



## Smart Market and Money

Hsinchun Chen, University of Arizona

**M**arket, a term frequently used in mass media and academic publications, is an elusive concept. In marketing, researchers and practitioners describe market as a place to exchange products and services (such as a retail market or real estate market). In economics and finance, financial and monetary concepts such as emerging markets, commodity markets, and the stock market are often mentioned. In all these areas, one of the most challenging research directions is modeling and predicting market movements. In recent years, the availability of diverse and voluminous market-related mass media and social media content (or Business Big Data) and the emergence of sophisticated, scalable text and social mining techniques present a unique opportunity for advancing research relating to smart market and money. This research area, at the intersection of computational and finance research, aims at developing intelligent (smart) mechanisms and algorithms for predicting market and stock performances.

### Market Modeling and Analysis

In economics, finance, accounting, and marketing, researchers have developed many sophisticated theories and analytical models. Most of these modeling and analytical techniques are quantitative in nature and rely on carefully constructed and highly relevant market- or firm-specific metrics, such as market return, market capitalization, and book-to-market ratio in the classical Fama-French three-factors model.<sup>1</sup> Numerous online, unstructured (text) Web-enabled business data sources considered by financial analysts in their expert industry and firm analysis, including 10K/10Q SEC reports, mass media news, local news, Internet news, financial blogs, investor forums, and tweets, are providing new opportunities for extracting market- or firm-specific

intelligence. Large-scale automated analysis of this text-based, qualitative content has only recently become possible via techniques such as topic extraction, named-entity recognition, sentiment and affect analysis, multilingual language analysis, social network analysis, statistical machine learning, and temporal data and text mining.

David Leinweber, an AI researcher by training and an early innovator in the application of modern IT in trading and investing, argued in his highly acclaimed book *Nerds on Wall Street* that future Wall Street development opportunities are in advanced electronic tools and understanding both quantitative and qualitative information.<sup>2</sup> Some practitioners believe that “stocks are stories and bonds are mathematics.” Advanced text analytics for mass and social media, when combined with well-grounded financial analytical models, could provide important new insights for understanding and predicting the markets.

In the IT community, this kind of data- and text-analytics-based approach to business analysis has gained significant attention and traction and is often referred to as business intelligence and analytics.<sup>3</sup> BI has been used as an umbrella term to describe concepts and methods to improve business decision making by using fact-based support systems. BI includes the underlying architectures, tools, databases, applications, and methodologies relevant to business decision making. As a data-centric approach, BI heavily relies on various advanced data collection, extraction, and analysis technologies. Since 2004, Web intelligence, Web analytics, Web 2.0, and user-generated content have begun to usher in a new and exciting era of Business Intelligence 2.0 research. Advanced information-extraction, topic-identification, opinion-mining, and time-series analysis techniques can be applied to traditional business information and the new BI 2.0 content for various accounting, finance, and marketing applications.

## Market Prediction

One of the most challenging areas of economics- and finance-related analytical research is in predicting stock performances. In the popular press and stock advisory columns, “beating the market” has often been considered the elusive Holy Grail for practitioners and the general public alike. American TV personality (on CNBC’s *Mad Money*), former hedge fund manager, and bestselling author Jim Cramer is one such example. According to an article on CNBC’s website titled “Mad Money Manifesto,” Cramer claims that the show’s mission and his job

is not to tell you what to think, but to teach you how to think about the market like a pro. This show is not about picking stocks. It’s not about giving you tips that will make you money overnight—tips are for waiters. Our mission is educational, to teach you how to analyze stocks and the market through the prism of events.

Can you really beat the market? Theoretical perspectives on stock behavior hold pessimistic assessments of the predictability of stock behavior. The famous Efficient Market Hypothesis argues that the price of a stock reflects all available information and the market reacts instantaneously, making it impossible to outperform the market. The Random Walk Theory states that the price of a stock varies randomly over time, so future prediction of the market is impossible.

Despite these theories, industry financial analysts have adopted two approaches to stock prediction. *Fundamentalists* utilize fundamental and financial measures of the economy, industrial sector, and firms to predict market and firm performances. For example, the Fama-French three-factor model considers market return, market capitalization, and book-to-market

ratio in its analysis.<sup>1</sup> On the other hand, *technicians* use historical time-series information of the stock and market behavior, such as historical price, volatility, and trading volume, to predict the market. In addition to standard regression-based analytical techniques, various machine learning methods have been adopted for financial analysis, including artificial neural networks, Bayesian classifiers, and support vector machines.

Recently researchers have incorporated firm-related news article measures. Computer scientists have developed trend-based language models, press release categorization (for example, good, bad, or neutral), and textual representation of news articles.<sup>4</sup> For example, using proper nouns and past stock price as representations and support vector regression (SVR) for analysis, the AZFin-Text system was able to outperform major quantitative (quant) funds during a five-week testbed period in 2005 based on 2,809 news articles and 10 million stock quotes.<sup>4</sup> Several studies have also attempted to correlate Web forums with stock behavior. Early studies focused on activity, without content analysis, and identified concurrent relationships. Subsequent research measured opinions in forum discussions and identified predictive relationships between forum discussion sentiment and subsequent stock returns, volatility, and trading volume.<sup>5,6</sup> With the widespread availability of Business Big Data and the recent advancement in text and Web mining, tremendous opportunities exist for computational and finance researchers to advance research relating to smart market and money.

## In This Issue

This T&C Department includes three article on smart market and money from distinguished experts in

information systems and business. Each article presents unique perspectives, advanced computational methods, and selected results and examples.

In “AZ SmartStock: Stock Prediction with Targeted Sentiment and Life Support,” Hsinchun Chen, Edward Chun-Neng Huang, Hsinmin Lu, and Shu-Hsing Li report on the design and testing of the AZ SmartStock system, which incorporates target sentiment and life support in a prototype stock-trading engine. We considered transaction costs and simulated trading performed using data collected for 129 trading days in 2008. The proposed trading model outperformed other benchmark models in the 10-day trading window. This article also suggests several directions for future research in predicting market movements.

In “A Stakeholder Approach to Stock Prediction Using Finance Social Media,” my colleague David Zimbra and I describe research that utilizes firm-related finance Web forum discussions to predict stock returns and trading of firm stock. Recognizing the diversity among forum participants, we segmented them into distinct stakeholder groups based on their interactions (posting activities) in the forums. By analyzing fine-grained stockholder groups, this system reported improved stock-return prediction versus a baseline system and aggregated forum model.

In the final article, “Computational Intelligence for Smart Markets: Individual Behavior and Preferences,” Paulo B. Goes argues that in today’s Web-enabled marketplaces, the economic environment is much more complex than the preference modeling used by experimental economists. The monitoring opportunities available with the Internet provide ample opportunities to build analytics

and computational intelligence to understand in real time the complexities of the preference structure and behaviors of today's heterogeneous market participants. Goes summarizes selected Web-based auction research that illustrates how to acquire computational intelligence on the preferences and behaviors of the participants in these new microeconomies.

### Acknowledgments

This material is based in part on work supported by the US National Science Foundation (NSF) under grant CNS-0709338 and CBET-0730908 and the Defense Threat Reduction Agency (DTRA) under award HDTRA1-09-0-0058.

### References

1. E. Fama and K. French, "Common Risk Factors in the Returns on Stocks and Bonds," *J. Financial Economics*, vol. 33, 1993, pp. 3–56.
2. D. Leinweber, *Nerds on Wall Street*, John Wiley & Sons, 2009.
3. H. Chen, "Business and Market Intelligence 2.0," *IEEE Intelligent Systems*, vol. 25, no. 1, 2010, pp. 68–71.
4. R. Schumaker and H. Chen, "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System," *ACM Trans. Information Systems*, vol. 27, no. 2, 2009, article no. 12.
5. W. Antweiler and M. Frank, "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," *J. Finance*, vol. 59, no. 3, 2004, pp. 1259–1295.
6. S. Das and M. Chen, "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web," *Management Science*, vol. 53, no. 9, 2007, pp. 1375–1388.

