LEARNING FROM INVESTOR ATTENTION

EXAMINING THE PREDICTIVE POWER OF INVESTOR ATTENTION ON MARKET RETURNS WITH MACHINE LEARNING

LUDVIG HARTLER

LUKAS UHRSTRÖM

Bachelor Thesis

Stockholm School of Economics

2022

Learning From Investor Attention: Examining the Predictive Power of Investor Attention on Market Returns with Machine Learning

Abstract:

We study the predictive properties of investor attention on time series market returns. Extending an earlier proposed index of investor attention aggregated from twelve popularly studied attention proxies, we show that it strongly predicts excess returns on the stock market. Adding to an inconclusive body of literature, our results suggest that when attention is high, the market earns higher returns in the subsequent months. Uniquely, we examine whether a deep learning method from the machine learning catalog, Long-Short Term Memory, can enhance the predictability. Our results show that the nonlinear patterns in the data studied in this paper are not strong enough to yield economic gains.

Keywords:

investor attention, machine learning, PLS, LSTM, RNN, time series, stock market, stock returns, nonlinearity

Authors:

Ludvig Hartler (24672)
Lukas Uhrström (24706)

Tutor:

Adrien d'Avernas, Assistant Professor, Department of Finance

Examiner:

Adrien d'Avernas, Assistant Professor, Department of Finance

Bachelor Thesis
Bachelor Program in Business and Economics
Stockholm School of Economics

Attention is inherently limited (Kahneman, 1973). While much research has been devoted to understanding its effects on stock-specific returns (Barber & Odean, 2008; Ben-Rephael, Da, & Israelsen, 2017; Da, Engelberg, & Gao, 2011; Dellavigna & Pollet, 2009; Peng & Xiong, 2006), recent literature suggests it has a market-wide impact (Chen, Tang, Yao, & Zhou, 2022). Given the presence of behavioral biases in the decision making of investors and the complex features of the market, this impact might not necessarily be linear in nature (Banerjee & Green, 2015; Hsieh, 1991). Despite the growing interest in nonlinear technical analysis since the entrance of machine learning in asset pricing, there is yet little empirical research on the nonlinear relationship between investor attention and time series market returns.

This article adds to the subfield of stock prediction studying the impact of investor attention on the time series of market returns. Our starting point is twelve proxies for investor attention put forth by Chen et al. (2022), with which we test for a combined predictive power on market returns in and out of sample. Subsequently, we broaden the understanding of nonlinearity in the data by applying a machine learning method capable of modeling complex relationships. In contrast with prominent cross-sectional literature claiming stock prediction to be enhanced by allowing for nonlinearity (Gu, Kelly, & Xiu, 2020; Kozak, Nagel, & Santosh, 2020), we find no robust evidence for the same applying to the predictive power of investor attention.

Our contributions are fourfold. First, we add to the evidence that investor attention has predictive power on market returns and uniquely show its maintained impact beyond 2017. Second, we show that a linear regression on components following a dimensionality reduction outperforms our nonlinear machine learning model as the chosen method for our dataset, which directly contributes to answering Chen et al. (2022) who called for machine learning applications as important future research on the topic. Third, we enrich the discussion surrounding the sign of the coefficient explaining the relationship between investor attention and stock returns. The literature is inconclusive (Campbell & Thompson, 2007), presenting evidence for positive (Gervais, Kaniel, & Mingelgrin, 2001; Li & Yu, 2012) and negative (Barber & Odean, 2008; Chen et al., 2022) slope signs depending on the examined measure used to proxy investor attention. By adopting identical methods and still finding results opposing those of Chen et al. (2022), we show that minor data differences can yield an impact on the direction of the slope. Lastly, we provide the most comprehensive investor attention index to date (to the authors' knowledge at the time of writing), extending the time horizon of the attention index developed by Chen et al. (2022). In aggregating the twelve individual proxies for the attention of investors, the index widely encapsulates market-level investor attention. Thus, its usage areas extend outside stock market return prediction, much like the more well-researched sentiment indices (Baker & Wurgler, 2007; Huang, Jiang, Tu, & Zhou, 2015).

The main empirical approach of this paper is as follows. Given the inherently unobservable nature of investor attention (Chen et al., 2022; Huang et al., 2015), we consider a number of frequently used proxies. In spite of their low individual contribution to predictive power, a combination of these proxies has shown to have predictive power on market returns (Chen et al., 2022). In order to separate the investor attention from these proxies and disregard noise, we use two separate approaches. In the linear benchmark, we extract the common components through a factor structure model before doing repeated regressions maximizing explanatory power and minimizing the error term. In our search for nonlinearity, we relax the assumption of a factor structure model to allow Recurrent Neural Networks (RNNs) to instead rid itself of noise by training itself on data before measuring its maximum predictive power on a separate test set.

Our model choice for the linear benchmark is Partial Least Squares (PLS), which since its creation by Wold (1966) has been subsequently developed into the version we apply (Kelly & Pruitt, 2015; Light, Maslov, & Rytchkov, 2017). In short, PLS was chosen due to being the best approach applied by Chen et al. (2022) since it elegantly deals with dimensionality reduction and autocorrelation, and because it has a strong history of successful applications in quantitative finance (Chen et al., 2022; Light et al., 2017).

We then apply a Long-Short Term Memory (LSTM) to identify potential nonlinearity in the relationship between investor attention and market returns through a comparison with our linear results (Hochreiter & Schmidhuber, 1997). LSTMs have a built-in memory making it apt for time series analysis (Hochreiter & Schmidhuber, 1997) and have been successful in financial modeling (Roondiwala, Patel, & Varma, 2017). Additionally, LSTM employs a rolling time window, allowing it to mitigate the issue of concept drift in public markets - a phenomenon that is well-known, unsolved as an issue and associated with market efficiency (Nagel, 2021).

We find that our PLS investor attention index, used as a single predictor, has significant predictive power of excess returns. Regressing excess returns of the stock market on our investor attention measure, we are able to predict stock returns on a forecasting horizon from one month up to two years. We find an in-sample R^2 of 12.23% for yearly excess returns. Interestingly, we find that high (low) investor attention predicts higher (lower) subsequent excess returns on the stock market. This finding contradicts earlier empirical findings within the investor attention domain (Chen et al., 2022). It is in line, however, with the economic logic and theory laid out by Gervais et al. (2001) and Li and Yu (2012). For all forecasting periods, we obtain highly significant and strictly positive slopes peaking at 0.90% for monthly predictions and shrinking in increasing prediction horizon to 0.28% at the two year forecast. This means that a 1% increase in investor attention today predicts a 0.90% increase in next month's excess return of the stock market. We turn to forecasting out-of-sample using our investor attention measure and find that the predictive power remains for prediction horizons up until a year. We obtain an R^2_{OOS} of 3.37% for monthly excess returns and 2.37% for yearly forecasts. The coefficients remain strictly positive, in line with our findings in-sample.

