9-2010

PREDICTING INTRADAY STOCK RETURNS BY INTEGRATING MARKET DATA AND FINANCIAL NEWS REPORTS

Tomer Geva
Tel-Aviv University, Israel, tomergev@post.tau.ac.il

Jacob Zahavi
Tel-Aviv University, Israel, JacobZ@tauex.tau.ac.il

PREDICTING INTRADAY STOCK RETURNS BY INTEGRATING MARKET DATA AND FINANCIAL NEWS REPORTS

Tomer Geva, Tel-Aviv University, Israel, tomergev@post.tau.ac.il

Jacob Zahavi, Tel-Aviv University, Israel, JacobZ@tauex.tau.ac.il

Abstract

Forecasting in the financial domain is undoubtedly a challenging undertaking in data mining. While the majority of previous studies in this field utilize historical market data to predict future stock returns, we explore whether there is benefit in augmenting the prediction model with supplementary domain knowledge obtained from financial news reports. To this end, we empirically evaluate how the integration of these data sources helps to predict intraday stocks returns. We consider several types of integration methods: variable-based as well as bundling methods. To discern whether the integration methods are sensitive to the type of forecasting algorithm, we have implemented each integration method using three different data mining algorithms. The results show several scenarios in which appending market-based data with textual news-based data helps to improve forecasting performance. The successful integration strongly depends on which forecasting algorithm and variable representation method is utilized. The findings are promising enough to warrant further studies in this direction.

Keywords: Data Mining, Text Mining, Financial

1 INTRODUCTION

Information systems (IS) play a vital role in the financial industry to deliver and process the vast amount of information created on a daily basis. However, IS capabilities are only marginally used to predict future stock prices, this, despite the fact that such prediction capability can have a major impact on investment decisions. In this research, we address the prediction problem focusing on predicting intraday stock prices.

Over time, the task of forecasting stock returns has been the subject of a myriad of data mining studies. The majority of these studies utilize the most common data source in the financial domain, primarily historical market data such as stock price and trading volume, to predict future stock returns.

Nevertheless, an alternative data source for explaining stocks behavior is textual financial news reports. These textual news reports are commonly available over internet web sites as well as via proprietary IS and other delivery systems. Most importantly, they provide significant information about influential real world events that historical market data can at best, only partially and indirectly represent. Surprisingly, only a significantly smaller number of studies consider the textual financial news reports in forecasting stock prices.

A potentially promising approach, proposed and assessed in this study, is to integrate the above two data sources: (1.) market-based numerical data, and (2.) news-based textual data, as predictors for forecasting stock returns. This integration enriches the prediction model with additional domain information that is not available in the market data itself. Furthermore, having both data sources enables capturing patterns that may not be identified by employing each data source separately. Thus, a financial forecasting system which utilizes both market data and data from financial news reports can benefit from the synergy between these two different data sources.

The small number of studies utilizing both sources of information to forecast the behavior of the stock market are lacking in the sense that they performed only limited comparative evaluations of different integration methods. Additionally, these studies did not explore the sensitivity of the integration methods' results to the usage of different forecasting algorithms.

In this work we aim to enhance existing knowledge concerning the extent to which augmenting numerical market-based data with data from textual news reports improves the forecasting of intraday stocks returns, seeking the conditions under which this integration yields the best performance. Our goal, as compared to previously reported studies, is to provide a more comprehensive evaluation of the major factors influencing the performance of the integration methods.

To this end, we have constructed a trading recommendation system which allows for the alteration of its main components. Using this system we have conducted a wide empirical comparative study involving a set of different integration methods and forecasting algorithms. We also explore how sensitive are the integration methods' results to the type of forecasting algorithm and variable representation employed and "drill down" to find possible explanations for the different performance levels. In addition, we make use of various pre-processing procedures, and utilize systematic feature selection and transformation procedures to convert the data into a form more amenable for data mining. The modeling results were evaluated by means of a simulation procedure that mimics the trading scheme that takes place in practice.

The rest of this paper is organized as follows: Section 2 provides a general background regarding forecasting stock returns, and intraday forecasting. Section 3 describes the forecasting system's data sources, pre-processing, modeling and implementation. In Section 4 we detail the evaluation method and analyze the results. In section 5 we provide conclusions and highlight future research.

2 BACKGROUND

Prediction of stock returns is undoubtedly a challenging undertaking. Key financial theory including the Random Walk model of stock prices (Fama, 1965) and Efficient Market Hypothesis (EMH) (Fama, 1970) critically question the ability to forecast future stock behavior. Additionally, it is recognized that using data

mining to forecast stock returns exhibits many challenges which have been widely discussed in the data mining literature. For example, (Dhar and Chou, 2001) mention that financial markets are inherently "noisy" with several types of nonlinearities. Additionally, (Dhar et al., 2000) report about: (a.) weak theory which results in a large number of variables thus increasing the dimensionality of the problem; (b.) relationships between variables being weak and nonlinear; and, (c.) the potential significance of variable interactions.

