Expert Systems With Applications 79 (2017) 153–163

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Stock market one-day ahead movement prediction using disparate data sources

Bin Weng[a], Mohamed A. Ahmed[a], Fadel M. Megahed[b,c,*]

[a] *Department of Industrial and Systems Engineering, Auburn University, AL 36849, USA*
[b] *Farmer School of Business, Miami University, OH, 45056, USA*
[c] *Center for Analytics and Data Science, Miami University, OH 45056, USA*

ARTICLE INFO

Article history:
Received 24 March 2016
Revised 25 February 2017
Accepted 26 February 2017
Available online 28 February 2017

Keywords:
Artificial intelligence
Feature selection
Financial expert system
Google news
R programming
Wikipedia

ABSTRACT

There are several commercial financial expert systems that can be used for trading on the stock exchange. However, their predictions are somewhat limited since they primarily rely on time-series analysis of the market. With the rise of the Internet, new forms of collective intelligence (e.g. Google and Wikipedia) have emerged, representing a new generation of “crowd-sourced” knowledge bases. They collate information on publicly traded companies, while capturing web traffic statistics that reflect the public’s collective interest. Google and Wikipedia have become important “knowledge bases” for investors. In this research, we hypothesize that combining disparate online data sources with traditional time-series and technical indicators for a stock can provide a more effective and intelligent daily trading expert system. Three machine learning models, decision trees, neural networks and support vector machines, serve as the basis for our “inference engine”. To evaluate the performance of our expert system, we present a case study based on the AAPL (Apple NASDAQ) stock. Our expert system had an 85% accuracy in predicting the next-day AAPL stock movement, which outperforms the reported rates in the literature. Our results suggest that: (a) the knowledge base of financial expert systems can benefit from data captured from nontraditional “experts” like Google and Wikipedia; (b) diversifying the knowledge base by combining data from disparate sources can help improve the performance of financial expert systems; and (c) the use of simple machine learning models for inference and rule generation is appropriate with our rich knowledge database. Finally, an intelligent decision making tool is provided to assist investors in making trading decisions on any stock, commodity or index.

1. Introduction

Stock market prediction has attracted much attention from both academia and business. The question remains: “To what extent can the past history of a common stock’s price be used to make meaningful predictions concerning the future price of the stock?” (Fama, 1965). Early research on stock market prediction was based on the Efficient Market Hypothesis (EMH) (Fama, 1965) and the random walk theory (Cootner, 1964; Fama, 1991, 1995; Fama, Fisher, Jensen, & Roll, 1969). These early models suggested that stock prices cannot be predicted since they are driven by new information (news) rather than present/past prices. Thus, stock market prices will follow a random walk and their prediction accuracy cannot exceed 50% (Bollen, Mao, & Zeng, 2011).

There has been an increasing number of studies (see e.g., Bollen et al., 2011; Malkiel, 2003; Nofsinger, 2005; Prechter Jr & Parker, 2007; Smith, 2003) that provide evidence contrary to what is suggested by the EMH and random walk hypotheses. These studies show that the stock market can be predicted to some degree and therefore, questioning the EMH’s underlying assumptions. Many within the business community also view Warren Buffet’s ability to consistently beat the S&P index (Kiersz, 2015; Loomis, 2012) as a practical indicator that the market can be predicted.

The significant stock market movements (i.e. spikes) over short horizons cannot also be explained by the EMH. These spikes are often driven by investor perceptions of a certain stock based on information (news) collected from disparate data sources. As an illustration, on April 23, 2013, at 1:07 p.m., Eastern Time, a tweet from the *Associated Press* (AP) account stated “Breaking: Two Explosions in the White House and Barack Obama is injured” (Megahed & Jones-Farmer, 2015). The fraudulent tweet, which originated from the hacked AP Twitter account led to an immediate drop in the

[*] Corresponding author.
E-mail addresses: bzw0018@auburn.edu (B. Weng), mza0068@auburn.edu (M.A. Ahmed), fmegahed@miamioh.edu (F.M. Megahed).

http://dx.doi.org/10.1016/j.eswa.2017.02.041
0957-4174/

Table 1
A review of financial expert systems that are used in stock movement prediction. ANN, GA, SVM and DT correspond to artificial neural network, genetic algorithm, decision tree and support vector machine, respectively.

Paper	Sources for knowledge base			
	Traditional	Crowd-sourcing	News	AI approach
Kimoto et al. (1990)	✓			ANN
Lee and Jo (1999)	✓			Time series
Kim and Han (2000)	✓			ANN, GA
Kim (2003)	✓			SVM
Qian and Rasheed (2007)	✓			ANN, DT
Li and Kuo (2008)	✓			ANN
Schumaker and Chen (2009)			✓	SVM
Vu et al. (2012)		✓		DT
Chen, Chen, Fan, and Huang (2013)	✓			ANN
Adebiyi, Adewumi, and Ayo, (2014)	✓			ANN, ARIMA
Nguyen, Shirai, and Velcin (2015)		✓		SVM
Shynkevich, McGinnity, Coleman, and Belatreche (2015)			✓	ANN, SVM
Chourmouziadis and Chatzoglou (2016)	✓			Fuzzy system
Our financial expert system	✓	✓	✓	ANN, SVM, DT

Dow Jones Industrial Average (DJIA). Although the DJIA quickly recovered following an AP retraction and a White House press release, this example illustrates the immediate and dramatic effects of perception/news on stock prices.

While the news may be unpredictable, some recent literature suggests that early indicators can be extracted from online sources (e.g., Google Trends and blogs) to predict changes in various economic indicators. For example, Google search queries have been shown to provide early indicators of disease infection rates and consumer spending (Choi & Varian, 2012). Schumaker and Chen (2009) showed that breaking financial news can be used to predict stock market movements. Bollen et al. (2011) used measurements of collective mood states derived from large-scale Twitter feeds to predict the daily up and down changes in the DJIA. In addition, Moat et al. (2013) observed that the frequency of views of Wikipedia's financially-related pages can be an early indicator of stock market moves. The authors hypothesized that investors may be using such pages as a part of their decision making process. This work was extended in Preis, Moat, and Stanley (2013) to include data from the number of relevant searches from Google Trends, and model the effect of search volume on trading behavior. Note that Mao, Counts, and Bollen (2011) indicated that *search* and *usage* are more predictive than *survey sentiment indicators*.

From an expert systems perspective, the stock market prediction problem can be divided into two components: (1) what information and predictors need to be tracked as a part of our "knowledge base"; and (2) what artificial intelligence (AI) algorithms can be used for effective rule generation and predictions. The literature discussed in the previous paragraph indicate that online sources that capture the "collective intelligence" of investors should be an integral component of a financial expert system's knowledge base. It is important to note that these online sources are not typically used in financial expert systems. Instead, the knowledge base of such systems typically rely on the historical prices of a stock and/or technical indicators extracted from a time-series analysis of stock prices (Chourmouziadis & Chatzoglou, 2016; Hassan, Nath, & Kirley, 2007; Kim, 2003; Kim & Han, 2000; Kimoto, Asakawa, Yoda, & Takeoka, 1990; Lee & Jo, 1999; Lin, Yang, & Song, 2011; Qian & Rasheed, 2007). We hypothesize that combining the expert's knowledge from online sources with features extracted from the price and technical indicators will offer a more accurate representation of the dynamics that affect a stock's price and its movement. Since these data sources were never combined in the context of financial expert systems, it is important to examine which AI algorithms are the most effective in translating the knowledge base into accurate predictions. Table 1 categorizes financial expert systems used for stock movement prediction based on their "knowledge base" and the AI approach used. From Table 1, it is clear that the all those papers relied on a single source for the knowledge base. The reader should note that there is a limited number of expert systems (e.g., Bollen et al., 2011) that combined traditional sources with crowd-sourced experts' data; however, they are not included in our table since they predicted price (i.e. a continuous outcome instead of our binary outcome). The integration of diverse data sources can improve the knowledge base (see Alavi and Leidner, 2001; Hendler, 2014 for a detailed discussion) and thus, improving the performance of the expert system.

