Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms

Al Nasseri, A.A., Tucker, A. and De Cesare, S.

This is the published version of the article Al Nasseri, A.A., Tucker, A. and De Cesare, S. (2015) Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms. *Expert Systems with Applications,* 42 (23), pp. 9192-9210.

It is available at https://dx.doi.org/10.1016/j.eswa.2015.08.008.

Expert Systems With Applications 42 (2015) 9192–9210

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms

Alya Al Nasseri[a,*], Allan Tucker[b], Sergio de Cesare[a]

[a] Brunel Business School, Brunel University London, UK
[b] Department of Computer Science, Brunel University London, UK

ARTICLE INFO

Keywords:
Decision tree
Filter approach
Text mining
Trading Strategy

ABSTRACT

Growing evidence is suggesting that postings on online stock forums affect stock prices, and alter investment decisions in capital markets, either because the postings contain new information or they might have predictive power to manipulate stock prices. In this paper, we propose a new intelligent trading support system based on sentiment prediction by combining text-mining techniques, feature selection and decision tree algorithms in an effort to analyze and extract semantic terms expressing a particular sentiment (sell, buy or hold) from stock-related micro-blogging messages called "StockTwits". An attempt has been made to investigate whether the power of the collective sentiments of StockTwits might be predicted and how the changes in these predicted sentiments inform decisions on whether to sell, buy or hold the Dow Jones Industrial Average (DJIA) Index. In this paper, a filter approach of feature selection is first employed to identify the most relevant terms in tweet postings. The decision tree (DT) model is then built to determine the trading decisions of those terms or, more importantly, combinations of terms based on how they interact. Then a trading strategy based on a predetermined investment hypothesis is constructed to evaluate the profitability of the term trading decisions extracted from the DT model. The experiment results based on 122-tweet term trading (TTT) strategies achieve a promising performance and the (TTT) strategies dramatically outperform random investment strategies. Our findings also confirm that StockTwits postings contain valuable information and lead trading activities in capital markets.

1. Introduction

Stock market prediction is an attractive and challenging area of research different methodologies that has been developed with the aim of predicting the direction of securities' prices as accurately as possible (Guresen, Kayakutlu, & Daim, 2011). The aim has been to create accurate models that have the ability to predict stock price behavioral movements in the stock market. However, predicting these changes is very challenging (and appealing to researchers to investigate), due to the fact that stock market data are noisy and time varying in nature (Atsalakis & Valavanis, 2009). To address the topic of future stock price predictions, several theories become relevant in this regard. Several works have attempted to study stock market prediction while providing an answer to the common question: can stock prices really be predicted? There are two theories that are mostly relevant in answering such a question: (1) Efficient Market Hypothesis and (2) Random Walk Theory. According to the Efficient Market Hypothesis (EMH), market prices reflect all publicly available information (Fama, 1970). This implies that past and current information is immediately incorporated into the stock prices, thus any price changes can only be explained by new information or "news". Due to the random arrival of new information, the stock price is said to follow a random walk pattern and it is impossible to predict the stock market, since prices are randomly determined. If this hypothesis is held; therefore the attempts to predict the stock market will be ineffective. The researchers continuing efforts in accurately forecasting stock markets using various methods and techniques have proved that underlying assumptions of the EMH and random walk turn out to be unrealistic and that some degree of predictability might be possible (Darrat & Zhong, 2000). A variety of machine learning techniques have been proposed to predict the future movement and trend of stock prices in capital markets. However, most of these studies focus on predicting the movement in stock prices rather than predicting the investment decisions that derive from and cause the movement itself, such as buying, selling and holding decisions. For example, Xue-shen, Zhong-ying, Da-ren, Qing-hua, and Hui (2007) adopted classification complexity of Support Vector Machines (SVM) as a feature selection

* Corresponding author. Tel.: +447703346765.
E-mail addresses: Alya.Al-Nasseri@brunel.ac.uk, aliaaa83@hotmail.com (A.A. Nasseri), Allan.Tucker@brunel.ac.uk (A. Tucker), Sergio.deCesare@brunel.ac.uk (S. de Cesare).

http://dx.doi.org/10.1016/j.eswa.2015.08.008

criterion to predict the Shanghai Stock Exchange Composite Index (SSECI). Huang, Yang, and Chuang (2008) employed a wrapper approach to select the optimal feature subset and apply various classification algorithms to predict the trend in the stock markets of Taiwan and South Korea. Lee (2009) proposed a prediction model based on a hybrid feature selection method and SVM to predict the trend of the stock market.

Investor sentiment has been proven significant in affecting the behavior of stock prices (Baker & Wurgler, 2006). Investors' expectations and their psychological thinking, from which the sentiments are derived, are considered the main factors that affect stock price movements in capital markets (Tan, Quek, & Ng, 2007). Behavioral finance theory suggested that the existence of different types of traders and the effect of their trading behavior has substantial effect on influencing price changes in financial markets (DeLong, Shleifer, Summers, & Waldmann, 1991). There are two types of traders in financial markets, the "irrational noise trader" or so-called "day trader" is one who does not possess fundamental information (Kyle, 1985) and the "rational trader" or "arbitrageur" who holds rational beliefs (DeLong, Shleifer, Summers, & Waldmann, 1990) and, in a way, is always updating their beliefs according to the new information available to them (Baberies and Thaler, 2003). The presence of noise traders in financial markets, who make irrational decisions regarding buying, selling or holding stocks, can then cause price levels and risks to deviate from expected levels, even if all other traders are rational (De Long et al., 1990). Noise traders are always taking part in the discussion and conversations related to financial information in capital markets. In the context of online investment forums, conversations among investors including the noise traders involve making predictions, exchanging opinions, asking questions, sharing analyses, and reporting financial information (Oh & Sheng, 2011). Therefore, the ability of noise traders to affect price changes will also appear in online investment forums, where information and opinion is widely spread among investors through the investment communication platforms (Zhang & Swanson 2010). It is therefore important to highlight the critical role played by trading decisions in the stock market. Trading decisions have a great effect on the profitability position of an investor in the capital market. Therefore, the ability to predict an intelligent trading support mechanism would help investors to make profitable investment decisions concerning a particular security in the capital market. Making correct investment decisions is a substantially difficult task for investors due to the problem of high nonlinearity embodied in the behavior of financial markets.

Many attempts have been made to provide investors and other financial professionals with consistently profitable autonomous trading support systems. Motivation for such trading systems comes from various fields of studies ranging from fundamental analysis and financial econometric modeling to evolutionary computation (Hu, Feng, Zhang, Ngai & Liu, 2015), machine learning (Booth Gerding, and McGroarty, 2014) and text mining (Gong, Zeng, & Zhang, 2011; Nuij, Milea, Hogenboom, Frasincar, & Kaymak, 2014). In this context, numerous financial researchers have progressively provided investors and their peers in capital markets with decision-making support systems in order to improve and enhance their ability to make a better-informed investment decision that will lead to greater return on their investments (Kodogiannis & Lolis, 2002; Li & Kuo, 2008; Skabar, 2005; Sun, Liang, Zhang, Lee, Lin et al., 2005; Chun & Park, 2005). Some of these studies are based on traditional time series predictions (Kodogiannis & Lolis, 2002; Skabar, 2005; Sun et al., 2005) and trend prediction (Cheng Wei, and Chen, 2009; Tsai & Hsiao, 2010) that mainly focused on historical past prices in predicting the future value of stocks. Most of the capital market players, however, are much more interested in time series predictions of future trends rather than exact future prices. In addition to the traditional time series approach, the application of artificial intelligence (AI), such as expert systems (Kee & Koh, 1994), fuzzy systems (Abraham Nath & Mahanti, 2001; Chang & Liu, 2008), and artificial neural networks (ANN) (Chiang, Urban, & Baldridge, 1996; Duan et al., 2009; Masoud, 2014), has received extensive attention by researchers with an attempt to make the forecast of future prices more reliable. Despite the effectiveness demonstrated by such methodologies, there are some drawbacks associated with their applications. For example, the main drawback with ANNs and other black-box techniques is that the results obtained from such methodologies are misleading and very difficult to interpret (Lai, Fan, Huang, & Chang, 2009). Another drawback is the lack of investigating the nature of interactions between technical indicators and stock market fluctuations. Methodologies that provide a greater insight into market procedures must therefore be developed (Chi, Chen, & Cheng, 1999; Zhang, 2007). However, most recent studies tend to provide accurate trading strategies by combining machine learning techniques (e.g., SVM) with all other techniques, namely robust feature selection, transactional volume incorporation, pattern models and technical analysis. The research community has had a long-standing argument on the effectiveness of technical analyses in stock trading. Some argue that stock prices are not predictable while others, such as Brock, Lakonishok and LeBaron, (1992) and Blume Easley, and O'hara. (1994), have presented positive empirical evidence on the effectiveness of technical analyses (Kaucic, 2010). Kara, Boyacioglu, and Baykan (2011) provide a comparable pattern whereby neutral networks and plain SVM were compared for the purpose of making stock price movement prediction with the extensive use of several technical indicators. Rosillo, Giner, and de la Fuente (2014) used Volatility Index and technical analysis with the aim to forecast weekly change in S&P 500. Dai, Shao, and Lu (2013) incorporated MARS splines for attribute selection, which then was used as an input for the Support Vector Regression model. Recently Żbikowski (2015) applied a modified Support Vector Machine (SVM) classifier (volume weighted SVM) with walk forward testing and the Fisher method for feature selection for the purpose of creating a stock trading strategy and forecasting short-term trends on the stock market. Hu et al. (2015) proposed a hybrid long-term and short-term evolutionary trend following algorithm (eTrend) that combines TF investment strategies with the eXtended Classifier Systems (XCS) for the purpose of providing effective trading guidance for investors in the capital market.

The provision of an accurate and timely trading support mechanism is the key success for traders to make a profitable decision in capital markets. This study presents a novel approach for developing a new decision support system based on tweet semantic terms extracted from the decision tree model (Quinlan, 1993) which then can be implemented as a trading strategy and constitute three different portfolios (sell, buy and hold). The decision tree proved successful in searching for rules hidden in large amounts of data. The visibility of the connected relationships between nodes branches and leaves in the tree makes it most suitable approach for feature selection and prediction of investment trading decisions in capital markets. It has also proved efficient for time series analysis. In addition, decision tree techniques have already been shown to be interpretable, efficient, problem independent and able to deal with large-scale applications. The decision tree model provides a visualized insight into the StockTwits data by highlighting the individual relationships with respect to the class as well as the combined associations of features with respect to the decision class. One would expect that the decision effect of individual terms (feature) appearing in a tweet posting would have a different decision effect than if it had appeared in combination with other terms. The ability of the decision tree model to explore the related interactions between the selected terms and their ability to predict trading decisions makes it a better and more suitable model for this research.

This research takes a different approach by proposing an automatic decision support system that integrates text mining techniques, feature selection and decision tree algorithm. This research

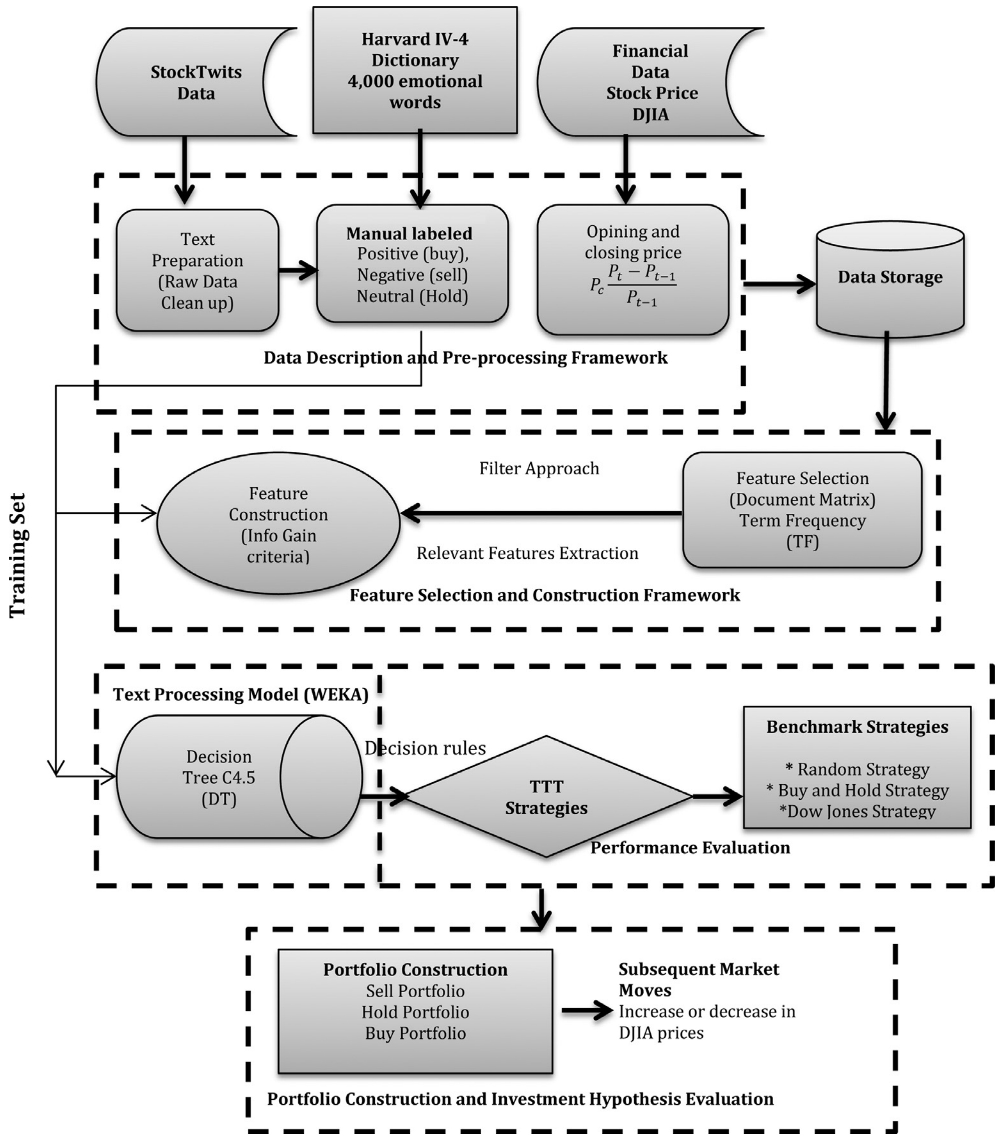

Fig. 1. The Framework Design of proposed model.

aims to predict an intelligent trading support mechanism to screen out the most significant and profitable trading terms or combination of terms from StockTwits data that may help investors to make correct and accurate (selling, buying or holding) decisions in capital markets. The attempt is to investigate whether the terms or combination of terms of trading decision rules extracted from the decision tree algorithm, may act as a trading decision guide to investors that may lead to a profitable investment decision while examining the predictive ability of each term or combination of terms in anticipating subsequent movement in the stock market.