Another challenge to data mining models in this domain has to do with the stability of the underlying process. Predictive models inherently assume that the process governing the relationships between future price changes and historical price changes and news items, are stable over time. However, as evident from the recent financial events, this process may be interrupted by unexpected major macroeconomic factors, abrupt changes in investors' preferences, and other factors. These changes may occur gradually or rapidly and can eventually render a forecasting model invalid. In the financial domain, perhaps more than in any other domain, it is therefore extremely important to detect pattern changes in market behavior as quickly as possible. One such approach for change detection in classification models induced from time series data is discussed in (Zeira et al., 2004).

2.1 Related work

Of the large number of previous studies exploring forecasting of stock returns the vast majority utilized data mining methods over predictors obtained from market data only. An alternative data source for explaining stocks behavior is textual financial news reports. These news reports provide important knowledge about influential real world events that historical market data can represent, at best, only partially and indirectly.

Surprisingly, only a significantly smaller number of studies consider financial news reports for forecasting stock prices. Of these studies we mention (Fung et al., 2005; Lavrenko et al., 2000) which create predictors from textual news data using a "bag of words" approach as well as using piecewise linear segmentation in order to calculate the dependent variable (stock trends); (Macskassy et al., 2001) utilize textual data to forecast whether stock returns will be more than one standard deviation away from the stock average hourly returns; (Robertson et al., 2007) utilize textual news data to forecast abnormal price volatility rather than stock returns; (Mittermayer, 2004) uses a "bag of words" approach to forecast short range intraday stock returns; (Mittermayer and Knolmayer, 2006) enhance Mittermayer's previous study to evaluate the effects of various filter based feature selection methods, different number of variables, and several classification algorithms on forecasting accuracy; (Wuthrich et al., 1998) predict major stock indexes following the identification of keywords (or groups of keywords), provided by an expert; and, finally (Schumaker and Chen, 2006) use various textual representation methods including "bag of words", noun phrases and named entities.

However, financial prediction which is solely based on textual news data has a downside too because it may miss patterns which could have otherwise emerged if both market data and arrival of news information were involved. Various such patterns are already known and have been described in financial literature e.g., (Busse and Green, 2002) for intraday patterns, and (Rendleman et al., 1982) for inter-day patterns. Therefore, applying data mining methods on data from the two information sources not only enriches the market-based data with additional domain knowledge, but also has the potential to exploit the synergy between the two sources to detect additional relevant patterns that may be missed by using either data source.

Among the few studies which integrate both types of explanatory information, we mention (Thomas and Sycara, 2000) that compared few forecasting methods as well as weighting them. (Thomas, 2003) used textual news reports as a gating signal to halt the activation of the previously discovered trading rules that were built using market data statistics. (Schumaker and Chen, 2008) consider an integration model in which trading rules based on news data were applied only on stocks that were previously selected by pre-determined strategies.[1] Nevertheless, these few studies only covered a small part of the

[1] We note that (Wuthrich et al., 1998), (Schumaker and Chen, 2006) might have also utilized both data sources as explanatory variables within their models. However, the exact way in which market data is appended to textual news data is not explicitly mentioned. We also note

major aspects and benefits pertaining to the integration of market data and news data.

2.2 Intraday Forecasting of Stock Return

This study is concerned with forecasting intraday stock returns, i.e., focusing on short range prediction of price changes during the trading day. Forecasting at the intraday level allows one to capture market signals and identify trading opportunities, as well as act before the information is fully assimilated and reflected at the stock price level. This process can be enhanced by continuous real-time monitoring of financial news reports and market data. The monitored data can then be input into a forecasting system which "interprets" existing conditions and produces immediate actionable trading recommendations.

This approach is supported on practical grounds. In accordance with EMH, when additional information is provided, traders' actions push the stock price toward efficiency. However, in practice this adjustment cannot happen instantaneously and there is still a short time interval during which the stock price remains non-efficient. It is during this time that an automated trading system can still take profitable action. The length of this interval in which the stock price remains non efficient has been researched by various financial studies. Studies such as (Busse and Green, 2002) and (Chordia et al., 2005) observed stock price inefficiency during time intervals ranging from less than one minute up to 30 minutes. (Busse and Green, 2002) also demonstrated the feasibility of acting to exploit short term market inefficiencies for financial gains following the broadcast of a financial news television program.

3 FORECASTING SYSTEM

3.1 Data

In this study we utilize textual news reports and intraday market data involving 48 S&P500 companies, collected during a period of ten and a half weeks. The textual news data was gathered by monitoring 19 leading web sites that provide financial news, financial commentary and public relations announcements. This large set was intended to provide a good coverage of relevant data sources and reduce the chance of missing important news.

The web sites were monitored using a Perl script that downloaded XML files obtained from the web sites' RSS feeds, twice every 15 minutes. RSS is a family of XML based formats used for Internet news syndication. This delivery method is supported by a large number of leading web sites. RSS feeds usually contain (among other data): (a.) a title field; (b.) a description field containing the gist of the news items; and, (c.) a link to a web page with the full news story. Relevant news items per company were selected using a lookup table including company's name, ticker, and known aliases. Overall, a total of 10,890 relevant news items were obtained and stored in a database.

Intraday market data for the respective time period was obtained from an online financial service provider. The data consisted of closing price and trading volume per one minute interval. This data was subsequently adjusted for dividends and stock splits.

