Practical Machine Learning Approach to Capture the Scholar Data Driven Alpha in AI Industry

1st Yunzhe Fang
Industrial Engineering and Operations Research
Columbia University
New York City, USA
yunzhe.fang@columbia.edu

2nd Xiao-Yang Liu
Dept. Electrical Engineering
Columbia University
New York City, USA
xl2427@columbia.edu

3rd Hongyang Yang
Dept. Statistics
Columbia University
New York City, USA
hy2500@columbia.edu

***Abstract*—AI technologies are helping more and more companies leverage their resources to expand business, reach higher financial performance and become more valuable for investors. However, it is difficult to capture and predict the impacts of AI technologies on companies' stock prices through traditional financial factors. Moreover, common information sources such as company's earnings calls and news are not enough to quantify and predict the actual AI premium for a certain company. In this paper, we utilize scholar data as alternative data for trading strategy development and propose a practical machine learning approach to quantify the AI premium of a company and capture the scholar data driven alpha in the AI industry. First, we collect the scholar data from the Microsoft Academic Graph database, and conduct feature engineering based on AI publication and patent data, such as conference/journal publication counts, patent counts, fields of studies and paper citations. Second, we apply machine learning algorithms to weight and re-balance stocks using the scholar data and traditional financial factors every month, and construct portfolios using the "buy-and-hold-long only" strategy. Finally, we evaluate our factor and portfolio in terms of factor performance and portfolio's cumulative return. The proposed scholar data driven approach achieves a cumulative return of 1029.1% during our backtesting period, which significantly outperforms the Nasdaq 100 index's 529.5% and S&P 500's 222.6%. The traditional financial factors approach only leads to 776.7%, which indicates that our scholar data driven approach is better at capturing investment alpha in AI industry than traditional financial factors.**

***Index Terms*—AI technology, scholar data, alternative data, AI in finance, quantitative investment, alpha research.**

I. Introduction

Investors trend to chase for "hot" things in the stock market, for example, hot topics and hot companies. Artificial intelligence (AI) is a buzzword in recent years, and companies who start using AI technologies are becoming the brightest stars in the investment universe. For example, Netflix, a media-services provider who uses AI technologies in almost every process of its business such as movie recommendation and even location scouting for movie production, is growing wild in recent years. Nvidia, a GPU manufacturer that powers many AI technologies such as self-driving cars, also experiences explosive growth in recent years. As of September 2018, Netflix's stock price has been increased to 80 times as compared with that of 2009, and 35 times for Nvidia. In addition, the cumulative return of Nasdaq 100 (a stock index that is well known for its technology stocks) during our backtesting period from 2009 to 2018 reached 529.5%, which is much higher than the S&P 500's 222.6% [1]. AI premium is one of the reasons for Nasdaq 100's high returns.

However, it is hard for investors to analyze the impacts of AI technologies on companies who use AI technologies in their business. Because there are no such metrics in companies' financial reports that directly describes how AI technologies are improving companies' business. Most of the companies do not generate revenues directly from selling AI technologies, instead, companies use AI technologies to facilitate their business. In other words, we need to create metrics to quantify the impacts of AI technologies on companies' financial performance and corresponding stock performance using other datasets.

How to evaluate and measure companies' growth premium that is rewarded by using AI technologies? How can we utilize the characteristics of AI industry to better predict the impacts of AI technologies on companies? How to connect AI technologies with companies' stock performance?

One traditional approach to answer these questions is to look at companies' business strategies related to AI, either through companies' website or earnings call and analyze the relationship between companies' utilization of AI technologies and trends of the earnings factors. However, this analysis is discretionary and may not be well-grounded because it is difficult to quantify the impacts of AI technologies.

Another approach is more systematic and challenging. A lot of investors start using alternative data sources so that they are able to quantitatively answer abstract questions based on the data [2]. A lot of alternative datasets are big data [3] that enables us to use machine learning techniques to extract new investment alphas. Alternative data such as text and audio in earnings calls can be used to predict company's future performance [4], news and blogs [5], [6] can be used to evaluate investors' sentiment that has a strong impact on short-term stock prices, credit card transactions and GPS traffic are quite popular in directly predicting company's revenue. Alternative data usually contain information different from traditional data, and the correlation between alternative data driven factors and traditional factors is usually low, which indicates that alternative data could be helpful to answer the questions that traditional data cannot answer.

Although alternative data solutions seem promising, our purpose is neither to predict news sentiment nor company's total performance and revenue. What we want to do is to capture the scholar data driven alpha in AI industry. The key to utilize alternative data for our task is to find proper datasets that contains the right information we need. The selection of datasets should be the most essential and important step for anybody who wants to use alternative data for investment questions, because most alternative data might be noisy, non-uniform in terms of quality [7] and may be of low information density. The selection of highly related alternative data to answer specific investment questions is sometimes more important than the selection of machine learning models. Because only the data that contains related information can be used to capture effective investment alphas.

The reasons we choose the scholar data from the Microsoft Academic Graph database [2] to explore alpha are based on three main characteristics of the AI industry. First, AI community has far more academic conferences than other areas, which contributes to the rapid development of the AI industry. Second, the large number of conferences accelerate the spread of knowledge and information within the AI industry. Third, AI-related marketing patents become the fastest growing global category. We believe this dataset with the records of conference/journal publications and patents reflect companies' research and development capacity of AI technologies. We use it to quantify the impacts of AI on companies' stock performance. There are plenty of researches show that patent citations and academic publications would bring financial benefits to companies [8].

In this paper, we utilize scholar data as alternative data to develop trading strategies and propose a practical machine learning approach to quantify the AI premium of a company and capture the scholar data driven alpha in the AI industry. First, by feature engineering on AI publications, patents and citations, we extract 40 features, and the 10 most popular financial indicators. Second, we apply machine learning models to predict the monthly returns of 115 stocks in our investment universe, and construct the equally-weighted AI portfolio using the top 25% stocks with the highest predicted monthly returns and then we trade the portfolio using the "buy-and-hold-long-only" strategy [9] and re-balance the portfolio monthly. Third, we evaluate our factor and portfolio performance in a robust framework including predictive factor analysis, portfolio performance analysis and risk analysis. We compare the performance of our strategy with the S&P 500 index, the Nasdaq 100 index, and the traditional financial factors only approach. The proposed scholar data driven approach achieves a cumulative return of 1029.1% during our backtesting period, which significantly outperforms the Nasdaq 100 index's 529.5% and S&P 500's 222.6%. The traditional financial factors approach only leads to 776.7%, which indicates that our scholar data driven approach is better in capturing investment alpha in AI industry than traditional financial factors.