The reminder of this paper is organized as follows. Section 2 presents the framework design built for data analysis. Section 3 describes the method used and presents the rationale of the model adopted for this research study. Section 4 outlines the design of the trading strategies. Section 5 describes the data and discusses the results of the empirical investigation. Section 6 provides the implications of the present study and discusses its contributions. Future improvements for extending the current work will be provided and suggested in Section 7. Finally, concluding remarks are given in Section 8.

2. Framework design

In this section we present the framework design for data analysis adopted for this research paper as shown in Fig. 1.

As can be seen from Fig. 1, the framework design is composed of five major components: Data Description and Pre-Processing Framework, Feature Selection and Construction Framework, Text Processing Model, Performance Evaluation and Portfolio Construction and Investment Hypothesis Evaluation Framework. These framework components are represented in dashed boxes identified with the relative component name. Each component framework consists of different procedures that are vital in performing the whole function of the relative component.

The *Data Description and Pre-Processing Framework* is the first component that appears at the top of the figure, which is accountable for data acquisition from various sources as well as pre-processing and filtering procedures to avoid irrelevancy of the data being collected. At this stage and after the text customization has been performed, the manual operation of sample tweet messages is preformed to manually classified tweets into three distinct classes, namely sell, buy or hold, by using the Harvard IV dictionary which then is used as a training set in the text processing model and feature construction stages.[1,2,3] A list of general rules applied when manually labeling the messages is provided in Appendix A.

The second component, *Feature Selection and Construction Framework*, represents the implementation of filter approaches of feature selection (based on Information Gain criteria (IG)) to extract the most relevant features from the datasets to build a features construction model. The construction model of relevant features (reduced features) is then used as input variables to the third component *Text Processing Model* where Decision Tree algorithm C4.5 is employed to process the text and detect relative sentiments. The trading decisions rules: sell, buy or holds of each term or combination of terms are extracted from the decision tree classifier. The proposed system treats each term (or combination of terms) as a trading strategy called Tweet Term Trading (TTT) Strategy and calculates the cumulative return from such strategies accordingly. These trading strategies are then evaluated by comparing its performance to a benchmark trading strategy (e.g. Random Strategies, Buy and Hold Strategy and Dow Jones Strategy), which is the task handled in the fourth component of the design *Performance Evaluation*.

In the final component *Portfolio Construction and Investment Hypothesis Evaluation*, investment portfolios for each decision class (sell, buy and hold) are constructed where each portfolio consisting of all possible terms and/or combination of terms belong to that class. Moreover, we empirically test the investment-trading hypothesis (short and long position) we adopted to calculate the cumulative return for each trading strategy.

The ultimate goal of this research, however, is to show whether the semantic terms trading strategies, extracted from an online stock forum (StockTwits), could earn abnormal returns and to evaluate whether these terms' strategies could act as a decision guide to help investors to make better-informed investing decisions regarding their traded securities discussed in such forums. This method is described in the following sections.

3. Method

The proposed methodology, followed in this research paper, combined text mining techniques, feature selection and decision tree algorithms.

3.1. Text mining

The nature of the data to be collected (StockTwits posts) and the purpose of the data analysis (to extract sentiment from online financial text) inherently proposes the need for text mining. The rationale of the model in this paper is that the models are trained from a corpus of manually labeled data to test the computational model, instead of using a sentiment lexicon, such as the SentiWordNet. Existing lexicons are not used in this paper mainly because this research is based on extracting sentiments from financial text, as the decision is to classify text into buy, sell or hold, not merely positive or negative. The vast majority of research papers in the sentiment analysis field focus mainly on domains, including emotional state (Kramer, 2010), product review (Turney, 2002) and movie review (Pang, Lee, & Vaithyanathan, 2002), in which case SentiWordNet is deemed a suitable lexicon. However, financial researchers have shown that dictionaries developed from other disciplines may not be effective for use in financial texts and may result in a misclassification of common words (Loughran & McDonald, 2011). We use the tm (text mining) package in R to preprocess the individual tweets (R project, 2012). Standard text mining procedures were employed for each tweet message in order to remove stop-words, white spaces, punctuation and numbers, and to stem all necessary words. This results in an n (terms) by m (documents) matrix for each post, where cells contain the number of times a term has appeared in the corresponding message.

3.2. Feature selection

Feature selection is an essential pre-processing step in the text mining process. Removing features (terms) that have no discriminatory power (John, Kohavi, & Pfleger, 1994) enables the classification performance to be obtained in a cost-effective and time-efficient manner, which often leads to more accurate classification results (Guyon & Elisseeff, 2003). In general, two methods are associated with feature selection: the filter and wrapper (Kohavi & John, 1997). The filter method evaluates the relevance of features and results in a subset of ranked features in accordance with relevancies. The wrapper method, however, assesses the relevance of features by selecting the most relevant (optimal) features from the original subset of features by using a special machine-learning algorithm. Therefore, the optimal features selected with the wrapper method are different and tailored to a particular classifier. Here however, we exploit the filter method, which assesses the relevance of features by ranking lists of subsets of features in accordance with relevancies where, the relevant features lie at the top of the list while the relevancy decreases toward the bottom of the list. The filter approach of feature selection is performed in the Weka machine-learning software, where the most relevant features selected are based on the information gain criterion.

Information gain criteria: Information Gain (IG) is the most commonly employed criterion to evaluate the goodness of the features in a machine-learning environment. It uses *Ranker* as a search method which ranks the attributes by their individual evaluations. Information gain is biased in favor of features with higher dispersion (Huang et al., 2008). IG measures the amount of information obtained for the predicted class within the data set by perceiving the absence and the presence of a feature (Yu & Liu, 2004). It is calculated based on the following formula:

$$IG\,(f_k) = \sum_{c\in(c_i,\bar{c}_i)} \sum_{f\in(f_k,\bar{f}_k)} \Pr(f,c) log \frac{\Pr(f,c)}{\Pr(f) X \Pr(c)} \tag{1}$$

where f_k means the presence of the features k and $\overline{f_k}$ indicates the absence of feature k. After the attribute selection is performed, a list of all subset attributes along with their relevance rank is shown in the output result. The output results rank attributes listed based on the relevant statistical score in which the attributes are arranged in accordance with the relevancy value. The top features in the list indicate the high relevant features, while the low relevant features are located at the bottom of the list. Performing feature selection by omitting the low relevant features down the list and retaining the most (best) relevant features will improve classification accuracy of different machine learning classifiers.

[1] Many psychological finance studies have widely used the Harvard IV-4 Dictionary for various text analysis tasks (Tetlock, 2007; Kothari, Li, & Short, 2009; Loughran and McDonald, 2009). The General Inquirer's Harvard-IV-4 classification dictionary of emotional words is used in this research to add each occurrence of emotional words in a message to the bag of words (Tetlock et al., 2008). From the domain knowledge of Harvard-IV dictionary, more than 4000 emotional words are being tagged and classified as either positive or negative.

[2] Many psychological finance studies have widely used the Harvard IV-4 Dictionary for various text analysis tasks (Tetlock, 2007; Kothari, Li, & Short, 2009; Loughran and McDonald, 2009). The General Inquirer's Harvard-IV-4 classification dictionary of emotional words is used in this research to add each occurrence of emotional words in a message to the bag of words (Tetlock et al., 2008). From the domain knowledge of Harvard-IV dictionary, more than 4000 emotional words are being tagged and classified as either positive or negative.

[3] Inputs for the model come from a training corpus of 2892, a representative sample of tweets, which are manually coded as either buy, hold or sell signals based on redefined dictionary (Harvard-IV-4 classification dictionary). A small sample was chosen to overcome the expected risk of over-fitting associated with text mining algorithms. The tweets are labeled as buy (1), hold (0), sell (−1).

3.3. Decision tree algorithm

Decision tree is one of the most frequently used techniques for classification problems. It is a tree structure consisting of nodes and branches. When decision trees are used for classification problems, they are often called a classification tree where each node represents the predicted class of a given feature. It is also used for regression problems where each node is indicated by an equation to identify the predicted value of an input feature. It applies the concept of information gain or entropy reduction, which is based on the selection of decision nodes and further splitting the nodes into sub-nodes. This function is performed by building decision trees or (decision nodes) from a set of training data.

This research uses a decision tree algorithm C4.5 that is an extension of Quinlan's algorithm ID3 that generates decision trees or nodes (Quinlan, 1993; Salzberg, 1994) by choosing the most effective attribute that splits each node into sub-nodes augmented in one class or the other. The normalized information gain is an impurity-based criterion that uses the entropy measure (Rokach & Maimon, 2005) to evaluate the effectiveness of an attribute for splitting the data. Therefore, these criteria state that the attribute with the greatest normalized information gain is chosen to make the decision. The process of splitting the decision nodes continues until no further split is possible. This means that the data has been classified as close to perfection as possible. This process safeguards maximum accuracy on the training data. In this research study, each tweet message in the training set is manually classified into one of three classes (e.g. $C = c_1, c_2$ and c_3) where c denotes a sell, buy or hold class. The tweets that have already been classified $T = t_1, t_2, t_3, \ldots t_n$ consist of different attributes or features 'x' so called vector (e.g. $t_1 = x_1, x_2, x_3, \ldots x_n$). A vector comprises the terms that have been selected using the filter approach of feature selection in Section 3.2, which will be classified later into either sell, buy or hold decisions using the decision tree algorithm. To form a decision tree, the following steps are required:

Step 1: Define the entropy of $\boldsymbol{x}$

$$\mathrm{H}(\boldsymbol{X}) = -\sum_j pj \log 2(p_j) \tag{2}$$

where $\boldsymbol{x}$ is a random variable with k discrete values, distributed according to probability value $P = (p_1, p_2, p_3, \ldots p_n)$.

Step 2: Calculate the weighted sum the entropies for each subsets.

$$H_S(\mathrm{T}) = \sum_{i=1}^{k} P_i H_S(T_i) \tag{3}$$

where Pi is the proportion of records in subset i.

Step 3: Calculate the information gain

$$Information\ gain\ IG\ (S) = H(T) - H_s(T) \tag{4}$$

The information gain is the criterion necessary to choose the most effective attribute to make the decision. Then the selection of the attribute at each decision node would be the one with the highest information gain, IG(s). The decision tree algorithm can yield a set of *classification rules* for classifying the whole attributes (terms) selected by the IG criteria. These classification rules are regarded as trading decision rules where the terms or combination of terms, extracted from the decision tree, indicated either a sell, buy or hold decision by investor.

4. Trading strategies design

Widespread evidence has been growing that stock prices overreact or underreact to information which suggests that a profitable trading strategy that selects stocks based on their past returns will probably exist. The concept of this research paper is built upon the previous research study of Tetlock, Saar-Tsechansky, and Macskassy (2008), who found a trading strategy based on negative words in firm specific news articles could earn abnormal annualized returns. To more thoroughly test the ability to earn abnormal profits based on specific terms in StockTwits messages, we designed a trading strategy as introduced in Preis, Moat and Stanley (2013) for some specific terms or set of terms that are believed to have an effect on the selling, buying or holding decisions in capital markets (as suggested by the feature selection method and the decision tree algorithm discussed earlier in this paper). Unlike the study of Tetlock et al. (2008), who used a simple quantitative measure of language to predict firms' accounting earnings and stock returns based on negative words alone, this study considers a collective use of the tweet language whereby positive, neutral and negative words are all considered in predicting tweet term trading strategies. To investigate whether the occurrence of a specific term or combinations of terms have the power to predict a trader's decision in a capital market, we analyzed closing prices $p(t)$ of the Dow Jones Industrial Average (DJIA) on a daily basis over a one year period. In this strategy, we use StockTwits data to obtain a volume frequency $n(t)$ of a term in day t. Then, we create a daily time series for the terms and/or the combination of terms based on the daily volume frequency of terms that appears in the tweet messages over the studied sample period. In the non-trading days, the volume frequency of a given term/combination of terms will be combined together with the volume frequency of the next immediate trading day. Note that there might be a silent period either because there were no messages posted or the terms might not have appeared in that particular tweet posting. In line with the study of Antweiler and Frank (2004) on the Internet message board, we place all silent periods with a value of zero. To minimize the effect of the silent period, we focus only on the terms with high volume frequency of appearance by ensuring that the minimum value of the term frequency considered is no less than 100, which represents a minimum volume frequency of the terms considered in this study. To compare the changes in term volume frequency to subsequent market moves, we implement a trading strategy for each of the 122 terms. The following section will explain the design of the proposed trading strategy followed in this research paper. To quantify changes in the appearance of a term in a tweet message, we use the relative change in volume frequency:

$$\Delta n(t, \Delta t) = n(t) - N(t-1, \Delta t) \tag{5}$$

where $n(t) =$ the volume frequency of a term appeared in a given day and $N(t-1, \quad \Delta t) = (n(t-1) + n(t-2) + \ldots + n(t-\Delta t)/\Delta t$ is the average number of term frequency of the previous 5 days. This method is called a simple moving average (MA) method where it is used to roll out the effect of the term appearance over the previous five days average. We average the term frequency over five realizations of its frequency value assuming that the effect of that term will last at least five trading days.