With the results of PLS as a benchmark, we then turn to predicting the excess stock market return using LSTM. We find that LSTM underperforms relative to PLS. Occasionally, we are able to obtain a positive R^2_{OOS} on par with PLS (2.23% for the six month forecast). However, the results are not robust over the prediction horizons and highly volatile depending on hyperparameter choices.

Our main points of analysis begin with the finding of a positive correlation between investor attention and market excess returns, as indicated by the sign of the prediction coefficient. The literature is inconclusive on what sign the coefficient should have, but our analysis shows that the timing of investor attention vis-a-vis market returns may be impactful. Our test examining the potential to enhance predictive power of investor attention on stock returns show that nonlinear patterns are not sufficiently detectable to yield economic gains. However, doubts pertaining to the compatibility between our dataset and LSTM means that we are unable to make general conclusions.

Our findings challenge the notion that increased investor attention predicts decreased future returns, mainly argued to stem from the reversal of temporary price pressure (Chen et al., 2022). Rather, the results we present suggest a stronger prevalence of increasing visibility driving shocks in trader interest, and thus increasing demand in the short term, aligned with the findings of Gervais et al. (2001).

In brief, our research question is:

Does investor attention have predictive power on market returns and can it be enhanced by allowing for nonlinearity?

This article builds on several contemporary themes in financial research. First, our analysis complements earlier literature studying the predictive power of investor attention on stock-returns (Barber & Odean, 2008; Ben-Rephael et al., 2017; Da et al., 2011; Dellavigna &

Pollet, 2009; Gervais et al., 2001; Peng & Xiong, 2006). Specifically, we expand on the body of work examining the impact of investor attention on the aggregate market. Li and Yu (2012) show that nearness to the Dow 52-week high, as a proxy for investor attention, has predictability of aggregate market returns. Yuan (2015) finds in-sample evidence of the relationship between investor attention and the stock market. We expand on the research by Chen et al. (2022). First, we provide a direct extension of their suggested investor attention index. Second, we put their twelve proxies for investor attention in a machine learning context. Our findings adds to the inconclusive literature examining the slope of coefficient in predicting returns from investor attention (Barber & Odean, 2008; Chen et al., 2022; Da et al., 2011; Gervais et al., 2001; Li & Yu, 2012).

We benchmark the application of machine learning models with linear models in asset pricing, in coherence with Gu et al. (2020) and Kozak et al. (2020). Their work, however, focuses not on time series but on the cross-section of returns, a subject attended to more extensively in previous research. As such, we add to the relatively sparsely researched field of machine learning applied to time series analysis of stock market returns (Cao, Lin, Li, & Zhang, 2019; Roondiwala et al., 2017; Zhang, Chu, & Shen, 2021).

The remainder of the paper is organized as follows. Section I describes data and the construction of proxies for investor attention. Section II explains the empirical approach. Section III presents the empirical results and analysis. Section IV concludes.

I Data and Construction of Attention Proxies

We gather data and construct a series of attention proxies, based on the assumption that true aggregated attention of investors is unobservable (Chen et al., 2022; Huang et al., 2015). Despite the literature suggesting weak results for many proxies used in isolation, Chen et al. showed investor attention to have strong predictive power once combined 2022 by aggregating twelve attention proxies. We follow their approach and use the following proxies: abnormal trading volume (Barber & Odean, 2008), extreme returns (Barber & Odean, 2008), past returns (Aboody, Lehavy, & Trueman, 2010), nearness to the Dow 52-week high and nearness to the Dow historical high (Li & Yu, 2012), analyst coverage (Hirshleifer & Hong Teoh, 2003; Hirshleifer, Hsu, & Li, 2013; Peng, 2005), changes in advertising expenses (Lou, 2014), media coverage (Barber & Odean, 2008; Fang & Peress, 2009), mutual fund inflow and outflow, Google search volume (Da et al., 2011), and the number of document downloads on EDGAR (Drake, Roulstone, & Thornock, 2015; Lee, Ma, & Wang, 2015). We follow Chen et al. (2022) in collecting and computing these measures. Here follows a detailed description of our data construction. Any discrepancies contra Chen et al. (2022) is noted and explained.

For abnormal trading volume A^{AVol}, we compute the abnormal trading volume as volume traded at the end of the month (EOM) for $firm_i$ during $month_t$ in relation to the average trading volume of $firm_i$ during the last twelve months (LTM). We define extreme returns A^{ERet} as the ratio of return for $firm_i$ during $month_t$ to its average during LTM. Past returns A^{PRet} are calculated as the cumulative monthly return for $firm_i$ over LTM at $month_t$.

We define nearness to the Dow 52-week high A^{52wH} as the ratio between the level of the Dow Jones EOM at $month_t$ to its highest level in the past 52 weeks at $month_t$, and Nearness to the Dow historical high A^{HisH} is computed as the ratio between the level of the Dow Jones at EOM on $month_t$ to its highest historical level up to $month_t$.

For analyst coverage $A^{\#AC}$, we count the number of analyst forecasts of earnings per share (1 year ahead) for each $firm_i$. For changes in advertising expenses A^{CAD} we calculate the log change in advertising expenses between $year_{t-1}$ and $year_t$. Due to restricted access to the Compustat database where the data is collected, we deviate from Chen et al. (2022) who collect monthly changes in advertising expenditure. We use the yearly change for each company across all months during the year.

To obtain the number of document downloads on EDGAR A^{EDGAR}, we follow Ryans (2017) in line with Chen et al. (2022). We download the pre-cleaned EDGAR Log File Dataset and choose "*Rpv*" as the measure for number of document downloads on EDGAR. Here, it is not clear which measure of EDGAR document downloads Chen et al. (2022) uses. However, "*Rpv*" is suggested by the author Ryans (2017) since it was shown to minimize bot downloads which can be considered noise (Ryans, 2017).

For mutual fund outflows $A^{Outflow}$ we calculate fund outflows using redemption of shares for each $mutualfund_i$ during $month_t$, and for fund inflows A^{Inflow} we aggregate total new shares sold and "other sales" for each $mutualfund_i$ during $month_t$.

In collecting media coverage data A^{Media}, due to restricted access to the RavenPack database used by Chen et al. (2022), we compute media coverage as the number of news articles covering the largest companies in North America with respect to market capitalization during the data time period, 2004 to 2021, simulating a value-weighted approach adopted by Ma, Wang, and Zhang (2017). As suggested by Da et al. (2011) we search for stock tickers in order to obtain results likely to contain news read by investors. See Appendix I for a step-by-step description of how we filter and collect the data.