**Hsinchun Chen** is the director of the Artificial Intelligence Lab at the University of Arizona. Contact him at [hchen@eller.arizona.edu](mailto:hchen@eller.arizona.edu).

## AZ SmartStock: Stock Prediction with Targeted Sentiment and Life Support

Hsinchun Chen, *University of Arizona*  
Edward Chun-Neng Huang, *Microsoft*  
Hsin-Min Lu and Shu-Hsing Li,  
*National Taiwan University*

Stock market prediction has long been a challenging research topic. One research stream focuses on using text information to predict stock market movements. Robert Schumaker and Hsinchun Chen investigated various representations of text data in breaking financial news, in conjunction with past stock returns, for intraday stock price prediction.<sup>1</sup> Their experiments suggest that proper nouns can deliver the best prediction accuracy in simulated trading. Their experiments, nonetheless, did not consider news sentiment (good or bad) and ignored transaction costs, which could significantly reduce profitability in high-frequency intraday trading. Sentiment extracted from financial text data (such as forums and news articles) has been examined as a proxy of public opinions and correlated to stock market activities by prior studies.

In other research, Paul Tetlock adopted GI Inquirer, a popular sentiment dictionary, to evaluate the pessimism level of the daily "Abreast of the Market" column in the *Wall Street Journal*.<sup>2</sup> His results showed that the pessimism level was correlated with the downward pressure on market prices. Werner Antweiler and Murray Frank examined the relationship between sentiment in Yahoo Finance Web forum discussions and stock behaviors.<sup>3</sup> They developed a naïve Bayes-based sentiment classifier to classify forum messages into buy, hold, or sell. To represent aggregated

sentiment, they proposed a measure of disagreement that they found to be associated with stock volatility and trading volume. Using the event study framework, Antweiler and Frank later found that news events observed from the *Wall Street Journal* had statistically significant effects on cumulative abnormal stock returns in five- to 40-day windows.<sup>4</sup> Their results suggest that stock markets might need some time to fully absorb the impact of new information. However, their experiment did not consider the effect of repetitive news from multiple news sources.

We identified several research opportunities from previous studies. News sentiment can contain valuable information, but one news article might cover multiple companies. Thus, associating sentiment with the right company (targeted sentiment) presents a challenge. Old, often-repetitive news might also contain significantly less value and could be processed differently if we consider an aging theory (life support). Lastly, prior studies have not systematically studied the impact of trading windows and often ignored transaction costs.

Our work addresses these challenges by developing a text-based stock-prediction engine with targeted sentiment and life-support considerations in a real-world financial setting. This article presents our system design and summarizes our findings.

### System Design

Our inter-day trading experiments follow Antweiler and Frank's work,<sup>4</sup> with five-, 10-, 20-, and 40-day trading windows. For a given day  $t$ , our system collects news articles published between 16:00 on day  $t - 1$  and 16:00 on day  $t$ . We use the information in this text collection to

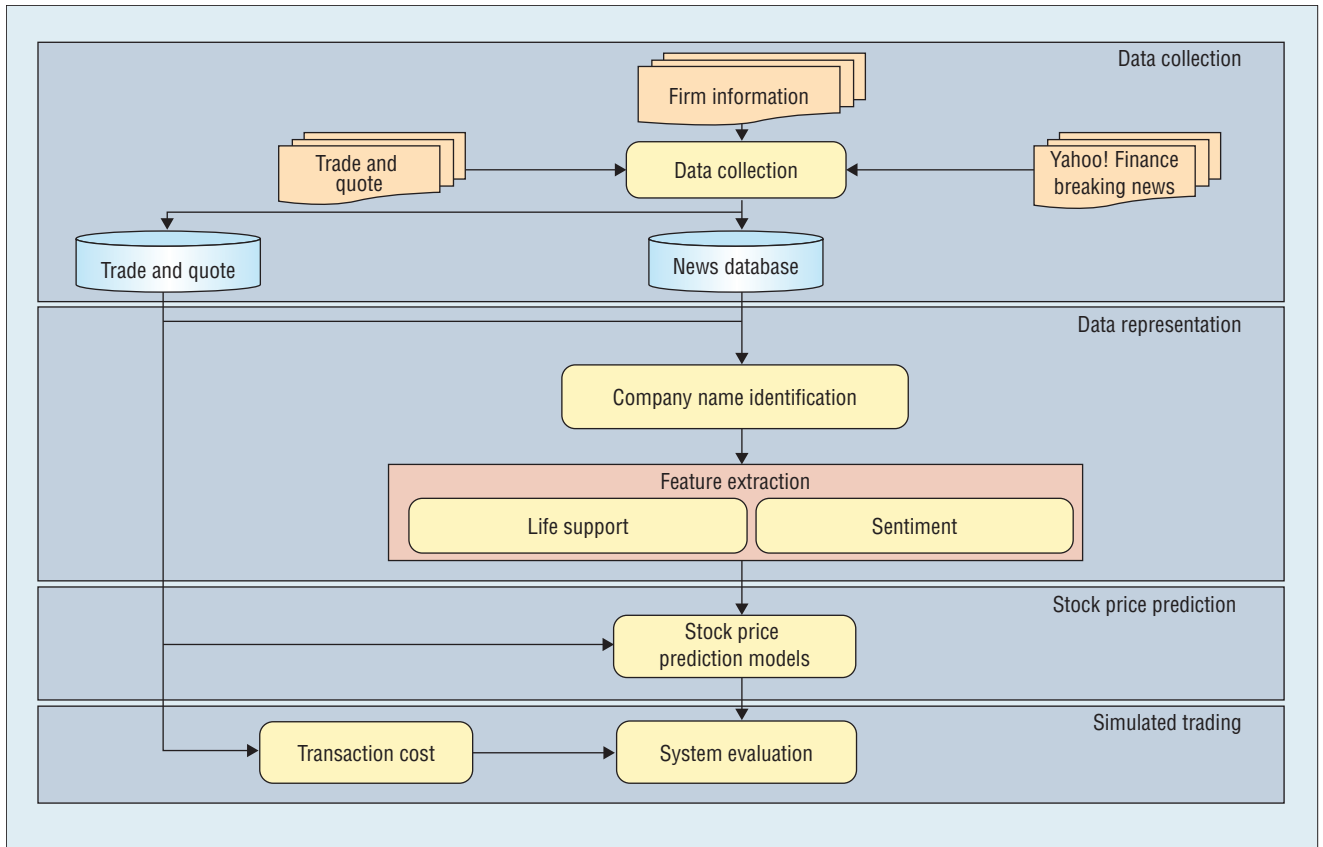


Figure 1. The AZ SmartStock system design. The four main components are data collection, data representation, stock price prediction, and simulated trading.

predict future stock prices. Trading decisions are made based on the predicted prices. Figure 1 presents our AZ SmartStock system design, which consists of four main components: data collection, data representation, stock price prediction, and simulated trading.

### Data Collection

Both historical news articles and stock market data need to be collected for our text analytics. We focus on S&P 500 firms to minimize the potential illiquid problem associated with thinly traded stocks. We extracted news articles from major newswires on Yahoo Finance. We collected high-frequency trading data from the Wharton Research Data Services (WRDS). This dataset was also used to estimate bid-ask spreads associated with the transactions suggested by our stock price prediction module.

### Data Representation

We identified company names in news articles using the well-established Stanford Named Entity Recognizer (NER) tool (<http://nlp.stanford.edu/software/CRF-NER.shtml>). The extracted named entities were then compared with the set of S&P 500 company names on that date. The matching process started from the tightest rule (exact match) to the loosest one (partial name match).