The proposed trading strategy presented in this paper is called tweet term trading (TTT) strategy. It simply evaluates the profitability of a tweet term strategy and is substantially effective for investors as it provides guidance in helping make a correct, accurate and profitable decision concerning a particular security in a capital market. As is well known, a trading strategy makes profit only if it could provide some predictability of future changes in stock prices, given the great variability of the data in the stock market. Therefore, we evaluate our investment strategy by hypothetically implementing it as follows:

$$\boldsymbol{Stat}(\boldsymbol{t}) = \begin{cases} Short\ position, & If\ \Delta n\ (t-1, \quad \Delta t) > 0 \\ Long\ position, & If\ \Delta n(t-1, \quad \Delta t)\ < 0 \end{cases} \tag{6a}$$

$$\boldsymbol{Sig}(\boldsymbol{t}) = \begin{cases} Short\ position\ then, \quad sell\, p(t)\ and\ buy\ p(t+1)\ and \\ \qquad Rtn = Ln\ p(t) - Ln\ p(t+1) \\ Long\ position\ then\ sell\, p(t+1)\ and\ buy\ p(t)\ and \\ \qquad Rtn = Ln\ p(t+1) - Ln\ p(t) \end{cases} \tag{6b}$$

Stat (t) denotes the current trading position of investors, while sig(t) indicates the trading instruction produced in this strategy design. According to this strategy, investors take a short position in the market following an increase in term volumes frequency ($\Delta n\ (t-1, \Delta t) > 0$) by selling the DJIA at the closing price $p(t)$ on the first trading day and buying back the DJIA at price $p(t + 1)$ at the end of the following day. If instead a long position has been taken following a decrease in term volume frequency ($\Delta n\ (t-1, \Delta t)\ < 0$) then investors buy the DJIA at the closing price $p(t)$ on the first trading day and sell the DJIA at price $p(t + 1)$ at the end of the next trading day. A cumulative return for each trading strategy therefore needs to be calculated. If investors take a 'short position', then the cumulative return R is $Ln\ p(t)\ -\ Ln\ p(t+1)$ whereas, if he/she takes a 'long position', then the cumulative return R then changes by $Ln\ p(t+1)\ -\ Ln\ p(t)$. Following this strategy, we assume that buying and selling activities will have a symmetric impact on the cumulative return R of a strategy's portfolio. As usual in this type of analysis, transaction costs are usually ignored (Zhang & Skiena (2010)). However, we cannot rule out the impact of such transaction costs on impacting profit in the real world implementation. Therefore, our study follows Hu et al. (2015) by considering the transaction cost to evaluate the performance of our (TTT) strategies. Clarkson, Joyce, and Tutticci (2006) argues that the level of transaction costs for online brokers are in the range of 0.15%–0.2%.[4]

4.1. Benchmark trading strategies

To assess the profitability of the tweet term trading strategies created in the previous section; the performance of these strategies has to be evaluated against benchmark trading strategies. Recall again that the purpose of this research is to find out whether the trading strategies based on the semantic terms in StockTwits forums could earn abnormal profits, while we are not emphasizing here that these strategies are the optimal and the best strategies for investors. In the present study, we consider three benchmark trading strategies as follows:

4.1.1. Random (RND) strategy

Random investment strategy is the simplest strategy where at time t the correspondent trader makes his/her prediction on trading completely at random. An investor following such a strategy makes decisions each day to sell or buy the market index in an uncorrelated, random manner. In any given day, there is an equal chance (probability = 50%) that the index will be bought or sold and this decision is independent and unaffected by decisions in the previous day. Statistically speaking, random strategy is a normal distribution strategy with the mean value of $\langle R \rangle_{Random\ Strategy} = 0$. In trading analysis, the means of any trading strategies developed are tested against the mean of the distribution curve that a random trading strategy would produce, which in statistics is assumed to be zero under the null hypothesis of no excess returns (Vanstone & Hahn, 2010). As with any standard normal random variable, the standard deviation of this strategy is derived from simulations of 1000 independent realizations of uncorrelated random strategy as shown in Fig. 2.

4.1.2. Buy and hold strategy

Buy and hold strategy is defined as a passive investment strategy in which investors take a passive role in the market with no active buying and selling of stocks from the time the portfolio is created until the end of the holding period (end of investment horizon). We implement the 'buy and hold' strategy by buying the index at the beginning of the period 3rd April 2012 and selling it at the end of the holding period of investment at 5th April 2013. This strategy yields

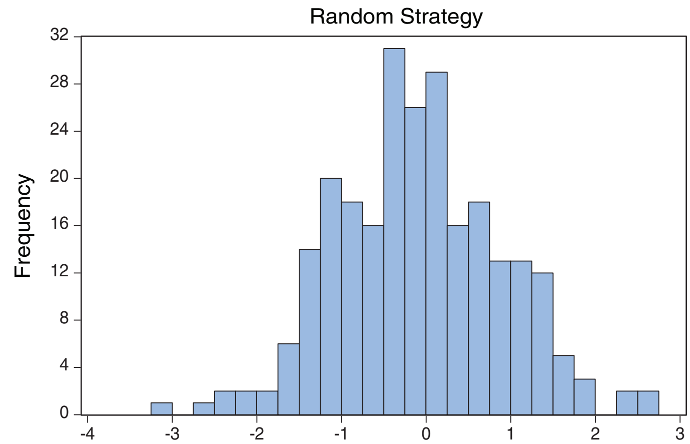

Fig. 2. The standard deviation of 1000 simulations of average returns using purely random investment strategy.

10.347% profit, which is equal to the overall increase in value of DJIA over the investment period of one year from 3rd April 2012 until 5th April 2013. The return obtained from this strategy is 0.0985 standard deviations of cumulative returns of uncorrelated random investment strategies.

4.1.3. Dow Jones (DJ) strategy

This strategy is based on changes in DJIA prices $p(t)$ instead of changes in the term related frequency data as the basis of buy and sell decisions. Implementing this strategy resulted in a loss of 6.177%, or when determined by the mean value of random strategy, results in a negative return of −0.0245.

5. Empirical test and analysis

5.1. Data preparation and pre-processing

StockTwits Data: We construct our analysis on different semantic terms of StockTwits data about the DJIA Index. One year of StockTwits data are downloaded from the website's Application Programming Interface (API) for the period of April 3rd 2012–5th April 2013. StockTwits postings were pre-processed where those posts were without any ticker, or had more than one ticker; those not in the DJIA index were removed, leaving 289,443 valid postings containing the dollar-tagged ticker symbol of the 30 stock tickers of the Dow 30. A random selection of a representative sample of 2892 tweets on all 30 stocks on the Dow Jones Index are hand-labeled as either buy, hold or sell signals. These hand-labeled messages constitute the training set, which is then used as an input for the decision tree model employed in this research.

Financial Data: The financial data is obtained from Bloomberg (the leading financial professional service provider), at a daily frequency covering the period from 3rd April 2012 to 5th April 2013. The daily closing prices of the DJIA index of the period of study from April 2012 to April 2013 are depicted in Fig. 3. There were no extraordinary market conditions reported during this period, so it represents a good base test for the evaluation. In this research paper, the focus will be placed on the DJIA index to adequately reflect the US stock market. The DJIA is a price-weighted average of 30 large 'blue-chip' stocks traded on the New York Stock Exchange (NYSE) and the NASDAQ. Regardless of the limitations in the composition and structure of the index, it is nevertheless the most widely followed and reported stock index (Lee, Jiang, & Indro, 2002). The DJIA is particularly suited for this study because it constitutes the large capitalization industrial companies of the US equity market. The 30 stocks making up the index comprise about 25% of the market value of all NYSE exchanges

[4] Tetlock, Saar-Tsechansky, and Macskassy (2008) even use only 10 bps to assume reasonable transaction costs.

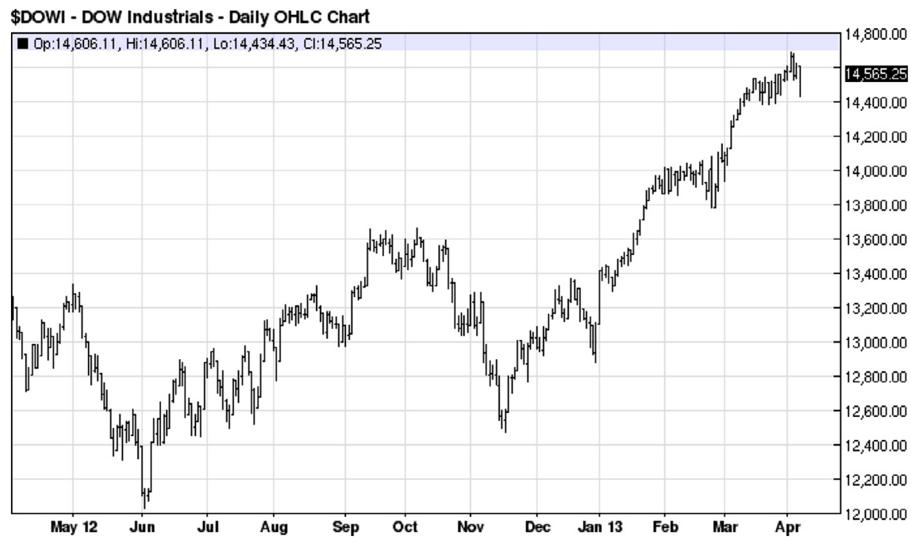

Fig. 3. Time series for the daily closing prices of DJIA from the period of 3rd April 2012 to 5th of April 2013.

Table 1
Features selected under the filter approach using information gain criteria.

Sr.	Feature	IG	Sr.	Feature	IG
1	short	0.0706	**24**	**utx**	0.0070
2	**cat**	0.0320	**25**	move	0.0070
3	**csco**	0.0251	26	stop	0.0066
4	bearish	0.0225	27	bull	0.0063
5	**aapl**	0.0205	28	**unh**	0.0059
6	bullish	0.0196	**29**	volum	0.0056
7	**cvx**	0.0170	30	**pfe**	0.0051
8	nice	0.0159	**31**	target	0.0047
9	breakout	0.0142	32	support	0.0046
10	lower	0.0141	33	**msft**	0.0044
11	**xom**	0.0123	**34**	bounc	0.0044
12	break	0.0121	35	entri	0.0043
13	look	0.0115	36	sell	0.0042
14	strong	0.0114	37	set	0.0042
15	quot	0.0099	38	weak	0.0042
16	current	0.0096	39	gap	0.0041
17	high	0.0089	40	head	0.0041
18	buy	0.0087	41	market	0.0040
19	**goog**	0.0082	42	flag	0.0039
20	post	0.0080	43	bottom	0.0038
21	report	0.0079	44	bought	0.0037
22	**spi**	0.0078	45	run	0.0034
23	**mrk**	0.0072			

(Lakonishok & Smidt, 1988). Therefore, focusing on large and highly traded firms would probably reduce the problems associated with non-concurrent trading (Rudd, 1979). This in fact makes the DJIA a reasonably valuable index for representing short-term market movements. In addition, since the companies that make up the DJIA are actively traded companies, their stocks generate a greater buzz on social media networks. Therefore, these stocks are heavily discussed in StockTwits and have very high volume of tweet messages, though this research study may still be valid to any companies/indices that generate high tweet volumes.

5.2. Feature selections and construction framework

Filter Approach: For extracting the filter subset, we used a ranker search method (Hall et al., 2009) in conjunction with the information gain criteria where the worth of an attribute is evaluated by measuring its information gain (IG) score with respect to the class. Table 1 presents the result of filter feature selection showing the list of terms ranked according to their IG values.

As can be seen from Table 1, 45 terms are retained from performing the filter approach using information gain criteria. The terms listed in the table are ranked according to their relevancies where the terms at the beginning of the list (indicated by the serial number) are most relevant, as the relevancy decreases as one goes down the list. The IG value is reported next to each term. For example, the term ‘short’ appears to be the most significant term among all listed terms with the IG value of 0.0706 while ‘run’ is the least important term with the IG value of 0.0034.

5.3. Text processing model

Decision Tree Model: Quinlan's C4.5 (DT) algorithm (Quinlan, 1993) is used to classify the tweet messages based on the reduced model of the features selected under a filter approach using the IG criterion. Information gain was used originally to measure splitting criteria for decision trees where it was used to find out how well each single feature separates the given data set. The decision tree model is structured in the form of a decision tree, where each node is a leaf node holding the class prediction of a specified attribute. Applying ten-fold cross validation reveals that the classification performance of the decision tree algorithm of the reduced features using IG, resulted in a slightly better classification accuracy of 60.9% than the accuracy achieved with the complete feature set of 60.7%, but with a reduction of about 55% of the feature space. The good performance of IG indicated that the reduced features were informative for this classification task.

As can be seen from Table 2 below the Decision Tree Classifier, based on IG results, is an overall sample of classification accuracy of 60.9%. This is considered a good percentage giving a random chance of 30% of the three classes (buy, sell and hold). The ROC (Receiver Operating Characteristic) Area measures the quality of the trade-off between true and false positives and shows accuracy of 73.7%. This validates the automatic classification of tweet messages for this research.

This table shows the classification accuracy by class. It demonstrates that the model used in this paper classified the majority of messages in the training set correctly. True positives (or precision) represent, for example, the share of messages classified as sell, which were labeled as such in the training set. False positives are messages incorrectly classified as sell. Recall represents the share of all messages of a particular class, which were classified correctly. The F-measure combines precision and recall. The ROC area measures the quality of the trade-off between true and false positives.

Table 2
The classification accuracy by class.

