Ensemble Application of Transfer Learning and Sample Weighting for Stock Market Prediction

Simone Merello, Andrea Picasso Ratto, Luca Oneto
DIBRIS, University of Genova
Via Opera Pia 11A, I-16145 Genova, Italy
{simone.merello, andrea.picasso}@smartlab.ws, luca.oneto@unige.it

Erik Cambria
SCSE, Nanyang Technological University
50 Nanyang Ave, Singapore
cambria@ntu.edu.sg

Abstract—**Forecasting stock market behavior is an interesting and challenging problem. Regression of prices and classification of daily returns have been widely studied with the main goal of supplying forecasts useful in real trading scenarios. Unfortunately, the outcomes are not directly related with the maximization of the financial gain. Firstly, the optimal strategy requires to invest on the most performing asset every period and trading accordingly is not trivial given the predictions. Secondly, price fluctuations of different magnitude are often treated as equals even if during market trading losses or gains of different intensities are derived. In this paper, the problem of stock market forecasting is formulated as regression of market returns. This approach is able to estimate the amount of price change and thus the most performing assets. Price fluctuations of different magnitude are treated differently through the application of different weights on samples and the scarcity of data is addressed using transfer learning. Results on a real simulation of trading show how, given a finite amount of capital, the predictions can be used to invest in high performing stocks and, hence, achieve higher profits with less trades.**

Index Terms—**Financial forecasting, Stock market prediction**

I. Introduction

The potential revenue and the possible impact on the society of accurate stock market prediction has attracted investors and researchers since long time [1]. Nevertheless, its properties of time dependence, high stochasticity and chaotic behavior lead to a challenging problem. The Efficient market hypothesis [2] states that stocks are always traded at their fair value but behavioral economics tell us that emotions can profoundly affect individual behavior in financial decision making [3], [4]. Effects of emotions have been taken into account through technical analysis, exploiting the existence of patterns or motifs that would repeat in the future due to the collective attitude of investors [5]. Recent works focus on the extraction of sentiment related to the market from several sources of textual information, e.g., tweets [6], [7], microblogs [8], [9] and news articles [10], [11].

In sentiment analysis research [12], sentences are decomposed into concepts and targets of the opinion expressed [13]. Vectorial representations of words are the starting point of many machine learning applications. As a consequence, several algorithms have been designed to compute word vectors. GloVe [14] and Word2Vec [15] focus on capturing the general meaning and the relations between words while AffectiveSpace [16] poses particular attention on concepts and opinions.

Unfortunately, in the world of finance, writings can be different from usual text [17] thus, specialized tools have been developed starting from specific dictionaries [17], [18] to word embeddings computed on economic writings, whose pretrained version is publicly available [19].

In the literature of stock market prediction, different approaches have been used to supply market participants with useful trading signals. Some work focus on the regression of the stock's future price [20], [21]. Other proposals focus on the optimization of a monetary gain through the training of machine learning [22] models or through the construction of a policy able to take investment decisions on the market [23].

Most of the recent works propose to map the trading decision of each asset in a binary classification task (either "buy" or "sell"). Some papers introduce specific new models for the classification [24], [25], while others focus on natural language processing [26], [27] or on the class balancing problem [28]. Ternary classification has been considered as well by adding a third class representing the financial decision "hold the current position" [10].

According to the Capital growth Theory [29], an optimal strategy for the optimization of financial profits is to always invest the whole capital in the most performing stock of the next period. As a consequence, the outputs of a good predictor should provide an estimation of the most performing stocks in a set so that assets related to lower returns can be disregarded during trading. Unfortunately, current approaches are not directly related to the optimal strategy. Predicting the future prices of a stock does not seem to be helpful since only the change (increase or decrease) in the values over time is related to the profits.

Classification approaches instead supply signals regarding a single stock independently from the other assets. The outcomes of the predictions cannot be used as an estimation of the stock's performance but two assets can be correctly predicted as "buy" even if the strength of their fluctuation differs significantly and investing the whole capital on the most performing would have generated higher returns.

In this paper, the stock price prediction problem is formulated as regression of market returns. The regression-based approach is able to estimate not only the direction but also the amount of the price change of each stock and thus the most performing assets.

During market trading, investments on fluctuations of different intensities have different impacts on the portfolio value. To address the issue, this paper discusses the application of different weights for each data point. The results are compared with the formulation of the task as a binary classification between 'up' and 'down' trends. In this setting, the problem of unbalanced classes is addressed with appropriate techniques during training and evaluation to avoid biased predictions.

With the purpose of generalizing the results, the two approaches are tested on several models which differ in input space and learning algorithm. Two different representations of textual information are used: the first is based on a simple tool specific for finance, the second is based on more complex but general purpose pretrained embeddings. Several widely adopted algorithms are taken in account: Kernel support vector machine (KSVM), kernel support vector regression (KSVR) and feed forward neural network (FFNN). As a deep learning approach, the latter requires huge amounts of data, especially in case of complex relations between input and output and significant noise. This paper shows how transfer learning technique can be effectively applied on stock market prediction so that a single model is trained on a bigger dataset.