Based on the insights from Table 1 and the discussion above, we outline a novel methodology to predict the future movements in the value of securities after tapping data from disparate sources, including: (a) the number of page visits to pertinent Wikipedia pages; (b) the amount of online content produced on a particular day about a company, the stock of which is publicly traded; and (c) commonly used technical indicators and company value indicators in stock value prediction. In the AI component of our expert system, we compare the performance of ANN, SVM and DT for stock movement prediction. We have chosen these three specific approaches since: (i) neural networks have been widely deployed in intelligent trading systems (Bollen et al., 2011; Guresen, Kayakutlu, and Daim, 2011; Kimoto et al., 1990; Li and Kuo, 2008; (ii) SVM was successfully used by Kim and Han (2000) and Schumaker and Chen (2009); and (iii) decision trees have been effectively used in crowd-sourced expert systems (Vu, Chang, Ha, & Collier, 2012). In those papers, the authors reported that these AI models outperformed the more traditional approaches. However, it is unclear whether: (1) such results will hold for our predictions since our knowledge base is more diverse, and (2) the results will hold when predicting different stocks and indices. Thus, our expert system will evaluate the performance of these models and select the best approach for a given prediction problem.

To demonstrate the utility of our system, we predict the one-day ahead movements in AAPL stocks over a three year period. Based on our case study, we show that the combination of online data sources with traditional technical indicators provide a higher predictive power than any of these sources alone. The remainder of the paper is organized as follows. In Section 2, we present a detailed description of the methodology we used to extract the data from the online sources, the variable selection techniques employed, and the corresponding predictive models. In Section 3, we highlight the main results and offer our perspective on their importance/interpretation. Our concluding remarks and recommendations for future work are provided in Section 4. In Appendix A–Appendix C, we explain how Google News data was captured, present the formulas for our generated features, and de-

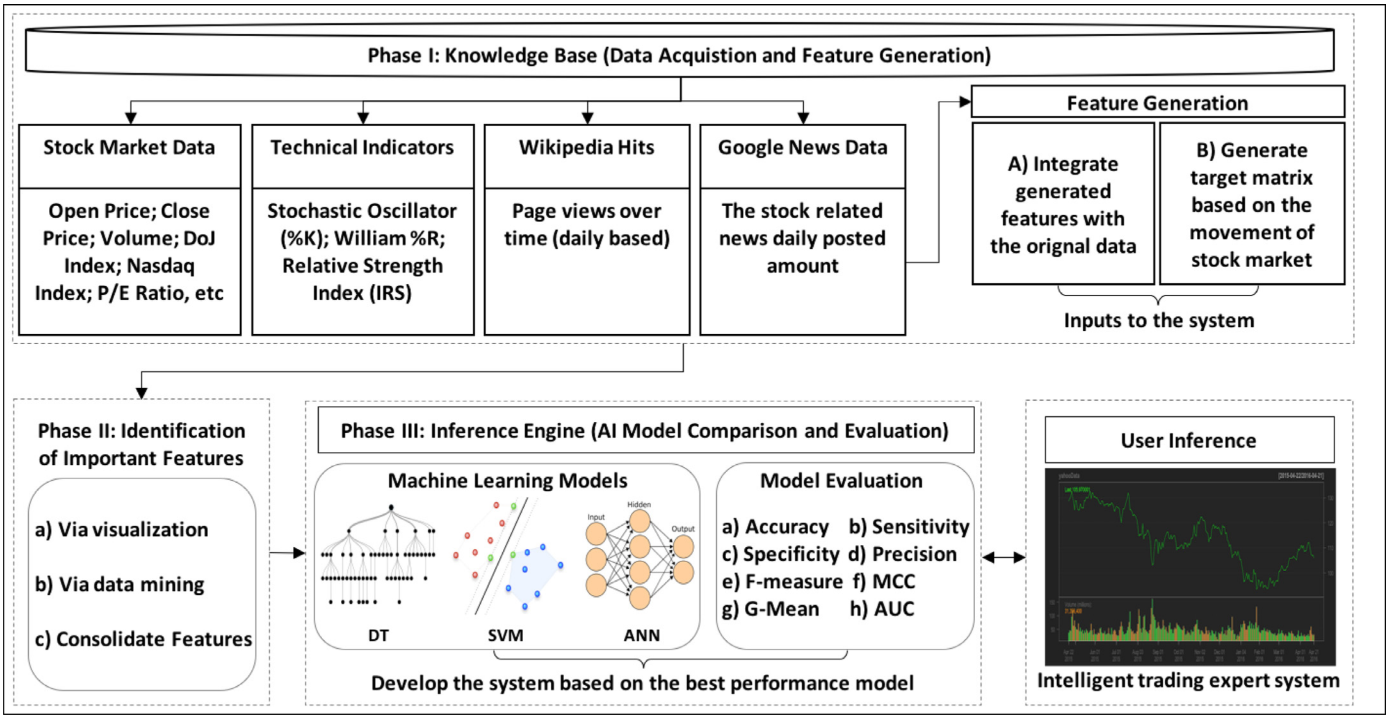

Phase I: Knowledge Base (Data Acquistion and Feature Generation)
Stock Market Data
Open Price; Close Price; Volume; DoJ Index; Nasdaq Index; P/E Ratio, etc
Technical Indicators
Stochastic Oscillator (%K); William %R; Relative Strength Index (IRS)
Wikipedia Hits
Page views over time (daily based)
Google News Data
The stock related news daily posted amount
Feature Generation
A) Integrate generated features with the orignal data
B) Generate target matrix based on the movement of stock market
Inputs to the system
Phase II: Identification of Important Features
a) Via visualization
b) Via data mining
c) Consolidate Features
Phase III: Inference Engine (AI Model Comparison and Evaluation)
Machine Learning Models
DT
SVM
ANN
Input
Hidden
Output
Model Evaluation
a) Accuracy b) Sensitivity
c) Specificity d) Precision
e) F-measure f) MCC
g) G-Mean h) AUC
Develop the system based on the best performance model
User Inference
Intelligent trading expert system

Fig. 1. An overview of the proposed method.

fine the predictors identified from our variable selection steps. We also present a copy of our full dataset, code and prediction tool at https://github.com/binweng/ShinyStock.

2. Methods

To predict stock movements, we propose a data-driven approach that consists of three main phases, as shown in Fig. 1. In Phase I, we scrape four sets of data from online resources. These datasets include: (a) publicly available market information on stocks, including opening/closing prices, trade volume, NASDAQ and the DJIA indices, etc.; (b) commonly used technical indicators that reflect price variation over time; (c) daily counts of Google News on the stocks of interest; and (d) the number of unique visitors for pertinent Wikipedia pages per day. We also populated additional features (i.e. summary statistics) in an attempt to uncover more significant predictors for stock movement. In Phase II, we use variable selection methods to select a subset of predictors that provide the most predictive power/accuracy. Then, in Phase III, we utilize three AI techniques to predict stock movement. These models are compared and evaluated based on a 10-fold cross validation sample using *the area under the operating characteristics curve (AUC)* and seven other metrics. Based on the evaluation, we select an appropriate model for real-time stock market prediction. We present the details for each of the phases in the subsections below.