3.2 Data Processing and Representation

Various pre-processing stages were carried out in order to translate the raw data to predictors and create a flat file for use by the forecasting models. Each record in the flat file represents the information for an individual stock at the end of each 15 minute interval, from 9:30 to 15:00, during trading days. The dataset consisted of ~56,000 records (instances) which were split into a training set containing ~37,000 records (first 7 weeks of data), and the validation set containing ~19,000 records (subsequent 3.5 weeks of data).

The dependent variable used in this study was a binary variable exhibiting whether stock returns at the end of the trading day exceeds the S&P500 index by at least one percent ("positive" in 5.2% of the

---

that a few studies mentioning that they concurrently use textual news data with market data are effectively utilizing the market data only in relation to defining dependent variable. E.g., (Lavrenko et al., 2000).

instances). Forecasts were performed for each 15 minutes interval from 9:30 to 15:00 on trading days. The S&P500 index was represented by the SPY fund which follows the S&P500 index.

News Items Count: In this study, we used two subsets of explanatory variables extracted from news-based textual data. The first involves news items counts only: number of news items during the five (15 minutes) time intervals preceding the current time interval, counts of news items from the beginning of the trading day and count of news items from the end of previous trading day. We also created simple transformations such as whether there was an increase or decrease in the number of news items during the last several intervals.

Since news items count accounts only for the occurrence of the news but not for its content, we also categorized the news according to the different types of web sites and RSS feeds in which they appeared (e.g., News reports, PR announcements, real-time headline news). This provides a proxy, although limited, regarding the nature of the news.

To account for the fact, which is typical of news data in the financial domain, that many news items contain repetitive information on previous stock price changes (approximately 25% in our case), we employed a classifier to highlight news reports that are not "repetitive". This approach is not reported in existing related data mining studies. The approach was supported by a process involving several stages:

1. A pre-processing stage that involves tokenizing the news items' text, stemming (Porter, 1980), and "stop list" filtering.
2. Picking a sample of 500 news items randomly selected from the first seven weeks of news data and manually classifying them into "repetitive" and "non-repetitive" news items.
3. Using 335 news items from this manually-categorized sample to train a Naïve Bayes classifier and the balance 165 news items to validate the classification results. Various settings were tested including: (a.) using single words as features; (b.) using pairs of adjacent words as features (as well as single and adjacent words combination); (c.) Information Gain (Witten and Frank, 2005) based feature selection choosing the top 1,000 and 200 features; and, (d.) manipulating the prior probabilities of the dependent variable classes (due to non-proportionate representation of the less common class).

The best results were obtained using a variable representation scheme that considers the top 1,000 features based on information gain ranking involving both single and pairs of adjacent words. Table 1 details the performance for this setting (For brevity we do not report the results for the other settings).

Class	Precision	Recall
"Does Not Contain Repetitive Stock Price Information"	89.6%	96.8%
"Contains Repetitive Stock Price Information"	86.7%	65.0%

Table 1: Intermediate Classification Results.
*Based on the 165 items used for validation. Total items correctly classified: 89.1%.

We implemented the resulting classification scheme over the entire textual dataset to highlight news items that belong to the class "Does Not Contain Repetitive Stock Price Information". The recall rate for this class turned out to be very high (96.8%). This translates into low chances of false dismissal of potentially valuable news. The high precision rate (89.6%) implies that we were able to create textual news-based variables while filtering out a large part of the news items which contain repetitive stock related information.

Finally, we aggregated the news items' count over each 15 minutes interval by the category of their original web site and by their classification results (i.e., whether they contain repetitive information) to render predictors for our forecasting models.

Bag of Words Representation: The second subset of explanatory variables was directly based on a "bag of words" representation. Using this approach, we use words within the news items' text as predictors in the forecasting model. This approach provides more insight regarding the nature of the news items than using news item count and was employed in various other related studies (section 2.1).

The "bag of words" variable representation was created using similar pre-processing stages (i.e., tokenizing, stemming and "stop list" filtering), and then aggregated over 15 minute intervals.

Typically, the "bag of words" representation results in a large amount (thousands) of potential word-based variables, rendering the prediction problem very high-dimensional, significantly complicating the modeling process and increasing the risk of overfitting. To address this problem, we performed an initial feature selection procedure by ranking the word-based features and then picking the top features from the list as predictors for the model. Two ranking procedures were used: (a.) Chi Square weighting scheme, keeping the 75 top ranked features; and (b.) an information gain ranking, selecting the top 500 features. Interestingly, we noticed that using the 500 word-based features only marginally affected the forecasting results (compared to the 75 Chi square ranked features). For brevity, the information gain approach is not reported in this paper.

Market Data Representation: Finally, we created an additional (third) subset of explanatory variables based on market information. This data was aggregated to create different predictors based on stock returns and trading volume during each 15 minutes time intervals and preceding time intervals.

The three variable subsets are summarized in Table 2. These subsets were utilized selectively in the different forecasting models. In addition, we have added "demographic" company related information to all forecasting models, such as: industry, sector, and membership in various indices.