Class	True positive	False positives	Precision	Recall	F-measure	ROC area
Buy	76.6%	43.4%	61.1%	76.6%	68.0%	71.7%
Hold	48.1%	10.9%	53.1%	48.1%	50.5%	79.0%
Sell	46.1%	11.1%	66.8%	46.1%	54.6%	73.2%
Weighted average	60.9%	26.2%	61.3%	60.9%	60.0%	73.7%

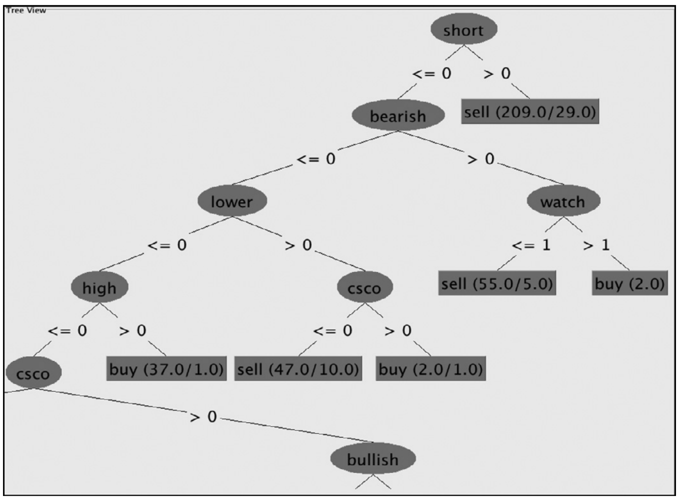

Fig. 4. An extracted version of the visualized decision tree model.

Performing feature selection using decision trees reveals that 45 attributes, indicated by the nodes in the tree model, are regarded as the most relevant features that can make a better prediction of the three decision classes (buy, sell, and hold). All of the selected features were deemed relevant in predicting the sentiment class, whether these feature nodes connected directly to the decision class or were connected through leaves with other decision nodes in the tree to the sentiment class. One of the main advantages of the decision tree model is that it naturally explores interactions between terms via the visualized connections between different nodes connected through leaves in the decision tree. To provide more understanding of the connected relationships between terms in the tree model, we provide an extracted version of the visualized tree while explaining the nature and type of these connected relationships for classifying StockTwits sentiment class. Fig. 4 shows the visualized output of some of the selected features of the decision tree near to the root node. Other aspects of the extracted versions of the tree model are shown in the Appendix A.

From the extracted visualized tree model graphed in Fig. 4, we can see that the decision node (sell) is connected, through leaves, to the words (short, bearish and lower). This indicates that these words are the most relevant words that best classify "sell" messages. Each term is indicated by a node in the tree and connected through leaves to one of the three decision classes (sell, buy or hold). In some cases, a set of terms might be connected together to one decision class, where this indicates that the combined appearance of these connected terms may have a different effect on the trading decisions than the decision when the term appears alone. For example, when the decision node "bearish" appears in a tweet message, it indicates a sell decision as it is connected through leaves to the decision class sell. However, when the term "bearish" is connected together with the decision node "watch" through leaves, it indicates a buying decision despite its individual independent appearance as a sell decision. Therefore, trading decisions can sometimes be affected inversely depending on whether each term appears independently or in combination. Due to the large size of the decision tree generated for StockTwits data in this research paper, we will provide another exemplary screen of a visualized decision tree in Appendix B.

A set of decision rules can be generated from the DT model by following the decision tree from top to bottom. These decision rules are based on the idea that the appearance of a term or a set of terms in tweet postings might inform investors about whether to buy, sell or hold a stock in a capital market. Therefore, it is worth pointing out at this stage the nature of the decision rules that can be extracted from the DT model. Table 3 shows the trading decision rules corresponding to each term that can be extracted from the decision tree model. Note that we exclude the individual appearance of terms indicating the company ticker symbols as shown in bold in Table 1 that is because including the single appearance of such terms might bias the volume frequency, which may result in misleading the strategy performance of such terms. However, we still consider the combined appearance of those terms, as they might be more informative when appearing together with other terms in the tweet postings.

The table above shows that there are some specific terms associated with the decision classes, sell, hold and buy where their appearance in a StockTwit message gives indications to financial market practitioners as to whether to sell, buy or hold the discussed stocks. For example, if terms like "bought", "bullish", "move" and "nice" appear in a tweet posting discussing a particular stock of DJIA that provides a buying signal to investors to buy that particular stock. While the appearance of terms like "bearish", "bottom", "lower" and "short" indicates a sell signal to investors and most probably recommends investors to take a sell decision concerning that particular stock. The appearance of terms such as "report", "market", "week" and "set" seems to inform investors to hold the discussed stocks. Looking closely at the nature of the terms associated with each decision class we find that StockTwits postings provide reasonable reflections of the linguistic bullishness of the three classes (buy, sell and hold). We find that positive emotional terms are more likely associated with the decision 'buy', which by nature reflects investors' optimism towards particular traded stocks in financial markets. On the other hand, negative emotional terms are likely to be associated with the decision 'sell' indicating investors' pessimism about that particular stock. Neutral terms are more likely to be found in a tweet message discussing a particular stock if a holding decision is to be made by investors.

Having discussed the decision rules associated with the individual occurrences of some terms in the StockTwits postings, it is important therefore to shed light onto the impact of the combined appearance of those terms with other terms in tweet postings. Table 4 shows the decision rules obtained from the DT model where it is the set of terms or combination of terms that constitute the decision rules rather than individual terms [5].

Table 4 provides the trading decision rules extracted from the DT model, where a set of rules based on the combined appearance of the terms in tweet postings are listed under the decision class where they belong. What is interesting in this table is that the companies'

[5] Note that in Table 4 to maintain unbiased results we only report the term and combination of terms that has a minimum total volume frequency of 100 over the period studied where the terms/combination of terms of less than 100 value frequency will be withdrawn from the analysis.

Table 3
The decision rules for the individual occurrence of the term in the StockTwits postings.

***Decision rule**: If the term*	*bearish, botoom, bounc, flag, lower, sell, short, stop, support, volume*	appears in a tweet message then the decision would be **Sell**
***Decision rule**: If the term*	*current, entri, market, post, report, set, week*	appears in a tweet message then the decision would be **Hold**
***Decision rule**: If the term*	*bought, break, breakout, bull, bullish, buy, head, high, look, move, nice, run*	appears in a tweet message then the decision would be **Buy**

Table 4
The decision rules for combinations of terms appeared in the StockTwits postings.

Decision rule: If the term"…", the term "… " and /or the term "…" appeared in a tweet message then the decision would be

Sell	**Hold**	**Buy**
unh + gap	appl + jpm + amzn	lower + csco
mrk + amp	csco + amzn + goog	bullish + csco
mrk + break	appl + jpm	csco + amzn
report + intc	appl + stock	csco +trend
report + wmt	appl +wmt	csco + break
break + mmm	appl + market	cvx + move
break + nke	appl + trade	cvx + entry
xom + bottom	appl + msft	cvx + xom
amp + head	goog + sell	cat + mmm
look + intc	sell + pfe	cat + break
stock + jpm	watch + jpm	cat + run
week + nke	watch + nke	cat + call
wmt + qout	watch + wmt	unh + hold
break + bounc	mcd + trade	unh + day
break + stock	stock + amzn	unh + look
break + support	jnj + wmt	appl + nke
break + weak	wmt + friday	report + jpm
bullish + market	chart + post	appl + ibm
chart + flag	market + time	spy + msft
chart + price	watch + follow	axp + ibm
day + expect	watch + list	axp + look
day + news	watch + news	amp + news
head + move		amp + stock
hold + gap		amp + daily
hold + look		amp + sold
look + close		amp + trade
look + daily		amp + dis
stock + current		stock + amzn + wmt
strong + support		jnj+ chart
support + break		bought + sell
week + daily		break + look
week + time		hold + bounc
yesterday + bought		hold + play
		report + low
		stop + current

ticker symbols such as "csco", "jpm", "mrk" …and many others, when combined with other terms, contain valuable information regarding investing decisions to be taken by investors not just merely a ticker symbol. The most surprising aspect of the decision rules presented in the table above is that the trading decision rules differ completely depending on whether the term independently appeared in a tweet message (see Table 3) or in combination with other terms. For example, while the appearance of the term "lower" in Table 3 indicates a purely sell decision, this term when combined with the company ticker symbol "csco" markedly indicated a buying position to be taken by investors (see last column of Table 3). Another example that demonstrates this finding is when considering the term "look" where its individual appearance indicated a buy signal to market participants, this term when mutually combined with other terms (i.e. "look + intc", "look + hold", "look + close" and "look + daily") excessively signifies a sell signal to investors.

5.4. Performance evaluation

Implementing the trading strategy explained in Section 3 for all of the 122 trading decisions reveals that the majority of the terms (95 terms trading strategies) outperform the random strategies indicated by positive returns. However, the remaining 27 terms show negative returns indicating that these strategies fail to perform better than random chance. Constructing the investment strategy defined in Eqs. (6) and (7) in Section 4 for each time series of all the terms/combination of terms presented in Table 4, we provide information not only about the cumulative average return but also about the number of the buy/sell signals per TTT strategy. Table 5 reports the number of trades per strategy along with its corresponding cumulative returns. As can be seen from Table 5, the first column reports the list of the tweet term, while the average returns and the number of trades of the corresponding term are shown in second and third columns respectively. The column reporting the number of trades indicates the total number of the buy and sell signals conducted for each term when implementing the investment strategy defined in Eqs. 6 and 7. The tweet term and/or combination of terms are listed

Table 5
The cumulative average returns and the number of sell/buy trades of the Tweet Term Trading (TTT) Strategies.

Term	Cumulative Average return	Number of trade		Term	Cumulative Average return	Number of trade	
		Buy	Sell			Buy	Sell
Report	0.153	136	116	lower+csco	0.061	163	89
Support	0.145	135	117	run	0.061	138	114
report+intc	0.129	160	92	post	0.061	136	116
break+support	0.113	142	110	break+look	0.061	143	109
chart+price	0.111	148	104	volume	0.060	131	121
Entri	0.110	131	121	move	0.058	124	128
amp+stock	0.109	126	126	sell	0.057	134	118
appl+trade	0.107	131	121	bullish+csco	0.057	171	81
amp+trade	0.107	137	115	bounc	0.056	130	122
yesterday+bought	0.102	175	77	axp+ibm	0.056	149	103
Bought	0.101	122	130	head	0.056	136	116
Gap	0.098	125	127	cvx+xom	0.055	126	126
day+expect	0.093	153	99	bullish+market	0.054	163	89
cat+call	0.093	135	117	appl+nke	0.054	144	108
bottom	0.089	125	127	qout	0.053	132	120
Market	0.088	129	123	current	0.050	136	116
Nice	0.084	138	118	axp+look	0.049	189	63
appl+msft	0.081	144	108	csco+amzn+goog	0.049	166	86
watch+follow	0.080	173	79	short	0.047	133	119
cvx+move	0.080	170	82	appl+jpm	0.046	139	113
Target	0.076	125	127	spy+msft	0.046	139	113
look+intc	0.074	140	122	buy	0.045	128	124
appl+market	0.073	126	126	csco+break	0.043	166	86
Stop	0.071	132	120	amp+daily	0.041	152	100
break+stock	0.071	148	104	goog+sell	0.040	149	103
amp+news	0.067	142	110	look+close	0.039	129	123
bearish	0.067	144	108	amp+head	0.039	150	102
mrk+amp	0.066	143	109	bullish	0.037	135	177
appl+ibm	0.065	138	114	market+time	0.036	143	109
report+low	0.065	151	101	appl+stock	0.036	125	127
Look	0.036	130	122	unh+look	0.008	181	71
report+wmt	0.036	171	81	mrk+break	0.006	187	65
cat+break	0.036	145	107	head+move	0.005	178	74
watch+jpm	0.036	148	104	bought+sell	0.005	154	98
stock+current	0.036	137	115	break	0.001	124	128
look+daily	0.035	139	113	appl+wmt	-0.004	131	121
xom+bottom	0.035	199	53	stop+current	-0.006	169	83
day+news	0.033	138	114	cat+run	-0.008	153	99
Set	0.033	136	116	sell+pfe	-0.010	184	68
Breakout	0.032	136	116	unh+day	-0.013	160	92
cvx+entry	0.031	204	48	stock+amzn+wmt	-0.014	184	68
wmt+Friday	0.031	173	79	flag	-0.017	133	119
stock+amzn	0.028	140	112	chart+post	-0.017	139	113
watch+list	0.027	143	109	strong+support	-0.019	142	110
hold+look	0.027	138	114	hold+gap	-0.019	160	92
hold+play	0.027	138	114	week	-0.020	130	122
Lower	0.027	114	138	break+mmm	-0.020	185	67
watch+wmt	0.024	156	96	break+weak	-0.022	136	116
watch+nke	-0.019	152	100	watch+nke	-0.022	148	104
High	0.023	133	119	csco+amzn	-0.022	156	96
chart+flag	0.022	163	89	hold+bounc	-0.023	157	95
unh+gap	0.020	204	48	week+daily	-0.024	134	118
jnj+chart	0.018	161	91	week+time	-0.025	148	104
Bull	0.017	129	123	stock+jpm	-0.031	137	115
report+jpm	0.015	152	100	appl+ibm+amzn	-0.033	147	105
unh+hold	0.015	186	66	break+nke	-0.034	163	89
csco+trend	0.013	160	92	cat+mmm	-0.039	149	103
Strong	0.013	130	122	amp+dis	-0.042	141	111
jnj+wmt	0.010	141	111	break+bounc	-0.046	163	89
wmt+qout	0.009	152	100	watch+news	-0.051	162	90
				amp+sold	-0.073	144	108
				mcd+trade	-0.079	147	105

in accordance with performance based on the cumulative average returns.