2.1. Data acquisition and feature generation for our “knowledge base”

In this paper, we focus on predicting the AAPL, Apple NASDAQ, stock movement based on a 37 month-period from May 1, 2012 to June 1, 2015. There are four datasets that were obtained, preprocessed and merged in Phase I. First, we obtain publicly available market data on AAPL using the *Yahoo Finance* website. We considered the following common predictors of stock prices (see e.g., Jasemi, Kimiagari, & Memariani, 2011; Lee & Jo, 1999; Li & Kuo, 2008; Wang, 2002): the daily opening and closing prices, daily high/low, and volume of trades of the AAPL stock. In addition, we included the day-to-day movements in the DJIA and NASDAQ composite indices as indirect measures of risk that the AAPL stock is subject to due to the general market movements. We also used the price to earnings ratio (P/E) as an estimate for the fundamental health of the company (Gabrielsson & Johansson, 2015).

The second set of predictors is comprised of three indicators that are used in technical analysis. Technical analysis is used to forecast future stock prices by studying historical prices and volumes (Chourmouziadis and Chatzoglou, 2016. Since all information is reflected in stock prices, it is sufficient to study specific technical indicators (created by mathematical formula) to predict price fluctuations and evaluate the strength of the prevailing trend (Bao & Yang, 2008). In this paper, we consider three technical indicators:

(A) Stochastic oscillator (%K), developed by George C. Lane as a momentum indicator that can warn of the strength or weakness of the market. When the market is trending upwards, it tries to measure when the closing price would get close to the lowest price in a given period. On the other hand, when the market is trending downwards, it estimates when the closing price would get close to the highest price in the given period. For additional details on the %K and its calculation, the reader is referred to: Bao and Yang (2008) and Lin et al. (2011).

(B) The Larry William (LW) % R indicator - It is a momentum indicator that facilitates the spotting of overbought and oversold levels. For its calculation, refer to Kim and Han (2000).

(C) The Relative Strength Index (RSI)- Similar to the LW %R, it compares the magnitude of recent gains to recent losses in an attempt to determine overbought and oversold conditions of an asset. RSI ranges from 0 to 100. In practice, investors sell if its value is $\geq$ 80 and buy if it is $\leq$ 20. For more details, see Bao and Yang, 2008 and Lin et al., 2011.

The reader should note that the values for these three technical indicators were calculated based on the market price data obtained from *Yahoo Finance*.

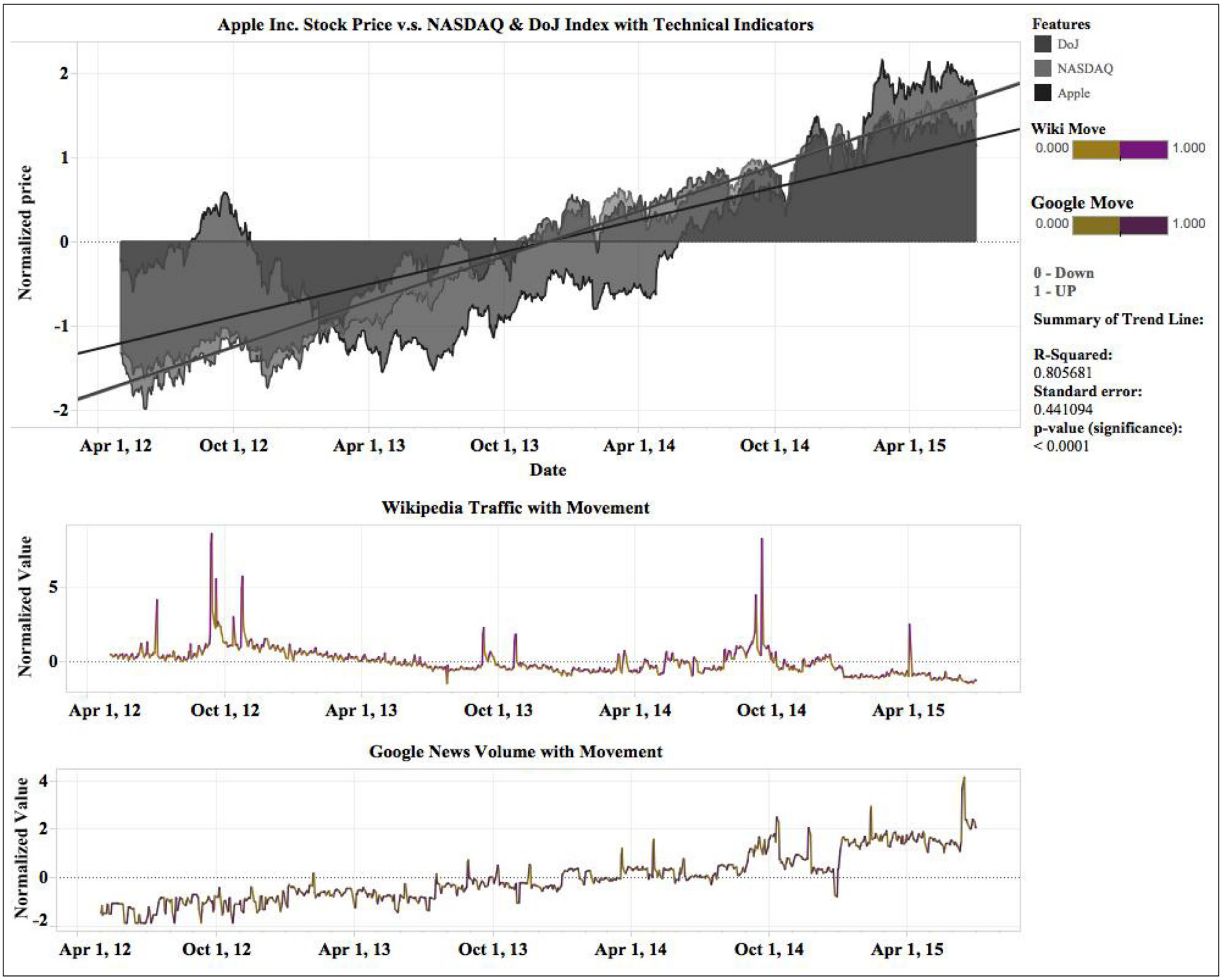


Fig. 2. A visual summary of the main predictors from the four data sources. An interactive version of this plot is available at: https://goo.gl/fZSQEy. Note that we rescaled the variables (by subtracting the mean and dividing by the standard deviation) to facilitate the visualization of the data.

Table 2
One-day-ahead targets used in this paper.

Target	Formula
Target 1	$Open(i+1) - Close(i)$
Target 2	$Open(i+1) - Open(i)$
Target 3	$Close(i+1) - Close(i)$
Target 4	$Close(i+1) - Open(i)$
Target 5	$Trade\ Volume(i+1) - Trade\ Volume(i)$

In the third data source, we scrape the amount of daily online content produced about a company, and its products/services. In this paper, we obtain a count for aggregated news and blogs based on the daily count of content on Google News. We detail this step in Appendix A. The fourth and final data source is based on the Wikipedia page view counts of terms related to Apple stock (AAPL, Apple Inc., iPhone, iPad, Macbook, and Mac OS). We queried the daily visits for these pages from www.wikipediatrends.com. A graphical summary of the second and third set of predictors is provided in Fig. 2.