Subset	Source	Description	Number of Features*
1	Textual News Reports	News item count during different time intervals considering web site categories and classification whether the news item contains repetitive information.	~240
2	Textual News Reports	Bag of Words representation. 75 top rated features based on Chi Square ranking feature selection. (Alternative representation used 500 variables using information gain ranking)	75
3	Market Data	Stock returns and trades during different time intervals	~80

Table 2: Variable Subset.
* In later stages, additional feature selection and transformation activities were performed (see section 3.4)

3.3 Integration Methods and Forecasting Models

In this study we employed several integration schemes to combine market-based and news-based data: (a.) "variable-based" methods - combining explanatory variables in a forecasting model; and (b.) "bundling" methods – combining the predictions of the two separate models, one based on textual data and the other on market data. The variable based integration method aims to generate knowledge (forecasts) by detecting patterns that can be observed in the presence of raw information from both data sources. On the other hand, the integration by "bundling" models method aims to generate knowledge (forecasts) by combining the knowledge (forecasts) generated from each data source separately. These integration methods yield several forecasting models, detailed as follows and summarized in Table 3:

Benchmark models - forecasting models based on a single type of data source (either news-based or market-based):
Model I *– Market Data*: Model I uses explanatory variables based on market data only (variable subset 3).
Model II *–Textual News Data*: Model II is based on the entire set of news-based explanatory variables (variable subsets 1 and 2).
These benchmark models have two purposes: (a.) to serve as a benchmark for the various forecasting models; (b.) to be used as a component in more elaborate "bundling" models (Models V and VI).

Variable-based integration
Model III *– Incorporating Market Data and Partial Textual News Data:* Model III integrates textual and market data by combining variable subset 1 (news item counts) and variable subset 3 (market

data) as explanatory variables in a single forecasting model.
Model IV *- Incorporating Market Data and Full Set of Textual News Data:* It is similar to Model III in that it incorporates both textual and market data in a single forecasting model, but uses a broader textual data that also involves variable subset 2 (word based features).

Integration by "bundling" models
Model V *– Weighing Models I and II:* Model V assigns a weight of α to the predicted probabilities of Model I (market data) and (1-α) to Model II (textual news data) to create a single probability ("score"). The cutoff probability for the weighted model was weighted using a similar scheme based on the cutoff probabilities of the separate models. The different weights (α, 1-α) were assigned to the two models in order to avoid cases in which one model dominates the other. This dominance could occur simply because one model may tends to produce predicted probabilities ("scores") in a wider range than the other. However, in later stages we observed that the different forecasting models produced very similar average and variance for the predicted probabilities of the different models so a simple average (using α=0.5) was eventually utilized.
Model VI *– Intersection between Model I and Model II:* Model VI uses the intersection of Model I and Model II and classifies as "positive" only instances that are classified as "positive" by both models.

Model	Data Sources	Integration Scheme
I	Market data (subset 3)	
II	Textual data – news items count (subset 1) Textual data – word based features (subset 2)	
III	Textual data – news items count (subset 1) Market data (subset 3)	Incorporating both variables types into one forecasting model
IV	Textual data – news items count (subset 1) Textual data – word based features (subset 2) Market data (subset 3)	Incorporating both variables types into one forecasting model
V	Textual data – news items count (subset 1) Textual data – word based features (subset 2) Market data (subset 3)	Weighing the outputs of Model I (market data) and Model II (full textual data)
VI	Textual data – news items count (subset 1) Textual data – word based features (subset 2) Market data (subset 3)	Employing the intersection of Model I (market data) and Model II (full textual data)

Table 3: Models' Data Sources and Integration Schemes

3.4 Implementation

The forecasting models above were implemented by means of GainSmarts software package (Levin and Zahavi, 2002). The GainSmarts software is noted for its preprocessing and feature selection procedures. It performs multiple automatic transformations on the explanatory variables to capture non linear relations between the dependent and the independent variables and then applies a rule-based expert system to select the most influential subset of predictors explaining the dependent variable. The objective is to increase the prediction accuracy while reducing overfitting, which is a known concern in financial forecasting (Thomas, 2003).

For a wider comparative evaluation of integration models for forecasting stock returns, we ran each model using three different algorithms implemented in the GainSmarts system, each represents a different "family" of classification algorithms: (a.) Stepwise Logistic Regression (SLR); (b.) CHAID tree; and (c.) Feed forward Neural Network (NN).[2]

We note that some studies on related research topics presented prototypes of trading systems which included hand-crafted trading strategies such as early exit strategies and a combination of long/short positions. However, the prime objective of this study is to evaluate the effectiveness of data mining

[2] Minimizing sum of squares error function. Using two layers of weights and three hidden units (following a trial and error procedure).

and integration methodologies. Therefore, we avoided manual "tweaking" of trading rules since we were concerned that it may obscure the actual performance of the data mining and integration methodologies which we utilized.