This paper proceeds as follows. Section II describes how we collect and process the scholar data, and the investment universe we use. Section III presents our practical machine learning approach that captures the scholar data driven alpha in the AI industry. Section IV presents the performance evaluation. Section V concludes the paper.

II. Scholar Data for AI Technologies

The investment question we are trying to answer is that can we find or create quantitative metrics to estimate the impacts of using AI technologies on companies' stock performance.

A. Utilizing Scholar Data as Alternative Data

Not every data contains useful information in a specific industry. For example, GPS traffic data is important to analyze retail industry because it can be used to predict the number of visitors to actual stores, while this data is less important to Netflix which is an online media firm. Every industry has its own characteristics which might give us some hint to find useful information.

As for the AI industry, there are three main characteristics. First, the AI community has far more academic conferences and journals than other areas. Scholars in both academia and industry exchange and share ideas in these conferences. We regard this as the main characteristics of the AI community that is different from other areas. Second, the large number of conferences make it much faster and easier for AI researchers to publish their research results and catch up with the cutting-edge technologies in the AI industry and utilize them. In another words, knowledge and information flows fast in AI industry [10]. Third, since 2000, more than 225,000 AI patents filed globally. AI-related marketing patents becomes the fastest growing global category, with an annual growth rate of 29.3% between 2010 and 2018 [11].

The large number of conferences and journals and high speed information flow give us much information about the trends of technologies as well as the persons and companies that invented these technologies in the AI industry, which is important for us to figure out the impacts of AI technologies on the companies. Also, the high growth rate of AI patents indicates that AI companies think AI technologies are valuable to their business and could generate revenue for them, so they filed patent applications to protect intellectual property rights. These all become the reasons why we decide to use scholar data, namely publication and patent data, to capture the investment alpha in the AI industry because these characteristics finally point to the scholar data.

Our scholar data comes from the Microsoft Academic Graph database [2], which is a public database containing records of conference and journal publications, patents as well as additional information such as organizations, citations, publishers, and fields of study. The Microsoft Academic Graph database now indexes more than 221 million publications, 241 million authors and 25 thousand institutions. The great academic coverage of this database builds our solid foundation to find the alpha in the AI industry.

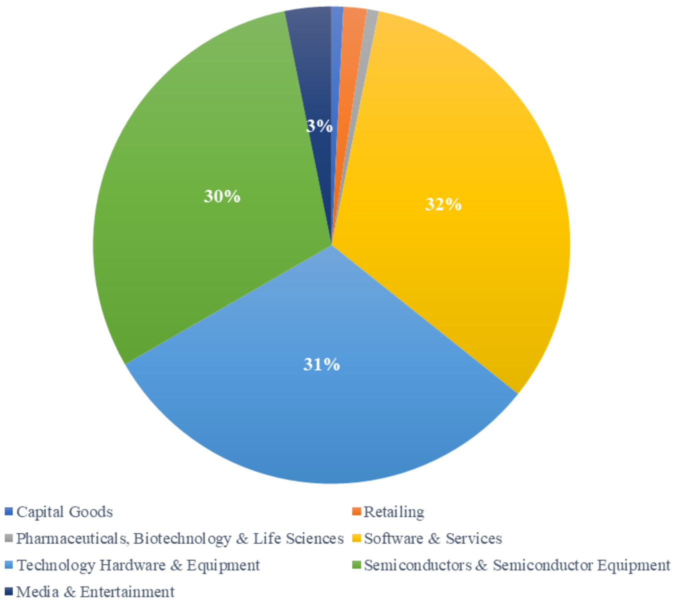

Fig. 1. Stocks in the investment universe by industry

B. Investment Universe

Before we start data collection, we select companies. Our investment universe contains 115 publicly traded companies from the New York Stock Exchange (NYSE) and Nasdaq with a broad coverage of industry companies.

As shown in Figure 1, most of the companies come from three industries, Software & Services industry (32%), Technology Hardware & Equipment industry (31%), and Semiconductors Semiconductor Equipment industry (30%). 3% of the stocks come from the Media & Entertainment industry. We also have stocks come from the Retailing, Pharmaceuticals, Biotechnology & Life Sciences, and Capital Goods industry.

The selection process of our 115 companies investment universe consists of two steps. First, we collect and combine every component stock from several AI related indexes such as Vanguard Information Technology Exchange-Traded Fund (ETF) and Global X Robotics & Artificial Intelligence Thematic ETF [12]. This gives us a basic company pool to select from. Then we get a list of companies including companies in non-US markets or do not have publications and patents. Second, we remove the stocks which are not in the US stock market and only keep the companies that have at least one publication or patent record during our backtesting period 2009-2018. Finally, we obtain our investment universe that contains 115 publicly trade companies.

C. Data Collection

Based on the investment universe and the database we choose, we start collecting data.

We firstly search for the 115 companies in our investment universe in the Microsoft Academic Graph database, for example, Google is recorded as an institution. Then, we collect all the papers and patents records of Google from the database including publication date, publisher and fields of studies. We get $782,294$ publication and patent records in total. Table 1 shows the schema of the original data.

TABLE I
AN EXAMPLE OF THE DATA SCHEMA.

Field	Type	Example
company	categorical	Apple Inc.
ticker	string	AAPL
type	categorical	Conference/Journal
fields_of_study	string	Machine Learning
publisher	string	ICML

But not every record in our data is within the fields of AI. Hence, we filter the records based on their fields of study tagged by Microsoft Academic Graph to get the final dataset. Microsoft Academic Graph tags papers with fields of study using artificial intelligence and semantic understanding of content. The criteria we used to select the fields of study is also based on this database. Microsoft Academic Graph will calculate the related field for each field of study. For example, within the web page for the field Artificial Intelligence [13], we collect all the fields in the "related fields of study" section that are related to Artificial Intelligence such as "Machine Learning", "Natural Language Processing", etc., and then we filter our original records using these fields.