To extract the *specific sentiments* toward a given company from the news articles, we used companies identified from the previous step as sentiment targets. We then adopted a sentiment dictionary developed by Tim Loughran and Bill McDonald specifically for financial text analytics, to classify words in news articles into positive and negative categories.<sup>5</sup> Six negation rules were implemented following Loughran and McDonald's work.<sup>5</sup>

Our sentiment-extraction approach works at the sentence level to link companies with specific-sentiment words. The algorithm was developed based on two assumptions:

- sentiments about a company will be expressed after the company has been mentioned and
- companies are more likely to be associated with sentiment words closer to them.

Unlike sentiment analysis, *life support* aims to extract the degree of topic novelty. Following previous work,<sup>6</sup> a company's life support comes from the aggregated energy (novelty) of terms used in the news articles where the company is mentioned. The energy is measured by the degree to which the distributions of terms have changed. We do this by first observing the changing proportion of individual terms in

**Table 1. Variables for determining the energy score of a term  $k$ .**

Day	In news sources	In other news sources
$t$	$a_s$	$b_s$
$t-1$	$c_s$	$d_s$

each news source. Specifically, the energy score of a term  $k$  on day  $t$  is given by summing the chi-square statistics of different news sources:

$$E_{k,t} = \sum_s \frac{(a_s + b_s + c_s + d_s)(a_s d_s - b_s c_s)^2}{(a_s + b_s)(c_s + d_s)(a_s + c_s)(b_s + d_s)}$$

Table 1 gives values of  $a_s$ ,  $b_s$ ,  $c_s$ , and  $d_s$ .

Then, we transform the energy scores of terms into life-support values using a sigmoid function. The life support score  $LS_{g,t}$  of a company  $g$  on day  $t$  are the aggregation of life supports of all terms in the news articles where  $g$  has been mentioned.  $LS_{g,t}$  will decay over time by a constant decay function,  $LS'_{g,t} = \text{Max}(LS_{g,t} - \beta, 0)$ , where  $\beta$  is the decay nutrition factor, an empirical constant.

We further propose *adjusted life support*, which combines life support and sentiment scores that represent the joint effect of topic novelty and opinion directions of news events. The adjusted life support score of a company  $g$  on day  $t$ ,  $\text{AdjLS}_{g,t}$ , is defined as  $LS_{g,t} * S_g$ . The total adjusted life support of company  $g$  on day  $t$  is the accumulative adjusted life support ( $\text{TotalAdjLS}_{g,t}$ ), which is calculated as  $\text{AdjLS}_{g,t} + \text{AdjLS}'_{g,t-1} + \dots + \text{AdjLS}'_{g,0}$ .

### Stock Price Prediction

Although there are many predictive approaches available, we adopted support vector regression (SVR) in this research, similar to a prior study.<sup>1</sup> SVR are an extension of support vector machines, which are known to deliver excellent prediction performance in classifying discrete outcomes. Our firm-level stock-prediction models

considered past stock returns and text data in past news articles when making predictions for future stocks. To study the effect of different text representations, we adopted several model specifications. The first model, a baseline (M-Reg), predicts future returns using past returns:

$$R_{t+n,t} = \langle \alpha, R_t \rangle + b$$

where  $\langle \bullet, \bullet \rangle$  denotes the dot product,  $R_t = (\text{Price}_t - \text{Price}_{t-1}) / \text{Price}_{t-1}$  denotes the stock return on day  $t$ , and  $R_{t+n,t} = (\text{Price}_{t+n} - \text{Price}_t) / \text{Price}_t$  denotes the stock return on the following  $n$  days.

The second model, sentiment model (M-Senti), includes additional sentiment-related information as input. We capture two aspects of news sentiment in the model. The first aspect is the absolute sentiment on day  $t$ :  $\text{Sentiment}_t = \text{POS}_t - \text{NEG}_t$ , where  $\text{POS}_t$  is the number of positive words appearing on day  $t$  and  $\text{NEG}_t$  is the number of negative words appearing on day  $t$ . The second aspect of sentiment is the volume of sentiment words:  $\text{SentiWords}_t = \text{POS}_t + \text{NEG}_t$ . Combining the two sentiment variables, our second model can be written as follows:

$$R_{t+n,t} = \langle \alpha, (\text{Sentiment}_t, \text{SentiWords}_t, R_t) \rangle + b$$

The third model, life-support model (M-LS), was included to investigate the effect of life-support scores. Instead of including sentiment scores, here we added life-support scores to the baseline regression model:

$$R_{t+n,t} = \langle \alpha, (\text{TotalAdjLS}_t, \text{SentiWords}_t, R_t) \rangle + b$$

All models were trained using 60 days of historical data with a linear kernel for SVR.

### Simulated Trading

We aggregated news articles by day and conducted prediction at the closing time (16:00) of a trading day. We experimented with the holding periods of five, 10, 20, and 40 days. This study considered transparent transaction costs and a bid-ask spread. The transparent transaction cost was set to \$4.95 per order (\$9.90 for a roundtrip) according to the online broker TradeKing. In addition, we also considered two regulatory fees that occur during transactions: a Securities and Exchange Commission fee, which is the principal amount times 0.0000169 and applied to the sale transactions of all equities, and a trading activity fee, which is 0.000075 per share for equity sells with a maximum charge of \$3.75 per sale transaction. We also considered the bid-ask spread by matching the quote records with the trade records (which we don't discuss in detail here due to space limitations).

Our study incorporates two simple trading strategies: buy and short. The *buy strategy* is to purchase and hold stocks for a period of time. The *short strategy* short-sells stocks first and buys them back later. Therefore, the buy strategy gets a positive return when the stock price increases, while the short strategy gains positive returns for stocks with decreasing prices.

At the end of each trading day, trading decisions of different holding periods were made independently. For each holding period, the top five stocks with the highest predicted positive and negative returns are traded with a \$1,000 investment for each stock. Positions were held for a pre-specified trading window and closed on the last day of the trading window. We used three measures to evaluate the performance of our prediction engine: mean square error (MSE), directional

accuracy (DA), and return of simulated trading. We report selected DA and return results here.

## Experimental Results

We evaluated the performance of our AZ SmartStock system using 114,347 news articles published from 1 January 2008 to 31 October 2008 (129 trading days during the 2008 US financial collapse) from the Yahoo Finance Breaking News section. We compared the performance of M-Reg, M-Senti, and M-LS prediction models.

Table 2 gives the overall DA results. The performances of the three models did not differ significantly from one another. Overall, M-LS achieved the best directional accuracy (52.05 percent) after holding the stock for 10 days. The 10-day holding window consistently performed better than the five-, 20-, and 40-day holding windows.

Simulated trading was conducted to evaluate each model's profitability. To compare the returns for different trading windows, we converted all the reported returns into returns of the entire testing time period, 129 days. Holding the S&P 500 during this period of time yielded a return of -30.10 percent (during the 2008 US financial collapse).

Table 3 and Figure 2 summarize the trading returns of the three models across different trading windows. The results show that M-LS had the best performance across all trading windows—albeit all the models achieved negative returns during the 2008 financial collapse. The best performance was obtained with M-LS for a 10-day trading window, resulting in a return of -0.35 percent. M-Senti came in the second place in most cases. In addition, all models outperformed the S&P 500 index in all trading windows. The 10-day trading

Table 2. Directional accuracy (DA) results.

Prediction model	Directional accuracy (%)			
	5 days	10 days	20 days	40 days
Baseline (M-Reg)	50.88	51.67	48.66	42.26
Sentiment (M-Senti)	50.79	51.76	48.44	42.38
Life support (M-LS)	50.90	52.05	42.98	44.01

Table 3. Trading returns across different trading windows.