The final goal of this work is to propose an approach whose predictions can be useful in a real scenario of trading where answering to the financial decision of which stocks to trade in the correct moment is crucial. For this reason, useful properties are observed using data science metrics but the final evaluation is based on financial measures.

The rest of the paper is organized as follows: Section II formalizes the problem; Section III defines the approach; Section IV provides an overview of the collected data; Section V explains in detail experiments and results; finally, Section VI points out conclusions and future work.

II. Problem Formalization

The problem is time dependent. The predictions $\hat{y}_t$ are made at fixed and discretized time steps t relying on text published on dates identified as $\bar{t}$. The labels y_t are defined according to the cumulative returns cr_t achieved by the market during trends of length w

$$cr_t = \frac{p_{t+w} - p_t}{p_t}$$

where $p_t \in \mathbb{R}^+$ is the open price at time step t and $cr_t \in [-0.2, 0.2]$ in our experiments. The task of stock market prediction regards forecasting in t the change in the future stock prices during $[t{+}1, t{+}1{+}w]$. Therefore for the regression problem, y_t can be defined as the value of the price change

$$y_t = cr_{t+1}$$

while for the classification task the aim of the predictions is only on its direction ('up' or 'down')

$$y_t = \mathbb{1}(cr_{t+1})$$

where $\mathbb{1} : \mathbb{R} \to \{0, 1\}$ represents the unit step function.

The information available at time t relative to technical indicators and news articles is considered as leading the trends. The single news article published in $\bar{t}$ is encoded in a feature vector defined as $\overline{n_{\bar{t}}} \in \mathbb{R}^d$. For each interval t, $I_t \in \mathbb{R}^f$ represents the value of technical indicators computed over the recent past data and $n_t \in \mathbb{R}^{d+1}$ represents the aggregation of news published in the previous interval $\bar{t} \in [t-1, t)$

$$n_t = [\quad \frac{\sum_{\bar{t} \in [t-1,t)} \overline{n_{\bar{t}}}}{m_t}, \quad m_t \quad]$$

where m_t is the number of news in the considered interval and $[\cdot, \cdot]$ represents the concatenation operation.

The potential relation of future price fluctuations with news articles is not limited only at the previous interval but time spans of different lengths ending in t are considered through $n_{t,\hat{w}}$.

$$n_{t,\hat{w}} = \frac{\sum_{i=0}^{\hat{w}-1} n_{t-i}}{\hat{w}}$$

Thus, the input of the models can be defined as:

$$x_t = [N_t, I_t]$$

where $N_t = [n_t, n_{t,5}, n_{t,10}, n_{t,15}, n_{t,20}, n_{t,30}, n_{t,50}]$ in the experiments.

According to the problem definition, the collected data is considered as a time series of samples not independent $(\exists (i,j) : p(x_j, y_j | x_i, y_i) \neq p(x_j, y_j))$. In particular, y_t implies the existence of a temporal dependency such that (x_t, y_t) is deterministically correlated with (x_{t-v}, y_{t-v}), $v \in [0, w]$. Particular care is taken during the experiments to avoid biased results due to the dependency of samples.

III. Proposed Approach

The main target of this work is to propose a regression approach for the problem of returns forecasting whose predictions can be used to invest in high performing assets during market trading. The labels y_t indicate the change in the future prices of a given stock and each input x_t summarizes the information derived by recent news and technical indicators available at time t. KSVR and FFNN algorithms are applied to the regression problem. Support vector regression search the optimal solution through the minimization of the cost $\mathcal{L}_\epsilon$.

$$\mathcal{L}_\epsilon = C \sum_t \ell_t^\epsilon + \frac{1}{2} ||w_m||^2$$

where C is a regularization parameter that controls the trade off between the dimension of the model's weights w_m and $\ell_t^\epsilon = |y_t - f(x_t)|_\epsilon$. The latter measures the differences bigger than ϵ between the truth y_t and the output of the model $f(x_t)$. In this paper, $f(x_t)$ is computed through the use of the Gaussian kernel. Nonetheless, FFNN models for regression are often trained using the back propagation of Mean squared error (MSE) cost $\mathcal{L}_m$.

$$\mathcal{L}_m = \frac{1}{n} \sum_t \ell_t^m, \quad \ell_t^m = (y_t - f(x_t))^2$$

where n is the number of samples.

In the formulation of $\mathcal{L}_\epsilon$ and $\mathcal{L}_m$ all samples contribute equally to the cost value but this assumption is not adequate for stock market prediction. In market trading, the loss or gain derived by financial decisions is not always equal but depends on the value of the price fluctuations. With small fluctuations such as 0.01%, even if the investment leads to a decrease in the value of the portfolio, it would be marginal. But, trading on high fluctuations is more risky because the loss or the gain derived can be considerable. In the proposed regression experiments, samples are weighted according to the price changes cr_t so that data points related to bigger returns contribute more to the cost value, thus are considered more important by the models. In particular, the weighted costs are defined as:

$$\ell_t^{w\epsilon} = cr_t^2 \cdot \ell_t^\epsilon, \qquad \ell_t^{wm} = cr_t^2 \cdot \ell_t^m$$

and are used respectively instead of ℓ_t^ϵ, ℓ_t^m for the optimization process. Moreover, ℓ_t^{wm} is used during model selection with the purpose of choosing the best values for the hyperparameters of the regressors.