Evaluating the overall trading strategies reveals that the term “report” appears to be the best performing term in our analysis followed by the term “support”. Fig. 5 shows the monthly average cumulative performance of the top four trading strategies: “report”, “ support” “report+intc” and “support+ break”. The blue bars in the graphs depict the cumulative return of our trading strategies where the spikes of these blue bars are more likely pronounced at the top half of the figure indicating positive returns. The red bars on the other hand indicate the standard deviation of the cumulative return from random strategy (in which buying and selling is done in an uncorrelated random manner) where more spikes of these red bars are pronounced at the bottom half of the graph indicating negative returns in general. From the figure below, we can see that the trading strategy of the best four performed terms is performing better than random

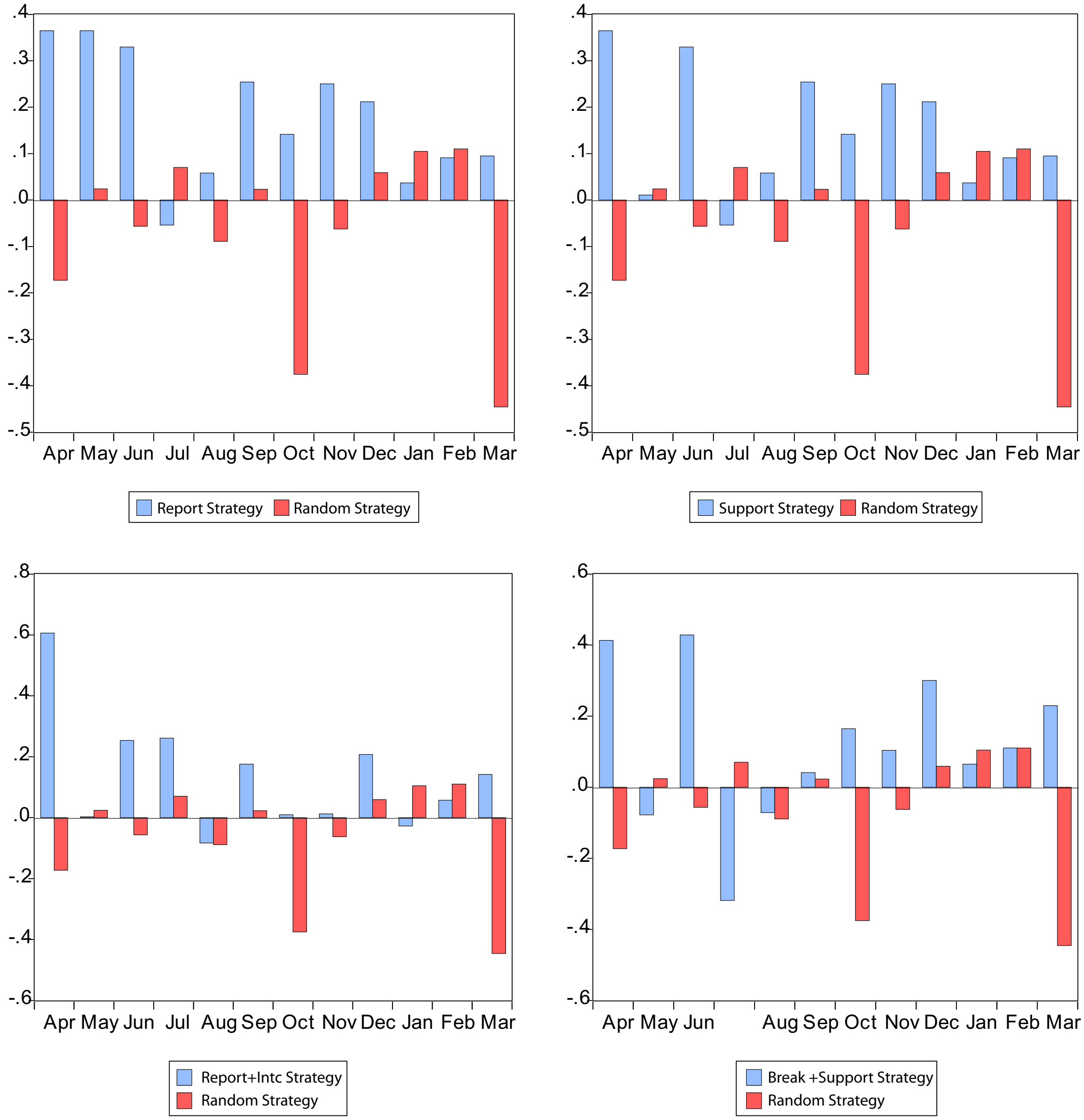

Fig. 5. A comparison of the monthly average cumulative performances of the trading strategy of the tweet terms “report”, “support”, “report+intc” and “break+support” with the random investment strategy.

strategy meaning that there are significant higher positive returns than the random investment strategies in all graphs. As can be seen from the four charts below that more spikes of blue bars are found in the upper area of positive returns in contrast with random strategy where the red bars spikes more in lower negative return area of the graphs.

We rank the full list of the 122 investigated tweet terms by their trading performance indicated by the cumulative average returns of each strategy. Fig. 6 depicts the cumulative return of the 122 (TTT) investment strategies based on their performance. From the figure below, we can see that the vast majority of the (TTT) strategies are profitable as it resulted in cumulative average returns greater than random strategy$\langle R\rangle_{Random\ Strategy} = 0$. The top half of the figure denoted by the red bars indicates the strategies with positive returns, while the bottom half of the figure signified by the white bars indicates the negative returns strategies. Taking the average return of all strategies, we find that returns from Tweet Terms Trading strategies tested are significantly higher overall than returns from random strategies ($\langle R\rangle_{TTTstrategies} = 0.0355, t = 8.705, df = 121, p < 0.001, \textit{one sample test}$). The t statistic would be calculated as follows:

$$\text{T-statistic} = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \tag{7}$$

where $\bar{x}$ is the average return of $\langle R\rangle_{TTT\ strategies} = 0.0355$, μ is the mean return of the random strategy $\langle R\rangle_{Random\ Strategy} = 0$, $S = 0.0450$ is the standard deviation of the 122 TTT strategies sample and $n = 122$ is the number of the TT trading strategies. Using a one-tailed test and 0.001 level of significance and n–1 degree of freedom (121 df) the result of t-statistics is 8.705 > 3.1589 (critical value), which leads to a rejection of the null hypothesis and concludes that the average returns of TTT strategy is statistically different than the mean return of the uncorrelated random strategy. This result indicates that our TTT strategies are successful and could produce potential return from implementing them in stock markets. Despite the small average returns of 3.55% of the TTT strategy, these returns exceed the frequently assumed levels of transaction costs for online brokers that range from 0.15% to 0.2% where the net profits produced by

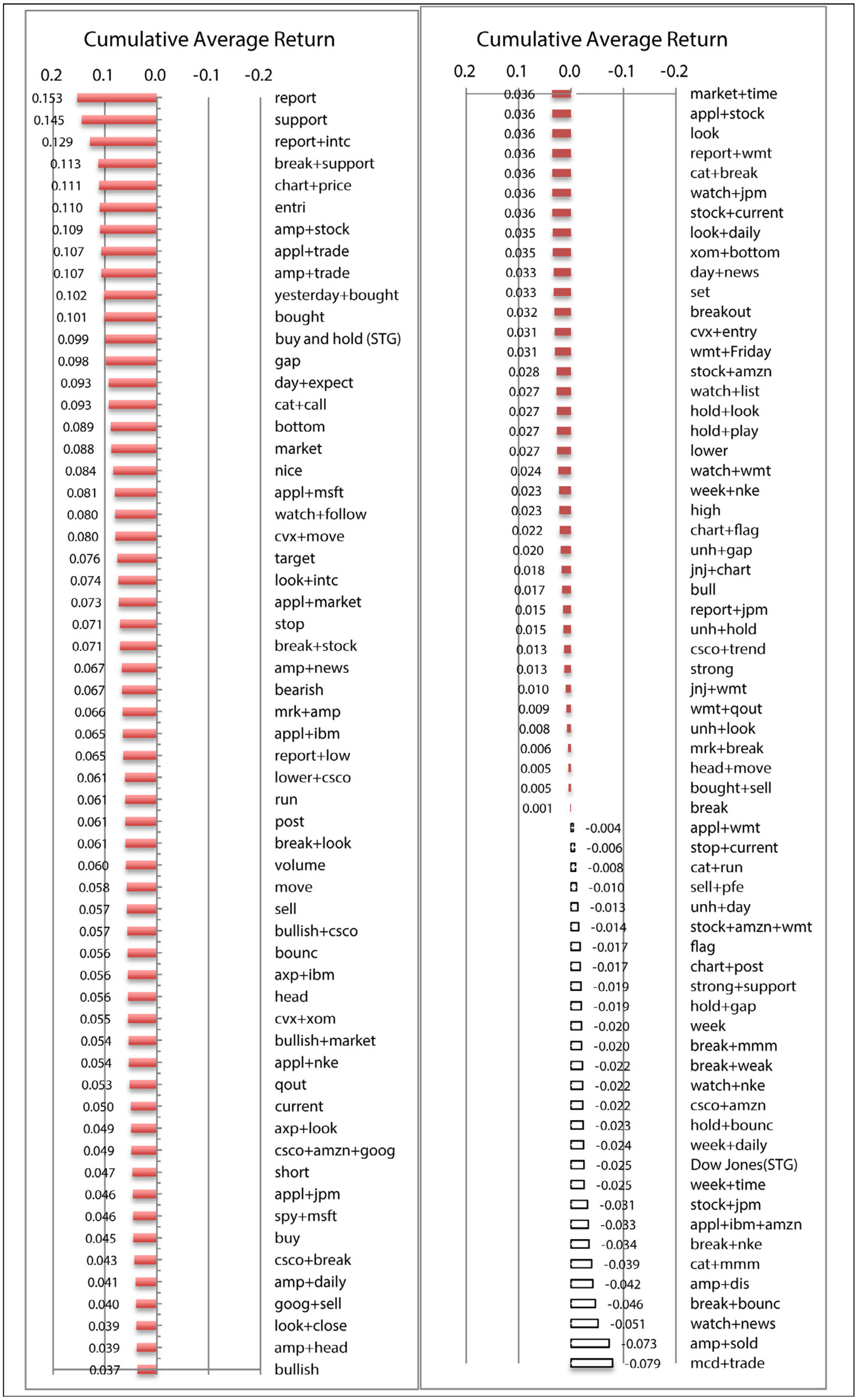

Fig. 6. Performance of TTT investment strategies based on term related frequency. Cumulative returns of 122 investment strategies based on tweet term volume frequency are displayed for the entire time period of the study from 3rd of April to 5th of April 2013. Two colors of bars are used to distinguish the positive return strategies from the negative returns. We use red bars for the positive returns and white bars for the negative returns. The cumulative performance of the "buy and hold" strategy and the "Dow Jones" strategy is also shown. Figures depicted next to the bars indicate the returns of a strategy, R, in standard deviation from the mean return of uncorrelated random investment strategy, $\langle R \rangle_{RandomStrategy} = 0$. The lines correspond to 0.2, 0.1, 0, –0.1, –0.1 standard deviations of random strategies. All strategies' returns fall between [0.2, –0.2] standard deviation of RND strategy.

Table 6
The mean-variance analysis.

Mean-variance analysis			
Investment strategy	Mean	Variance	(Mean-variance)
TTT	0.035	0.002	0.033
Random strategy	0	1	-1
Buy and hold	0.099	0.558	-0.459
Dow Jones	-0.025	0.559	-0.584

our TTT strategy are between 3.4% and 3.35%. On the other hand, the ‘Buy and Hold’ strategy resulted in a return of 0.09845 that is a slightly higher return than the overall average of TTT strategies, $(\langle R\rangle_{TTTstrategies} = 0.0355, t = -15.4791 < 3.1589, p < 0.001,$ *onesampletest*) which concluded that a ‘buy and hold’ strategy is considered more profitable than the $R_{TTT\,strategies}$. However, considering the performance of the individual term or term combination trading strategy our results show that there are some terms and/or combination of terms trading strategies that outperform the ‘buy and hold’ strategy. Those strategies are: ‘bought’, ‘yesterday and bought’, ‘amp and trade’, ‘appl and trade’, amp and stock’, ‘entri’, ‘chart and price’, ‘break and support’, ‘report and intc’, ‘support’, ‘report’. In contrast to ‘buy and hold’, the ‘Dow Jones’ strategy underperformed the average returns of TTT strategies where the ‘Dow’ strategy resulted in negative returns of –0.0245 compared to 0.0355 of the mean returns of TTT strategies.

5.5. Mean-variance analysis

Mean Return should not be the only evaluation factor to consider when evaluating profitability of an investment strategy. A trading strategy is considered superior over another strategy if the risk factor is also involved in the benchmarking process. Mean variance analysis is an element of modern portfolio theory whereby a more efficient investment strategy is made by a rational investor through the process of weighting the variance against expected returns of an asset (Markowitz, Todd, & Sharpe, 2000). Table 6 shows the resulting analysis of the mean-variance of each of our studied trading strategies. Note that the Random Strategy is derived from simulations of 1000 independent realizations of uncorrelated random variables that have a mean of zero and a variance of one whereby at any number of realizations of uncorrelated variables this strategy will always have a mean of zero and a standard deviation of one ($\mu = 0$ and $\sigma = 1$).

As can be seen from Table 6, our TTT strategies outperform the other benchmark strategies when the risk factor is taken into consideration. All other benchmark strategies (Random, Buy and Hold and Dow Jones Strategy) show a high risk compared to their expected returns indicated by the negative value in the mean-variance column in Table 5. While the buy and hold strategy showed better performance when the mean return was the only factor in the evaluation process, it does not show any good performance when the risks are considered. The TTT strategies are considered the superior strategy among all other benchmark strategies where it exhibits positive returns while maintaining the same level of profitability with a lower level of risk. Although the buy and hold strategy is a more profitable investment choice, it however involves much more risk than our TTT strategy.

5.6. Portfolio constructions and investment hypothesis

This paper aims to investigate the predictability between the TTT decisions obtained from the decision tree algorithm and the market behavior of stocks of the DJIA index. To start the analysis, we construct three portfolios namely sell, buy and hold portfolio. Each portfolio consists of all possible terms and/or combination of terms belonging to a particular decision. For example, all sell decision rules extracted from the decision tree corresponding to the sell class will be listed under sell portfolio. The same with the buy and hold portfolios, where all the decision rules belonging to the buy or hold class will constitute the buy and hold portfolios respectively. Table 7 shows the list of terms constituting the sell, buy and hold portfolios.