Prediction model	5 days (%)	10 days (%)	20 days (%)	40 days (%)
M-Reg	-10.32	-3.09%	-5.67	-17.07
M-Senti	-12.51	-0.71	-5.28	-15.72
M-LS	-9.95	-0.35	-4.02	-13.77

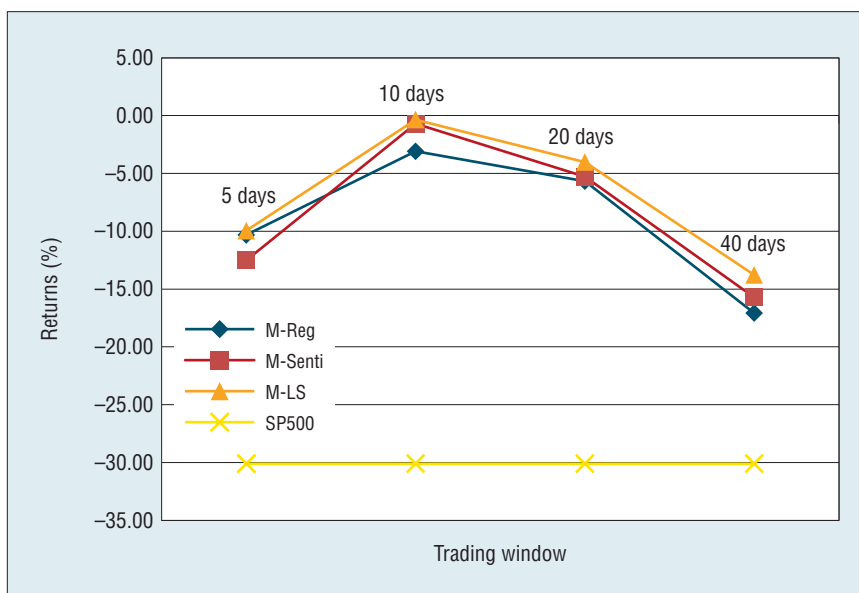


Figure 2. Comparison of different prediction models: the baseline model (M-Reg), sentiment model (M-Senti), and life support model (M-LS). M-LS had the best performance across all trading windows. In all trading windows, all the models outperformed the S&P 500 index, which yielded a -30.10 percent return during this period.

window resulted in significantly better returns than other trading windows, which is consistent with the DA results.

Despite our promising results, there are several caveats thus far. Our results might have been heavily influenced by the highly volatile and disastrous financial events in 2008. A more systematic experimentation of other, longer time periods involving ups and downs in the financial

markets are necessary. In addition, it would be interesting to perform an additional sensitivity analysis of each model's selected best and worst stock recommendations to better understand the reasons for the success and failure of different techniques. We also hope to better interpret the decision rules or knowledge derived from the AZ SmartStock text analytics so we can make more informed decisions about

future market movements and firm performances.

## Acknowledgments

This material is based in part upon work supported by the US National Science Foundation (NSF) under grants CNS-0709338 and CBET-0730908 and the US Department of Defense (DOD) under award HDTRA1-09-0-0058. We thank all members of the Business Intelligence research team at the AI Lab.

## References

1. R. Schumaker and H. Chen, "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System," *ACM Trans. Information Systems*, vol. 27, no. 2, 2009, article no. 12.
2. P. Tetlock, "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *J. Finance*, vol. 62, no. 3, 2007, pp. 1139–1168.
3. W. Antweiler and M. Frank, "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," *J. Finance*, vol. 59, no. 3, 2004, pp. 1259–1295.
4. W. Antweiler and M. Frank, "Do US Stock Markets Typically Overreact to Corporate News Stories?" working paper, 2006, [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=878091](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=878091).
5. T. Loughran and B. McDonald, "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *J. Finance*, vol. 66, no. 1, 2011, pp. 35–65.
6. K. Chen, L. Luesukprasert, and S.T. Chou, "Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling," *IEEE Trans. Knowledge and Data Eng.*, vol. 19, no. 8, 2007, pp. 1016–1025.

**Hsinchun Chen** is the director of the Artificial Intelligence Lab at the University of Arizona. Contact him at [hchen@eller.arizona.edu](mailto:hchen@eller.arizona.edu).

**Edward Chun-Neng Huang** is a software design engineer in test (SDET) at Microsoft. Contact him at [edhuang@microsoft.com](mailto:edhuang@microsoft.com).

**Hsin-Min Lu** is an assistant professor at the National Taiwan University. Contact him at [lu@im.ntu.edu.tw](mailto:lu@im.ntu.edu.tw).

**Shu-Hsing Li** is the dean of the College of Management at the National Taiwan University. Contact him at [shli@management.ntu.edu.tw](mailto:shli@management.ntu.edu.tw).

## A Stakeholder Approach to Stock Prediction Using Finance Social Media

David Zimbra and Hsinchun Chen,  
*University of Arizona*

The prediction of a firm's stock return has long been of interest to researchers in the finance and machine learning disciplines. Traditionally, the firm's financial measures and historical stock behavior information are utilized for prediction. More recently, researchers have demonstrated improved performance by including professional news articles on the firm.<sup>1</sup> Finance Web forums on the firm have also been revealed as valuable sources in explaining subsequent stock behavior,<sup>2</sup> although few studies have leveraged them in a true predictive context performing simulated trading based on the extracted information. This article examines the inclusion of measures extracted from the discussion of a firm within a finance Web forum in the simulated trading of firm stock for one year.

### Stock-Return Prediction and Finance Web Forums

Although the two prominent theoretical perspectives on stock market behavior, the efficient market hypothesis and the random walk theory, provide pessimistic assessments of its predictability, researchers have demonstrated in empirical studies that

stock-return prediction might be possible. Stock trading philosophies and the prediction of return have generally followed the fundamentalist and technician approaches. According to the fundamentalist approach, a stock's price is determined by the fundamental and financial measures of the economy, industry, and firm. Technicians utilize historical time-series information of the stock and market behavior to predict future returns. In simulated trading, fundamentalist strategies correspond to longer waiting times before reacting to new information on the firm, while technicians respond quickly to capture profits before the market fully absorbs the new information into the stock price.<sup>3</sup>

Following the technician philosophy, researchers have integrated professional news articles into the predictive models since new information on the firm is often released through the press, improving model performance by capitalizing on the time lag before investors react.<sup>1</sup> Approaches to the automated analysis of the news articles generally take two forms, where the article content is represented as textual features directly applied to learn the relationship with stock return or by performing sentiment analysis on the article, such as classifying good, bad, and neutral news, and using the derived sentiment measures to predict return.

With similar motivations, researchers have examined the relationships between discussions in firm-related finance Web forums, such as Yahoo Finance, and subsequent stock behavior.<sup>2</sup> Studies have revealed these forums provide significant explanatory power of subsequent firm stock return, with unique information not covered in the professional news. In addition to considering forum activity, such as message-posting volume,

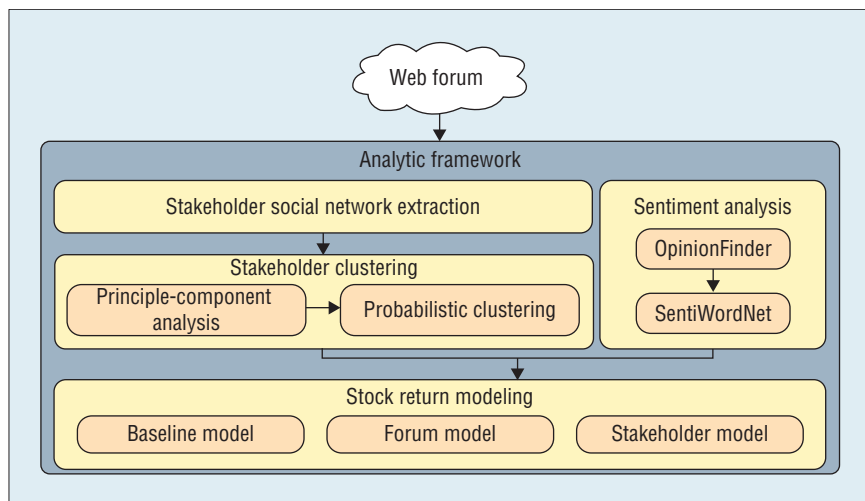
researchers have developed “bullishness” classifiers to perform sentiment analysis and classify messages representing investors’ buy, hold, and sell positions. Bullishness classifiers typically performed modestly, with 60 to 70 percent accuracy attributed to a noisy forum environment.<sup>2</sup> Although investors and shareholders are specifically targeted by these classification schemes, forums hosted on major Web platforms can attract a diverse collection of the firm’s stakeholders. Thus, the diversity of the forum population might be responsible for the unimpressive performance in classifying investment positions.