Figure 1. presents the main components and processes used in the forecasting system and the simulation procedure (detailed in sections 3-4)

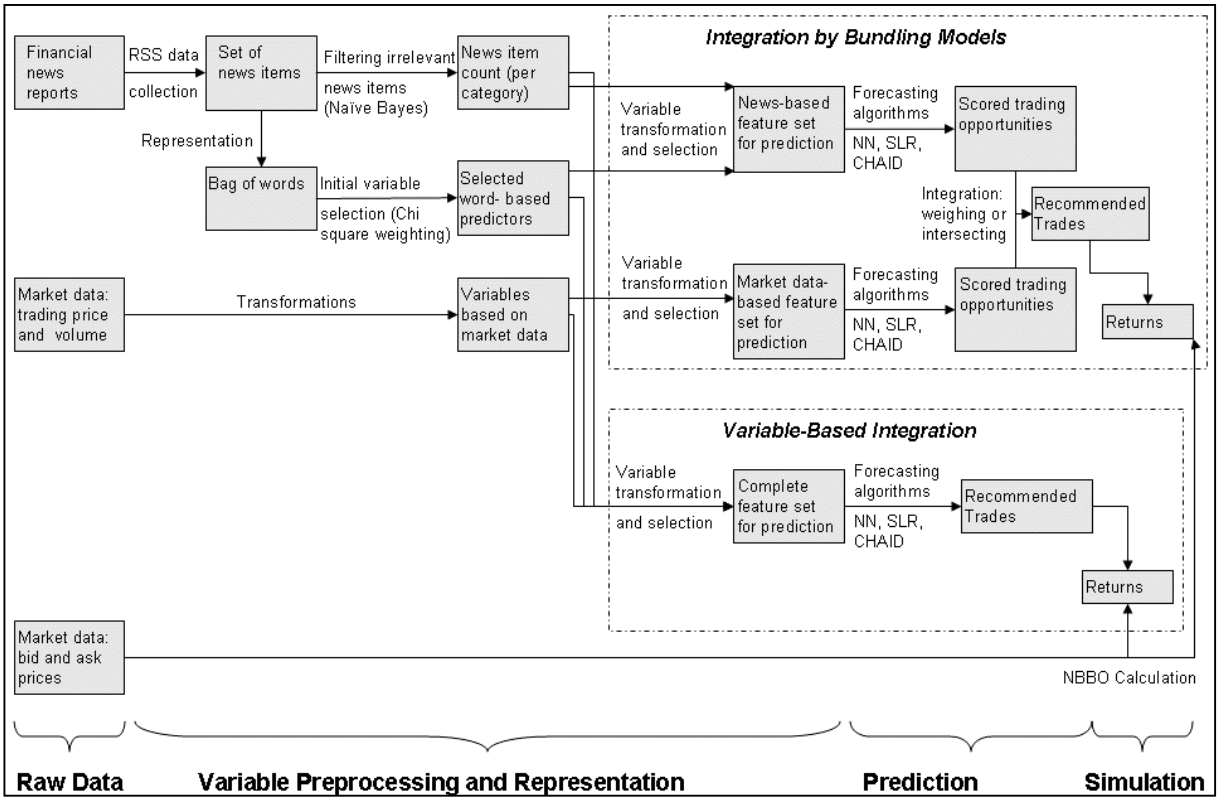

Figure 1: Main components and processes used in the forecasting system and the simulation procedure

4 MODEL EVALUATION

4.1 Performance Measure

We use the average returns above the S&P500 index as our primary criteria.[3] This measure was obtained by running a trading simulation over the validation dataset, "buying" a stock when the model categorizes it as "positive" and "selling" it at the end of the trading day. The returns of each transaction were calculated based on the bid and ask prices - "buying" the stock at the ask price and "selling" it at the bid price. Subsequently, we calculate the returns difference between the trading transaction and the S&P500 during the same time interval. The final stock performance measure is given by the average return difference over the simulation period.

The bid and ask prices were obtained from the Quote file of the NYSE Trades and Quotes database. which provides historical best bid and ask prices from various exchanges and market centers. We first filtered out irregular data using procedures described by (Bessembinder, 2003). We then calculated the momentary best bid and ask prices (also known as the NBBO - National Best Bid and Offer) over all exchanges. Last, we registered each trade initiated by our models according to the relevant NBBO bid or ask prices.

[3] Many measures to evaluate financial prediction models exist, including: Sharpe ratio, precision and recall. For brevity, we report and analyze average returns, a measure commonly reported in related data mining literature. We evaluated additional measures reaching similar conclusions.

It should be noted that as opposed to using trading prices (used by previous studies in this domain), using the bid and ask prices yields much more accurate estimation of model performance because bid and ask prices inherently reflect the transaction cost.[4] By and large, using trading prices often results in overly optimistic returns since it does not reflect the impact of transaction costs.

4.2 Performance Evaluation

Table 4 exhibits the modeling results. Overall, our analysis yields encouraging performance of intraday stock returns prediction with 16 out of the 18 forecasting model/algorithm combinations resulting in positive average returns above the S&P500 index, even while accounting for transaction costs (the two model/algorithm combination with negative returns involved benchmark models). Furthermore, 11 out of the 18 combinations obtained statistically significant average returns, using a P-value of 0.05 or better.