We have $88,749$ records in total back to 1970 for all the companies in our investment universe in our final dataset. 55% of the records are patents, 28% are conference publications, 12% are journal publications, and 5% are books and articles. Notice that only 2% and 22% among all the records in Microsoft Academic Graph are conference publications and patents. Conference publications and patents have a much larger proportion in our dataset than the whole Microsoft Academic Graph database, which is consistent with our observations that AI industry has more conference publications and higher growth rate of patents than other industries. This means our collection process is robust and our final dataset is reasonable because it confirms the trends in AI industry.

In order to verify whether our company selection and record selection process is appropriate and our final dataset is in consistent with the trends of the AI industry, we calculate the correlation between the number of records in our final dataset and the number of record under the Artificial Intelligence topic all over the world in the Microsoft Academic Graph database. Figure 2 shows that the corresponding correlation and first difference correlation between the two time series are pretty high, which shows that our scholar data is typical and in consistent with the trends of the AI industry.

D. Feature Engineering

Firstly, we count the total number of publications and citations every month. As R&D expenditures usually have cumulative effects [14], [15], we calculate the cumulative summation. Then we have two features based on the number of publications and two features based on the number of

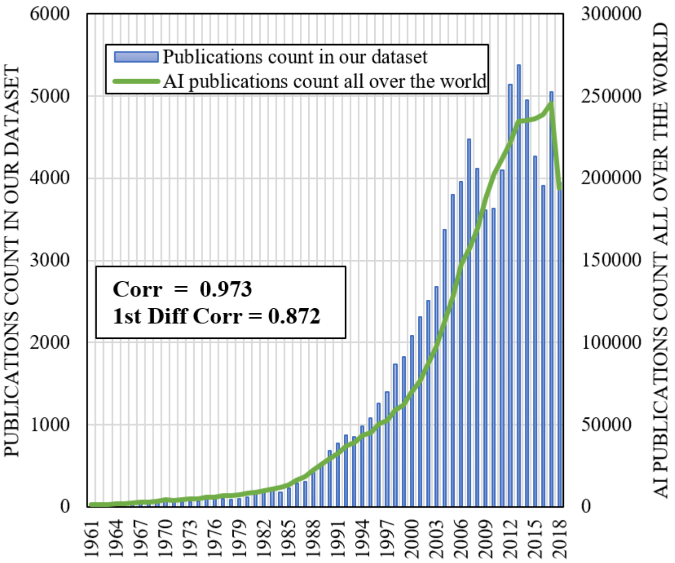

Fig. 2. Trends of publication count.

citations. Next we split the features into four types, which are conference publications, journal publications, patents or books. As we have four features for four types, we get another 16 features and 20 features in total. Besides, we create relative number of publications and its cumulative summation to reduce the potential bias of absolute values caused by unbalanced development and popularity of different fields of study. For each paper, its relative count equals to one divided by the total number of publications last year under the most popular topic within this paper's fields of study. By calculating the relatives values and the cumulative relatives values, we add another 10 features our features. Secondly, we calculated the average and highest number of citations among the records within each month by and across four types. Then we get another 10 features. Because in most cases a high citation number may indicate high quality works or a breakthrough in a field, which might have significant impact on companies' performance [8]. So far, we have calculated 30 features based on number of publications and number of citations and 10 features based on the max and average number of citations, which is 40 features in total.

E. Financial Indicators

Except for the scholar features, we select 10 traditional financial indicators that best represent company's fundamental conditions, which can be used to predict company's stock price [16]–[18]. We use financial indicators as a baseline to create an investment strategy, and then compare it with the investment strategies based on scholar data only as well as scholar data and financial indicators together.

We select 10 financial factors based on different aspects of company's fundamental conditions. We use Price-to-Earnings ratio (P/E), Price-to-Sales ratio (P/S), and Price-to-Book ratio (P/B) as price ratios to evaluate whether a company's current stock price is reasonable or not; Earnings-per-Share (EPS), Return-on-Assets (ROA), Return-on-Equity (ROE), and Net Profit Margin (NPM) as profitability ratios to see how profitable a company is; Current Ratio (CR) and Quick Ratio (QR) to represent a company's short term liquidity; Debt-to-Equity Ratio (D/E) to check a company's long-term health. These financial factors curve the fundamental situation of companies. The financial data is mainly taken from Compustat database accessed through Wharton Research Data Services [19].

F. Dealing with Citation Causality

Causality is the relationship between cause and effect, which is very important in the process of investment strategies development and backtest. One can only make investment decisions based on information in the past, using any future information in any format would break causality and lead to invalid results.

During our investigation, we do two things to avoid using future data and generate more reliable results. First, we calculate monthly number of citations for each paper based on the actual citation relationships instead of using cumulative citation counts from the Microsoft Academic Graph database. If paper A in our dataset was cited by n papers at month t, then the citation number for paper A at month t is n. To compute historical citation numbers for each paper, we firstly search for all papers which cited papers in our final dataset. Then we count the number of citations based on the publication date of the cited-by papers. More than one million citation relationships are used during this process.

Second, to construct our investment universe, we select companies based on every component stock including historical components from several AI related indexes. The reason why we also include historical components is to avoid survivorship bias. Survivorship bias happens when we only select stocks existing in the market or in the index. Index changes its component stocks either add stocks or remove stocks sometimes. Stocks may be removed from an index due to bad financial performance or bankruptcy. If we only select stocks existing in the market or in the index today for backtesting, then we actually filter out the companies with potential bad performance and bankruptcy, which contains future information about companies financial performance. Survivorship bias will lead to the overestimation of strategy performance and make the results of back testing invalid.

III. Practical Machine Learning Approach

We are essentially facing a regression problem which uses scholar data as features to predict individual stock's monthly return. We need algorithms that can deal with the high dimensions, capture linear as well as non-linear relations between different variables, and memorize patterns and dependencies in time-series data.