The results are compared with a classification approach pursued through similar models. KSVM is used to select the optimal maximum-margin hyperplane that separates the classes while FFNN algorithm is optimized through the back propagation of the binary cross entropy cost. In this setting, balancing the classes is fundamental to avoid biased results. In accordance with [28], SMOTE algorithm [30] is applied on the training and validation set since weights or hyperparameters optimized on screwed classes often lead to predictions biased towards the most frequent class. For the same reason, the model selection of classifiers is held with Matthew correlation coefficient (MCC).

$$MCC = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}}$$

where tp: True positive, fp: False positive, tn: True negative, fn: False negative extracted from the Confusion matrix.

The architecture of the FFNN is depicted in Fig. 1. Four dense layers are applied whose parameters were trained with Adam optimizer [31] algorithm. In each layer, batch normalization [32] is inserted to stabilize the distribution of the internal representation of samples and speed up the training phase.

Batch normalization guarantees to output samples distributed with mean zero and unit variance so that the parameters of the next layer are not required to adapt to the changing distribution of the input during training. In the first three layers, the normalized samples are directly fed into a Leaky ReLu activation function. Dropout [33] is applied on the output of the first and the second layer together with the use of Max-norm regularization over all layers to avoid co-adaptation of neurons and improving the generalization power of the network. The activation function of the last layer is the only change in the structure between classification and regression task. The Sigmoid function $\sigma(\cdot) \in [0, 1]$ is selected to perform the classification.

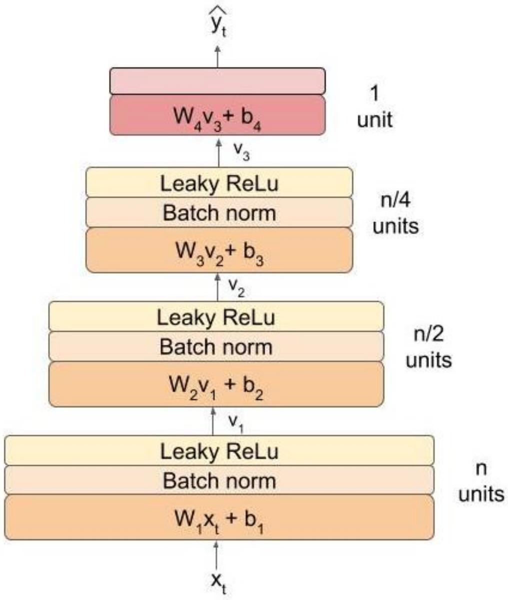

Fig. 1. FFNN architecture used for the experiments. The last layer is the only one affected by transfer learning and his activation function is the only change in the structure between classification and regression task.

Instead, the regression task is pursued through the Hyperbolic activation $tanh(\cdot) \in [-1, 1]$. The latter is considered reasonable since it behaves linearly in the domain of the regression labels $y_t = cr_{t+1} \in [-0.2, 0, 2]$. Furthermore, $tanh(\cdot)$ is used to infer the prior knowledge that predictions above $|\hat{y}_t| > 1$ must be avoided since these are not likely and the trading decision derived would be the same.

The last layer is the only one affected by transfer learning. Firstly, a single FFNN is trained over different stocks to capture the general relations between news, technical indicators and price fluctuations. Training a model over multiple stocks implies the assumption that the samples behave similarly and are drawn from the same distribution. This assumption is considered feasible since the features x_t are made up of the same technical indicators and news published by same sources. Moreover, y_t, defined in accordance with the returns cr_t, can be considered related to the same market portfolio returns in accordance to Capital asset pricing model theorem [34]. Secondly, the model is fine tuned on the specific stocks separately since some alteration between news of different companies are expected, e.g., “AAPL” news will probably more related to words such as “Apple”, “iPhone” and “Tim Cook” with respect to “FB”.

Our experiments propose two different settings from the point of view of the input. Firstly, a representation of each news article $\overline{n_t}$ is obtained through the use of Loughran/McDonald dictionary as it represents a simple source specific for finance. Secondly, the concepts present in news articles are extracted and used with AffectiveSpace to obtain a representation of the concept-related sentiment contained inside the financial writings.

The results are evaluated in three steps on a set of different stocks. Firstly, the length w of the trend cr_t is chosen accordingly to the best average performance of the models. Secondly, The behavior of the two approaches in predicting fluctuations of increasing strength is discussed.

A threshold is applied to consider Accuracy and MCC metrics over $[100\%, 80\%, 60\%, 40\%, 20\%]$ of returns according to their intensity. The evaluation over different subsets is necessary since a model that has low performance on small fluctuations but good results on higher returns can lead to considerable profits. Finally, the performance of the predictions is tested with a trading simulation performed with a specific tool[1].