Departing from the prior literature, we adopt a stakeholder, rather than shareholder, perspective of forum participants and segment them into stakeholder groups to be assessed individually. According to the stakeholder theory of the firm, satisfying the interests of a diverse collection of stakeholders, beyond those of the shareholders, is essential to success. Following that theory, a stakeholder approach to forum analysis recognizes that various stakeholder groups might have distinctive information in explaining firm stock return. Although few studies have leveraged firm-related finance Web forums in a true predictive context, this research performs simulated trading of firm stock based on the measures extracted from online discussions.

### Web Forum Analysis and Stock-Return Modeling

Figure 3 shows the analytic framework we applied in this research. The framework’s four stages include stakeholder social network extraction, stakeholder clustering, sentiment analysis, and stock modeling.

Web forum participants form a social network through their interactions in discussions. Subgroups are a



**Figure 3. Analytic framework.** The framework consists of four stages: stakeholder social network extraction, stakeholder clustering, sentiment analysis, and stock modeling.

nontrivial structural feature of social networks, and identifying such subgroups in the forum social network might reveal distinctive stakeholder perspectives within the population. To perform the stakeholder segmentation through clustering, we first extract the forum social network and represent it using by an interaction matrix. In this research, posting messages in the same discussion thread constitutes a relationship between participants. Each time participants interact in a discussion thread, their relationship is strengthened.

Because stakeholders often belong to more than one stakeholder group, we apply a probabilistic clustering approach to group-related forum participants. Specifically, a finite mixture model is utilized, with the expectation-maximization algorithm for estimating parameter values. We determine the number of clusters to represent the stakeholder groups in the forum using maximum likelihood estimation and cross validation. To ensure the independence of attributes included in the mixture model, we extract principle components from the interaction matrix prior to the clustering. Principle-component analysis also serves to reduce the highly dimensional feature space.

After clustering, stakeholders are represented by their probabilistic assignments to each of the identified clusters throughout the stock modeling.

Similar to prior studies on firm-related finance Web forums and stock behavior,<sup>2</sup> we perform sentiment analysis on the forum messages. However, unlike the bullishness classifiers devised to interpret investor communications, we apply a more general approach to evaluate the sentiments expressed by various stakeholder groups. Specifically, we use the Opinion Finder (OF) system for subjectivity analysis<sup>4</sup> and Senti WordNet (SWN) lexicon<sup>5</sup> in a combined fashion for sentiment analysis. Lexical approaches to sentiment analysis are general, but they lack contextual knowledge of the specific usage of terms, which might result in erroneous application. To incorporate contextual information and ensure the correct SWN entry is applied based on the specific usage of the term in a subjective statement, we apply OF prior to assigning an SWN score. OF enables targeted application of SWN and provides measures of subjectivity to enrich the analysis.

To evaluate the application of firm-related finance Web forums in the prediction of daily stock return, we



Table 4. Summary of the Yahoo Finance Forum data.

Forum	Messages	Discussion threads	Stakeholders	Messages per thread	Messages per stakeholder
Yahoo Finance (finance.yahoo.com), Wal-Mart (WMT)	134,201	40,633	5,533	3.3	24.25

Table 5. Major stakeholder groups in the March 2006 prediction model.

Stakeholder group	Forum participants (%)	Top word bigrams	Top word trigrams
1	53	wall street	http news yahoo
		long term	http finance messages
		bottom line	moneycentral msn com
		holiday season	long-term sentiment strong
		wmt stock	quote profile research
2	11	sales growth	higher gas prices
		middle class	wal-mart super center
		small town	www nlpc org
		shop wal-mart	ap wal mart
		people want	pay labor costs
4	23	million people	pay health care
		mom pop	choose spend money
		work wal-mart	wal-mart distribution center
		united states	employee health care
		past years	everyday low prices

compare the performance of a baseline model consisting of well-established fundamental and technician explanatory variables to models that incorporate forum measures. The dependent variable in all models is the log-difference in the daily close price of stock. The independent variables included in the baseline model are the Fama-French factors: market return; book-to-market ratio; market capitalization; two lagged terms of prior daily stock return, volatility, and trading volume; and dummy variables for the day of the week. In addition to the baseline variables, the forum-level model includes six forum measures characterizing the forum discussions during the prior day (using the trading day definition of 16:00 to 16:00): the number of messages posted, average message length,

and average and variance in sentiment and subjectivity.

Instead of forum-level measures, the stakeholder-level model incorporates six measures for each of the stakeholder groups identified in the clustering. All stock-return models utilize support vector regression, calibrated with five months of daily historical information and applied for the prediction of daily return for one month of trading days. For each month during the year of trading, we perform stakeholder clustering to assess the current state of the forum and estimate new stock-return models. To identify the most relevant variables and ensure parsimonious prediction models, feature selection is performed prior to model calibration using a correlation-based approach similar to step-wise regression;

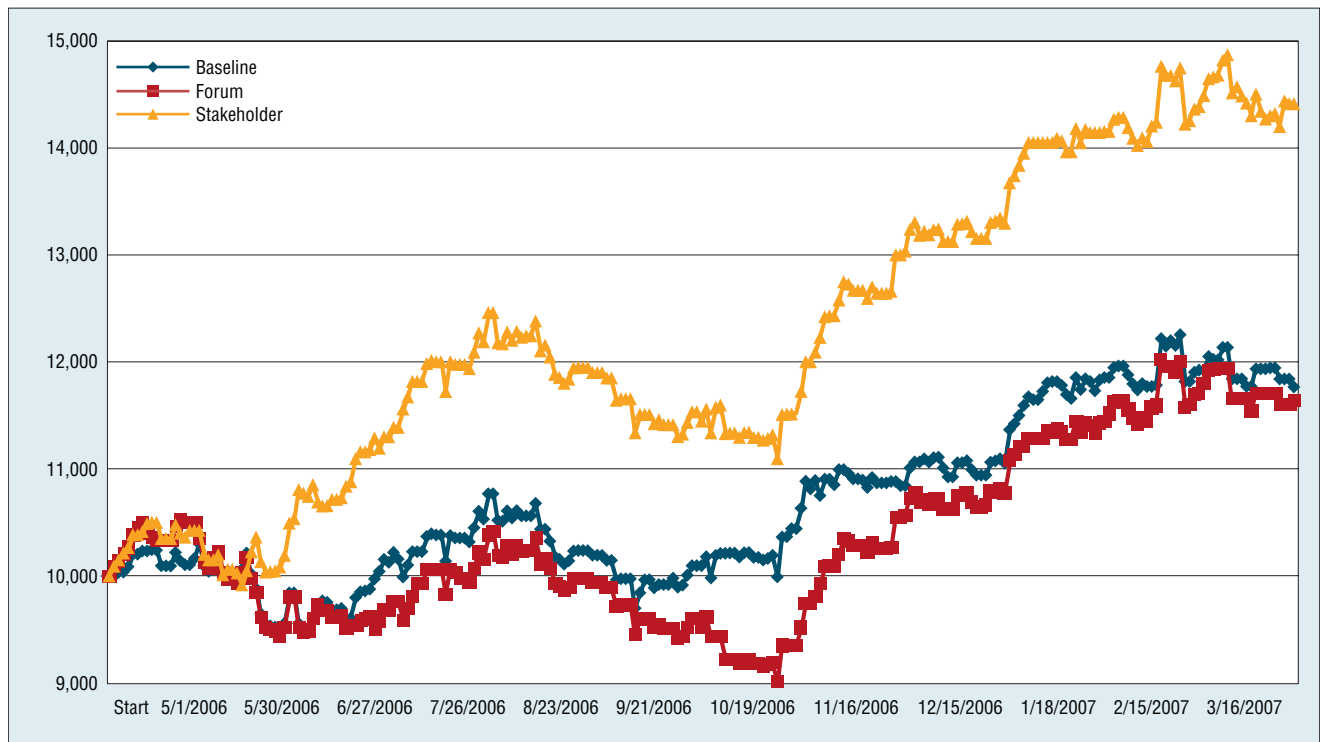
the algorithm seeks subsets of input variables with high correlation with the dependent variable, and low inter-correlation. During the prediction period, participants in the forum are classified according to the stakeholder groups established in the clustering. In the simulated trading, firm stock is bought or shorted and sold on a trading day based on each model's daily stock-return predictions.