As can be seen from the table above, a total of 122 trading decisions were returned from the decision tree algorithm C4.5. The sell portfolio consists of 49 terms, while 44 terms indicated buying decisions and 30 terms represented holding decisions.

5.6.1. Cumulative performance of the sell, buy and hold portfolios

This section documents the strategies’ returns of the portfolio constructed in Section 5.5. The returns of all terms constituting each portfolio are calculated based on the trading strategy described in Section 4. Fig. 7 shows the average returns of the 122 different terms distributed based on their trading decisions in the sell, buy and hold portfolio. The most successful strategies are those terms composing the sell portfolios that yielded higher average returns of 0.0408 compared to 0.0369 and 0.0366 for the sell, buy and hold portfolio respectively. All portfolios returns are statistically significant and higher overall than returns of the random investment strategy. The individual t-statistics of each portfolio are sufficiently large to be significant to reject the null hypothesis that the mean portfolio returns are equal to the mean return of the random strategy.

5.6.2. Investment hypothesis evaluation

This section evaluates the effectiveness of the trading strategies in anticipating subsequent moves in financial markets. Our results show that performance of the Tweet Term Trading (TTT) strategies varies with cross terms or (combination of terms) that appeared in tweet postings. We additionally found that the different buy, sell and hold portfolios produce different average cumulative returns suggesting that each of these portfolios would have different roles in affecting our strategy returns. We implement our empirical result based on a two-part investment hypothesis, which are:

Table 7
The term trading strategies in the sell, hold and buy portfolios.

Portfolio	Term trading strategies
Sell portfolio	“bearish”, “bottom”, “bounc”, “flag”, “gap”, “lower”, “sell”, “short”, “stop”, “support”, “volume”, “unh+gap”, “mrk+amp”, “mrk+break”, “report+intc”, “report+wmt”, “break+mmm”, “break+nke”, “xom+bottom”, “amp+head”, “look+intc”, “stock+jpm”, “week+nke”, “wmt+qout”, “break+bounc”, “break+stock” “break+support”, break+weak”, “bullish+market”, “chart+flag”, “chart+price”, “day+expect”, “day+news”, “head+move”, “hold+gap”, “hold+look”, “look+close”, “look+daily”, “stock+current”, “strong+support”, “support+break”, “week+daily”, “week+time”, “yesterday+bought”
Hold portfolio	“current”, “entri”, “market”, “post”, “qout”, “report”, “set”, “week”, “appl+ibm+amzn”, csco+amzn+goog”, “appl+jpm”, “appl+stock”, “appl+wmt”, “appl+market”, “appl+trade”, “appl+msft”, “goog+sell”, “sell+pfe”, “watch+jpm”, “watch+nke”, “watch+wmt”, “mcd+trade”, “stock+amzn”, “jnj+wmt”, “wmt+friday”, “chart+post”, “market+time”, “watch+follow”, “watch+list”, “watch+news”,
Buy portfolio	“bought”, “break”, “breakout”, “bull”, “bullish”, “buy”, “head”, “high”, “look”, “move”, “nice”, “run”, “strong”, “target”, “lower+csco”, “bullish+csco”, “csco+amzn”, “csco+trend”, “csco+break”, “cvx+move”, “cvx+entry”, “cvx+xom”, “cat+mmm”, “cat+break”, “cat+run”, “cat+call”, “unh+hold”, “unh+day”, “unh+look”, “appl+nke”, “report+jpm”, “appl+ibm”, “spy+msft”, “axp+ibm”, “axp+look”, “amp+news”, “amp+stock”, “amp+daily”, “amp+sold”, “amp+trade”, “amp+dis”, “stock+amzn+wmt”, “jnj+chart”, “bought+sell”, “break+look”, “hold+bounc”, “hold+ply”, “report+low”, “stop+current”

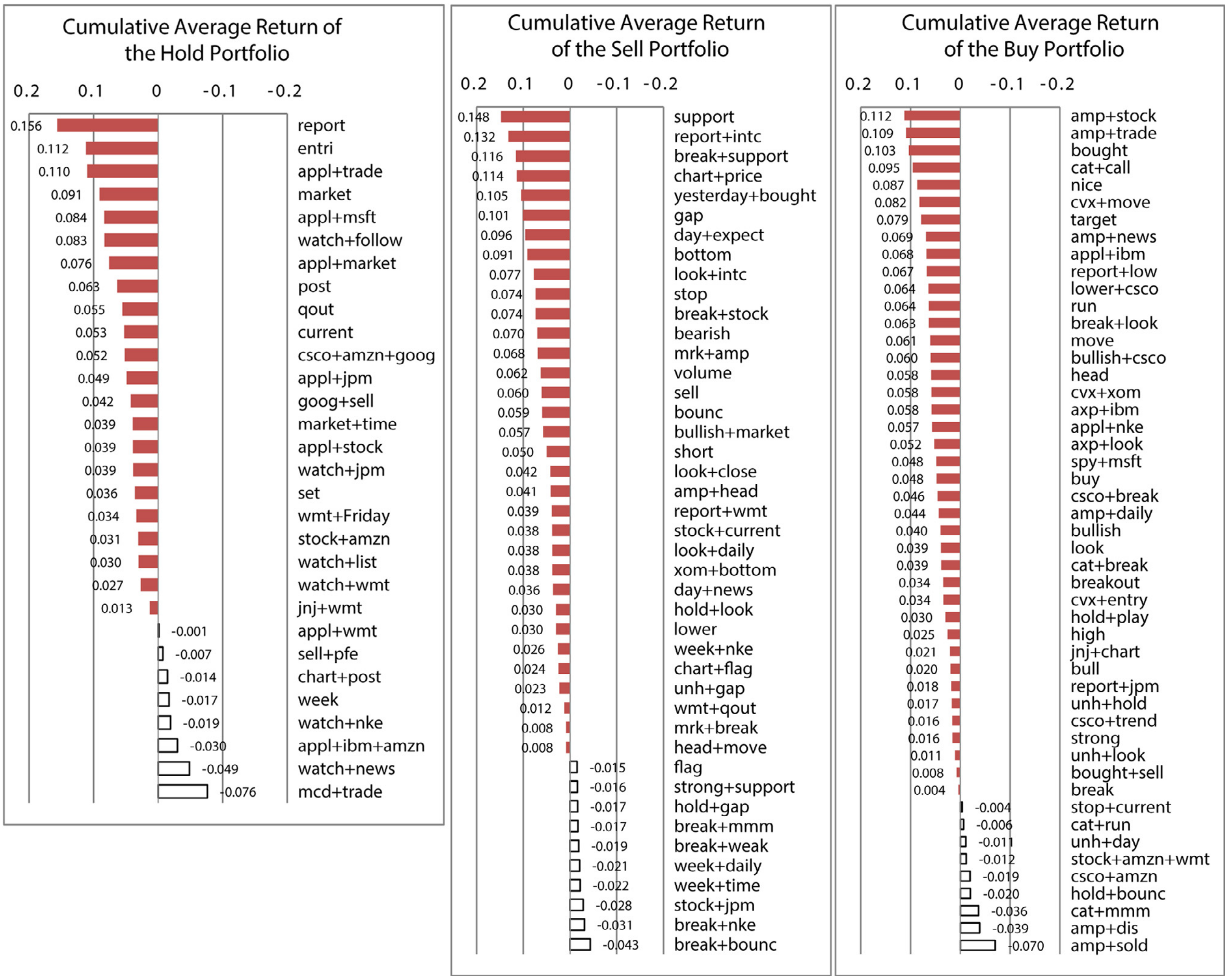

Fig. 7. Performance of sell, buy and hold portfolios strategies. Cumulative returns of 122 investment strategies distributed based on their trading decision into the sell (43 terms), buy (49 terms) and hold (30 terms) portfolios. Two colors of bars are used to distinguish the positive return strategies from the negative returns. We use red bars for the positive returns and white bars for the negative returns. The cumulative performance of the "buy and hold" strategy and the "Dow Jones" strategy is also shown. Figures depicted next to the bars indicate the returns of a strategy, R, in standard deviation from the mean return of uncorrelated random investment strategy, $\langle R\rangle_{RandomStrategy} = 0$. The lines correspond to 0.2, 0.1, 0, –0.1, –0.1 standard deviations of random strategies. All strategies' returns fall between [0.2, –0.2] standard deviation of RND strategy. The average returns of all of our portfolios (sell, buy and hold) are positive. The t-statistics of the portfolios' returns using one tailed test are ($\langle R\rangle_{sell\,port\,folio} = 0.0408$, $t = 5.600 > 3.2959$ $df = 42$, $p < 0.001$); ($\langle R\rangle_{buy\ port\,folio} = 0.0369$, $t = 6.506 > 3.2689$ $df = 48$ $p < 0.001$); ($\langle R\rangle_{hold\,port\,folio} = 0.0366$, $t = 3.997 > 3.3969$ $df = 29$, $p < 0.001$) for the sell, buy and hold portfolio respectively.

- Increases in the prices of the DJIA were preceded by a decrease in the volume frequency of related terms, which prompt us to sell or take a short position.
- Decreases in the prices of the DJIA were preceded by an increase in the volume frequency of related terms, which prompt us to buy or take a long position.

It is therefore important to test and verify these two strategy components. To validate the significant role each part of this hypothesis plays, we implement these two strategy components by examining the asymmetric effects of the increase and decrease of the mean relative change in the tweet term frequency. At each day t we calculate the mean relative change in the term frequency, $x_{i,t} = \Delta n(t, \Delta t)/N(t-1, \Delta t)$ for the sell, buy and hold portfolios over the previous five days average.

In order to test each part of our hypothesis, we would expect that the sell portfolio terms would confirm our first part, in which the appearance of such terms signify a sell signal in the stock market (short position), while the buy portfolio terms would be used to explain and verify the second part of our investment hypothesis, fueling the fact that the appearance of those terms in tweet messages indicates a buy signal to other market participants (long position). Whereas the holding decision would have a limited effect on the profitability position of an investor in a capital market and we would expect that the returns of the hold portfolio may have equal feedback to the effect of the increases and decreases of the mean relative frequencies of the tweet terms. We now formally investigate whether the language of StockTwits provides new information about investment decisions in stock markets and whether stock market prices efficiently incorporate this information. This approach also permits us to explore relationships between the magnitudes of the increases and decreases in volume frequency of the related terms and the magnitude of the subsequent returns of our trading strategies.

To isolate the effects of an increase or decrease in the mean relative change of a term, we compute the following indicator variables,

$$I^{+} = \begin{cases} 1 & if\ x_{i,t} > 0 \\ 0 & otherwise \end{cases} \quad ; \quad I^{-} = \begin{cases} 1 & if\ x_{i,t} < 0 \\ 0 & otherwise \end{cases}$$

Table 8
Predicting portfolio's trading strategy returns based on the asymmetric effects of the increase and decrease in the mean relative changes of the term related frequencies. On data measured on daily frequency, panel regressions with term fixed effects are estimated separately for each portfolio j = (sell, buy and hold) where trading strategy returns are used as a dependent variable. The independent variables were obtained from the mean relative change in volume of a term appeared in StockTwits postings in a particular day (t): The positive (increase)x_t^+ and negative (decrease) x_t^- variation in mean relative change of volume of a term (i) in portfolio (j). This table shows the predictive power of the positive and negative variation in tweet term volume in explaining the subsequent change of trading strategy returns of different portfolios. In all regressions, Market return is added as a control variable. Market return denotes the log difference of DJIA price. To control for Monday return anomaly, dummy variable for first day of the week is added in all portfolio returns regressions.

Subsequent return $\boldsymbol{R}_{(t+1)}$	Increase in mean RCHG $\boldsymbol{x_{it}}^+$	Decrease in mean RCHG $\boldsymbol{x_{it}}^-$	Market	Dummy
Sell portfolio	0.0060	0.0324**	−0.0185*	0.0026
	(1.5421)	(2.1687)	(−1.9185)	(0.1514)
Buy portfolio	0.0052**	0.00415	−0.0211**	−0.0108
	(1.9915)	(0.3064)	(−2.3420)	(−0.6719)
Hold portfolio	0.0087**	0.0371*	−0.0215*	−0.0171
	(1.9919)	(1.9073)	(−1.8563)	(0.3927)

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, t-statistics in parenthesis below the coefficients.

For the increase in the mean relative frequency the indicator variable I^+ takes the value of one when $x_{i,\,t}$ is positive, and the value of zero otherwise. Likewise, for the decrease in the mean relative frequency the indicator variable takes the value of one when $x_{i,\,t}$ is negative, and the value of zero otherwise. Accordingly, we create those two variables for each term undertaken in this study analysis.

We focus on Ordinary Least Square (OLS) regression estimates of the effect of increases and decreases of the mean relative frequency of terms of different portfolios on the subsequent returns of our investment strategy relative to the occurrence of the terms in StockTwits postings. Therefore in this section, panel regression with cross section fixed effect for each term i is employed to estimate the contemporaneous regressions for each portfolio j (sell, buy and hold) separately. Regressions will be estimated using standard ordinary least squares (OLS) techniques, where the return from our trading strategy is treated as a dependent variable and regressed on two independent variables; the increase in the mean relative change of the terms frequency indicated by $x_{i,t}{}^+$, and $x_{i,t}{}^-$ indicates the decreases in the mean relative change of the terms frequency. The market return of DJIA index return is added in the regressions as a control variable to control for overall market wide effects. Dummy variables for the first day of the trading week (NWK) are also added in return regressions to control for the potential Monday return effect in line with Antweiler and Frank (2004). The OLS subsequent return regression equations for each of the three portfolios are shown in Table 7 and can be expressed as:

$$R_{it,SellPortfolio} = \alpha_1 + \ \beta_1{}^+x_{it}{}^+ + \beta_2{}^-x_{it}{}^- + \beta_3 MRKT_{it} + \varepsilon_{it} \tag{8}$$

$$R_{it,BuyPortfolio} = \alpha_2 + \ \beta_4{}^+x_{it}{}^+ + \beta_5{}^-x_{it}{}^- + \beta_6 MRKT_{it} + \varepsilon_{it} \tag{9}$$

$$R_{it,HoldPortfolio} = \alpha_3 + \ \beta_7{}^+x_{it}{}^+ + \beta_8{}^-x_{it}{}^- + \beta_9 MRKT_{it} + \varepsilon_{it} \tag{10}$$

The OLS estimates of the coefficients β_s in Eqs. (8)–(10) are the primary focus of these regression equations. These coefficients describe the dependence of the positive (increase) and negative (decrease) variation in mean relative change of volume of a term that appeared in a tweet message on the subsequent change of returns (returns of our investment strategy calculated in an earlier section). Table 7 summarizes the estimates of β_s.