For Google search volume A^{Google} we obtain the Search Volume Index (SVI) from Google Trends. It is indexed so that the $month_t$ where search volume peaked for $firm_i$ during the time period takes a value of 100 for $firm_i$. Here we obtain data for the same tickers used in A^{Media}, a deviation from Chen et al. (2022) who uses all tickers on NYSE, NASDAQ and AMEX. The lower quantity of data from Google Trends is countered by the higher quality reached from manual downloads, since neither an API nor web crawler can select the correct topic ID. We thus assure that a search of e.g. "COP" refers to the searches interpreted by Google as pertaining to the stock of ConocoPhillips (including misspellings) instead of the exact word "cop". The manually collected data is available for free public use on our github[1] and might itself constitute a small contribution to further studies since the problem is noted in the literature (Da et al., 2011).

Lastly, we obtain the data for our target variable, excess stock returns R, which we define as monthly value-weighted aggregate stock return minus the U.S Treasury Bill-rate.

All cross-sectional equity data is collected from companies listed on the three major American stock exchanges NYSE, AMEX or NASDAQ during each respective period. We obtain data for A^{AVol}, A^{ERet}, A^{PRet}, A^{52wH}, A^{HisH}, $A^{\#AC}$ and A^{CAD} for the time period January 1980 to October 2021. Data for $A^{Outflow}$, A^{Inflow}, A^{Media} and A^{Google} is obtained from January 2004 to October 2021. Finally, for A^{EDGAR} the only available data extends from January 2004 to June 2017[2].

Data for A^{AVol}, A^{ERet}, A^{PRet}, $A^{Outflow}$ A^{Inflow} are obtained from the Center for Research in Security Prices (CRSP) database. We collect data for A^{52wH} and A^{HisH} from the Capital IQ database. The data for $A^{\#AC}$ is obtained through the Institutional Brokers Estimate System (IBES) database, data for A^{CAD} collected from the Compustat database and the news counts for A^{Media} are obtained from the Dow Jones Factiva. Lastly, we extract data for A^{Google} from Google Trends by manual downloads and data for A^{EDGAR} from the EDGAR log file dataset provided by Ryans (2017). Table 1 reports a summary of the construction of our twelve investor attention proxies.

[1] github.com/lukuhr

[2] We contacted the Structured Data division at the SEC who claimed they can provide no information on why the EDGAR log file dataset only extends to 2017. The same limit is imposed on the research conducted by Chen et al. (2022).

Table 1: The Twelve Proxies

Name	Source	Sample period
A^{AVol}	CRSP	Jan 1980 - October 2021
A^{ERet}	CRSP	Jan 1980 - October 2021
A^{PRet}	CRSP	Jan 1980 - October 2021
A^{52wH}	Capital IQ	Jan 1980 - October 2021
A^{HisH}	Capital IQ	Jan 1980 - October 2021
$A^{\#AC}$	IBES	Jan 1980 - October 2021
A^{CAD}	Compustat	Jan 1980 - October 2021
A^{Inflow}	CRSP	Jan 2004 - October 2021
$A^{Outflow}$	CRSP	Jan 2004 - October 2021
A^{Media}	Factiva	Jan 2004 - October 2021
A^{Google}	Google Trends	Jan 2004 - October 2021
A^{EDGAR}	Ryans	Jan 2004 - June 2017

Table 1 summarizes all variable names, the source from where the data was obtained and the sample periods for our twelve proxies of investor attention.

We compute monthly measures at the firm level before aggregating up to market-level for all attention proxies. For all measures where available, we use equal weighting in order to capture investor attention from a variety of different stocks and thus avoid biasing the attention towards firms with high market capitalization. Equal weighting is also applied by (Chen et al., 2022; Jondeau, Zhang, & Zhu, 2019; Rapach, Ringgenberg, & Zhou, 2016). Table 2 describes the median, first- and third- quartile, skewness, first-order autocorrelation and sample period of the twelve investor attention proxies.

Table 2: Summary Statistics

Name	1st Quartile	Median	3rd Quartile	Skewness	$p(1)$
A^{AVol}	-0.65	-0.14	0.45	1.40	0.50
A^{ERet}	-0.47	0.12	0.60	-0.80	0.17
A^{PRet}	-0.64	0.05	0.52	0.08	0.92
A^{52wH}	-0.20	0.41	0.64	-2.42	0.86
A^{HisH}	-0.36	0.40	0.77	-1.71	0.93
$A^{\#AC}$	-0.83	0.07	0.81	-0.04	0.98
A^{CAD}	-0.42	-0.13	0.43	1.67	0.90
A^{Inflow}	-0.37	-0.37	-0.37	2.46	0.94
$A^{Outflow}$	-0.37	-0.36	-0.36	2.79	0.90
A^{Media}	-0.47	-0.13	0.25	1.92	0.60
A^{Google}	-0.73	-0.22	0.31	1.74	0.93
A^{EDGAR}	-0.56	0.14	0.68	-0.43	0.82

Table 2 reports the 1st- and 3rd quartiles, median, skewness and first-order autocorrelation of the twelve attention proxies A^{AVol}, A^{ERet}, A^{PRet}, A^{52wH},A^{HisH}, $A^{\#AC}$, A^{CAD}, A^{Inflow}, $A^{Outflow}$, A^{Media}, $A^{Google}$and A^{EDGAR}. All variables are standardized to normal.

We restrict the data set to the time period 1980 to 2017 in order to compare the distribution of our data proxies vis-à-vis those constructed by Chen et al. (2022).We conclude that our data aligns closely to the measures presented by Chen et al. (2022) for the data obtained through CRSP, Capital IQ, IBES and Compustat. As expected, there are discrepancies in the distributions of A^{CAD}, A^{Media}, A^{Google}, and A^{EDGAR} where we are restricted by database access to reproduce identical data. In particular, we are cautious of the discrepancies in collecting Google Search- and Media Coverage data. The proxies could pick up different signals due to the value-weighted nature of our approach. Table A.2 reporting the summary statistics of our data during the time period January 1980 to December 2017 can be found in the Appendix.

II Empirical Approach

A Partial Least Squares

Methodologically, we begin by applying Partial Least Squares (PLS) as a benchmark to our nonlinear approach. Since introduced by Wold (1966), PLS has been refined and popularized as a method related to both principal component analysis (PCA) and ordinary least squares (OLS), but different from both. Instead of maximizing variation of the independent variables, they are maximized according to their explanatory power vis-a-vis the dependent variable. Its dimensionality reduction tackles the problem of autoregression and makes it particularly useful for time series analysis, and the method has found plenty of popularity in finance research (Chen et al., 2022; Kelly & Pruitt, 2015; Light et al., 2017).