An investment strategy is considered through the application of thresholds on the value of the predictions. Stocks are bought or sold in t only if the associated value is above the threshold. Since at time step t all the previous predictions are available, the trading actions in t are based not only on the predictions in t but also on the previous k values $[\hat{y}_t, \hat{y}_{t-1}, .., \hat{y}_{t-k}]$ through a simple average. Considering multiple predictions represents a careful behavior which avoids financial decisions based only on one signal. k is empirically chosen as 2 in the experiments. Sharpe ratio [35] and Annualized gain are examined as results of the trading performance.

IV. Available Data

In this section, the two kinds of data used for the experiments are described: stock data and textual data. The problem of stock market forecasting was mapped to the prediction of the price movements of the top 10 stocks in capital size of NASDAQ, respectively 'AAPL', 'AMZN', 'GOOGL', 'MSFT', 'FB', 'INTC', 'CSCO', 'CMCSA', 'NVDA' and 'NFLX'. Previous work had already focused on similar experiments [24]. Some of the stocks traded there are well known for their popularity and consequently a huge amount of textual information regarding them is constantly published.

Textual data were extracted from aggregated news services that gather the information from various professional periodicals[2,3]. The stock prices were collected from public sources[4] and only the periods in which the market was open were considered as time steps t. The time span between two subsequent samples was selected as hourly. One hour represented a lower bound from the point of view of the available news published for each time step and allowed the creation of a significant amount of samples for the experiments.

The information regarding prices and published news of all the selected stocks was available during the time span considered. Approximately eleven months of data were used for training, from 2017-04-03 to 2018-02-23 and four months were used for testing, from 2018-02-24 to 2018-06-21.

TABLE I
Technical Analysis indicators

Name	Formula
Momentum	$\frac{p_t}{p_{t-s}} - 1$
SMA	$\frac{p_t}{\overline{p_{t,s}}} - 1$
Bollinger Bands	$\begin{cases} 1 & \text{if } p_t > u_t^{bb} \\ 0 & \text{if } p_t \in [u_t^{bb}, d_t^{bb}] \\ -1 & \text{if } p_t < d_t^{bb} \end{cases}$
Differentiated v_t	$v_t - \overline{v_t}$

Different features were extracted from the raw data. For what concerns technical analysis, the extraction of meaningful values from past prices p_t and volumes v_t is complicated by the non-stationarity of these. Therefore, several differentiated metrics computed on the recent past were used as measures of the changes in the past financial values. $I_t \in \mathbb{R}^f$ $(f = 10)$ was made up of Momentum indicator, Simple moving average (SMA) based on the average value of the previous s prices $\overline{p_{t,s}}$, Bollinger Bands $[u_t^{bb}, d_t^{bb}]$ crossing signal and a differentiated measure of v_t based on the average of the previous values $\overline{v_t}$. In the experiments, s was chosen as $s \in \{30, 50, 100, 150\}$.

Different representations for the news $\overline{n_t}$ were considered. Firstly, features were extracted using the Loughran/McDonald dictionary [17]. It was constructed on annual reports of US enterprises considering also companies quoted in NASDAQ and since the focus of this work is on the same market, it was considered highly related to our task.

The dictionary includes positive, negative, litigious, interesting, uncertainty and other categories of words specific for finance and recently updated. Accordingly, $\overline{n_t} \in \mathbb{R}^d$ was made up of numerical counts of the words spotted in a news belonging to the different categories of the dictionary $(d = 7)$. Secondly, $\overline{n_t}$ was defined according to the embeddings of AffectiveSpace to obtain a representation of the concept-oriented sentiment contained in financial writings $(d = 100)$. The news articles were parsed to retrieve the contained concepts and the AffectiveSpace embeddings of different concepts found in a news were averaged to obtain its representation.

V. Experiment

A. Experiment Setup

Model selection was performed using cross validation on time dependent data. As depicted in Fig. 2, the folds were selected so that the last fold took into account all the training points, but the validation points were left out to ensure the temporal independence between Train and Test.

In our experiments, points used for evaluation were always chosen ahead the training set in time so that a possible look ahead bias was avoided since the information contained in the past samples cannot depend on the future.

[1] http://backtrader.com

[2] http://finance.yahoo.com

[3] http://nasdaq.com/news

[4] http://finance.google.com/finance

For the FFNN models, the number of epochs of training was considered as an hyperparameter optimized on the validation set during model selection. The computation of the optimal value involved only the last fold since the dimension of the others was different from the training set used for testing.

In a first test Accuracy and MCC achieved by the models were evaluated on average above all the considered stocks. Averaging on different stocks from the finance point of view means considering the overall performance of the predictions over a portfolio of different financial assets as if earnings or losses were obtained by all of them betting on their performance. Furthermore, considering different stocks appear fundamental to generalize the result and to avoid specific conclusions for the single stock.