## Experimentation and Results

For our experiment, we selected Wal-Mart for stock-return modeling due to its prominence in the market and diverse and active collection of stakeholder groups. Prior studies have examined Yahoo Finance firm-related Web forums, which we use in this research. We performed Wal-Mart stock trading based on our models' stock-return predictions for one year (250 trading days) from 1 March 2006 to 28 February 2007. In total, we used 17 months of data, including the 5 months required to calibrate the first prediction models prior to trading. Table 4 summarizes the Yahoo Finance forum data in the study.

To construct the stakeholder-level stock-return prediction models, we performed clustering each month during the year of trading to identify the current stakeholder groups in the forum and assign participants to their appropriate mixture of groups. For example, to develop the first prediction model for March 2006, the stakeholder clustering was based on the previous five months of forum data (from 1 October 2005 to 28 February 2006). Eight stakeholder groups were identified, three of which were considered major with more than 10 percent of forum participants.

Table 5 gives the top word bigrams and trigrams representing the three



**Figure 4. Investment values during one year of simulated trading. The stakeholder model produced a 27 percent increase over the baseline annual return in simulated trading.**

major stakeholder groups (ranked by their term frequency). Stakeholder group 1 seemed to consist of technical investors, heavily engaged in the exchange of news. Group 2 primarily discussed Wal-Mart’s growth and its impact on consumers and communities. Stakeholder group 4 seemed to be made up of employees conversing on work and healthcare related issues.

In the simulated trading of Wal-Mart stock, each model began with an initial investment of \$10,000. According to each model’s daily predictions, if the anticipated daily stock return was greater than 0.1 percent (or less than -0.1 percent), the Wal-Mart stock was bought (shorted). Unchanged positions on consecutive trading days were held; otherwise, the stock was sold. Although prior studies typically disregard trading costs, for additional realism, we incorporated an \$8 charge per transaction.

Table 6 presents the results of the experiment. For each of the prediction models, we report the final value of the

**Table 6. Results of stock-return prediction models and simulated trading.**

Model	Directional accuracy (%)	Ending investment value (%)	Annual return (%)
Baseline	56.8	\$11,767	17
Forum	58.0	\$11,643	16
Stakeholder	61.2*	\$14,413	44

\*Pair-wise t-test; improvement over baseline model at  $p < 0.10$ .

investment after completion of trading during the year as well as directional accuracy, the performance in correctly predicting the direction of the daily stock return, positive or negative.

As an additional point of reference, holding the Wal-Mart stock for the year would have resulted in an ending investment value of \$10,096, providing an annual return of less than 1 percent. Results from the experimentation and simulated trading revealed each prediction model performed well, with better than 50 percent directional accuracy, and earned substantial profit in simulated trading of Wal-Mart stock. Forum-level variables incorporated into the model, however, provided little

improvement over the baseline. Only after stakeholder segmentation and extraction of group-level measures from the forum was the improvement in directional accuracy over the baseline statistically significant. The stakeholder model also produced a 44 percent annual return in simulated trading, an impressive 27 percent increase over the baseline. Figure 4 depicts the investment values for each model over the year of trading.

This research shows that recognizing the true diversity among forum participants, segmenting them into stakeholder groups based on their interactions in the forum social network, and assessing them independently

refined the measures extracted from the forum and improved stock-return prediction. The impressive performance of the stakeholder-level model represented a statistically significant improvement over the baseline in directional accuracy and provided an annual return of 44 percent in simulated trading of firm stock.

This study contributes to the emerging trend in stock-return prediction research to incorporate additional sources of text-based information (professional news articles or online social media) into the models to improve performance. However, the study has several limitations, including the generality of the findings since the analysis covered only one firm for one year and the consistency of the approach under different market conditions and industries. In addition to expanding the number of firms, industries, and time periods included in our future research, we intend to examine shorter time windows for stock-return prediction and higher-frequency trading.

### Acknowledgments

Funding for this research was provided by US National Science Foundation grants CNS-0709338 and CBET-0730908 and Defense Threat Reduction Agency (DTRA) grant HDTRA1-09-1-0058.

### References

1. R. Schumaker and H. Chen, "Textual Analysis of Stock Market Prediction using Breaking Financial News: The AZFinText System," *ACM Trans. Information Systems*, vol. 27, no. 2, 2009, article no. 12.
2. S. Das and M. Chen, "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web," *Management Science*, vol. 53, no. 9, 2007, pp. 1375–1388.
3. B. LeBaron, W. B. Arthur, and R. Palmer, "Time Series Properties of an Artificial Stock Market," *J. Economic*

*Dynamics and Control*, vol. 23, no. 9–10, 1999, pp. 1487–1516.

4. T. Wilson et al., "OpinionFinder: A System for Subjectivity Analysis," *Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing*, Assoc. for Computational Linguistics, 2005, pp. 34–35.
5. A. Esuli and F. Sebastiani, "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining," *Proc. Conf. Language Resources and Evaluation*, 2006.

**David Zimbra** is a doctoral student in management information systems at the University of Arizona. Contact him at zimbra@email.arizona.edu.

**Hsinchun Chen** is the director of the Artificial Intelligence Lab at the University of Arizona. Contact him at hchen@eller.arizona.edu.

## Computational Intelligence for Smart Markets: Individual Behavior and Preferences

Paulo B. Goes, *University of Arizona*

In an influential work for the experimental economics field, Nobel Prize winner Vernon Smith established the concept of a microeconomic system as consisting of an economic environment, together with an economic institution (or economic mechanism).<sup>1</sup> The economic environment is the preferences of the people in the economy. In the context of experimental economics, Kevin A. McCabe, Stephen J. Rassenti, and Vernon L. Smith introduced the concept of "smart computer-assisted markets."<sup>2</sup> The "smartness" of the markets came from the real-time computation achieved by the optimization algorithms to derive prices and determine winners in the auction experiments.

In the years that followed, the term "smart market" continued to be applied to computational and optimization efforts of price and winner determination involved in combinatorial auctions, usually formulated through complex math programming allocation problems. In a recent research commentary, Martin Bichler, Alok Gupta, and Wolfgang Ketter repositioned the term "smart markets" in the more general context of computational intelligence for decision making by market participants.<sup>3</sup>

In today's Web-enabled marketplaces, the economic environment is more complex than the preference modeling used by experimental economists. The monitoring opportunities available with the Internet provide ample opportunities to build analytics and computational intelligence to understand in real time the complexities of the preference structure and behaviors of today's heterogeneous market participants. This article uses the context of Web-based auctions to illustrate how to acquire computational intelligence on participants' preferences and behaviors in these new microeconomies.

### Online Auctions Bidding Behavior

Over the last two decades, online auction sites, such as eBay and samsclub.com, have developed sophisticated user interfaces that let bidders search the entire marketplace of all available auctions, compare posted price alternatives, place bids, and monitor bidding on auctions of interest. These enhanced user interfaces, the long duration of the auctions, the heterogeneous population of bidders who join and leave an auction at will, and the knowledge acquisition opportunities of the Internet have pushed the boundaries and redefined Smith's economic environment.

Bidders' preference models must account for all the additional behavioral characteristics of the bidder's interaction with the economic institution (mechanism) that could impact an auction's outcome.