To enhance the performance of the predictive models, we generate some additional features from the four predictor sets. We incorporate some of the underlying principles behind technical analysis (see e.g., Bao & Yang, 2008) to generate our feature set. Therefore, our generated features include: Wikipedia Momentum, Wikipedia Rate of Change, Google Momentum, Google Relative Strength Index, and three moving averages of stock prices (where $n = 3, 5,$ and 10, respectively). For the sake of completion, we explain how each of these features are calculated in Appendix B.

2.2. Variable/feature selection

The end goal of this phase is to have the data processed for the artificial intelligence models. This phase is comprised of two steps. First, we define different one-day-ahead outcomes (hereafter targets). Then, we use *recursive feature elimination* (RFE) to select the features/variables that offer the highest predictive power.

There are several one-day-ahead outcomes that can be of interest to investors. We examine five different targets. These targets are defined in Table 2. Target 1 compares the opening stock price of day $i+1$ with the closing price of the previous trading day. In Target 2, we compare the opening stock price of day $i+1$ with the opening price of the previous trading day. Targets 3 and 4 follow a similar logic with the closing price used for day $i+1$ instead of the opening price. In Target 5, we examine the differences in trade volume between day $i+1$ and day i. It is important to note that

Table 3
The twenty most predictive variables/features for each target.

Target	Variables/features selected				
Target 1	Close	Open	High	Low	P/E ratio
	Wiki_3_day_disparity	Wiki_5_day_disparity	Wiki_10_day_disparity	Wiki_Momentum_1	Wiki_ROC
	Google_MA_5	Google_EMA_3	Google_3_Day_disparity	Google_5_day_disparity	RSI
	Stochastic Oscillator (%K)	Wiki_RSI	Google_MA_4	William %R	Google_MA_3
Target 2	Close	Open	High	Low	P/E ratio
	Wiki_5_day_disparity	Wiki_Move	Wiki_MA3_Move	Wiki_EMA5_Move	Wiki_5day_disparity_Move
	Google_EMA5_Move	Google_3day_disparity_Move	Google_ROC_Move	Google_RSI_Move	Wiki_3_day_disparity
	Stochastic Oscillator (%K)	RSI_Move	Wiki_RSI_Move	Google_MA_6	Google_Move
Target 3	Close	Open	High	P/E Ratio	Stochastic_Move
	Wiki_Monentum_1	Wiki_Move	Wiki_MA3_Move	Wiki_EMA5_Move	Wiki_ROC_Move
	Google_EMA5_Move	Google_3day_disparity_Move	Google_ROC_Move	Google_RSI_Move	Wiki_10_day_disparity
	RSI_Move	Wiki_RSI_Move	Wiki_3_day_disparity	Google_Move	Google_MA5_Move
Target 4	Close	Open	High	Low	P/E ratio
	RSI_Move	Wiki_10_day_Disparity	Wiki_Move	Wiki_MA3_Move	Wiki_EMA5_Move
	Google_Move	Google_3day_disparity_Move	Google_ROC_Move	Google_RSI_Move	William %R
	Stochastic Oscillator (%K)	Stochastic_Move	Wiki_3day_disparity_Move	Wiki_ROC_Move	Wiki_RSI_Move
Target 5	Close	Open	High	Low	William %R
	Wiki_Monentum_1	Wiki_RSI	Google_MA_2	Google_MA_3	Google_MA_4
	Google_MA_9	Google_3_day_disparity	Google_5_day_disparity	Google_10_day_disparity	Wiki_10_day_disparity
	Wiki_3_day_disparity	Wiki_5_day_disparity	Google_MA_6	Google_MA_7	Google_MA_8

we only calculate these targets for the AAPL stock as a case study. In addition, we have transformed all targets to a binary variable where 0 → no increase in target, and 1 → an increase in the target value from the previous day.

In Step 2, we selected the significant features using the SVM *recursive feature elimination* (RFE) algorithm. RFE is implemented through backwards selection of predictors based on predictor importance ranking. The predictors are ranked and the less important ones are sequentially eliminated prior to deploying our three predictive models. The goal of this step is to find a subset of predictors that can result in accurate predictions without overfitting. It should be noted that we used the SVM-RFE algorithm for each target. Thus, we have obtained different predictor sets for each target. These predictors are presented in the Table 3 in Section 3. For more details on how RFE can be deployed using open-source programming languages, the readers are referred to the **R** Package *Caret* (Kuhn, 2008) and to the *sklearn.feature_selection* module in *Python* (Scikit-Learn-Developers, 2014).

2.3. The inference engine: AI model comparison and evaluation

In this phase, we compare the effectiveness of artificial neural networks (ANN), decision trees (DT), and support vector machines (SVM) for predicting movements in the AAPL stock based on the predictors identified in Section 2.2. In the paragraphs below, we first introduce how we used a 10-fold cross validation approach to minimize the sampling bias. Then, we provide a very short overview of the three classification approaches, and introduce the performance evaluation metrics used to identify the most suitable approach. The reader should note that, in this paper, we deploy the described methodology for each of the five targets. Hereafter, we use the term *dataset* to reflect each set of features/variables with its associated target for the AAPL stock over the 37 months of the study.

The k-fold cross-validation approach is used to minimize the bias associated with the random sampling of the training and test data samples (Kohavi, 1995). The entire dataset is randomly split into k mutually exclusive subsets of approximately equal size. The prediction model is tested k times by using the test sets. The estimation of the k-fold cross validation for the overall performance criteria is calculated as the average of the k individual performances as shown in Dag, Topuz, Oztekin, Bulur, and Megahed (2016). In our analysis, we use the stratified 10-fold cross validation approach to estimate the performance of the different classification models. Our choice for $k = 10$ is based on literature results (see e.g., Dag et al., 2016; Kohavi, 1995) that show that 10-folds provide an ideal balance between performance and the time required to run the folds.

ANNs are widely employed in a wide variety of computational data analytics problems that include classification, regression and pattern recognition. In the context of stock market prediction, ANNs have been extensively applied in predicting stocks and indices at different markets (see Atsalakis and Valavanis, 2009; Bollen et al., 2011; Dase and Pawar, 2010; Guresen et al., 2011; Hassan et al., 2007; Kim and Han, 2000; Zhang and Wu, 2009, and the references within). We assume that the reader is familiar with ANNs and their construction (otherwise, refer to Hastie, Tibshirani, & Friedman, 2011). In this paper, we use the sigmoid function as the activation function for our ANN. We have also used the Multilayer Perceptron (MLP) learning model with a back-propagation algorithm due to its superior performance to the radial basis function (RBF) in our preliminary analysis.

Decision trees are widely used in several data mining and stock market prediction problems (Atsalakis & Valavanis, 2009; Lai, Fan, Huang, & Chang, 2009; Qian & Rasheed, 2007) since they are very easy to interpret. The modeling procedure starts with splitting the dataset into several subsets each of which consists of more or fewer homogeneous states of the target variable (Breiman, Friedman, Stone, & Olshen, 1984). Then the impacts of each independent variable on the target variable are measured. This procedure takes place successively until the decision tree reaches a stable state. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5 (Quinlan, 1986; 2014) and C&RT (Breiman et al., 1984). In our data analysis, the C5 algorithm was used since it: a) is computationally efficient; and b) has outperformed the other methods examined in our preliminary analysis.