Assume current time t, our goal is to predict stocks' return r_{t+1} given feature vector X_t constructed in Section II. The return r_{t+1} is calculated as:

$$r_{t+1} = (S_{t+1}/S_t) - 1, \tag{1}$$

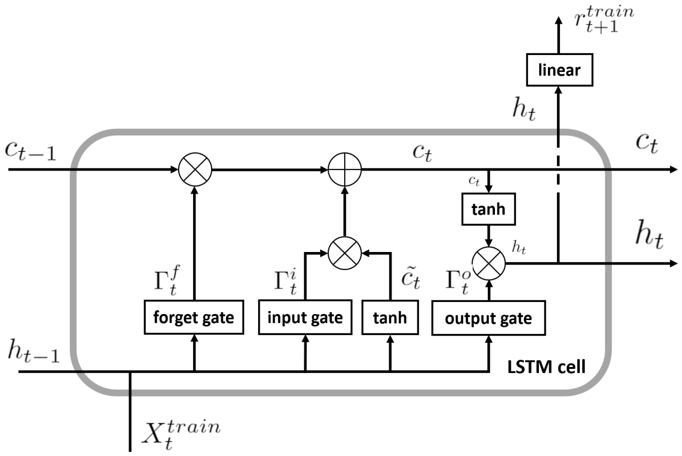

Fig. 3. Illustration of an LSTM-cell.

where S_t is stocks' price in time t. We model the relationship between r_{t+1} and X_t as follows:

$$r_{t+1} = g_\theta(X_t) + \epsilon, \tag{2}$$

where we use machine learning and deep learning algorithms g_θ to model the relationship [17], [18], [20], [21].

A. Machine Learning Models

1) LSTM [22]: The development and evolution of technology are accumulating. Past technology paves the way for future technology. We need to memorize the long term dependency to capture the cumulative effect of the technology evolution [23]. Recurrent neural network (RNN) is a type of neural networks that are specializing in modeling sequential and time-series data. In order to address the vanishing gradient problem in long term dependencies of RNN, Hochreiter and Schmidhuber proposed an effective gradient-based method called Long Short-Term Memory (LSTM) [22]. LSTM is then widely used to predict stock market prices and returns in recent studies [24]–[26].

LSTM has the ability to manipulate its memory state to caputre the cumulative effect of technology evolution by using the gate mechanism. A typical LSTM network contains input layer, one or more hidden layers, and an output layer; a hidden layer consists different memory cells; a memory cell uses cell state to pass information; LSTM has three gates to control the cell state by keeping or removing information [22], [27]. Figure 3 illustrates an LSTM-cell.

We want to use scholar data as features X_t^{train} to predict stocks' return r_{t+1}^{train}. We will use LSTM as a regressor in (2). First, we need to find a way to remove our previously stored memory state of the return values. We use the forget gate:

$$\Gamma_t^f = \sigma(W_f \cdot [h_{t-1}, X_t^{train}] + b_f), \tag{3}$$

where W_f are weights that control the forget gate, b_f is the bias of the forget gate, $[h_{t-1}, X_t^{train}]$ is the concatenate of h_{t-1} and X_t^{train}, σ is the sigmoid gating function that outputs values between 0 and 1, so that it can let no flow or complete flow of information pass through the gate, h_{t-1} is the previous hidden state, X_t^{train} is the training data at time step t. The equation above results in a vector Γ_t^f with values between 0 and 1, which is the forget state at time step t. Next, we need to find a way to update the information from new input data to current memory cell, the input gate is:

$$\Gamma_t^i = \sigma(W_i \cdot [h_{t-1}, X_t^{train}] + b_i), \tag{4}$$

similar to the forget gate, here W_i are weights that control the input gate, b_i is the bias of the forget gate, Γ_t^i is also a vector of values between 0 and 1. We need to update new information to the cell so we create a new vector that we can add to our previous cell state. Here we use:

$$\tilde{c_t} = \tanh(W_c \cdot [h_{t-1}, X_t^{train}] + b_c), \tag{5}$$

where W_c and b_c are the weights and bias to control new information. So we can get the new cell state:

$$c_t = \Gamma_t^f \cdot c_{t-1} + \Gamma_t^i \cdot \tilde{c_t}. \tag{6}$$

where c_t is calculated based on previous cell state c_{t-1}, the forget gate Γ_t^f in (3), the input gate Γ_t^i in (4) and new information $\tilde{c_t}$ in (5). Last, to control the output information from the memory cell, we use the output gate:

$$\Gamma_t^o = \sigma(W_o \cdot [h_{t-1}, X_t^{train}] + b_o), \tag{7}$$

$$h_t = \Gamma_t^o \cdot \ \tanh(c_t), \tag{8}$$

where W_o and b_o are the weights and bias for the output gate, the hidden state at time step t is then calculated by multiplying the output gate in (7) by the tanh of the new cell state in (6).

We choose the Mean Absolute Error (MAE) as the loss function for our LSTM network, because it is more robust to outliers. The MAE is calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |r_{t+1}^{train} - g_\theta(X_t^{train})|, \tag{9}$$

where t is the current time step, n are the number of training data points, r_{t+1}^{train} are the true values of stocks' next month return, $g_\theta(X_t^{train})$ are the predicted returns, g_θ in here is our LSTM network.

2) Linear Regression and its improvements [28]: We use linear regression because it is simple and easy to train. It minimizes the residual sum of squares:

$$\text{RSS} = \sum_{i=1}^{n} (r_{t+1}^{train} - g_\theta(X_t^{train}))^2, \tag{10}$$

where $g_\theta(\cdot)$ in here is linear regression. By adding regularization penalty term in (10) and (11) to RSS, Linear Regression is improved to Lasso and Ridge, respectively,

$$\text{minimize}\Big\{\text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|\Big\}, \tag{11}$$

$$\text{minimize}\Big\{\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2\Big\}, \tag{12}$$