In this subsection, we consider a model in which excess returns on the stock market is a linear function of the true but unobservable investor attention A^* plus some unpredictable noise-term ε unrelated to A^* as depicted in equation (1).

$$r_{t+1} = \alpha + \beta A_t^* + \varepsilon_{t+1} \tag{1}$$

where r_{t+1} is the realized excess stock return at time $t + 1$. Due to the unobservable nature of A^*, equation (1) describes the optimal but infeasible best forecast (Kelly & Pruitt, 2015). In our case, the true but unobservable investor attention A^* is a latent factor that drives the systematic variation of both our target variable, excess returns on the stock market, and our predictors, the twelve investor attention proxies. Intuitively, we are in need of a factor estimation step where we can identify the true investor attention driving excess returns from our proxies and remove all noise of the individual attention proxies unrelated to excess returns. To do so, we assume the data can be described by an approximate factor model. Specifically, we assume a linear factor structure for our investor attention proxies. Let A_t represent an $N \times 1$ vector of our attention proxies at time t such that $A_t = (A_{1,t}, ..., A_{N,t})^T$, where N is the number of attention proxies. The structural model for $A_{i,t} (i = 1, ..., N)$ is given by

$$A_{i,t} = \eta_{i,0} \ + \ \eta_{i,1} A_t^* \ + \eta_{i,2} E_t \ + \ e_{i,t} \tag{2}$$

where A^* is the true and unobservable investor attention in equation (1), $\eta_{i,1}$ is the factor loading that captures the sensitivity of attention proxy $A_{i,t}$ to A^*, E_t is the approximation-error common for all proxies that are unrelated to excess return, and $e_{i,t}$ is the noise idiosyncratic to attention proxy A_i .

In order to make predictions, we extract the subset of factors that influences our target variable, stock market excess returns, by implementing Partial Least Squares (PLS). Technically, the method we apply is a special case of the Three-Pass Regression Filter (3PRF) pioneered by Kelly and Pruitt (2015) and further developed by Light et al. (2017). It serves as a way to replicate a PLS forecast by running three passes through separate OLS regressions, given three assumptions: 1) the predictors are standardized in a preliminary step 2) the first two OLS regression passes are run excluding constants and 3) proxies are automatically selected (Kelly & Pruitt, 2015). We will refer to them both as PLS throughout this paper since they are used synonymously in adjacent research (Chen et al., 2022). The first pass is a time series regression of our attention proxies $A_{i,t}$ on realized excess return r_{t+1} as a proxy for future excess returns on the stock market.

$$A_{i,t} = \pi_0 + \pi_i r_{t+1} + u_{i,t}, \tag{3}$$

where π_i is the slope to be estimated and serves as each individual attention proxy's loading on the true target variable, excess returns. Consider equation (1), where r_{t+1} is driven by the true investor attention A^{*t}. Then follows that equation (3) describes how $A_{i,t}$ is related to A^{*t} instrumented by future excess returns r_{t+1}. The second pass is a cross-sectional regression of $A_{i,t}$ on the estimated loadings in equation (3) $\hat{\pi} i$ for each time period t

$$A_{i,t} = c_t + A_t^{PLS} \hat{\pi}_i + v_{i,t}, \tag{4}$$

where $A_t^{Attention}$ is the estimated measure of true investor attention at time t. The collective output of equation (4) is our investor attention index and predictor used to estimate equation (1) and forecast stock market excess returns. Note that if the true loadings π_i are unknown, and thus the slope $\hat{\pi}_i$ in equation (3) approximates the loadings π_i. Naturally, if the true relationship between our attention proxies $A_{i,t}$ and true investor attention A_t^* were known, we could consistently estimate our investor attention measure $A_t^{Attention}$ by running cross-sectional regressions of $A_{i,t}$ on π_i for each period. Figure 1 describes the time series index of market-level investor attention between January 1980 and October 2021. All numbers are standardized to have a mean of zero and standard deviation of zero. Investor attention was abnormally low during the 2008 financial crisis. Sicherman, Loewenstein, Seppi, and Utkus (2015) finds that attention falls by almost 10% after market declines. This phenomenon was attributed to the ostrich effect (Karlsson, Loewenstein,

& Seppi, 2009) coined after the metaphor of sticking one's head in the sand. Notably, this psychological bias is not detectable in the data pertaining to the COVID-19 crisis. Investor attention reaches its all-time high in March 2020 and continues on a high-level throughout the data period's end in October 2021.

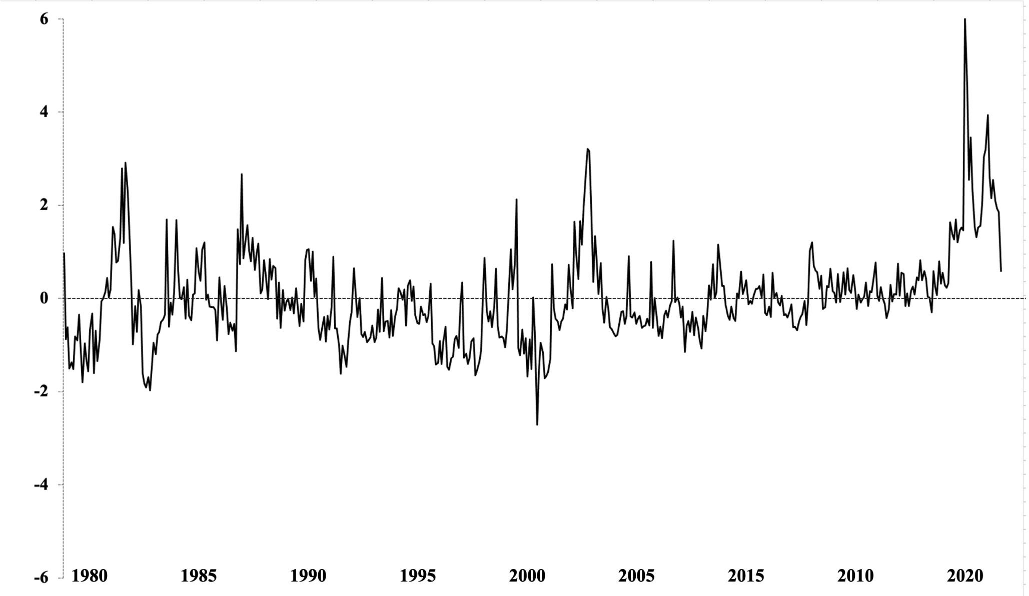

Figure 1 displays the market-level Investor Attention Index $A^{Attention}$ over the the time period Jan 1980 to Oct 2021. The index is standardized to normal.

In the empirical implementation described in Section I, we apply the third pass OLS to examine the forecasting power of the attention index $A^{Attention}$. Crucially, we complement the assessment of investor attentions' predictive power by testing our model out-of-sample.