Similar to the previous two other classification approaches, SVM is a popular approach for stock market prediction (Kim, 2003; Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014; Schumaker & Chen, 2009; Yang, Chan, & King, 2002). More interestingly, SVMs are favored in applications where text mining is used for market prediction (Nassirtoussi et al., 2014). SVMs can be used for both linearly and non-linearly separable datasets. When the data is linearly separable, SVMs construct a hyperplane on the feature space to distinguish the training tuples in the data such that the margin between the support vectors is maximized. For nonlinear cases, the data is typically mapped into a higher-dimensional space so that the new dataset in higher-dimension becomes linearly separable. This problem can be handled efficiently by using a Kernel function (see Han, Kamber, and Pei, 2011 for more details). Based on our preliminary analysis, we have used the Radial Basis Function (RBF) Kernel function in our SVM classification algorithm since it has resulted in the best performance.

To evaluate the performance of the three classification methods, we present eight commonly used metrics in the literature: a) *accuracy*, b) *area under the receiver operating characteristic curve (AUC)*, c) *F-measure*, d) *G-mean*, e) *MCC*, f) *precision*, g) *sensitivity*, and h) *specificity*. In addition, we provide our code and the confusion matrix for the sake of completion. Our selected measures are all suitable for our binary classification problem. For details on how any of the above metrics can be calculated, we refer the reader to Han et al. (2011), and Hastie et al. (2011). We use the AUC as our primary evaluation metric for the reasons explained in Dag et al. (2016).

3. Results and discussion

In this section, we first highlight the results from the variable/feature selection phase of our methodology. Then, we present the results from the prediction accuracy of our expert system with respect to the five potential targets. This is followed by some preliminary analysis to evaluate the impact of the information attained from the five different data sources on our prediction power. For the sake of completion and to allow for the replication of our results, we present our code and a detailed tabular view of our results as supplementary documents to this manuscript.

3.1. Variable/feature selection

As mentioned in Section 2.2, the end goal of this phase is to prepare the data for the three machine learning models. Here, we employed the SVM RFE model five times (once for each target). This resulted in five different sets of twenty variables/features that offer the most predictive power for each of the five respective targets. We list these sets in Table 3. There are several additional observations to be made from Table 3:

(A) For any of the five targets, the selected variables/features span all predictor sets. This implies that there are non-redundant, useful information that can be captured from each data source.

(B) The previous day's closing, opening and high prices were significant predictors for all five targets. The previous day's low price is a strong predictor for four of the five targets (with the exception of Target 3).

(C) The *Price to Earnings (P/E) Ratio* is predictive for the four price targets, but not for Target 5 (i.e., trade volume target). In our opinion, this makes sense since the P/E Ratio measures the current share price relative to the per-share earnings. Thus, it may not be suited for predicting trade volume since it does not capture any movements.

(D) Target 5 had the highest number of Google features of 10. This was a somewhat expected result since *Google Trends* should reflect interest more than price fluctuations. The number of Google features selected for any of the other targets varied between 4 and 6.

(E) Perhaps the most important observation has to do with the order of the variables/features selected. We have arranged the items in a descending order (left to right and then to the next row). For all targets, variables selected from the *first set of predictors* were the most significant predictors. They were followed by one or more *technical indicators*. Then, the list would include several *Wikipedia* features, which were followed by some *Google News* features. The final grouping included a mixture of *technical indicators* and *Google/Wikipedia features*.

Note that we provide the definition for each of the features listed in Table 3 in Appendix C.

3.2. Predictive modeling outcomes

As explained in Section 2.3, we use the AUC as the primary evaluation criterion to evaluate the performance of the ANN, DT, and SVM models in predicting the five different day-ahead outcomes (while presenting the 7 other metrics for completion). In Fig. 3, we present the best-case, worst-case and the mean performance of the three machine learning models for each of the five targets. Note that the best-case, worst-case, and mean performances are determined based on the 10-fold cross validation step of our approach. The reader is encouraged to visit the interactive version of this plot at https://goo.gl/L06FSA. Based on Fig. 3, there are several interesting observations that can be made:

(A) Based on the AUC metric, SVM outperforms the ANN model for all five targets, DT outperforms the ANN model for Targets 2–4, DT outpeforms the SVM model in predicting Targets 2–3 (while having a similar performance in Target 4), and the DT model failed to predict one of the classes for both Targets 1 and 5.

(B) For Targets 2–4, the recommended models have an AUC value greater than 0.89. The AUC is the probability that the model will rank a randomly chosen positive instance (i.e. increase in price) higher than a randomly chosen negative one (i.e., decrease in stock price).

(C) The acquired data may not be capturing the underlying factor's for changes in trade volume (i.e., Target 5). The DT model could not predict decreases in trade volume (i.e., all its predictions were "1"s), and the ANN has a similar prediction to that of a random predictor (i.e., flipping a coin). The SVM had a somewhat reasonable mean AUC value of 0.632.

(D) Based on the eight evaluation metrics' values, our disparate data sources and machine learning models can best predict Target 2. Recall that Target 2 compares *next day's opening price* with *today's opening price*. This is a somewhat surprising result since we expected Target 1 to have the best results.

(E) Perhaps more importantly, our results (especially for Target 2) are more accurate than those typically reported in the literature. Our model resulted in $\approx$ 85% accuracy/hit ratio with an average AUC of $>$ 0.874 for *SVM and DT for Target 2*. In the literature, the previous predictions did not exceed an accuracy of 83% (see Table 5 in Nassirtoussi et al. (2014), which summarizes the outcomes of 24 text-mining-based financial expert systems).

(F) Building on the previous result, it is also clear that the addition of data from disparate data sources have resulted in improved accuracy. For example, Kim (2003) used the SVM model with only technical indicators as inputs, and obtained an accuracy rate of 65% accuracy for their best performance model. Our $>$ 20% accuracy improvement (when SVM or DT are used) is significant and justifies the effort needed to include new data sources.

From the above discussion, we have established that we can predict Targets 1–4 reasonably well through the deployment of an adequate machine learning model with inputs identified in Table 3. To formally understand the usefulness of the four disparate data sources and our generated features, we consider several scenarios that are summarized in Table 4. Note that *Scenarios 1–2* involve