A second test was performed to evaluate the predictions in a real trading simulation. The output of the trading simulation depended on the selection of which stocks to trade (trading strategy) but also from the the amount of capital to invest in each trade (sizing strategy). The trading strategy was defined according to the trading signal s_t computed as average of the last three predictions $s_t = \frac{1}{3}\sum_{i=0}^{2}\hat{y}_{t-i}$. A threshold T was used to trade only on the most performing stocks for which $|s_t| > T$. Every trade started in t lasted until time step $t+w$ accordingly to the prediction target. The sizing strategy was selected so that given the portfolio value P_t each time step t an amount $\frac{P_t}{w \cdot n_{a,t}}$ was invested on the single asset. $n_{a,t}$ denotes the number of actions (buy or sell) suggested by the model in t. Thus, if in t many stocks are traded, only a small amount of capital can be invested in each asset.

A commission rate of 0.01% was considered and the risk free rate required by the computation of the Sharpe ratio was set as the interest rate of three-month U.S. Treasury bill in the period of evaluation.

B. Optimal trend classification

The definition of the labels y_t depend on the length w of the trend cr_t considered thus, our first achievement was needed to fix an optimal value for it. Several experiments were done to predict in t trends starting at $t+1$ and ending $[1, 7, 28, 35, 49]$ hours later.

That means, predicting the trend of the next hour, the next trading day but also the next four, five and seven trading days. The optimal window was selected according to the average MCC score achieved by all models described in Section III.

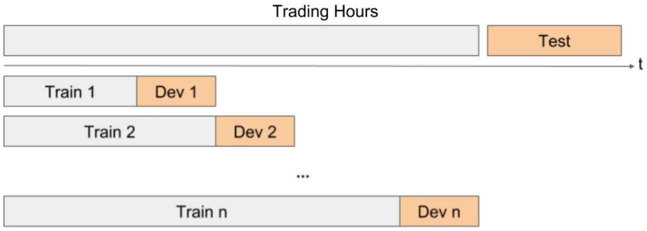

Fig. 2. Dataset division.

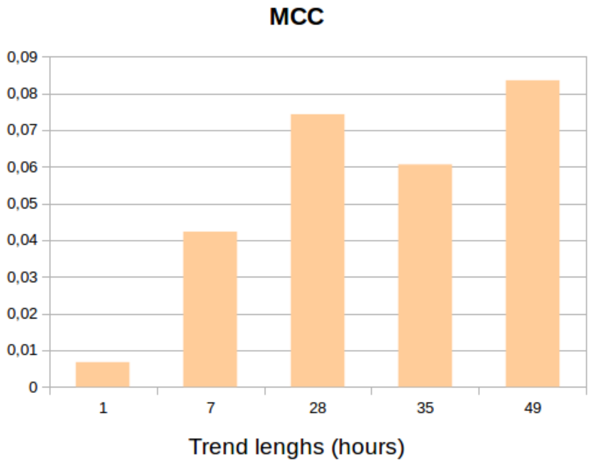

Fig. 3. MCC values averaged over the tested models. Different experiments regard the prediction of different trends cr_t. Trends lasting 1 hour, 1, 4, 5 and 7 days are evaluated.

According to Fig. 3, the MCC value relative to the trend length of seven trading days (49 hours) achieved the best score. This result is considered significant since most of the state-of-the-art papers on return classification focus on daily trend prediction [10], [24] without a proper explanation. The results of this experiment highlight that daily prediction is sub-optimal with respect to other choices of the trend length.

According to the average performance of the tested models, predicting trends lasting seven or four days allowed approximately to double the MCC score (0.084 and 0.074, respectively) in comparison to the daily prediction (MCC 0.042). For the rest of the experiments, the trend length at seven open market days was used for the evaluations.

C. Comparison Regression-Classification

The comparison between regression and classification approach was divided in two steps. During the first step, the comparison was held on MCC and Accuracy scores computed on different subsets of fluctuations. A threshold was used to select only the highest returns, respectively different evaluations took in account the 100% 80%, 60%, 40%, and 20% of the highest returns. To highlight the tendency of the evaluation scores, Fig. 4 shows the evaluation metrics computed considering the value relative to the 100% of returns set as zero and subtracted from the other results. Fig. 5 shows the MCC and Accuracy metrics.

During the second step, the predictions of the models were used in a trading simulation and were evaluated using financial measures. The Annualized gain and the Sharpe ratio were computed considering to perform buy or sell actions in t only if the trading signal s_t was bigger than a threshold $|s_t| > T$. T represents the threshold set so that in different experiments the 100% 80%, 60%, 40%, and 20% of the highest predictions were taken in account for trading.

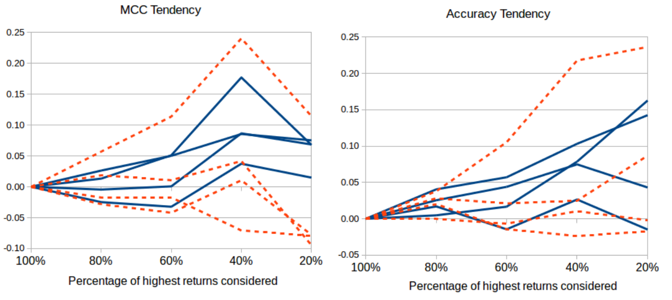

Fig. 4. MCC and Accuracy values of the models are evaluated on different subsets of the returns. Dashed lines represent classification models while continuous lines are relative to regression models.