B Long-Short Term Memory

The Long-Short Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) and differs from traditional neural networks by its ability to store an internal state generally considered as a memory (Hochreiter & Schmidhuber, 1997). As demonstrated by Zhang et al. (2021) this makes LSTM models well suited for handling complex non-linear relationships in time series data. Additionally, its application of a rolling window is a technique appropriate for mitigating concept drift (Nagel, 2021). The model is restricted at each month t to only consider data w months back, where w is the size of the time window.

The following description of our LSTM method assumes the reader possesses prior knowledge of the basic mechanisms of a traditional neural network. Appendix II provides a more detailed description of Neural Networks and LSTM. For a thorough guide on Neural Networks and Deep Learning, see Goodfellow, Bengio, and Courville (2016) and for a well-written book on machine learning applications in finance, consult Nagel (2021).

The units, often referred to as 'neurons' in the machine learning literature, in an LSTM hidden layer are linked together by the network's feedback connections. At each time step the LSTM neuron passes information downstream in the network for a prediction at time t. However, the same information is also passed sideways to the next neuron in line to be considered in the prediction at time $t + 1$.

Each LSTM neuron consists of four essential parts: A cell state C_{t-1},passed on from previous LSTM units, acts as the network's long term memory. Second, a forget gate ft controls what should be removed from the cell state. Third, an input gate determining what should be added to the cell state Ct. Finally, the output gate ht executes the LSTM's output at time t that serves two purposes: as the network's prediction at time t and its short-term memory at time t+1. Below follows a condensed description of the processes inside an LSTM unit. The first process takes place in the forget gate. The process defining what part of the cell state C_{t-1} to forget is given by

$$f_t = \sigma\ (W_f\ [x_t,\ h_{t-1}] + b_f), \tag{5}$$

where σ is a sigmoid function with output between 0 and 1; W_f a matrix of parameters commonly referred to as weights; x_t the vector of inputs at time t; h_{t-1} the short term memory and prediction in the previous time step; and b_f the noise-term of the forget gate f_t. The second process determines what new information to combine with the remainder of old cell state C_{t-1} and subsequently form the updated cell state C_t. The process of updating is given by

$$i_t = \sigma\ (W_i\ [\ x_t, h_{t-1}] + b_i), \tag{6}$$

$$\tilde{C}_t = tanh\ (W_C\ [\ x_t, h_{t-1}] + b_C), \tag{7}$$

$$C_t = C_t \cdot i_t + C_{t-1} \cdot f_t, \tag{8}$$

where σ is a sigmoid function; tanh a hyperbolic tangent function with output between -1 and 1; W_i and W_C are weight matrices; and b_i and b_C the noise-terms of i_t and $\tilde{C}_t$ respectively. The third and final process constructs the prediction at time t by means of two subsequent calculations considering both the new information x_t combined with the short term memory ht-1 and the updated cell state C_t. The output process is given by

$$o_t = (W_o[x_t, h_{t-1}] + b_o), \tag{9}$$

$$h_t = o_t \cdot tanh(C_t), \tag{10}$$

where o_t is the part of the output at time t contributed to by x_t and $h_t - 1$; W_o a weight matrix; and b_o the noise-term of o_t ; h_t the prediction at time t; and C_t the cell state formed in equation (9).

Next, preparatory choices are considered in order to optimize the model. LSTMs and other neural networks have the capacity to approximate any underlying model and thus make predictions with arbitrary accuracy in-sample (Cybenko, 1989). This is by virtue of their nonlinear nature and vast number of trainable parameters (Cybenko, 1989). Thus, a main challenge while applying an LSTM is how to avoid overfitting the sample data and optimize for out-of-sample accuracy (Gu et al., 2020).

First, the tuning of hyperparameters controls a model's complexity and works as machine learning's major antidote to overfitting. There is no universally applicable method of how to best tune hyperparameters (Gu et al., 2020; Nagel, 2021). We follow prominent research by Gu et al. (2020) and dedicate a set of our data as a validation sample in order to choose hyperparameters adaptively. In chronological order, we split our data into a training, validation and test set. The first data set is dedicated to training the model. The second part serves as a benchmark in a simulated out-of-sample test in the current state. The validation data is thus being restricted from the model during training, and used to evaluate prediction accuracy before the model adjusts its weights again. The test data is left untouched for true out-of-sample prediction evaluation. To simulate real-time out-of-sample testing while predicting, it is crucial that the network is set up so that it is not informed at time t by information occurring in time $t+1$. Thus, we normalize the three data sets separately according to its own mean and standard-deviation.

We apply Mean Squared Errors (MSE), also used by Gu et al. (2020) to evaluate the model's accuracy on validation data, and Adam optimization algorithm to minimize the loss, also applied by Yadav, Jha, and Sharan (2020) and Zhang et al. (2021). To avoid overfitting, we add two regularization terms to the loss function called elastic. Elastic net combines lasso regularization (L1) and ridge regularization (L2). L1 regularization serves to penalize the number of weights used in the model, incentivizing the model to rid non-essential parameters. The L2 regularization term penalizes the total absolute values of the weights in the network which prompts to shrink the size of any individual parameter. In effect, the network is incentivized by the elastic net regularization to find the best balance between minimizing in-sample errors, the number of parameters and shrinking the size of parameters in use. The extended loss function is given by

$$L_enet(\hat{\beta}) = \; MSE(\hat{\beta}) \; + \; \lambda \left(\frac{1-\alpha}{2} \, \Sigma \hat{\beta}_i^2 \; + \; \alpha \Sigma |\hat{\beta}_i| \right) \tag{11}$$

where $\hat{\beta}$ are the trained parameters , MSE the Mean Squared Error-term, λ the learning rate and α the parameter balancing the presence of added penalty terms. Remaining preparatory steps include selecting the number of epochs, the time window size and the network's depth and width. The number of epochs refers to the number of passes of training and validating the model taken before choosing the best performing model. Thus, an increasing number of epochs allows more parameter updates. The depth and width of a network describes how many parameters each layer contains and secondly how many layers of neurons the network should include. Here again, no universal best-practices apply (Gu et al., 2020; Nagel, 2021) and we exercise recursive testing to obtain the best prediction accuracy.

Given the aforementioned opaque nature and strong ability of LSTM networks to model complexity, in-sample predictions are rarely of interest while applying machine learning models of akin character. Thus, our focus while applying LSTM is solely to compare its out-of-sample accuracy to our PLS model. Next, we conduct empirical testing and benchmark the prediction accuracy of our linear and nonlinear models.

III Empirical Results and Analysis

In this section we forecast stock market returns using our PLS. Sequentially we compare the methods in order to bring light to our hypothesis of nonlinearity in the relationship between investor attention and excess market returns.