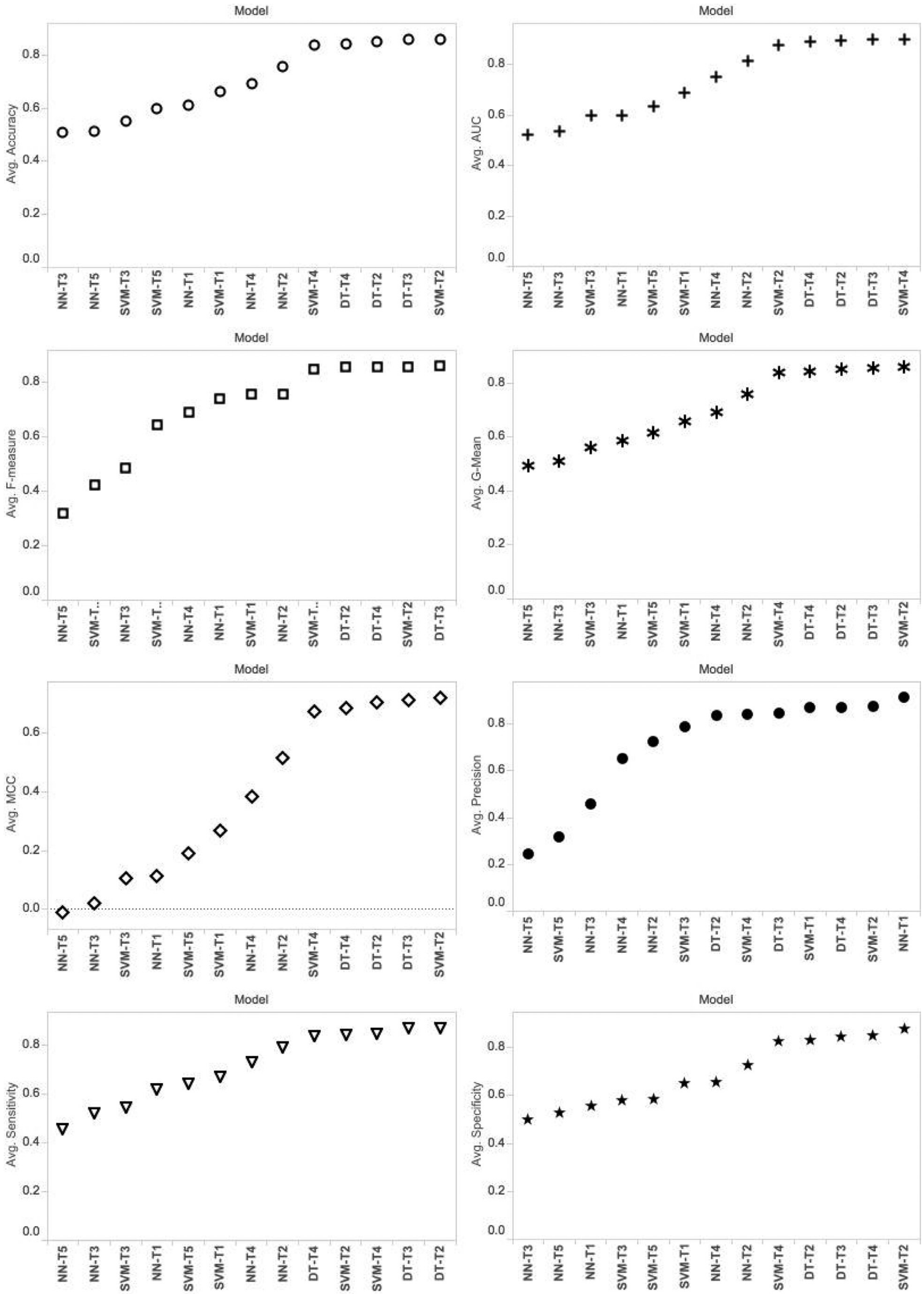


Fig. 3. A visual summary of the performance of the 3 data mining models for each of the five targets. An interactive version of this plot can be found at: http://goo.gl/L06FSA.

Table 4
Examining the impact of the non-traditional data sources.

Scenario #	Description
1	Market data
2	Market data, technical indicators
3	Market data, technical indicators, Wikipedia Traffic
4	Market data, technical indicators, Google news counts
5	Market data, technical indicators, Wikipedia Traffic, generated features
6	Market data, technical indicators, Google news counts, generated features
7	Market data, technical indicators, Wikipedia traffic, Google news counts, and generated features

the data sources most commonly used in traditional stock market prediction. *Scenarios 3–4* build on *Scenario 2* with the additional of one online data source. In *Scenarios 5–6*, we add the generated features to *Scenarios* 3 and 4, respectively. Scenario 7 include all five data sources.

As an example, consider the SVM model for Target 2. Let us examine how the inclusion of data sources, according to the seven scenarios presented in Table 4, impact the eight evaluation metrics. We present the results in Table 5. From the results, it is clear that the best performance is obtained when all data sources are included. In addition, by comparing S5 and S6 (or alternatively S3 and S4), one can see that the Wikipedia data is more informative than the Google Data for Target 2. Note that we consider the results presented in Table 5 as a formal way to evaluate our observation in Point (F) above.

4. Conclusions and future work

4.1. An overview of the impacts and contributions of our proposed expert system

In this paper, we developed a financial expert system to predict movements in the one-day ahead stock price/volume. To construct our “knowledge base”, we scrapped four different data sets: (i) historical stock market data, (ii) commonly used *technical indicators*, (iii) *Wikipedia* traffic statistics pertaining to the company’s pages (i.e. general company profile, stock page, and pages pertaining to the company’s main products), and (d) *Google News*. Since these data sources were never used in combination in an expert system, we generated features from the *two online sources* to further improve our knowledge base. Our AI framework consisted of two major phases: (1) variable/ feature selection, which helps improve the performance of our AI algorithms by reducing the dimensions of the data without the loss of information; and (2)the incorporation of ANN, SVM and DT for prediction, which allows us to select the “best” model for a given target and stock. We provide a web-based user interface (see https://github.com/binweng/ShinyStock) to promote the adoption of our expert system by investors and financial planners.

From an *Expert and Intelligent Systems* research perspective, our system is innovative and novel. Specifically, the related literature on stock movement prediction (shown in Table 1) primarily considered the use of traditional data sources (i.e. market data and technical indicators) and none, to our knowledge, combined multiple data sources. Our system utilizes disparate data sources in an attempt to have a more holistic representation of the factors and conditions that precede stock movement. The proposed expert system is tested using a large and feature-rich *Apple Inc.* dataset collected for a period of 37 months (May 1, 2012 to June 1, 2015), providing a hit ratio of 85% (which exceeds the reported results in the literature). Perhaps more importantly, we have addressed the following theoretical questions that relate to the design of expert and intelligent systems:

(a) What is the value of using online sources (specifically Wikipedia and Google News) when predicting the one-day ahead stock movement? In contrast with the majority of literature, we analyze this question through combining variables/features from these online sources with more traditional predictors. This allows us to quantify the value added rather than just obtaining a predictive model.
(b) Does the added value of these online sources differ with different targets? We chose five different *one-day-ahead* targets to examine if the value obtained from these sources changes according to different prediction questions.
(c) Which targets are most suitable for prediction based on the aforementioned five data sources?
(d) Which AI models provide the best predictive performance for each of the five targets?

From our case study, we have learned that the addition of these online sources are useful (especially for Targets 1–4). In addition, based on the Apple stock, it seems that Wikipedia has more predictive power than Google News. That being said, the addition of Google News indicators improve the predictive accuracy the AI models utilized by our expert system (see Fig. 3 and Table 5). From our seven scenarios of data aggregation, it is clear that the addition of online data sources and our generated features can significantly improve the prediction accuracy. This can imply that there are *news* hidden in these sources according to the followers of the *Efficient Market Hypothesis*. Alternatively, one can say that changes in these data sources precede changes in the stock market. Our analysis also indicates that all five targets can be predicted (using the best model) better than a coin-flip. Our intelligent system’s prediction performance is better than the results reported in the literature (see Section 3.2).

4.2. Implementing our expert system in practice

From an *Expert and Intelligent Systems* practical implementation perspective, our proposed system can be used in a number of different ways. First, on a basic level and through our interface, an investor who does not have a strong programming background can use our “knowledge base” to capture the total number of “Google News” articles and visitors of relevant Wikipedia pages. Through our plotting tools, that investor can visualize the crowd’s percep-

Table 5
Comparison of seven scenarios using eight evaluation metrics.