Fig. 6 shows the evaluation metrics computed considering the value relative to the 100% of predictions set as zero and subtracted from the other results. Fig. 7 shows the Annualized gain and the Sharpe ratio.

For what concerns MCC and Accuracy, Fig. 4 shows how the performance of the models trained with the regression approach increased considering higher fluctuations. Nonetheless, this property seems to hold with only one model trained with the classification approach. According to Fig. 5, the MCC and Accuracy values computed on the whole returns (100%) were often higher for the classifiers rather than the values of the regressors. Augmenting the threshold and thus considering only higher fluctuations the performance of models trained with the regression approach increased and frequently overcome the classification scores. It is opinion of the authors that the application of weights to samples proportional to cr_t was fundamental to feed the model the information that higher fluctuations were more important and thus, to achieve a growth on the performance relative to higher fluctuations.

During the trading simulation, the differences between regression and classification approach were even more stressed. Fig. 6 shows how the performance of regressors increased trading only on higher predictions. Nonetheless, the performance of the classification approach increased less or not at all.

During a generic time step t, the optimal investment and sizing strategy would bet the whole capital on the stock corresponding to the highest future return. The regression models were able to supply an estimation of the strength of the fluctuation thus, augmenting the threshold over the predictions allowed to select more carefully on which stocks to trade. Since during the trading simulation the amount of capital was limited, trading on less stocks implied to invest more on the single stock that the regressor models estimated as the most valuable. With the classification approach instead, this behavior could not be exploited since, higher values of the predictions were not necessarily related with higher fluctuations thus, during trading classification models did not supply enough information to choose the most worth stocks in which to invest. As depicted in Fig. 7, the highest Sharpe ratio and the three highest annualized gains achieved were related to the regression approach.

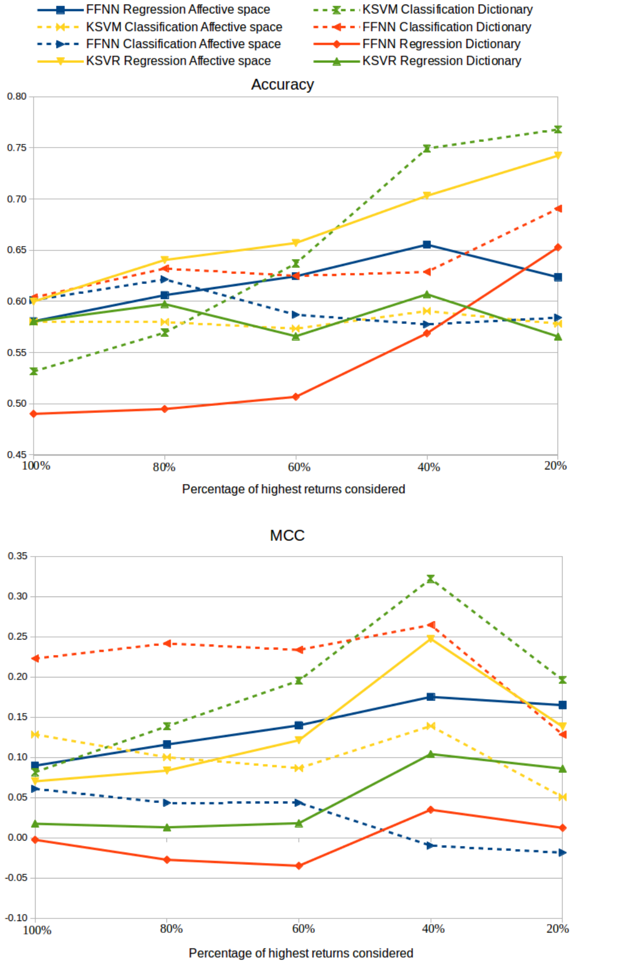

Fig. 5. MCC and Accuracy values of the benchmarked models evaluated on different subsets of the returns.

In our experiments, the model able to achieve best performance from both the points of view of Sharpe ratio and Annualized gain was the FFNN that exploited the features extracted though AffectiveSpace, the regression approach and that was trained through transfer learning. The application of this technique is discussed in the next Section.

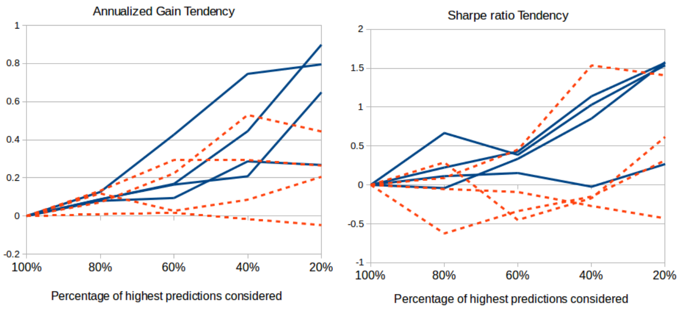

Fig. 6. The Annualized gain and the Sharpe ratio of the trading simulations considering different subsets of the predictions. Dashed lines represent classification models while continuous lines are relative to regression models.