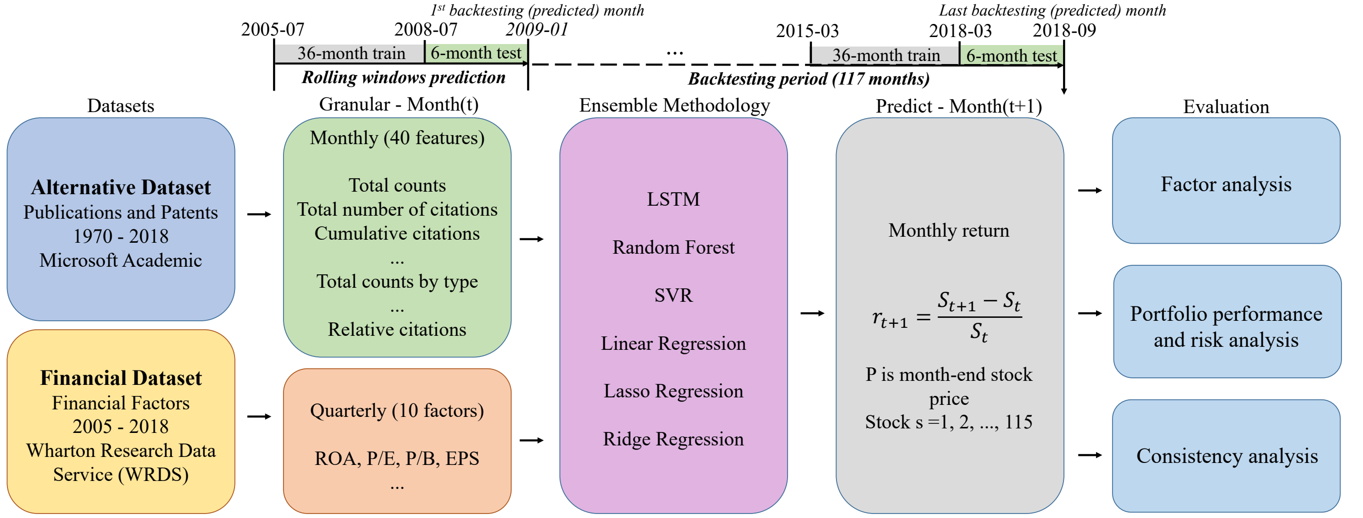

Fig. 4. Backtesting process.

where β is the vector of coefficients, p is the number of features in the training data. Lasso also has feature selection characteristics that can better filter out the insignificant coefficients in our scholar features [28].

3) Support Vector Regression [29]: Support Vector Regression (SVR) scales well to high dimensional data by using Radial Basis Function (RBF) as kernel. SVR solves an optimization problem by using the approximation function:

$$\sum_{i=1}^{l}(-\alpha_i+\alpha_i^*)K(x,x')+b, \tag{13}$$

where $K(x,x')$ is a kernel function, here we use RBF:

$$K(x,x')=\exp\Big\{-\gamma||x-x'||_2\Big\}, \tag{14}$$

where $||x-x'||_2$ is the squared Euclidean distance between two data points x, x' and γ is a parameter that sets the spread of the kernel [29].

4) Random Forest [28]: To reduce out-of-sample variance, we consider Random Forests because it can prevent overfitting. Random Forests are an ensemble learning method for a bunch of decision trees, in our case we use regression trees. A regression tree also use RSS in (10) as the cost function [28]. Random forests are extremely flexible and usually have very high accuracy.

B. Hyperparameter Tuning

We use Python's deep learning library keras to implement our LSTM network [30]. We tune the hyperparameters such as the number of hidden layers, the number of neurons in the hidden layers, activation function, optimizer, batch size, and epoch during our in sample period. As noted by Reimers and Gurevych, applying a regularization method called dropout will reduce over-fitting and increase performance significantly. So we add dropouts at every time step of the LSTM-layer [31]. We choose linear activation function in the output layer because we are facing a regression problem instead of classification problem which usually uses sigmoid or softmax as activation function.

We use Python's machine learning library scikit-learn to implement Linear Regression, Lasso, Ridge, SVR, and Random Forest [32]. We tune the parameters using grid-search every time we train. For example, Lasso and Ridge have alpha to balance RSS and magnitude of coefficients. SVR has C and Gamma as hyperparameters for RBF, where Gamma is a parameter to control non linear hyperplanes; C is the penalty parameter of the error term, it trades off between smooth decision boundary and classifying the training data points correctly. Random Forest has hyperparameters such as number of trees, maximum features, maximum depth, minimum samples leaf, minimum samples split, and bootstrap.

C. Ensemble Method

Our purpose is to build a highly robust trading model. So we use an ensemble method to choose the best algorithm from LSTM, Linear Regression, Lasso Regression, Ridge Regression, Random Forest, and SVR to trade.

We train and validate the 6 algorithms concurrently with a rolling-window based train and validation set. We only select the best performing model at each trading period after the train-validate process by minimizing the evaluation metrics Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|r_{t+1}^{val} - g_\theta(X_t^{val})|, \tag{15}$$

for 6 models, we select the model with minimum MAE as our trading model.

D. Evaluation Metrics

Mean Absolute Error (MAE) is a general metric to compare the out-of-sample forecast accuracy. MAE is the most natural

TABLE II
MODEL USAGE FOR TRADING

Model	Usage	Percentage
LSTM	41	35%
Lasso	37	31%
SVR	27	23%
Linear Regression	6	5%
Random Forest	3	3%
Ridge	3	3%

measure of average error magnitude over other measurement such as the Root Mean Square Error (RMSE), because MAE is more appropriate to measure dimensional evaluations and inter-comparisons of average model-performance error [33]. Furthermore, we choose MAE over Mean Square Error (MSE) and Mean Squared Logarithmic Error (MSLE), because our goal is to predict stocks' returns which are between 0 and 1, MSE will result in a very small number that is hard to compare between models, and we can't use MSLE because returns have negative number. Therefore, we select MAE as our model evaluation metrics.

E. Backtesting

We follow a train-validate-trade process. We use rolling window for both training and validation by Figure 4.

The rolling window length for our training period is 36 months after that is our validation period which is 6 months, and the number of increments between successive rolling windows is 1 month as we trade and re-balance the stocks monthly. We also lag the scholar factors by 1 month because usually it takes time for research outputs go into production [8]. At the end of each month, we will get a set of predicted returns for each stock. Finally, we will use the set of predicted returns to construct the equally-weighted AI portfolio by selecting the top 25% performance stocks and then we use the buy-and-hold-long-only strategy to trade the AI portfolio.

The train-validate-trade process is described as follows:

Step 1. We train the 6 models based on the same features and monthly returns data concurrently on a 36-month-train-rolling window.

Step 2. We validate all 6 models by using a 6-month-validate-rolling window followed by the 36-month-train-rolling window. We calculate the MAE of each model.

Step 3. After validation, we select the best model which has the lowest MAE to predict and trade. So we will have a set of predicted returns for all current stocks. We rank the stocks by the predicted return and only select top 25% stocks with highest predicted returns to form our portfolio. We use LSTM as our trading model 41 times out of 117 backtesting months, as listed in Table II. From the result we can see that LSTM is the best model for our topic.

