of the semantic forecast approach. The estimated statistics have a standard normal distribution. The results are reported for each firm. The values marked with one asterisk are significant at the 5% level, and those with two asterisks are significant at the 1% level.