The regression results of the return equations of the sell, buy and hold portfolios as shown in Table 8 are largely as we would expect. That is, that the term trading strategies constituting each portfolio were generating positive returns indicated by the positive β_s coefficients (regardless of not being statistically significant) of the asymmetric effects of both the increase and decrease in the mean relative change of the terms frequency in all three portfolios regressions. These results are in line with what we found previously in Section 5.6.1. However, to test the two parts of our investment hypothesis we need, therefore, to investigate in depth analysis of the effect of the increase and decrease in the mean relative changes of the terms frequency in each portfolio separately. Looking at the sell portfolio regression, we found a statistical significant coefficient of the decrease in the mean relative change in the terms frequency (β_2 = +0.0324, p-value <0.05) while an increase on the other hand, exerted no statistical significance in forecasting our portfolio trading strategy returns. This suggests that the decrease in the mean relative change of the sell terms that appeared in tweet postings have a proportionally larger impact on the subsequent returns of our TTT strategies of DJIA index than an increase. Since the appearance of the terms in the sell portfolio signifies a sell signal to market participants, a decrease in the appearance of such terms conveys a good signal before market rises. One possible explanation of these results could be interpreted from a psychological viewpoint. The most common words that are more likely to appear in sell messages in StockTwits are negative words like “break” and “lower”, “bottom” as well as words like; “sell”, “bearish” and “short” which give a clear sign indicating that investors expected the discussed stocks to fall. Therefore, a decreased appearance of such negative words/terms is an indication of a decrease in an investor's bearishness, which implies good signals to their relative peers in the market that prices will start to recover and move upwards. These findings strengthen the first part of our trading hypothesis that an increase in DJIA prices was proceeded by a decrease in the volume frequencies of the sell terms, which prompts us to sell or take short position.

The buy portfolio regression however shows inverse results to what we have found in the sale portfolio regression. There is no statistical significant effect in the relation between the decrease in the mean relative change of the related terms frequency and the buy portfolio returns. But we have found that the increase in the mean relative change of the term frequency exerts a statistically significant influence in forecasting the buy portfolio trading strategies returns of DJIA index. Despite the statistical significant effect of the increase in the mean relative change, the estimated effect of +0.0052 is very small in magnitude. However, even here such tiny price effects would be difficult to take advantage of because this potential gain would likely to be offset even by transaction costs resulting in relatively trivial gain if not negative. Hence, an increase in the mean relative change of the buy terms is more likely to be followed by a decrease in DJIA prices where people see a buying opportunity and tend to take a long position in the market. Since the buy terms that appeared in StockTwits messages indicates an investor's optimism and provides a “buy” signal to the market participants, an increase in such terms will increase bullishness of investors where they are more likely to see a buying opportunity of stocks expecting prices to fall. Our evidence supports the “bargain shopper” hypothesis: the market speculators who see stocks becoming a bargain, they see a buying opportunity and become bullish (Brown & Cliff, 2004). These results however confirm the second part of our trading hypothesis that decreases in the prices of the DJIA were proceeded by an increase in the volume frequency of related terms, which prompts us to buy or take the long position in capital markets.

Looking at the hold portfolio regression in Table 5, we found that both the increase and decrease in the mean relative change of term frequency exerts a statistically significant positive effect in explaining our strategy returns indicated by the significant coefficients of ${\beta_7}^+ =$ + 0.0087, p-value < 0.1 and ${\beta_8}^- =$ + 0.0371, p-value< 0.05 for the increase and decrease in the mean relative change respectively. The estimated coefficients of both effects are economically small, which is in the context of our investment hypothesis; H_a: an increase in DJIA prices preceded by a decrease in the term volume frequency (which recommends investors to sell and take short position). This will be offset by the inverse effect of the second part of our hypothesis; H_b: a decrease in DJIA prices preceded by an increase in the term related frequency (which prompts investors to buy and take the long position). This result is not surprising, however, where in real life economics a holding decision has taken place where an investor is not optimistic enough to buy a stock, but not pessimistic enough to sell a stock. This is also true if one gets closer to investigate the nature of the words/terms comprising the hold portfolio, where an equal balance of positive and negative terms/combinations of terms is more likely to be found. It also contains neutral words like "report", "qout", "entri" as well as the name of the companies like; "cat", "jpm" and "wmt". The appearance of these kinds of terms in tweet messages would cause an investor to hold a neutral opinion about particular traded stocks where they most probably take holding decisions rather than buy or sell. The coefficients β_3, β_6 and β_9 of the market return (DJIA) index were statistically and negatively significant in all portfolios' regressions whereas, the dummy variable of the first day of the week effect, reported insignificant in all regression equations in Table 8.

6. Implications

This research study has implications to the following two groups of stakeholders, namely researchers and practitioners.

6.1. Contribution to the research community

This study contributes a decision support artifact using emerging social networks to research communities. The models and approach constructed herein may become the groundwork for future research where researchers and practitioners alike may find it a fruitful area of research to pay attention to the boom of financial blogs in understanding the significant role of sentiment, especially micro-blogging sentiments, in predicting stock price behavioral movement in stock markets. This study specially contributes to two different groups of a research community: financial research community and data mining community. The theoretical investigation presented in this paper contributes to the finance literature in strengthening ties with reference disciplines in tackling and addressing the ongoing debate between efficient market hypothesis (EMH) (Fama, 1970), random walk theory and behavioral finance theory. We provide support to behavioral finance theory in the existence of different types of investors in financial markets and their sentiment effect of their trading behavior in influencing price changes (DeLong et al., 1990). With regard to the Data Mining community, this paper proposes a novel method of performing feature selection to effectively extract the most relevant words and terms to provide better predictions of investors' decision-making. This method helps in the selection of an accurate set of relevant features, thus providing an insight into the relevancies present within the financial information used.

6.2. Contribution to practitioners

This study contributes to two groups of practitioners, investors and managers. Institutional and individual investors have both been long demanding an effective and efficient mechanism to predict prices in financial markets. This research paper offers a primary contribution by providing real-time investing ideas by utilizing stock micro-blogging sentiments. This assists investors with the potential for practical applications that provide investors and their peers with an investment decision support mechanism. We present a nascent approach by providing them with the robust methodology that could provide guidance to investors and other financial professionals for constructing and rebalancing their investment portfolios. This potentially offers guidelines to help investors and traders determine the correct time to invest in the market, what type of stocks or sectors to invest in, and which ones yield maximum returns on their investments. Moreover, companies and managers may choose to disseminate their financial reporting information and or advertise with postings deemed with the higher predictive value.

7. Future improvements

Regarding future improvements, one extension to our study is to explore the interactions between StockTwit terms by utilizing the wrapper approach of feature selection. Performing the wrapper approach may result in an interesting set of term combinations. Since the selected terms are tailored with the machine algorithm used, this might provide another pertinent extension to explore and compare different sets of combinations from different machine learning algorithms.

Another extension of this study is to consider training the model on StockTwits of each company ticker make up of the Dow index, separately over a longer period of time. By doing so one could explore and investigate the firm-specific terms and how different terms and or, combinations of terms may interact and interrelate in each ticker rather than considering the market index as whole. This is expected to help improve the performance of ticker sentiment prediction.

8. Conclusion

In this research paper, we proposed a novel approach by combining text mining, feature selection and a decision tree model to quantify and predict investor sentiment from a stock micro-blogging forum (StockTwits) of DJIA companies. The experiments reported in this paper provide quantifications of the StockTwits semantic terms trading decisions extracted from the decision tree algorithm, while providing a linkage between changes in the volume of semantic terms and subsequent stock market moves. The findings of this paper proved the success of our investment-trading hypothesis implemented for the different semantic terms trading strategies of StockTwits. We suggest two subsequent stages in the decision making process of investors using both StockTwits semantic terms and stock market data. Trends to sell short a stock at higher prices resulted from a decrease in the volume appearance of negative words (terms constituting the sell portfolio) in the tweet postings, while the trends to buy or take long positions resulted from an increase in the volume appearance of positive words (terms constituting the buy portfolio) in tweet postings.

Overall, our results indicate the existence of the asymmetric effect of StockTwits sentiments indicated by the (sell, buy and hold) portfolios on the subsequent moves in the stock market. We confirm that StockTwits postings contain valuable information and precede trading activities in capital markets. Changes in the average occurrences of different semantic terms in StockTwits postings informed decisions on whether to buy or sell the DJIA stocks. The findings of this research paper may yield promising insights into the potential

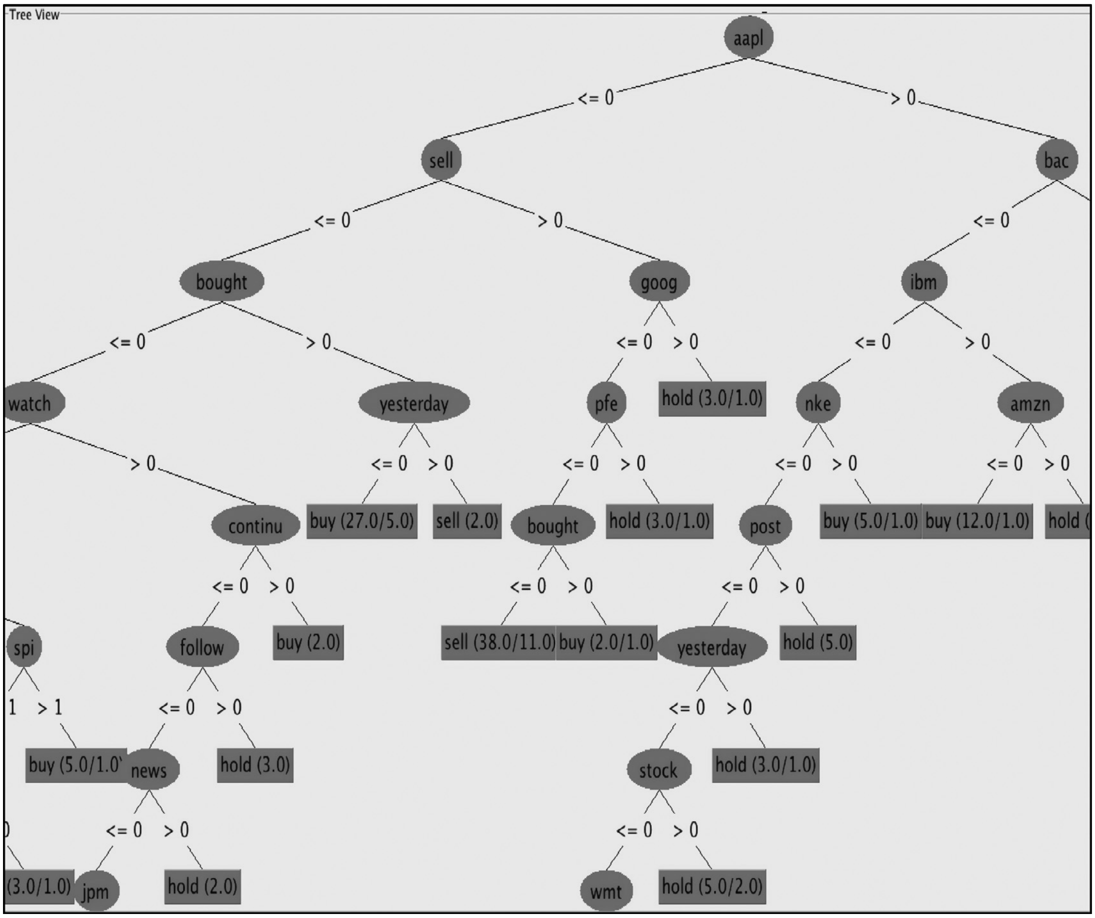

Fig.B1. A screen shot of visualized DT model: focus on the decision nodes: “appl” and “sell”.

provision of an investment support mechanism for analysts, investors and their peers. Practically, this could be used to determine the precise time when stocks are to be held, added (buy) or removed (sell) from a portfolio, thus yielding the maximum return on the investment for the investor. This could save time and effort and will lead to making a better-informed investment decision in the capital market.

Appendix A. The general rules for manually hand labeled tweet

There are a number of general rules applied in labeling the StockTwits data that are used as input (training set) in the text-processing model. These rules are as follows:

(i) If the tweet post contains external links of long articles or numerical charts about the stocks, it is generally marked as neutral. The content of the article and the information revealed by the chart are not taken into account.
(ii) Buy, hold or sell labels are only given when the sentiment can be explicitly speculated from the tweet.
(iii) Tweets with question marks are generally marked as neutral.
(iv) Simple summarizations of the stock performance by the end of the day are not taken into consideration.
(v) If the user reports a loss in a subjective way instead of reporting numbers, it is fair to assume that the user has a negative feeling towards the stock and vice versa.
(vi) If a tweet post contains company names (Apple, Google, Microsoft) or any other neutral words like; day, report, look, watch, etc.), it is generally marked as a hold message.
(vii) All positive words/emotions in a tweet message gives indication of linguistic bullishness (e.g. strong, high, happy, earn, etc.) will therefore be marked as a buy message.
(viii) Sell messages contain corresponding bearish words (e.g., loss, weak, low, fall, decline, down, etc.); therefore all negative words/emotions in a tweet message give an indication of linguistic bearishness and are commonly marked as a sell signal.
(ix) Normally tweet posts containing a balance of positive and negative words will be classified as a hold message.
(x) A tweet post containing a mixture of positive and negative emotional words will be assigned to the class with the highest probability. For example, if a tweet message contains 65% positive words, 20% negative and 15% neutral words, the message will be classified as a buy message since positive words are more likely associated with the buy signal.”