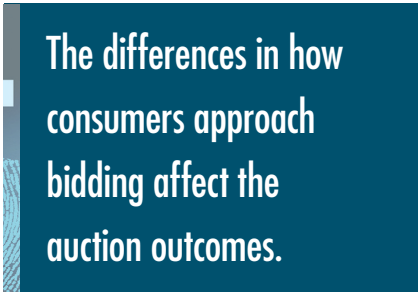
Unlike the constrained experimental economics environment, where bidders are assumed to be homogeneous, a great deal of heterogeneity exists among bidders in the online environment. Bidders are smart consumers or resellers who approach the bidding process in different ways, affected by how they process information, their participation cost structure, and their purchasing intentions and demand. For example, in the same auction, we can find experienced small retailers who are purchasing items for resale and individuals who are buying for personal consumption. Insights from bidding behavior of heterogeneous consumers are critical to designing auction mechanisms, which directly affect the auctioneers' revenue, the appeal of the marketing channel to the consumers, and the efficiency of the goods-consumers allocation.

### Determining Bidders' Participation Profiles

A typical consumer-oriented online auction lasts from several hours to several days. This helps attract more bidders, who are unknown beforehand to the auctioneer, and lets the auctioneer promote the auction to multiple time zones. Ravi Bapna and his colleagues used clustering techniques on detailed bidding data available from the interactions between bidders and auctioneers.<sup>4</sup> *Time of entry* (TOE) captures the time of the first bid placed by the bidder in the auction, while *time of exit* (TOX) clocks the time of the last bid by the bidder in the auction. NOB is the *total number of bids* placed by the bidder in the auction between TOE and

TOX. Together, these variables capture bidders' degree of participation in an auction and can also be associated with the bidders' cognitive preference of how they approach the interactive bidding process. Through a k-means cluster analysis on a large dataset from Yankee auctions, the authors were able to identify the following general participation profiles or strategies:

- A *participator* is a bidder with low TOE, high TOX, and high NOB. This bidder monitors the auction throughout its duration and bids actively.
- An *evaluator* is a bidder with relatively low or medium TOE, which generally coincides with the TOX.



The differences in how consumers approach bidding affect the auction outcomes.

This bidder places close to one bid (NOB = 1). This is the profile of bidders who know how much they want to pay, probably have good knowledge about the item and its common value, and do not want to spend time actively participating throughout the auction.

- An *opportunist* is a bidder with a NOB close to 1, but who also has a high TOE and high TOX. These bidders place their only bids toward the end of the auction, looking for a bargain. This general strategy has also been identified in the eBay environment as “sniping.”

The differences in how consumers approach bidding affect the auction

outcomes. In the Yankee auction that utilizes a popular online multi-item, ascending, discriminatory mechanism, evaluators tend to realize the lowest levels of surplus, while opportunists have a higher probability of winning. The actual mix of profiles present in any given auction dictates the competitive nature of the bidding and the final outcome.

This bidding-strategy classification, which was found using data mining, has proven robust. Using data from various online auctions in different environments—consumer to consumer (C2C), business to consumer (B2C), and even B2B—similar clusterings were obtained.

The Internet has also opened opportunities for retailers to increasingly explore the concept of sequential auctions to dispose of a large inventory of items. Dell and Sam's Club, for example, routinely conduct a series of auctions selling identical items through their own auction sites to liquidate large inventories. These sequences of auctions often span several weeks. Bidders in such sequences have the opportunity to participate in multiple auctions of the same item, and as they do, they learn from the experience and fine tune their bidding behavior to maximize their payoffs. This evolution in bidding behavior affects the nature of demand in these auctions.

In an earlier work, my colleagues and I introduced the variable *time since last auction* (TLA) to measure how long the repeat bidder waited between auctions of the same sequence to bid again.<sup>5</sup> With the enhanced tuple [TOE, TOX, NOB, TLA], we performed cluster analyses on extensive datasets of sequential auctions and confirmed the original categorization of evaluators, participators, and opportunists, but with a recurrence dimension. Table 7 shows the results. New clusters with high TLA

Table 7. Mean (standard deviation) of behavioral clusters of bidders participating in sequential auctions.

Clusters/strategies	Bidders adopting strategy				
	Percentage	No. of bids	Time of entry	Time of exit	Time since last auction
<b>Recurrent strategies</b>					
Early evaluators (EE-R)	22.76	1.17 (0.53)	1.83 (1.08)	1.92 (1.13)	0.66 (2.49)
Middle evaluators (ME-R)	24.56	1.20 (0.48)	5.55 (1.06)	5.64 (1.07)	0.62 (1.60)
Opportunists (O-R)	36.89	1.15 (0.41)	9.04 (0.88)	9.20 (0.77)	0.77 (2.71)
Participators (P-R)	5.45	2.56 (0.84)	2.83 (1.72)	8.97 (1.13)	0.49 (1.44)
<b>Intermittent strategies</b>					
Early evaluators (EE-I)	2.73	1.15 (0.50)	1.90 (1.05)	1.97 (1.14)	6.40 (7.97)
Middle evaluators (ME-I)	2.82	1.40 (0.70)	5.41 (1.49)	6.22 (1.49)	8.86 (10.56)
Opportunists (O-I)	4.79	1.11 (0.38)	9.17 (0.82)	9.30 (0.71)	9.04 (12.11)

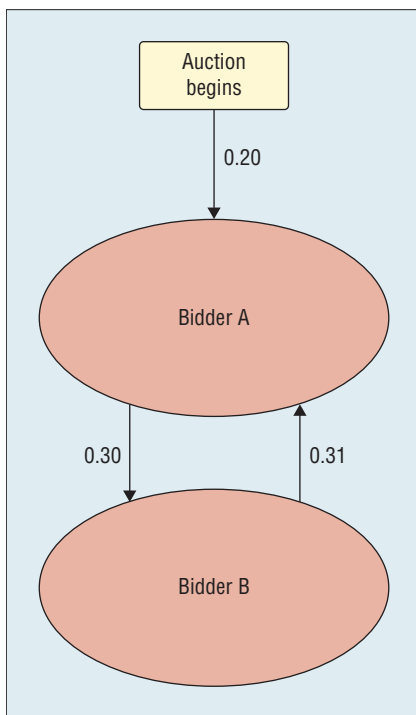


Figure 5. The outbid-by graph. Example with two bidders. In the bidding sequence, Bidder A eventually outbids Bidder B at time 0.31.

levels were identified as “intermittent strategies,” in which evaluators and opportunists intermittently opted to skip auctions. The vast majority of repeat bidders were assigned to clusters

that reflected their recurrent participation, not skipping auctions in the sequence.

By applying longitudinal data analytics, we were able to capture the effect of learning and experience as bidders repeat their participation throughout the auction sequence.<sup>5</sup> We also revealed that there is demand heterogeneity among the participants of sequential auctions. Individual bidders who want to purchase only one item coexist with resellers who are participating to purchase multiple items. These two types of consumers learn from their experiences differently. Individual buyers start bidding using early or middle evaluator strategies and tend to try out various strategies over time. As they accumulate more experience, they shift into using opportunist behavior. Resellers tend to stick with opportunist or late-bidding strategies throughout the auction sequence.

### Understanding Bidders’ Behaviors

Additional analysis of the actual competitive behavior of bidders within each auction is possible using a network structure called the *outbid-by graph*. Each node of the

graph is a bidder who is a winner at some point in the auction. There is a direct arc between bidder  $i$  and bidder  $j$ , if bidder  $j$  directly outbid bidder  $i$ . Cycles in the graphs indicate competition activity. The numbers along each arc display the normalized time of the bids, so we can track the timing of the competitive interactions. For example, suppose bidder A places the first bid in an auction at time 0.2 (when 20 percent of the auction duration has elapsed). At time 0.3, bidder B bids higher and is now the current winner. At time 0.31, bidder A reacts and outbids bidder B. This bidding sequence defines a cycle in the graph (see Figure 5).