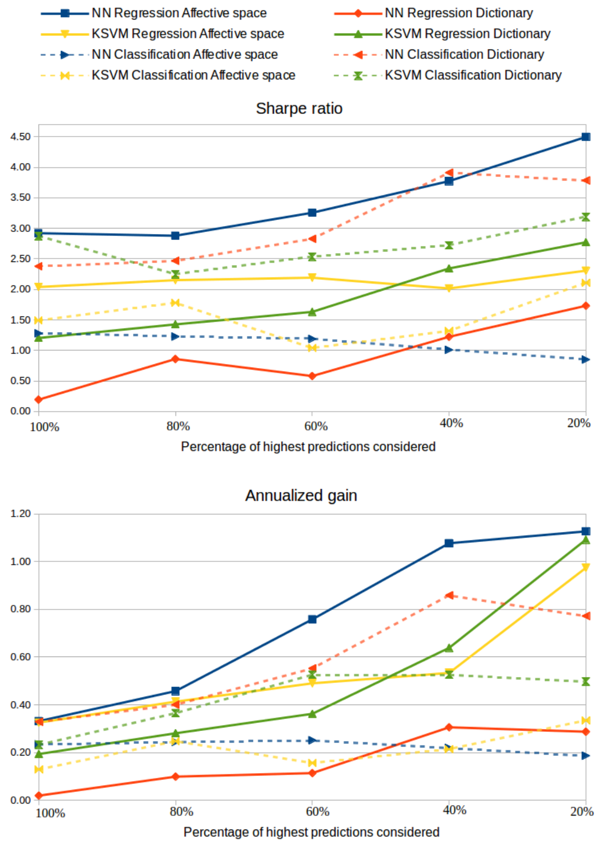

Fig. 7. Annualized gain and Sharpe ratio values of the benchmarked models evaluated on different subsets of the returns.

D. Transfer learning evaluation

Transfer learning can be considered effective if during the pretraining phase the model is able to reach a good minimum which result a useful starting point for further optimization process during the fine-tuning. Table II shows for each FFNN and for each stock the number of epochs of fine-tuning optimized on the validation set in the range $[0, 200]$.

TABLE II
NUMBER OF EPOCHS OF FINE-TUNING FOR EACH STOCK OF THE FFNNS MODELS, RESPECTIVELY CLASSIFICATION AND REGRESSION APPROACHES WITH FEATURES BASED ON AFFECTIVESPACE (AS) AND LOUGHRAN/MCDONALD DICTIONARY (LM-DICT).

	classification		**regression**	
	AS	**LM-Dict**	**AS**	**LM-Dict**
AAPL	0	14	200	0
AMZN	0	200	79	200
GOOGL	0	44	19	0
MSFT	0	54	200	0
FB	149	0	200	0
INTC	26	0	32	0
CSCO	185	173	190	200
CSMA	29	0	0	47
NVDA	200	24	0	0
NFLX	193	55	0	200

For all the models that used transfer learning, the optimal number of epochs of training was estimated as 0 for several stocks. Thus, fine-tuning of these stocks was not helpful in increasing the performance on the validation set but the optimal minimum was reached with the model trained over all the stocks.

To quantify the contribution of the transfer learning on the problem of stock market prediction an additional experiment was based on the model that achieved the highest trading results. Two FFNN AffectiveSpace regressors trained with and without transfer learning were compared. Since in this experiment the financial decision was not the main interest and the evaluated models were based on regression, the comparison was based on the Normalized MSE. Fig. 8 shows how using transfer learning FFNN AffectiveSpace regressor was able to achieve better performance improving on all the considered subsets of the returns.

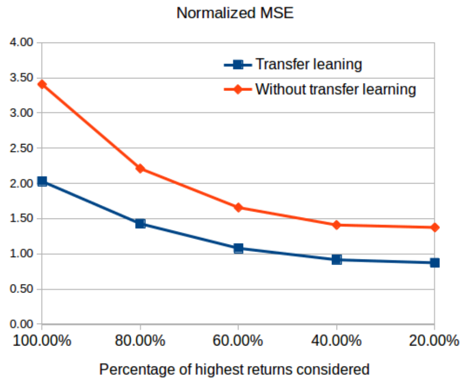

Fig. 8. Normalized mean square error of the FFNN AffectiveSpace regressor with and without transfer learning. Results are evaluated on different subsets of the returns.

VI. CONCLUSION AND DISCUSSION

In this paper, we have shown that the predictions of returns generated by a regression approach are more meaningful with respect to 'buy' or 'sell' signals provided by classification approaches during trading. This gap of information can be used to augment financial profits through an investment strategy able to focus only on the most performing assets.