IV. PERFORMANCE EVALUATION

In this section, we present performance evaluation of our trading strategy from different aspects. Section IV.A examines robustness of the scholar data driven features and our thesis about the relationship between companies AI research capacity and companies' stock performance. Section IV.B presents the performance comparison between our portfolio and different benchmarks. Section IV.C presents the performance comparison between our portfolio and the financial-factor-only portfolio. Section IV.D analyzes the consistent performance of our portfolio.

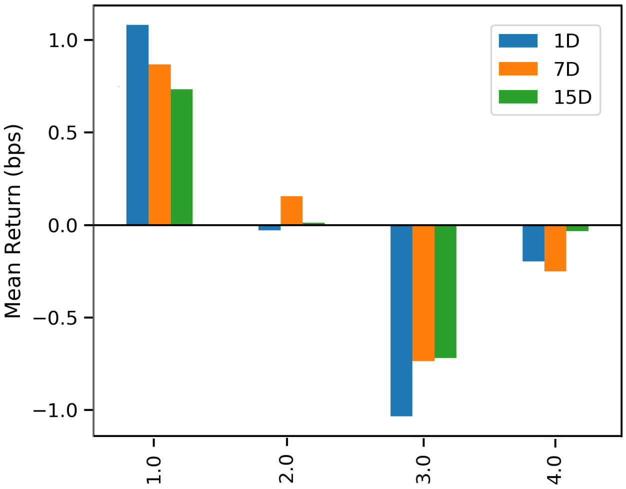

Fig. 5. Mean period wise return by factor quantile.

A. Factor Analysis

We firstly use the features generated only from the scholar data to predict stocks' monthly returns. When we construct this portfolio using the scholar data, we take the predicted return as our factor value, which means the higher the predicted return, the higher return we expect from this stock, in other words, we tend to buy stocks with the highest predicted returns and not to buy stocks with the lowest predicted returns. Hence, the predicted return can be viewed an alpha factor for us to construct our portfolios.

We use a popular approach in factor analysis to analyze the scholar data driven factor's predictive power, that is, to look at the period wise mean return of factor's different quantiles. We simply rank the value of factors and divide them into four groups. The stocks in the top quantile (top 25%) have the highest factor scores and the stocks in the bottom quantile (bottom 25%) have the lowest factor scores. We expect the top quantile to have the highest returns because those companies published more papers and patents than the companies in another three quantiles and thus have more research capacity to utilize the premium of AI technologies. Figure 5 shows the period wise mean return for each of the scholar-data-driven factors' quantiles. The top quantile group has the highest positive returns, the third quantile group has the highest negative returns and the bottom quantile group also has nagative returns. This result is consistent with our thesis about the positive correlation between companies AI research capacity and companies' stock performance.

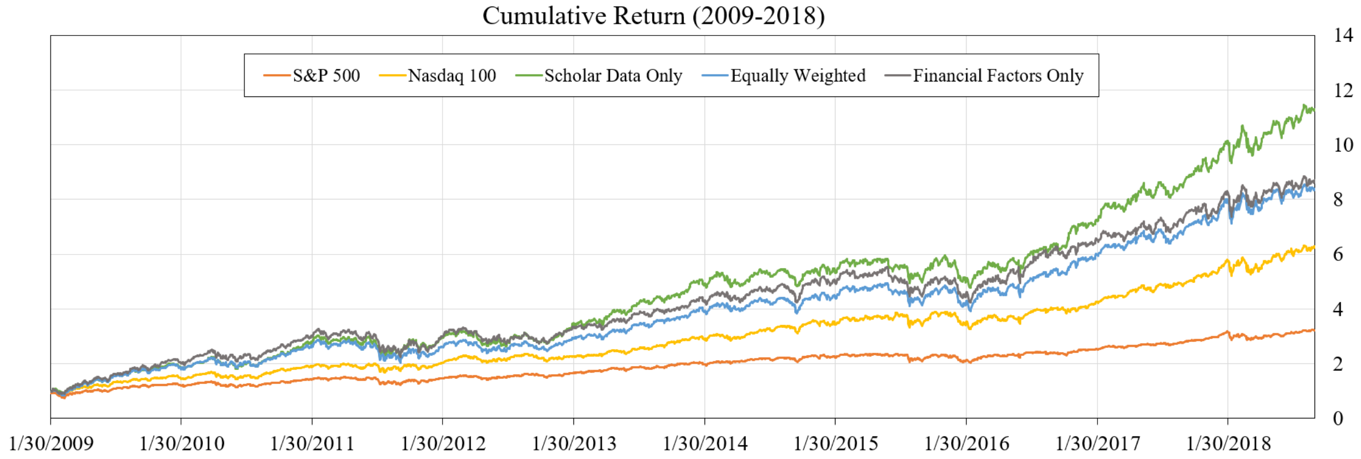

Fig. 6. Cumulative returns for different portfolios.

TABLE III
PERFORMANCE EVALUATION COMPARISON.

(2009-2018)	**Scholar Data Only**	**Equally Weighted**	**Financial Factors Only**	**Nasdaq 100 Index**	**S&P 500 Index**
Annual Return	28.5%	24.2%	25.0%	20.8%	12.8%
Cumulative Return	1029.1%	723.6%	776.7%	529.5%	222.6%
Sharpe Ratio	1.24	1.09	1.10	1.14	0.82
Annual Volatility	22.2%	22.1%	22.7%	18.0%	16.4%
Max Drawdown	-27.3%	-29.1%	-31.2%	-18.6%	-27.6%
Daily Value at Risk	-2.7%	-2.7%	-2.8%	-2.2%	-2.0%

In this paper, *we use long-only portfolios instead of long-short portfolios and other portfolio allocation techniques because we want to highlight that the performance of the scholar data driven approach is due to the predictive power of our alpha factor but not portfolio allocation techniques.* We can still achieve higher return if we apply asset allocation rules such as long the top quantile group and short the third quantile group.

B. Performance Comparison: Scholar Data Factors Only Portfolio and Benchmarks

First, we construct a portfolio that is only based on scholar data factors, which means during the model training process, we do not use any financial factors to predict stock returns. As described in Section III, we use scholar data factors to predict monthly returns of each stock and only long the top 25% stocks which have the highest predicted returns to construct this scholar data factors only portfolio. And we also construct a baseline portfolio by equally weighted all the stocks in our investment universe.