Figure 6 shows two real eBay auctions using outbid graphs. In Figure 6a, bidder “bizzywildcat” places the first bid at time 0.35. He or she is outbid by “ivannada” at time 0.55 but regains the item at time 0.58. Three bidders subsequently become winners at times 0.63, 0.77, and 0.86, respectively. A bidding war takes place at the end of the auction between “tikitude11” and “terrible88,” who ends up winning the auction.

In Figure 6b, most of the bidding activity takes place toward the end

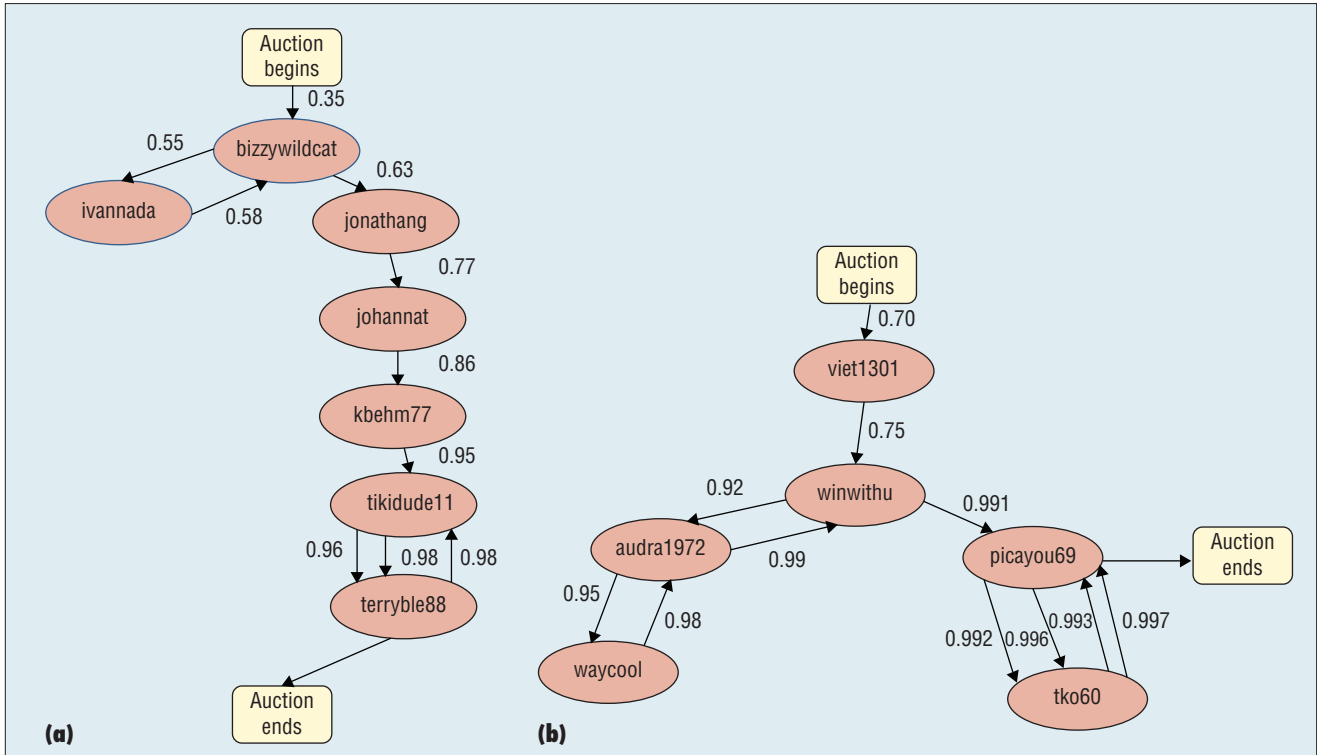


Figure 6. Outbid-by graphs of actual eBay auctions. (a) In the first auction, various bidders outbid one another early on, and then a bidding war between “tikidude11” and “terryble88” ultimately determines the winner. (b) In the second example, most of the bidding activity takes place toward the end of the auction.

of the auction (after 92 percent of total duration), with several bidders engaging in a bidding war. The graph in Figure 6a shows a more linear bidding process with fewer competitive flare outs, while the one in Figure 6b displays a much more competitive auction, especially toward the end.

Constructed in real time, the outbid graph can reveal patterns of competition and detailed characterization of the bidding process. It provides the basis for advanced data mining analyses such as graph clustering and sequence mining, with the idea of identifying and characterizing bidding patterns and whole auction patterns that are useful for decision making, both for the auctioneers and potential bidding agents that can aid the bidder.

Overall, in wide-open, online, consumer-oriented auctions, it is important to realize that heterogeneous bidders with different levels of experience,

different purposes, and different bidding behavior coexist throughout online auctions. The specific mix of behavior clusters present at any individual auction together with the competitive interactions will determine the outcome of the auction.

**A**uctioneers have the tools to monitor participants’ interactions in real time, perform analytics, and use the knowledge obtained to set the marketplace and auction parameters to optimize their objectives. Bidders or third-party agents can also closely monitor an auction’s progress. I anticipate more advanced bidding tools will be created that can help bidders with the process of selecting bidding strategies that optimize their desirable outcomes. In the eBay environment, for example, third-party vendors are already offering sniping agents to regular bidders.

Future environments where all participants are armed with computational intelligence to guide their decisions will generate a sophisticated computational game. We are not far from such an environment today. In small focused B2B environments such as energy and telecommunication marketplaces, this is already happening. Still, there will always be a heterogeneity of consumers with different levels of expertise and experience in general, open consumer-oriented markets.

In just two decades, the Internet has developed into a massive collection of interconnected virtual sensors that capture all kinds of information about consumers—what they do, where they are, what they think, and how they communicate with each other. Connectivity is everywhere, so the monitoring capabilities of tweets, social media, websites, blogs, news feeds, and mobile

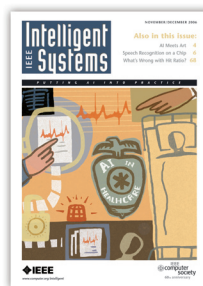
# Call for Articles

Be on the Cutting Edge of Artificial Intelligence!

Publish Your Paper  
in IEEE Intelligent Systems

IEEE Intelligent Systems  
seeks papers on all aspects  
of artificial intelligence,  
focusing on the development  
of the latest research into  
practical, fielded applications.

For guidelines, see  
[www.computer.org/mc/  
intelligent/author.htm](http://www.computer.org/mc/intelligent/author.htm).



devices are limitless. Monitoring platforms can be built that lead to computational intelligence of the underlying economic environment with accurate modeling of participants' behaviors and intentions. Designers of smart marketplaces and their participants should carefully consider the existence of these capabilities. ■

## References

1. V. Smith, "Microeconomic Systems as an Experimental Science," *Am. Economic Rev.*, vol. 72, no. 5, 1982, pp. 923–955.
2. K. McCabe, S. Rassenti, and V. Smith, "Smart Computer-Assisted Markets," *Science*, vol. 254, no. 5031, 1991, pp. 534–538.
3. M. Bichler, A. Gupta, and W. Ketter, "Research Commentary: Designing Smart Markets," *Information Systems Research*, vol. 21, no. 4, 2010, pp. 688–699.
4. R. Bapna et al., "User Heterogeneity and its Impact on Electronic Auction Market Design: An Empirical Exploration," *Management Information Systems Q.*, vol. 28, no. 1, 2004, pp. 21–43.
5. P. Goes, G. Karuga, and A. Tripathi, "Bidding Behavior Evolution in Sequential Auctions: Characterization and Analysis," working paper, Univ. of Arizona, 2011.

**Paulo B. Goes** is the Salter Distinguished Professor of Management and Technology and head of the Management Information Systems Department at the University of Arizona. Contact him at [pgoes@email.arizona.edu](mailto:pgoes@email.arizona.edu).

**cn** Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.

The #1 AI Magazine  
[www.computer.org/intelligent](http://www.computer.org/intelligent)

IEEE  
Intelligent  
Systems