A Partial Least Squares

To examine the in-sample predictive power of investor attention on excess market returns we apply the third pass regression of PLS, which is a univariate predictive regression given by

$$R_{t+h} = \alpha + \beta A_t^{Attention} + \varepsilon_{t+h}, \tag{12}$$

where R_{t+h} is the average excess stock return over the forecast horizon h, where h = 1, 3, 6, 12 and 24 months; $A_t^{Attention}$ is the estimated slope in equation (4) and measure of investor attention at time t; β the slope to be estimated; and ε_{t+h} the noise-term unrelated to $A_t^{Attention}$.

For in-sample predictions, we use the full data set spanning from January 1980 to October 2021 for attention measures A^{AVol}, A^{ERet}, A^{PRet}, A^{52wH} and A^{HisH}; from January 2004 to October 2021 for A^{Inflow}, $A^{Outflow}$, A^{Media} and A^{Google}; and January 2004 to June 2017 for A^{EDGAR} to approximate the loadings $\hat{\pi}_i$ in equation (3) and sequentially estimate $A_t^{Attention}$ in equation (4). We forecast excess returns R_{t+h} in the predictive regression displayed in equation (12) over the full time period. To evaluate the in-sample prediction we use the commonly applied R^2-metric, also used by Chen et al. (2022).

We find that the market-level investor attention has in-sample predictive power over stock returns. Our results are significant for all forecasting horizons. In predicting next month's return between January 1980 and October 2021 we find an in-sample R^2 of 3.21% and the explanatory power increases until the yearly forecast horizon where R^2 peaks at 12.23%, implying that $A^{Attention}$ can explain 12.23% of the time variation in the yearly excess return of the stock market. Furthermore, the betas are significant for all prediction horizons. The largest coefficient slope is obtained for monthly forecasts and declines as the prediction horizon increases. The coefficient under monthly predictions is 0.90%. The corresponding for 24 months is 0.28%. Since all our data is standardized, this means that 1 standard-deviation increase in investor attention at time t predicts a 0.90% increase in excess returns in the consecutive month. Annualized, this equates to 10.8%, which is in the higher ranges of common macro economic predictors. An equal increase in the dividend–price ratio, and the net payout ratio approximately increases the risk premium by 3.60% and 10.2% per year, respectively (Jacob Boudoukh & Roberts, 2007; Lettau & Ludvigson, 2001).

Our in-sample prediction performs significantly better than Chen et al. (2022) who's R^2 results are lower for all forecast horizons. However, the time periods over which we conduct in-sample predictions differ, and thus we proceed by restricting the data set to the same time period, January 1980 to December 2017. Notably, our in-sample R^2 increases to 3.31% at the monthly horizon and reaches 14.5% at the yearly horizon. All coefficient slopes are significant and follow the same pattern - positive slopes decreasing in the prediction horizon. Table 3 summarizes the in-sample forecast R^2 and betas over both time periods.

Table 3: In-Sample Results

Time Period	Statistic	h = 1	h = 3	h = 6	h = 12	h = 24
Jan 1980 - Oct 2021	In-sample $R^2(\%)$	3.24	3.69	6.96	12.23	5.82
	β (%)	0.90***	0.58***	0.57***	0.55***	0.28***
Jan 1980 - Dec 2017	In-sample $R^2(\%)$	3.21	5.96	10.69	14.55	9.52
	β (%)	0.88***	0.73***	0.71***	0.56***	0.30***

Table 3 reports the in-sample results and β over the two time periods Jan 1980 - Oct 2021 and Jan 1980 - 2017. The results follow from the predictive regression in equation (12) where $R_t + h$ is regressed on $A^{Attention}$. *** indicate a p-value below 0.001.

In-sample prediction allows for usage of all available data and thus offers more accurate forecasting. However, prominent financial research such as Campbell and Thompson (2007) and Welch and Goyal (2008) widely argues that out-of-sample predictions more authentically reflect real-time predictive power. Accordingly, we proceed by applying PLS on out-of-sample forecasting.

In forecasting out-of-sample, we are restricted to use only the data available up until time t to make a prediction of stock returns at time $t + 1$. Sequentially, the latest time period used in the first pass given by equation (3) is r_t, and as such the latest observations of our attention proxies A_i are at time $t - 1$. In the second pass $A_t^{Attention}$ is estimated using data from month 1 through t. We apply the latest available loadings π_i at every month t in the second pass of estimating the model. Using all the sample data we estimate the initial predictive regression in equation (12). Thus the out-of-sample forecasting regression is given by equation (13) where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated coefficients of equation (12). As we predict the stock return later in time we repeat the passes and re-estimate equation (13) for each month using the latest available data. For instance, to forecast excess returns at time $t + 3$ we use data up until $t + 2$ to estimate our predictive regression.

We follow Chen et al. (2022) and devote 40% of the data to fit the initial predictive regression[3]. However, we add two months to have the sample data end at the end of a calendar year. Our sample data thus spans from January 1980 to December 1996 and we forecast monthly out-of-sample stock market returns between January 1997 and October

[3] Chen et al. (2022) uses 180 out of 456 months as their initial sample which is 39.5%, circa 40% of the data rounded to end at the calendar year.

2021. To measure the accuracy of our out-of-sample predictions we apply the commonly used R^2_{OOS} metric proposed by Campbell and Thompson (2007), also used by Chen et al. (2022). The R^2_{OOS} metric benchmarks the prediction against the historical average from the first month in the sample until time t. Logically, the model's prediction is deemed to be insignificant if the performance is worse than the historical average.

We find that our predictions are still significant for all forecasting horizons but 24 months, in line with the findings of Chen et al. (2022). We obtain an R^2_{OOS} of 1.09% on the monthly horizon and 1.76% for predictions 3 months ahead.

We continue by imposing theoretically motivated restrictions and adjustments to the predictive model. First, we force predictions of excess returns - the market premium - to be non-negative in line with economic theory suggested by Campbell and Thompson (2007). We find that non-negative restriction improves prediction accuracy to 1.74 % at the monthly horizon, and peaking at 2.29% for quarterly forecasts. Second, Light et al. (2017) suggests that the loadings π_i can be substituted by the average loadings π_i of the previous periods to estimate the coefficient in equation (4) more precisely. This assumes that time does not influence the relationship between investor attention and excess returns on the stock market (Chen et al., 2022; Light et al., 2017). We use both the historical average of the loadings π_i at each month t as well as a 5 year averaging scheme. In contrast to Chen et al. (2022) we observe that applying the most recently estimated loadings π_i yields the most accurate predictions. This suggests that the relationship between investor attention and stock returns changes over time (Light et al., 2017) and hints at the potential of LSTM to improve prediction accuracy.