As shown in Figure 6 and Table III, compared with S&P 500 and Nasdaq 100, the performance of the scholar data factors only portfolio as well as the equally weighted portfolio is very impressive. The cumulative return of the scholar data factors only portfolio is 1029.1% during our backtesting period from 2009 to 2018 (Table III), which is 42.3% better than the equally weighted portfolio, 94.4% better than the Nasdaq 100 index and 362.3% better than the S&P 500 index. The scholar data factors only portfolio also reaches the highest Sharpe ratio of 1.24.

The result shows that by using the scholar data, we are able to capture the efficacy of the scholar data driven alpha in AI industry. First, both of the scholar data factors only portfolio and the equally weighted portfolio generate better return than the two market indexes. Notice that every company in our investment universe has at least one publication or patent record during our backtesting period in the Microsoft Academic Graph database. In other words, we also utilized scholar data to construct the equally weighted portfolio as every stock was selected based on the scholar data. The better results of the two scholar data related portfolios indicate that we have successfully captured the scholar data driven alpha in AI industry. Second, the performance of the scholar data factors only portfolio is still much better than the equally weighted portfolio, which indicates that our practical machine learning approach is quite efficient to process massive information and capture the complex relationship between the scholar data factors and the stocks' returns. Cumulative return and Sharpe Ratio shows us overall performance of the investment strategies. To gain more details about our scholar data factors only strategy, we need to look at the more granular return data.

Figure 7 shows the details of return distribution for the scholar data factors only portfolio. Although the annual returns seem fluctuating, most of the annual returns are still quite steady and consistent in terms of direction. Time stable portfolio usually has more value than unstable factors as unstable factors are more risky and would lower down the actual Sharpe Ratio of strategy.

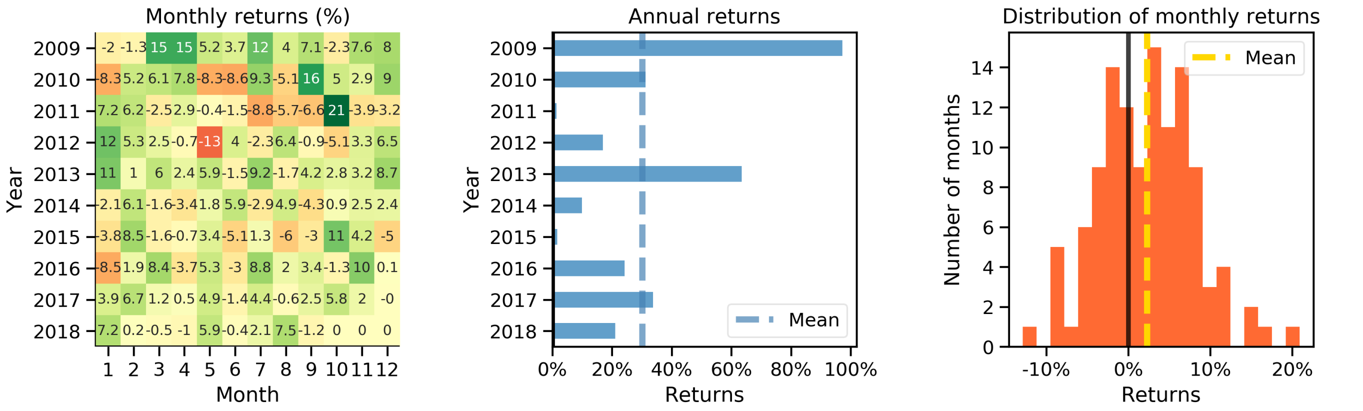

Fig. 7. Return distribution.

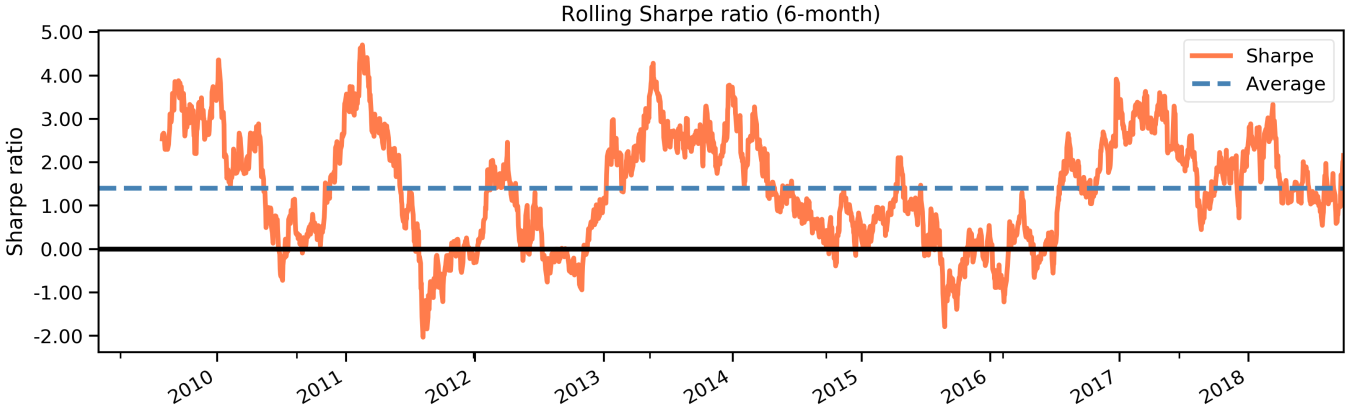

Fig. 8. Sharpe Ratio of the combined factors portfolio over 6 months.

C. Performance Comparison: Scholar Data Factors Only Portfolio and Financial Factors Only Portfolio

Except for the scholar data factors only portfolio, we construct another portfolio only based traditional financial factors. The comparison between these two portfolios is actually the comparison between scholar data and traditional financial data.

Figure 6 and Table III show the performance comparison between these two portfolios. The cumulative return of the scholar data factors is 34.5% better than the financial factors only portfolio (776.7%). The Sharpe Ratio of the scholar data factors only portfolio is still better than the financial factors only portfolio. The result shows that scholar data is much better to capture the AI growth premium in AI industry than the traditional financial data. As discussed in Section II, we choose scholar data based on the three main characteristics in AI industry. We believe that scholar data contains much more information than the traditional financial data in terms of companies' AI research capacity and AI growth premium.