Model	Algorithm	Average Returns*	Number of Trades	P-value Returns**
I	SLR	0.20%	570	0.002
	NN	-0.01%	691	
	CHAID	0.10%	514	0.048
II	SLR	-0.03%	519	
	NN	0.21%	521	0.000
	CHAID	0.17%	485	0.001
III	SLR	0.07%	572	0.163
	NN	0.20%	657	0.000
	CHAID	0.16%	498	0.000
IV	SLR	0.07%	550	0.159
	NN	0.26%	680	0.000
	CHAID	0.16%	498	0.001
V	SLR	0.11%	527	0.071
	NN	0.16%	486	0.012
	CHAID	0.17%	170	0.088
VI	SLR	0.20%	347	0.009
	NN	0.62%	185	0.000
	CHAID	0.18%	158	0.085

Table 4: Performance (validation set).
* Average returns per trade, above S&P500. Buying the stock at the "ask" price, Selling at the "bid" price.
** T-test: H0: average returns=0 H1: average returns>0. Value of "0.000" denotes a P-value lower than 0.0005.

Notwithstanding, the primary interest of this study is to evaluate whether integrating both type of explanatory variables yields better results, and if so, which integration method and algorithm is most suitable. Figure 2 provides a graphical comparison of the models employed in this study, highlighting the two salient model/algorithm combinations. Model VI (intersection of Model I and Model II) coupled with a NN algorithm, obtained the highest average returns (with a wide margin) but produced a reduced number of trades. Model IV ("variable-based" integration, full set of explanatory variables) coupled with a NN algorithm produced the second highest average returns and a high number of trades.

Model IV/NN empirically demonstrated that there was indeed merit in integrating both variable types into a single forecasting model as it obtained both: (a.) superior average returns per trade in comparison to both benchmark models (Model I -market data, and Model II - textual data) combined with any of the forecasting algorithms; (b.) the highest number of trades among all model/algorithm combinations that produce positive average returns (demonstrating that returns were not exchanged for reduced trading).

However, we note that this conclusion is not necessarily guaranteed for every variable representation or forecasting algorithm. In our study, for example, only when employing NN, did Model IV manage to outperform all model/algorithm combinations involving the benchmark models. The fact that the NN algorithm outperformed SLR and CHAID may be attributed to several factors: First, NN has an inherent ability to handle non-linear relationships between the dependent and the independent variables. Such non-linear relationships are commonplace in a financial forecasting environment (Dhar

[4] We note some limitations of the way we account for transaction costs: 1. Our data does not include the entire "order book" making it less suitable for simulating large trades; 2. Brokerage commissions are not considered. However, they are relatively low, especially for financial institutions.

and Chou, 2001) and are less suitably handled by the linear SLR algorithm. Second, the fact that SLR and CHAID employ greedy feature selection procedures (as opposed to the non greedy NN algorithm) could be particularly detrimental to successful integration of market and textual data because it may "push out" the weaker textual news-based variables (see findings in section 4.3 below) from the model, thus yielding a model which is incapable of capturing patterns that involve interrelations between the two variable types.

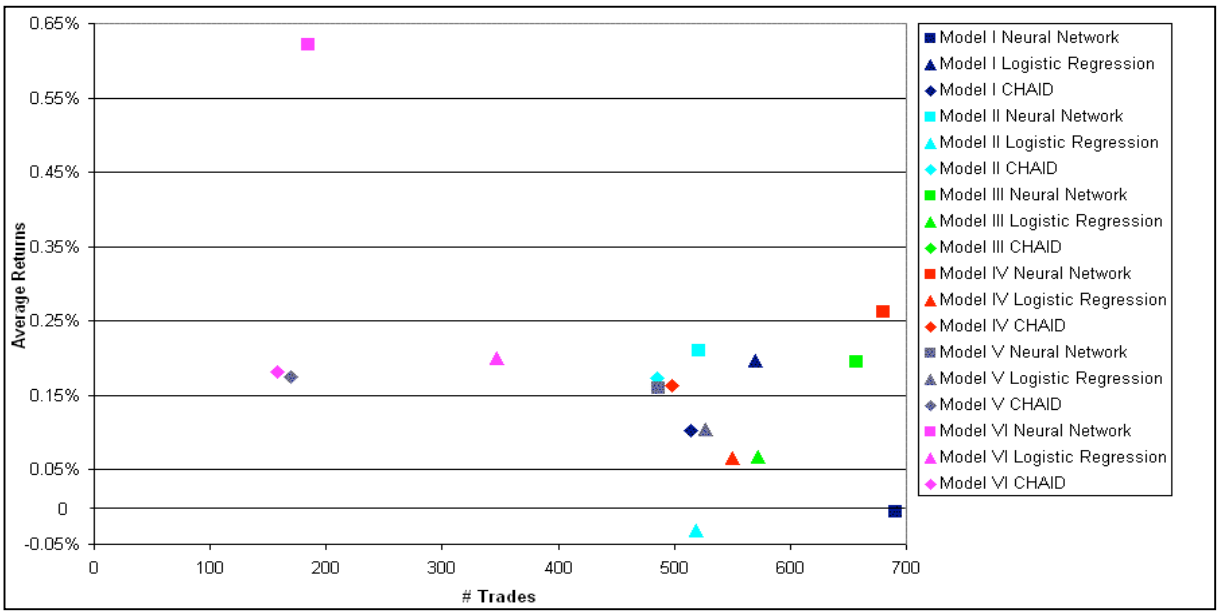

Figure 2: Models I-VI. Average Returns (above S&P500) Vs. Number of Trades.