We observe that Chen et al. (2022) obtains a higher R^2_{OOS}, and thus we proceed by examining what differs between our tests. First, we change the data time period to January 1980 to December 2017, and split our in-sample and out-of-sample data according to Chen et al. (2022). In shortening the time period, we find a higher R^2_{OOS} for all forecast horizons culminating during monthly predictions at 3.37%. Notably, we find a significant R^2_{OOS} for the 24-month forecast of 0.55%, not found by earlier research. Still, the results during shorter prediction horizons are still not quite in line with Chen et al. (2022).

Second, we impose a non-negative return premium, as reported by Chen et al. (2022). Interestingly, this deteriorates the prediction performance, albeit still yielding significant results. Chen et al. (2022) applies different averaging schemes of loadings π_i including 5 year average, 10 year average and the average of all available loadings at time t. They do not, however, report which averaging scheme yields their presented R^2_{OOS}. We implement all their reported averaging schemes, but do not observe an improved R^2_{OOS}. All else equal this should yield equal results. Thus, we have replicated all implementations of out-of-sample forecasting reported by Chen et al. (2022). We conclude that the one source left to explain the discrepancy lies in our data differences for the variables A^{CAD}, A^{Media}, A^{Google} and A^{EDGAR}. Table 4 summarizes the out-of-sample predictions using the best performing models for each respective time period.

Table 4: Out-of-sample Results

R^2_{OOS} (%) for Data Period: 1980 - 2021		
Forecast Horizon	Unrestricted Forecast	Restricted for Non-negative Prediction
h = 1	1.09	1.74
h = 3	1.76	2.29
h = 6	1.48	2.24
h = 12	0.01	1.60

R^2_{OOS} (%) for Data Period: 1980 - 2017		
Forecast Horizon	Unrestricted Forecast	Restricted for Non-negative Prediction
h = 1	3.37	1.81
h = 3	2.10	1.11
h = 6	2.67	1.82
h = 12	2.37	1.80

Table 4 shows the out-of-sample results and β for the two time periods Jan 1980 - Oct 2021 and Jan 1980 - 2017 over the prediction horizon h, where $h = 1, 3, 6$ and 12 months. In the Restricted for Non-Negative Prediction, we impose the economic restriction of a non-negative market premium, substituting negative forecasts by 0. No economic restrictions are imposed on the Unrestricted Forecast. *** indicate a p-value ¡ 0.001.

A notable aspect of our results is the positive sign of the prediction coefficient, which is different from the negative coefficient identified by Chen et al. (2022) with a similar array of proxies. However, the literature suggests that the sign of the slope is not definite (Andrei & Hasler, 2020; Campbell & Thompson, 2007). The positive predictability of investor attention is in line with the theory presented by George and Hwang (2001) and Li and Yu (2012) present evidence that attention, captured by nearness to the 52-week high, positively predicts excess market returns. Our findings support the economic phenomenon suggested by Gervais et al. (2001) who show that high attention, proxied by trading volume, leads to future increases in stock returns driven by shocks in trader interest from its increased visibility. However, we conclude that minor data differences yielded an impact on the direction of the slope contra Chen et al. (2022), and thus we echo the inconclusiveness regarding the coefficient previously addressed in the literature (Andrei & Hasler, 2020; Campbell & Thompson, 2007).

We observe that the effect is strongest in the short term (one month's horizon) and declines over time, in line with economically expected phenomena such as reversal from temporary price pressure (Barber & Odean, 2008; Da et al., 2011) harmonious with the theory of mean reversion to a stocks' fundamentals, first presented by De Bondt and Thaler (1985).

B Long-Short Term Memory

We begin our application of LSTM by testing the out-of-sample prediction accuracy, which is then compared with the results of the PLS analysis.

In order to find the best model setup for out-of-sample predictions, we make repeated tests where a single variable is changed while the rest are held constant. We cycle through: learning rates of 0.01 and 0.001; the number of epochs set to 100, and 200; window step size of 6, and 12; and the elastic net penalization constant α to 0.1. Importantly, we reset the layer weights after each test. A summary of the model setups examined is reported in Table 5.

Table 5: LSTM Model Setup

LSTM Architecture	*Hyperparameters and Regularization*	*Data Split*
Input Layer (11 and 12 inputs)	Learning Rates: 0.01 and 0.001	Training 80%
LSTM Layer (8 Neurons)	Dropout: 0.2 and 0.4	Validation 5%
Dense Layer (4 Neurons, ReLU)	Window Step Size: 6 and 12	Test 15%
Output layer (1 Neuron, Linear)	Alpha: 0.1 and 0.01	

Table 5 summarizes the LSTM model setup used in our tests. The left-most column describes the network architecture. The middle column describes hyperparameter and regularizer choices. The right-most column describes the share of data dedicated to training, validation and test data.

For comparison with PLS, we predict excess returns R_{t+h} over the same prediction horizons where h = 1, 3, 6, 12 and 24 months. LSTM requires that all predictors, the twelve attention proxies in our case, have the same time period. We restrict the dataset from January 2004 to June 2017 and call this $Data_{04-17}$. We also run the model omitting the predictor A^{EDGAR} (which only has data until 2017) in order to extend all other predictors to October 2021 and refer to it as $Data_{04-21}$. We attribute the first 80% of the months in each data set to training, 5% for the validation and use the remaining 15% as test data. Running each respective model described in Table 5 we find that the best performing model applies a learning rate of 0.01, 100 epochs and a window step size set to 6 months and yields an R^2_{OOS} of 2.23% for $Data_{04-21}$ which is on par with what we obtain using PLS (2.24%). The data set in the LSTM is notably shorter however, considering all proxies begin in 2004.

We find that our LSTM model fails to outperform the PLS model in out-of-sample predictions. While the is able to impressively fit in-sample data, we obtain negative R^2_{OOS} scores for all horizons except for the half-year forecast using $Data_{04-21}$. Adjusting hyperparameters to mitigate the bias-variance trade-off (Kozak et al., 2020) improves the forecast but yields no significant results.

As neural networks is far from closed-form mathematics and more often likened to a black box, identifying shortcomings is similarly more about trial-and-error debugging than the transparent inference commonly possible in traditional statistics (Chollet, 2021). That is why we turn towards finding and rectifying any eventual shortcomings in our model with alternative means below.

In order to draw conclusions on the absence of nonlinearity in our data, we intend to set up a test where we can expect our LSTM to perform well out-of-sample. If our LSTM accurately predicts a nonlinear relationship we know exists, we can discard that the failure to find predictive power on excess returns in our investor attention data is due to a computing problem. First, we measure the first-order autocorrelation of our excess return data and find that it increases drastically as the prediction horizon expands. The autocorrelation for average 24-month excess returns reaches 0.97 .

We add a variable containing past returns up to time t to the input data that the LSTM can use to predict returns at time $t + 1$. Given the high degree of autocorrelation in the 24 month average excess return, we would expect to observe a substantial increase in prediction performance, unless there is an issue in the model or data setup. We run the LSTM prediction horizon and we instantly find a considerable spike in R^2_{OOS}. Predicting average excess returns over the coming 24 months yields an R^2_{OOS} of 40% without tuning any hyperparameter. This supports our ability to make conclusions about our data.