Scenario	fAccuracy	Sensitivity	Specificity	Precision	F-measure	MCC	G-Mean	AUC
S1	0.565	0.577	0.551	0.601	0.589	0.127	0.564	0.634
S2	0.616	0.618	0.614	0.648	0.633	0.232	0.616	0.711
S3	0.618	0.634	0.601	0.629	0.632	0.235	0.617	0.703
S4	0.618	0.639	0.595	0.639	0.639	0.233	0.616	0.708
S5	0.822	0.835	0.807	0.824	0.830	0.642	0.821	0.800
S6	0.813	0.821	0.804	0.805	0.813	0.625	0.813	0.856
S7	0.858	0.838	0.879	0.873	0.854	0.719	0.858	0.874

tion of a given stock or index. We have shown that these perceptions can be predictive of stock movement. It is important to note that this information is not available by current commercial products. Second, on a more advisory level, our expert system can be used to provide a data-driven recommendation for investors; an informed short term buy, or sell strategy of stocks can be made relative to whether the investors portfolio carries the stock. From that viewpoint as well, investors can use our system to construct an ensemble of predictions (with at least two models - their current approach and our expert system's recommendation). In the case of a two-model scenario, our expert system can indirectly help with quantifying risk/uncertainty (i.e., if both models agree, the likelihood of a correct outcome increases). If the investor already had access to multiple forecasting systems, then our expert system will present a new perspective on a stock since our model combines both traditional and nontraditional sources. In such a case, the investor can make his/her decision through a simple voting procedure. Third, our code, which provide through a link in this article, can be deployed in an existing fully automated short term trading system to make its decision-making process more comprehensive.

4.3. Limitations and future research

There are several limitations and opportunities for future work that arise from this study. First, we have only examined Apple stock over a certain time-period. Thus, it is not clear if our results and/or conclusions can extend prior or past this period. More generally, it would be interesting to examine if our conclusion would differ if a different type of commodity stock is chosen and/or if a stock index is desired. Second, we did not attempt to include other online data sources. It is not clear if the relevance of our sources would change if, for example, Facebook data is used. Therefore, there are several opportunities to extend our work by the inclusion of additional data sources. We expect a diminishing return with the inclusion of new data sources, since we expect some redundancy in the information captured from online data sources. That being said, it would be interesting to rank the value obtained from the different online data sources (for different stocks and indices). A third direction can be to consider the stochastic nature of the prediction. Our "inference engine" presented a deterministic prediction; however in practice, it might be interesting to have a level of certainty that is associated with the prediction. This can be accomplished through the incorporation of fuzzy systems, Bayesian Belief Networks (BBN), and ensemble approaches. The fourth, and perhaps the largest improvement on this financial expert system, is to attempt to predict the actual price rather than the movement. From an investor's point of view, a 20% increase in stock price is very different than a 1% increase. In our analysis, these two scenarios are identical since they are both coded as an increase in stock price.

In conclusion, this paper presented a financial expert system to predict the movement of a stock on a daily basis. We have shown that taking into consideration predictive factors from multiple sources can improve its predictive performance. We have also shown that the performance of the AI models can change significantly depending on the target used. To encourage future research, we provide our code and data in https://github.com/binweng/ShinyStock.

Appendix A

A1. Process to acquire data from google news

Google News data is acquired from *Alphabet Inc.'s Google Search Engine*. The use of *Google News* allows us to gather all sources of news produced over a particular time period based on some search keywords. From a stock market perspective, this allows an end user to search for a publicly traded company's stock, and obtain a number for the amount of news produced for that stock. The steps to obtain the amount of news produced are presented below and depicted in Fig. 4.

1. Go to www.google.com.
2. Input the search keyword, such as "AAPL, Apple Stock".
3. Click "News" and then "Search tools".
4. Custom the date range to the date you want to search.
5. Click "Search tools" again to show the result.

B1. Formulas for the generated features

In this study, we generated seven different types of features for Wikipedia traffic data and Google news data. The formulas are shown below. In the formula, n means the time periods, V_t means the data point at period t.

1. *Moving average:*

$$MA(n)_t = \frac{V_t}{n} + \frac{V_{t-1}}{n} + \cdots + \frac{V_{t-n+1}}{n} \tag{1}$$

2. *Exponential moving average:*

$$EMA(n)_t = (V_t - MA(n)_{t-1}) \times (\frac{2}{n+1}) + MA(n)_{t-1} \tag{2}$$

3. *Disparity:*

$$Disparity(n)_t = \frac{V_t}{MA(n)_t} \times 100 \tag{3}$$

4. *Momentum1:*

$$Momentum1_t = \frac{V_t}{V_{t-5}} \times 100 \tag{4}$$

5. *Momentum2:*

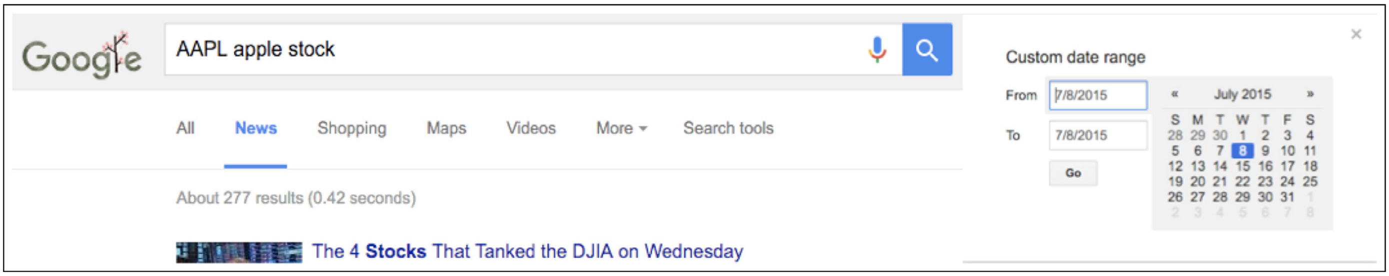
$$Momentum2_t = (V_t - V_{t-5}) \times 100 \tag{5}$$

6. *Rate Of Change:*

$$ROC_t = \frac{V_t}{Momentum2_t} \times 100 \tag{6}$$

7. *Relative Strength Index:*

$$RSI(n) = 100 - \frac{100}{1 + \frac{AverageGain(n)}{AverageLoss(n)}} \tag{7}$$



AAPL apple stock

About 277 results (0.42 seconds)

The 4 Stocks That Tanked the DJIA on Wednesday

Custom date range

From 7/8/2015

To 7/8/2015

Go

« July 2015 »

S	M	T	W	T	F	S
28	29	30	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	1
2	3	4	5	6	7	8

Fig. 4. Screen-shot of an illustration of acquiring data using Google's search engine.

Table 6
Definition of the most predictive variables/features.

Variable	Definition
Close	Closing price of the day
Google_x_day_disparity	Ratio of Google news volume to its x day moving average
Google_x_day_disparity_Move	Movement of Google_x_day_disparity as previous day
Google_EMA_x	x day exponential moving average of Google news volume
Google_EMA_x_Move	Movement of Google_EMA_x as previous day
Google_MA_x	x day moving average of Google news volume
Google_MA_x_Move	Movement of Google_MA_x
Google_Move	Movement of Google news volumes as previous day
Google_ROC_Move	Movement of the rate of change for Google news volume as previous day
Google_RSI_Move	Movement of relative strength index for Google news volume as previous day
High	Highest price of the day
Low	Lowest price of the day
Open	Opening price of the day
P/E ratio	Price-earning ratio
RSI	Relative strength index of the stock price
RSI_Move	Movement of RSI
Stochastic oscillator	Compares a security's closing price to its price range over a given time period
Stochastic_Move	Movement of stochastic oscillator
Wiki_x_day_disparity	Ratio of Wikipedia traffic to its x day moving average
Wiki_x_day_disparity_Move	Movement of Wiki_x_day_disparity
Wiki_EMA_x_Move	Movement of x day exponential moving average for Wikipedia traffic
Wiki_MA_x_Move	Movement of x day moving average for Wikipedia traffic
Wiki_Momentum_1	Ratio of current close price to the price three day's ago
Wiki_Move	Movement of Wikipedia as previous day
Wiki_ROC	Rate of change (ROC) of Wikipedia traffic
Wiki_ROC_Move	Movement of Wiki_ROC
Wiki_RSI	Relative strength index of Wikipedia traffic
Wiki_RSI_Move	Movement of Wiki_RSI
William %R	The level of the close price relative to the highest high

C1. Definition of variables/features in Table 3

We define the variables/features used in our model in the Table 6.