According to our results, the application of transfer learning and sample weighting over different market fluctuations has been effective to enhance the performance, especially on the biggest and most important returns.

Our paper does not contemplate some aspects that will be undertaken in future research. Firstly, the application of sample weights should be studied more in depth starting from their application with the classification approach to a comparison of different formulations. Secondly, the benefits of the regression approach should be benchmarked on state-of-the-art methods to better quantify the improvements of the proposed technique.

References

[1] F. Xing, E. Cambria, and R. Welsch, "Natural language based financial forecasting: A survey," *Artificial Intelligence Review*, vol. 50, no. 1, pp. 49–73, 2018.

[2] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *The journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[3] G. Loewenstein, "Emotions in economic theory and economic behavior," *American economic review*, vol. 90, no. 2, pp. 426–432, 2000.

[4] F. Xing, E. Cambria, L. Malandri, and C. Vercellis, "Discovering bayesian market views for intelligent asset allocation," in *ECML*, 2018.

[5] C.-H. Park and S. H. Irwin, "The profitability of technical analysis: A review," *AgMAS Project Research Report 2004-04*, 2004.

[6] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[7] F. Xing, E. Cambria, and R. Welsch, "Intelligent asset allocation via market sentiment views," *IEEE Computational Intelligence Magazine*, vol. 13, no. 4, pp. 25–34, 2018.

[8] S. R. Das and M. Y. Chen, "Yahoo! for amazon: Sentiment extraction from small talk on the web," *Management science*, vol. 53, no. 9, pp. 1375–1388, 2007.

[9] L. Malandri, F. Xing, C. Orsenigo, C. Vercellis, and E. Cambria, "Public mood–driven asset allocation: the importance of financial sentiment in portfolio management," *Cognitive Computation*, 2019.

[10] Z. Hu, W. Liu, J. Bian, X. Liu, and T. Y. Liu, "Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction," in *ACM International Conference on Web Search and Data Mining*, 2018.

[11] F. Xing, E. Cambria, and R. Welsch, "Growing semantic vines for robust asset allocation," *Knowledge-Based Systems*, 2019.

[12] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intelligent Systems*, vol. 32, no. 6, pp. 74–80, 2017.

[13] S. Poria, E. Cambria, G. Winterstein, and G.-B. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis," *Knowledge-Based Systems*, vol. 69, pp. 45–63, 2014.

[14] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[16] E. Cambria, J. Fu, F. Bisio, and S. Poria, "AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis." in *AAAI*, 2015, pp. 508–514.

[17] T. Loughran and B. McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.

[18] E. Henry and A. J. Leone, "Measuring qualitative information in capital markets research," *The Accounting Review*, vol. 91, no. 1, pp. 153–178, 2009.

[24] Y. Xu and S. B. Cohen, "Stock movement prediction from tweets and historical prices," in *Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2018.

[19] P. Saleiro, E. M. Rodrigues, C. Soares, and E. Oliveira, "Feup at semeval-2017 task 5: Predicting sentiment polarity and intensity with financial word embeddings," *arXiv preprint arXiv:1704.05091*, 2017.

[20] P.-F. Pai and C.-S. Lin, "A hybrid arima and support vector machines model in stock price forecasting," *Omega*, vol. 33, no. 6, pp. 497–505, 2005.

[21] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, "Comparison of arima and artificial neural networks models for stock price prediction," *Journal of Applied Mathematics*, vol. 2014, 2014.

[22] Y. Bengio, "Training a neural network with a financial criterion rather than a prediction criterion," in *Decision Technologies for Financial Engineering: Proceedings of the Fourth International Conference on Neural Networks in the Capital Markets (NNCM'96), World Scientific Publishing*, 1997, pp. 36–48.

[23] J. W. Lee, "Stock price prediction using reinforcement learning," in *Industrial Electronics, 2001. Proceedings. ISIE 2001. IEEE International Symposium on*, vol. 1. IEEE, 2001, pp. 690–695.

[25] T. H. Nguyen and K. Shirai, "Topic modeling based sentiment analysis on social media for stock market prediction," in *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2015.

[26] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi, "Sentiment analysis of twitter data for predicting stock market movements," in *International Conference on Signal Processing, Communication, Power and Embedded System*, 2016.

[27] Y. Peng and H. Jiang, "Leverage financial news to predict stock price movements using word embeddings and deep neural networks," *arXiv preprint arXiv:1506.07220*, 2015.

[28] A. Picasso, S. Merello, Y. Ma, L. Oneto, and E. Cambria, "Ensemble of technical analysis and machine learning for market trend prediction," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018.

[29] J. L. Kelly Jr, "A new interpretation of information rate," in *The Kelly Capital Growth Investment Criterion: Theory and Practice*. World Scientific, 2011, pp. 25–34.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[34] R. C. Merton, "An intertemporal capital asset pricing model," *Econometrica: Journal of the Econometric Society*, pp. 867–887, 1973.

[35] W. F. Sharpe, "The sharpe ratio," *Journal of portfolio management*, vol. 21, no. 1, pp. 49–58, 1994.