Appendix B. An extracted screen of visualized decision tree

These two figures show an exemplary screen of visualized decision tree (DT) model of StockTwits data. The decision tree is structured where each node in the tree is connected through leaves to another decision nodes where both connected to a leaf node holds the class prediction. The prediction class may take three possible states c= {Sell, Buy, Hold}. An inductive (If-Then) rule is created for each path from the root to leaf by which the trading decision is predicted. The visualized tree displayed all possible trading decisions represented either by an individual term or pair wise combinations of terms. Looking more deeply into Fig. B1, one could find that there are numbers of trading decision guidelines can be extracted from the DT model, based on (if-then) rule.

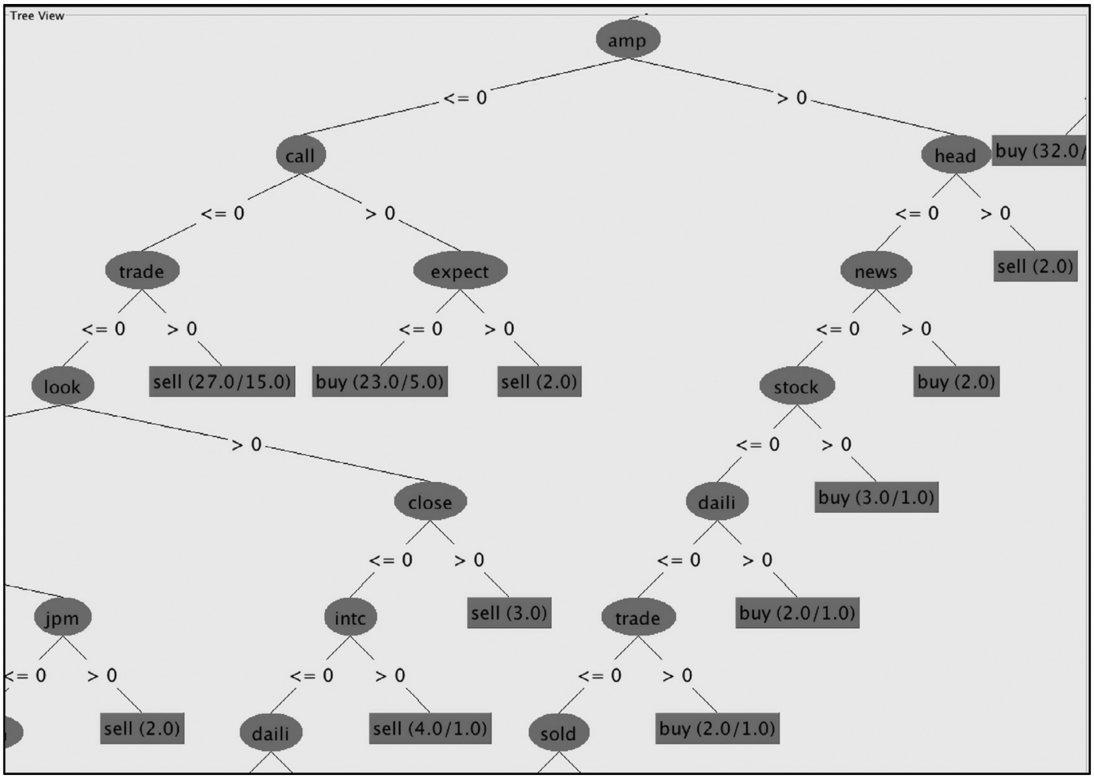

Fig. B2. A screenshot of visualized DT model: focus on the decision nodes: "amp" "look" and "call".

From the extracted visualized tree model graphed in Fig. B1, we can see that the decision class (hold) is connected, through leaves, to the words (appl, bac, ibm, nke, post, yesterday and stock). This indicates that these words are the most relevant words that best classify "hold" messages. Each term is indicated by a node in the tree and connected through leaves to one of the three decision classes (sell, buy or hold). For example, *if* the combined terms such as "appl+ ibm+ amzn", "appl+ post" "appl+ yesterday" and "appl+ stocks" appears in a tweet message, *then* the decision that investor recommends to take is to hold that discussed stock. The decision node of an attribute might be effected differently depending on whether that attribute appease individually or in combinations with other attributes in the tree. For example, by viewing another path in the tree such as the decision node "sell" where it connected to another decision nodes in the tree such as: "goog, pfe and bought". If the term "sell" appears alone in a tweet message it will excessively signifies a sell decisions indicated by the predicted decision class {sell} at the end of the leaf nodes that holds a sell class. Whereas the decision node "sell" shows exactly inverse decision when it connected to the decision nodes "bought" that is if the combined terms "sell + bought" appears in a tweet it will shows a buying decisions indicated by the decision class "buy" at the end of the tree root. While another decision might probably recommended when the term "sell" combined with the term "pfe" where it show a holding position.

Exploring different interactions of the combined appearance of the terms in tweet postings, another screenshot is presented in Fig. B2. The decision node "amp" is connected through leaves to different decision nodes such as: (head, news, stock, daily, trade and sold). Interestingly that all pair wise combinations of connected attributes with the term "amp" shows a buying decisions (refer to Table 4 in the text) except for the combined appearance of "amp+ head" where the predicted class hold a sell decision. Looking at the decision class "call", it is connected to another decision node "expect" where their combined appearance signifies a sell decisions despite the dominated buying decisions of the individual appearance of the term "call" in a tweet message. Another branch of tree might be explored by viewing the term " look" where its combined appearance with the other terms on the tree (i.e. close and intc) indicate a sell signal to the market participant in the stock market.

Reference

Abraham, A., Nath, B., & Mahanti, P. K. (2001). Hybrid intelligent systems for stock market analysis. *Computational science-ICCS 2001* (pp. 337–345). Berlin, Heidelberg: Springer.
Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance, 59*(3), 1259–1294.
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.
Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance, 61*(4), 1645–1680.
Blume, L., Easley, D., & O'hara, M. (1994). Market statistics and technical analysis: The role of volume. *The Journal of Finance, 49*(1), 153–181.
Booth, A., Gerding, E., & McGroarty, F. (2014). Automated trading with performance weighted random forests and seasonality. *Expert Systems with Applications, 41*(8), 3651–3661.
Brock, W., Lakonishok, J., & LeBaron, B. (1992). Simple technical trading rules and the stochastic properties of stock returns. *Journal of Finance*, 1731–1764.
Brown, G. W., & Cliff, M. T. (2004). Investor sentiment and the near-term stock market. *Journal of Empirical Finance, 11*(1), 1–27.
Chang, P., & Liu, C. (2008). A TSK type fuzzy rule based system for stock price prediction. *Expert Systems with Applications, 34*(1), 135–144.
Cheng, C., Wei, L., & Chen, Y. (2009). Fusion ANFIS models based on multi-stock volatility causality for TAIEX forecasting. *Neurocomputing, 72*(16), 3462–3468.
Chi, S., Chen, H., & Cheng, C. (1999). A forecasting approach for stock index future using grey theory and neural networks. In *Proceedings of the* 1999. IJCNN'99. IEEE International Joint Conference on Neural Networks (pp. 3850–3855).
Chiang, W., Urban, T. L., & Baldridge, G. W. (1996). A neural network approach to mutual fund net asset value forecasting. *Omega, 24*(2), 205–215.
Chun, S., & Park, Y. (2005). Dynamic adaptive ensemble case-based reasoning: application to stock market prediction. *Expert Systems with Applications, 28*(3), 435–443.
Clarkson, P. M., Joyce, D., & Tutticci, I. (2006). Market reaction to takeover rumour in Internet Discussion Sites. *Accounting & Finance, 46*(1), 31–52.
Dai, W., Shao, Y. E., & Lu, C. (2013). Incorporating feature selection method into support vector regression for stock index forecasting. *Neural Computing and Applications, 23*(6), 1551–1561.
Darrat, A. F., & Zhong, M. (2000). On testing the random-walk hypothesis: A model-comparison approach. *Financial Review, 35*(3), 105–124.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1991). The survival of noise traders in financial markets. *Journal of Business*, 1–19.
De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of political Economy*, 703–738.
Duan, J., Wang, W., Zeng, J., Zhang, D., & Shi, B. (2009). A prediction algorithm for time series based on adaptive model selection. *Expert Systems with Applications, 36*(2), 1308–1314.
Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work*. *The Journal of Finance, 25*(2), 383–417.
Gong, L., Zeng, J., & Zhang, S. (2011). Text stream clustering algorithm based on adaptive feature selection. *Expert Systems with Applications, 38*(3), 1393–1399.
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.
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research, 3*, 1157–1182.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10–18.
Hu, Y., Feng, B., Zhang, X., Ngai, E., & Liu, M. (2015). Stock trading rule discovery with an evolutionary trend following model. *Expert Systems with Applications, 42*(1), 212–222.
Huang, C., Yang, D., & Chuang, Y. (2008). Application of wrapper approach and composite classifier to the stock trend prediction. *Expert Systems with Applications, 34*(4), 2870–2878.
John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant Features and the Subset Selection Problem. ICML., 121–129.
Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications, 38*(5), 5311–5319.
Kaucic, M. (2010). Investment using evolutionary learning methods and technical rules. *European Journal of Operational Research, 207*(3), 1717–1727.
Kee, W. Y., & Koh, A. (1994). Technical analysis of Nikkei 225 stock index futures using an expert system advisor. In *Proceedings of the CBOT conference*.
Kodogiannis, V., & Lolis, A. (2002). Forecasting financial time series using neural network and fuzzy system-based techniques. *Neural computing & applications, 11*(2), 90–102.
Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1), 273–324.
Kothari, S., Li, X., & Short, J. E. (2009). The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review, 84*(5), 1639–1670.
Kramer, A. D. (2010). An unobtrusive behavioral model of gross national happiness. In *Proceedings of the 28th international conference on Human factors in computing systems* (pp. 287–290). ACM.
Lai, R. K., Fan, C., Huang, W., & Chang, P. (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications, 36*(2), 3761–3773.
Lakonishok, J., & Smidt, S. (1988). Are seasonal anomalies real? A ninety-year perspective. *Review of Financial Studies, 1*(4), 403–425.
Lee, M. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications, 36*(8), 10896–10904.
Lee, W. Y., Jiang, C. X., & Indro, D. C. (2002). Stock market volatility, excess returns, and the role of investor sentiment. *Journal of Banking & Finance, 26*(12), 2277–2299.
Li, S., & Kuo, S. (2008). Knowledge discovery in financial investment for forecasting and trading strategy through wavelet-based SOM networks. *Expert Systems with Applications, 34*(2), 935–951.
Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35–65.
Markowitz, H. M., Todd, G. P., & Sharpe, W. F. (2000). *Mean-variance analysis in portfolio choice and capital markets*. John Wiley & Sons.
Masoud, N. (2014). Predicting direction of stock prices index movement using artificial neural networks: The case of Libyan financial market. *British Journal of Economics, Management & Trade, 4*(4), 597–619.
Nuij, W., Milea, V., Hogenboom, F., Frasincar, F., & Kaymak, U. (2014). An automated framework for incorporating news into stock trading strategies. *IEEE Transactions on Knowledge and Data Engineering, 26*(4), 823–835.
Oh, C. and Sheng, O. (2011) "Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement".
Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing: 10* (pp. 79–86). Association for Computational Linguistics.
Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific reports, 3*.
Quinlan, J. R. (1993). *C4. 5: programs for machine learning*. Morgan kaufmann.
Rokach, L., & Maimon, O. (2005). Decision trees. *Data mining and knowledge discovery handbook* (pp. 165–192). Springer.
Rosillo, R., Giner, J., & de la Fuente, D. (2014). The effectiveness of the combined use of VIX and support vector machines on the prediction of S&P 500. *Neural Computing and Applications, 25*(2), 321–332.
Rudd, A. T. (1979). The revised dow jones industrial average: New wine in old bottles? *Financial Analysts Journal, 35*(6), 57–63.
Salzberg, S. L. (1994). C4. 5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning, 16*(3), 235–240.
Skabar, A. (2005). Application of bayesian techniques for MLPs to financial time series forecasting. *AI 2005: Advances in artificial intelligence* (pp. 888–891). Springer.
Sun, Y., Liang, Y., Zhang, W., Lee, H., Lin, W., & Cao, L. (2005). Optimal partition algorithm of the RBF neural network and its application to financial time series forecasting. *Neural Computing & Applications, 14*(1), 36–44.
Tan, T. Z., Quek, C., & Ng, G. S. (2007). Biological brain-inspired genetic complementary learning for stock market and bank failure prediction. *Computational Intelligence, 23*(2), 236–261.
Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139–1168.
Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance, 63*(3), 1437–1467.
Tsai, C., & Hsiao, Y. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems, 50*(1), 258–269.
Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417–424). Association for Computational Linguistics.
Vanstone, B., & Hahn, T. (2010). *Designing stock market trading systems: with and without soft computing*. Harriman House Limited.
Xue-shen, S., Zhong-ying, Q., Da-ren, Y., Qing-hua, H., & Hui, Z. (2007). A novel feature selection approach using classification complexity for SVM of stock market trend prediction. In *International Conference on Management Science and Engineering, 2007. ICMSE 2007.* (pp. 1654–1659). IEEE.
Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research, 5*, 1205–1224.
Żbikowski, K. (2015). Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy. *Expert Systems with Applications, 42*(4), 1797–1805.
Zhang, G. P. (2007). Avoiding pitfalls in neural network research. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 37*(1), 3–16.
Zhang, W., & Skiena, S. (2010). Trading strategies to exploit blog and news sentiment. In *Fourth International Conference on Weblogs and Social Media* (p. 186).
Zhang, Y., & Swanson, P. E. (2010). Are day traders bias free?—Evidence from internet stock message boards. *Journal of Economics and Finance, 34*(1), 96–112.