4.3 Explanatory Power of Individual Features

An interesting aspect of this research is to asses the prediction power of textual-based features as compared to market-based features. This comparison was performed by assessing the information gain values of the textual and market variables to the best of our knowledge such a comparison was not previously reported. Information gain (Witten and Frank, 2005) is a basic measure commonly used to assess the explanatory power of features. It is utilized as splitting criteria for the induction of decision trees (Quinlan, 1986) and in feature selection procedures.

However, information gain ranking usually tends to favor variables with a large number of possible values. Thus, using this measure "as is" may distort the explanatory power of the textual data because they contain far less possible values as compared to continuous variables which are dominant in the market data. One possible remedy is to use the Gain Ratio measure instead of information gain. However, the Gain Ratio can sometimes overcompensate and favor variables just because they have lower intrinsic information (Witten and Frank, 2005). To compensate for this effect and bring both variable types to a common ground, we converted all variables into integer values by discretizing the continuous market-based variables into up to 20 sections (comparable to the midpoint of the textual-based variables that, in our dataset, had a range of 0-40).

While this compensation method can somewhat reduce the explanatory power measure of our market based variables, it is still possible to notice that the majority of textual-based variables resulted in lower information gain than market based variables. Figure 3 demonstrates this point for the 50 top ranked variables within each one of the three variable subsets. The figure also shows that the word-based variable subset (variable subset 2) possesses an even lower explanatory power than the variable subset based on textual news items count (variable subset 1). These observations were also confirmed by means of Chi-Square variable rankings, which are not presented here for the sake of brevity.

Another evidence of the "weakness" of individual word-based variables as compared to the other variables subsets, is provided by Model IV (considering full textual information and market data), which yields a similar classifier to that of Model III (which does not include word-based variables)

when both used CHAID. Evidently, the CHAID algorithm employing greedy heuristics in order to choose which variables to split nodes by, gave little importance to the word-based variables due to their low individual explanatory power.

Therefore, it is interesting to note in our study that while each individual textual based feature does not have much explanatory power, when we combine textual-based data with market-based data and using the "right" algorithm, the collection of textual based features significantly improves forecasting performance. (E.g., this is clearly visible when comparing the combined model IV/NN, with the separate models I/NN and II/NN)

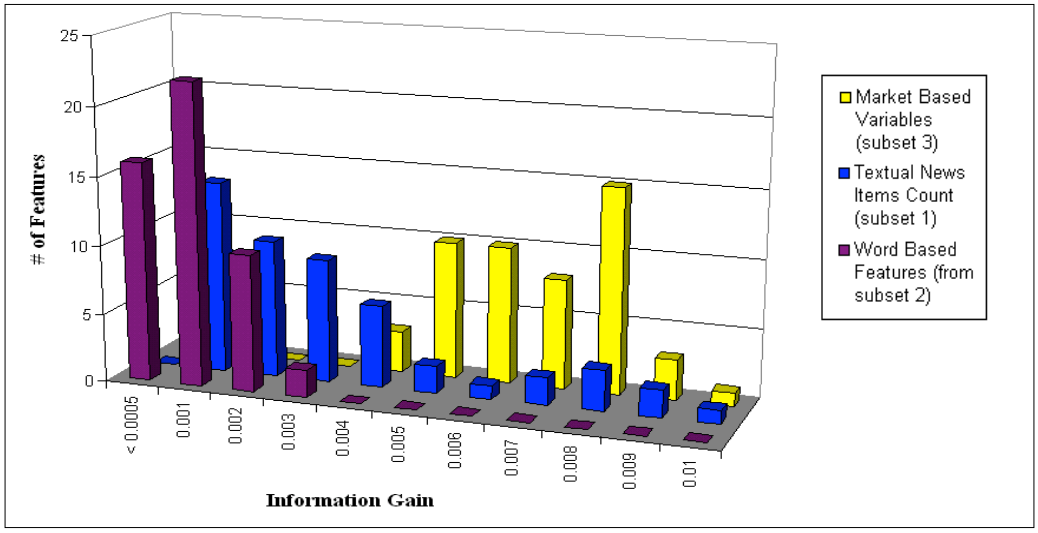

Figure 3. Distribution of the Top 50 Features for Each of the Three Feature Subsets, by Information Gain. Information gain values for the market based variables (yellow) are generally distributed toward the right side of the graph with higher values. Among the textual based variables, word based features (purple) are concentrated to the left side of the graph with lower values, while information gain values for variables based on news items count (blue) are more widely distributed

5 CONCLUSIONS AND FUTURE WORK

In this study we offer several contributions and enhancements to forecasting in the financial domain. This study performed a wide empirical evaluation of integration methods that append information from textual news-based data and market-based data to predict intraday stocks returns. Each integration method was implemented using three different forecasting algorithms by means of GainSmarts software, which is noted for its variable transformation and feature selection methods. A Naïve Bayes classifier was used to characterize news items as being "repetitive" or "non-repetitive". Also, to the best of our knowledge, this is the first study that uses bid and ask price data to inherently account for transaction costs, among data mining studies utilizing textual data to forecast stock returns.