References

Adebiyi, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Comparison of arima and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics, 2014*.

Alavi, M., & Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly*, 107–136.

Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques–part ii: Soft computing methods. *Expert Systems with Applications, 36*(3), 5932–5941.

Bao, D., & Yang, Z. (2008). Intelligent stock trading system by turning point confirming and probabilistic reasoning. *Expert Systems with Applications, 34*(1), 620–627.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Chen, M.-Y., Chen, D.-R., Fan, M.-H., & Huang, T.-Y. (2013). International transmission of stock market movements: an adaptive neuro-fuzzy inference system for analysis of taiex forecasting. *Neural Computing and Applications, 23*(1), 369–378.

Choi, H., & Varian, H. (2012). Predicting the present with google trends.. *Economic Record, 88*, 2–9.

Chourmouziadis, K., & Chatzoglou, P. D. (2016). An intelligent short term stock trading fuzzy system for assisting investors in portfolio management. *Expert Systems with Applications, 43*, 298–311.

Cootner, P. H. (1964). The random character of stock market prices,.

Dag, A., Topuz, K., Oztekin, A., Bulur, S., & Megahed, F. M. (2016). A preoperative recipient-donor heart transplant survival score. *Decision Support Systems*.

Dase, R., & Pawar, D. (2010). Application of artificial neural network for stock market predictions: A review of literature. *International Journal of Machine Intelligence, 2*(2), 14–17.

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

Fama, E. F. (1991). Efficient capital markets: Ii. *The Journal of Finance, 46*(5), 1575–1617.

Fama, E. F. (1995). Random walks in stock market prices. *Financial Analysts Journal, 51*(1), 75–80.

Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The adjustment of stock prices to new information. *International Economic Review, 10*(1), 1–21.

Gabrielsson, P., & Johansson, U. (2015). High-frequency equity index futures trading using recurrent reinforcement learning with candlesticks. In *Computational intelligence, 2015 ieee symposium series on* (pp. 734–741). IEEE.

Guresen, E., Kayakutlu, G., & Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications, 38*(8), 10389–10397.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques*. Elsevier.

Hassan, M. R., Nath, B., & Kirley, M. (2007). A fusion model of hmm, ann and ga for stock market forecasting. *Expert Systems with Applications, 33*(1), 171–180.

Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2011). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Hendler, J. (2014). Data integration for heterogenous datasets. *Big data, 2*(4), 205–215.

Jasemi, M., Kimiagari, A. M., & Memariani, A. (2011). A modern neural network model to do stock market timing on the basis of the ancient investment technique of japanese candlestick. *Expert Systems with Applications, 38*(4), 3884–3890.

Kiersz, A. (2015). Here's how badly warren buffett beat the market. http://www.businessinsider.com/warren-buffett-berkshire-hathaway-vs-sp-500-2015-3.

Kim, K.-j. (2003). Financial time series forecasting using support vector machines. *Neurocomputing, 55*(1), 307–319.

Kim, K.-j., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications, 19*(2), 125–132.

Kimoto, T., Asakawa, K., Yoda, M., & Takeoka, M. (1990). Stock market prediction system with modular neural networks. In *Neural networks, 1990., 1990 ijcnn international joint conference on* (pp. 1–6). IEEE.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International joint conference on artificial intelligence, 1995* (pp. 1137–1143).

Kuhn, M. (2008). Caret package. *Journal of Statistical Software, 28*(5).

Lai, R. K., Fan, C.-Y., Huang, W.-H., & Chang, P.-C. (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications, 36*(2), 3761–3773.

Lee, K., & Jo, G. (1999). Expert system for predicting stock market timing using a candlestick chart. *Expert Systems with Applications, 16*(4), 357–364.

Li, S.-T., & Kuo, S.-C. (2008). Knowledge discovery in financial investment for forecasting and trading strategy through wavelet-based {SOM} networks. *Expert Systems with Applications, 34*(2), 935–951. http://dx.doi.org/10.1016/j.eswa.2006.10.039.

Lin, X., Yang, Z., & Song, Y. (2011). Intelligent stock trading system based on improved technical analysis and echo state network. *Expert Systems with Applications, 38*(9), 11347–11354.

Loomis, C. J. (2012). Buffett beats the sp for the 39th year. http://fortune.com/2012/02/25/buffett-beats-the-sp-for-the-39th-year/.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *The Journal of Economic Perspectives, 17*(1), 59–82.

Mao, H., Counts, S., & Bollen, J. (2011). Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv preprint arXiv:1112.1051,.

Megahed, F. M., & Jones-Farmer, L. A. (2015). *Frontiers in statistical quality control 11* (pp. 29–47)). Cham: Springer International Publishing.

Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. (2013). Quantifying wikipedia usage patterns before stock market moves. *Scientific Reports, 3*.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications, 41*(16), 7653–7670.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications, 42*(24), 9603–9611.

Nofsinger, J. R. (2005). Social mood and financial economics. *The Journal of Behavioral Finance, 6*(3), 144–160.

Prechter Jr, R. R., & Parker, W. D. (2007). The financial/economic dichotomy in social behavioral dynamics: The socionomic perspective. *The Journal of Behavioral Finance, 8*(2), 84–108.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using google trends. *Scientific reports, 3*.

Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence, 26*(1), 25–33.

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

Quinlan, J. R. (2014). *C4. 5: Programs for machine learning*. Elsevier.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS), 27*(2), 12.

Scikit-Learn-Developers (2014). 1.13 feature selection - scikit-learn documentation. http://goo.gl/GDedwn.

Shynkevich, Y., McGinnity, T. M., Coleman, S., & Belatreche, A. (2015). Stock price prediction based on stock-specific and sub-industry-specific news articles. In *Neural networks (ijcnn), 2015 international joint conference on* (pp. 1–8). IEEE.

Smith, V. L. (2003). Constructivist and ecological rationality in economics. *The American Economic Review, 93*(3), 465–508.

Vu, T.-T., Chang, S., Ha, Q. T., & Collier, N. (2012). An experiment in integrating sentiment features for tech stock prediction in twitter,.

Wang, Y.-F. (2002). Predicting stock price using fuzzy grey prediction system. *Expert Systems with Applications, 22*(1), 33–38. http://dx.doi.org/10.1016/S0957-4174(01)00047-1.

Yang, H., Chan, L., & King, I. (2002). Support vector machine regression for volatile stock market prediction. In *Intelligent data engineering and automated learningIDEAL 2002* (pp. 391–396). Springer.

Zhang, Y., & Wu, L. (2009). Stock market prediction of s&p 500 via combination of improved bco approach and bp neural network. *Expert Systems with Applications, 36*(5), 8849–8854.