Using alternative datasets can sometimes generate better performance for investors, because those datasets might contain different or more granular information about companies' fundamental conditions. We should also notice that the performance of financial factors only portfolio is still better than all the other portfolios, which indicates that traditional financial data can still generate a baseline alpha for investors, and machine learning techniques can help investors capture the complex relationships between different financial factors and make better investment decisions.

D. Consistent Performance of the Scholar Data Factors Only Portfolio

Figure 8 shows the rolling Sharpe Ratio of our scholar data factors only portfolio which is fluctuating around 1.24. During most of the backtesting period, our portfolio has a positive Sharpe Ratio which is also larger than one, which indicates that the returns on the portfolio are larger than the risk taken can be considered good by investors. Especially after mid-2016, the performance of the portfolio is considered to be quite good and consistent, which shows the alpha does not decay across so many years. The consistent performance of the scholar data factors only portfolio indicates that investing in AI industry would become a great choice for investors, and scholar data could be the dataset to provide unique insights about the industry and companies.

V. Conclusion

Scholar data give us a novel and unique perspective to analyze and predict AI companies' stock price performance by exploiting the characteristics of AI industry to better predict the impacts of AI technologies on companies, and quantifying companies' growth premium that is rewarded by AI technologies. In this paper, we choose the publication and patent data as the scholar data to generate investment insights into AI industry. The scholar-data-factors only portfolio significantly outperforms the benchmarks as well as the traditional financial factors only portfolio in cumulative return and Sharpe ratio, indicating that we successfully extract the effective investment alpha from the scholar dataset. We firmly believe that in the era of big data, the rising of AI and machine learning techniques makes it possible for us to utilize various scholar datasets, expand the boundary of traditional financial data and generate higher returns. Future works would be more granular feature engineering based on the scholar data such as utilizing abstract of publication or patent to get more accurate topic classification, and constructing a portfolio hedged by common risk factors to filter out noise and get more accurate estimations about the scholar data driven alpha.

References

[1] YahooFinance, "Yahoo finance," 2018. [Online]. Available: https://etfdb.com/themes/artificial-intelligence-etfs

[2] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 243–246.

[3] X.-Y. Liu and X. Wang, "Ls-decomposition for robust recovery of sensory big data," *IEEE Transactions on Big Data*, vol. 4, no. 4, pp. 542–555, 2017.

[4] N. Ahmad and A. Zinzalian, "Predicting stock volatility from quarterly earnings calls and transcript summaries using text regression," *CS224N Final Report*, 2010.

[5] W. Zhang and S. Skiena, "Trading strategies to exploit blog and news sentiment," in *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[6] X. Liu, A. Nourbakhsh, Q. Li, S. Shah, R. Martin, and J. Duprey, "Reuters tracer: Toward automated news production using large scale social media data," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 1483–1493.

[7] F. Kilburn, "Ubs asset management cools on alternative data," 2019. [Online]. Available: https://www.risk.net/asset-management/6552756/ubs-asset-management-cools-on-alternative-data

[8] K. R. Harrigan and Y. Fang, "The financial benefits of persistently high forward citations," *The Journal of Technology Transfer*, pp. 1–29, 2019.

[9] E. F. Fama and K. R. French, "Multifactor explanations of asset pricing anomalies," *The Journal of Finance*, vol. 51, no. 1, pp. 55–84, 1996.

[10] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Transactions on Big Data*, vol. 3, no. 1, pp. 18–35, 2017.

[11] L. Columbus, "Microsoft leads the ai patent race going into 2019," 2019. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2019/01/06/microsoft-leads-the-ai-patent-race-going-into-2019

[12] ETFdb, "Artificial intelligence etf list," 2019. [Online]. Available: https://etfdb.com/themes/artificial-intelligence-etfs

[13] MicrosoftAcademic, "Artificial intelligence — topic — microsoft academic," 2019. [Online]. Available: https://academic.microsoft.com/topic/154945302

[14] Z. Griliches and J. Mairesse, "Productivity and r & d atthe firm level," 1984.

[15] M. Heeley, D. King, and J. Covin, "R&d investment level and environment as predictors of firm acquisition," *Journal of Management Studies*, 2006.

[16] M. C. Scott, "Value investing: A look at the benjamin graham approach," in *Stock Analysis Workshop*, 1996, pp. 12–15.

[17] H. Yang, X.-Y. Liu, and Q. Wu, "A practical machine learning approach for dynamic stock recommendation," in *IEEE International Conference on Trust, Security And Privacy (TrustCom)*. IEEE, 2018, pp. 1693–1697.

[18] J. J. Gerakos and R. Gramacy, "Regression-based earnings forecasts," *Chicago Booth Research Paper No. 12-26.*, 2013.

[19] WRDS, "Compustat industrial [daily and quarterly data," 2019. [Online]. Available: https://wrds-web.wharton.upenn.edu/wrds/

[20] K. Hou, C. Xue, and L. Zhang, "Digesting anomalies: An investment approach," *Fisher College of Business Working Paper No. WP 2012-03-021*, 2014.

[21] G. Ritter, "Machine learning for trading," *SSRN Electronic Journal*, 01 2017.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] K. R. Harrigan and Y. Fang, "Financial implications of technology-class code popularity and usage among industry competitors," *Scientometrics*, pp. 1–27, 2019.

[24] D. M. Nelson, A. C. Pereira, and R. A. de Oliveira, "Stock market's price movement prediction with lstm neural networks," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 1419–1426.

[25] B. Zhou, "Deep learning and the cross-section of stock returns: Neural networks combining price and fundamental information," *SSRN Electronic Journal*, 03 2019.

[26] X. Li, Y. Li, X.-Y. Liu, and C. D. Wang, "Risk management via anomaly circumvent: Mnemonic deep learning for midterm stock prediction," *KDD Workshop on Anomaly Detection in Finance*, 2019.

[27] J. S. F. A. Gers and F. Cummins, "Learning to forget: continual prediction with lstm," in *1999 Ninth International Conference on Artificial Neural Networks*. ICANN 99, 1999, pp. 850–855, vol.2.

[28] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[29] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[30] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[31] N. Reimers and I. Gurevych, "Optimal hyperparameters for deep lstm-networks for sequence labeling tasks," *ArXiv*, vol. abs/1707.06799, 2017.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[33] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate Research*, vol. 30, p. 79, 12 2005.