The study demonstrated empirically that under certain conditions, combining textual-based and market-based predictors in a single forecasting model improves the forecasting results. In particular, "best" results were obtained when using a full set of textual based explanatory variables coupled with an advanced non linear algorithm such as NN. We note that the fact that non-linear algorithms are more successful in handling prediction in the financial domain have already been widely discussed in literature. Notwithstanding, we show here that this usage of nonlinear algorithms may be even more important in the case of employing domain data from both textual news and market data sources. Furthermore, by analyzing the explanatory power of individual textual and market based variables we shed light on the reasons why non-linear algorithms are more successful than linear and "greedy" algorithms in handling the relatively "weak" textual based variables.

Not surprisingly, the study also shows that the bundling method which uses the intersection of forecasting models based on either textual data or market data, significantly improves forecasting performance. But this is obtained at the price of reducing the number of trading opportunities. This simple integration approach may appeal to risk averse traders.

Overall, the results of this study are promising enough to warrant further research on augmenting numerical market-based data with data from textual news reports. For example, exploring more advanced textual data representation methods such as classifying news items into business events, incorporating measurement of trends over time in textual news data, performing sentiment analysis, using larger datasets, and others.

References

Bessembinder, H. (2003). Quote-based competition and trade execution costs in NYSE-listed stocks. Journal of Financial Economics 70(3): 385-422.

Busse, J. A. and Green, T. C. (2002). Market efficiency in real time. Journal of Financial Economics 65(3): 415-437.

Chordia, T., Roll, R. and Subrahmanyam, A. (2005). Evidence on the speed of convergence to market efficiency. Journal of Financial Economics 76(2): 271-292.

Dhar, V. and Chou, D. (2001). A comparison of nonlinear methods for predicting earnings surprises and returns. Neural Networks, IEEE Transactions on 12(4): 907-921.

Dhar, V., Chou, D. and Provost, F. (2000). Discovering interesting patterns for investment decision making with glower—a genetic learner overlaid with entropy reduction. Data Mining and Knowledge Discovery 4(4): 251-280.

Fama, E. F. (1965). The behavior of stock-market prices. The Journal of Business 38(1): 34-105.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance 25(2): 383-417.

Fung, G. P. C., Yuy, J. X. and Luz, H. (2005). The predicting power of textual information on financial markets. IEEE Intelligent Informatics Bulletin 5(1).

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. and Allan, J. (2000). Mining of concurrent text and time series. In Proc of the 6th Int Conf on Knowledge Discovery and Data Mining.

Levin, N. and Zahavi, J. (2002). Case studies: Commercial, multiple mining tasks systems: Gainsmarts. Handbook of data mining and knowledge discovery 601-609.

Macskassy, S. A., Hirsh, H., Provost, F. J., Sankaranarayanan, R. and Dhar, V. (2001). Intelligent information triage. Research and Development: 318-326.

Mittermayer, M. A. (2004). Forecasting intraday stock price trends with text mining techniques. Proc. of the 37th Annual Hawaii Int Conf on System Sciences.

Mittermayer, M. A. and Knolmayer, G. F. (2006). Newscats: A news categorization and trading system. Proc. of the Sixth Int Conf on Data Mining.

Porter, M. F. (1980). An algorithm for suffix stripping. Program 14 (3): 130-137.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning 1(1): 81-106.

Rendleman, R. J., Jones, C. P. and Latan, H. A. (1982). Empirical anomalies based on unexpected earnings and the importance of risk adjustments. Journal of Financial Economics 10(3): 269-87.

Robertson, C., Geva, S. and Wolff, R. C. (2007). Can the content of public news be used to forecast abnormal stock market behaviour? Seventh IEEE Int Conf on Data Mining, 2007.

Schumaker, R. P. and Chen, H. (2006). Textual analysis of stock market prediction using financial news articles. 12th Americas Conf on Information Systems (AMCIS-2006).

Schumaker, R. P. and Chen, H. (2008). Evaluating a news-aware quantitative trader: The effect of momentum and contrarian stock selection strategies. Journal of the American Society for Information Science and Technology 59(2): 247-255.

Thomas, J. D. (2003). News and trading rules, Carnegie Mellon University. PhD Thesis.

Thomas, J. D. and Sycara, K. (2000). Integrating genetic algorithms and text learning for financial prediction. Data Mining with Evolutionary Algorithms: 72-75.

Witten, I. H. and Frank, E. (2005). Data mining: Practical machine learning tools and techniques, Morgan Kaufmann.

Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K. and Zhang, J. (1998). Daily stock market forecast from textual web data. IEEE Int Conf on Systems, Man, and Cybernetics.

Zeira, G., Maimon, O., Last, M. and Rokach, L. (2004). Change detection in classification models induced from time series data. Data mining in time series databases, m. Last, a. Kandel, and h. Bunke (editors), world scientific, series in machine perception and artificial intelligence.