LSTMs primary advantage over PLS is that it allows for nonlinear relationships and interactions between the input variables and the predicted target. Hence, our null hypothesis that there is insufficient nonlinearity in the data to improve its predictive performance cannot be rejected. Our tests conclude that the existence of nonlinear interplay between the twelve attention proxies and market returns is not great enough to enhance prediction performance, at least within the data set used in this paper. However, we are careful about generalizing our conclusion to an external setting.

IV Conclusion

Behavioral patterns and cognitive constraints have long been known to impact investors decision-making on public markets, yet its impact on market returns has been sparsely examined. This paper makes several contributions to this relatively thin body of research. Chief of our findings is that the relationship between investor attention and market returns is strong enough for economically meaningful gains to prediction. While the result has an intuitive theoretical underpinning in the limitations of the human mind, no consensus has been reached on whether the correlation is or ought to be positive or negatively correlated. Our findings show that when attention is high, subsequent market returns are predicted to increase in the following months.

Uniquely, we examine whether the predictive abilities of investor attention are enhanced through a Long-Short Term Memory, a machine learning model capable of predicting complex nonlinear relationships. We find that closed-form dimensionality reduction remains triumphant, and conclude that the existence of nonlinear interplay between the twelve attention proxies and market returns is not great enough to enhance prediction performance, at least within the data considered in this paper.

Future research would do well to compare this predictive performance with a series of cross-sectional and macro predictors to evaluate whether investor attention would improve all-encompassing predictive models. With professional quantitative finance steering towards character rich modeling, investor attention shows the potential to improve prediction models similar to Kozak et al. (2020). Additionally, researchers could further widen the understanding of investor attention by allowing for larger datasets and the testing a broader series of machine learning methods. This could entail higher statistical power in testing.

References

Aboody, D., Lehavy, R., & Trueman, B. (2010). Limited attention and the earnings announcement returns of past stock market winners. *Review of Accounting Studies*, *15*, 317–344.

Andrei, D., & Hasler, M. (2020). Dynamic attention behavior under return predictability. *Management Science*, *66*(7), 2906–2928.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*(2), 129–152.

Banerjee, S., & Green, B. (2015). Signal or noise? uncertainty and learning about whether other traders are informed. *Journal of Financial Economics*, *117*(2), 398–423.

Barber, B. M., & Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies*, *21*, 785–818.

Ben-Rephael, A., Da, Z., & Israelsen, R. D. (2017). It depends on where you search: Institutional investor attention and underreaction to news. *The Review of Financial Studies*, *30*(9), 3009–3047.

Campbell, J. Y., & Thompson, S. B. (2007). Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *The Review of Financial Studies*, *21*(4), 1509–1531.

Cao, H., Lin, T., Li, Y., & Zhang, H. (2019). Stock price pattern prediction based on complex network and machine learning. *Complexity*, *2019*, 1–12.

Chen, J., Tang, G., Yao, J., & Zhou, G. (2022). Investor attention and stock returns. *Journal of Financial and Quantitative Analysis*, *57*(2), 455–484.

Chollet, F. (2021). *Deep Learning with Python* (1st edition). Manning Publications.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *Journal of Finance*, *66*, 1461–1499.

De Bondt, W. F. M., & Thaler, R. (1985). Does the stock market overreact? *Journal of Finance*, *40*, 793–805.

Dellavigna, S., & Pollet, J. M. (2009). Investor inattention and friday earnings announcements. *The Journal of Finance*, *64*, 709–749.

Drake, M., Roulstone, D., & Thornock, J. (2015). The determinants and consequences of information acquisition via edgar. *Contemporary Accounting Research*, *32*, 1128–1161.

Fang, L., & Peress, J. (2009). Media coverage and the cross-section of stock returns. *The Journal of Finance*, *64*(5), 2023–2052.

George, T. J., & Hwang, C.-Y. (2001). Information flow and pricing errors: A unified approach to estimation and testing. *Review of Financial Studies*, *14*, 979–1020.

Gervais, S., Kaniel, R., & Mingelgrin, D. H. (2001). The high-volume return premium. *The Journal of Finance*, *56*(3), 877–919.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Hirshleifer, D., & Hong Teoh, S. (2003). Herd behaviour and cascading in capital markets: A review and synthesis. *European Financial Management*, *9*(1), 25–66.

Hirshleifer, D., Hsu, P.-H., & Li, D. (2013). Innovative efficiency and stock returns. *Journal of Financial Economics*, *107*(3), 632–654.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–80.

Hsieh, D. A. (1991). Chaos and nonlinear dynamics: Application to financial markets. *Journal of Finance*, *46*, 1839–77.

Huang, D., Jiang, F., Tu, J., & Zhou, G. (2015). Investor Sentiment Aligned: A Powerful Predictor of Stock Returns. *The Review of Financial Studies*, *28*(3), 791–837.

Jacob Boudoukh, M. R., Roni Michaely, & Roberts, M. R. (2007). On the importance of measuring payout yield: Implications for empirical asset pricing. *The Journal of Finance*, *62*, 877–915.

Jondeau, E., Zhang, Q., & Zhu, X. (2019). Average skewness matters. *Journal of Financial Economics*, *134*(1), 29–47.

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall Inc.

Karlsson, N., Loewenstein, G., & Seppi, D. (2009). The ostrich effect: Selective attention to information. *Journal of Risk and Uncertainty*, *38*, 95–115.

Kelly, B., & Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors [High Dimensional Problems in Econometrics]. *Journal of Econometrics*, *186*(2), 294–316.

Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, *135*(2), 271–292.

Lee, C. M., Ma, P., & Wang, C. C. (2015). Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics*, *116*(2), 410–431.

Lettau, M., & Ludvigson, S. (2001). *Measuring and modelling variation in the risk-return trade-off* (Vol. 3105). Centre for Economic Policy Research.

Li, J., & Yu, J. (2012). Investor attention, psychological anchors, and stock return predictability [Special Issue on Investor Sentiment]. *Journal of Financial Economics*, *104*(2), 401–419.

Light, N., Maslov, D., & Rytchkov, O. (2017). Aggregation of Information About the Cross Section of Stock Returns: A Latent Variable Approach. *The Review of Financial Studies*, *30*(4), 1339–1381.

Lou, D. (2014). Attracting investor attention through advertising. *The Review of Financial Studies*, *27*(6), 1797–1829.

Ma, Q., Wang, H., & Zhang, W. (2017). Trading against anchoring. *Review of Behavioral Finance*, *9*.

Nagel, S. (2021). *Machine learning in asset pricing* (Vol. 8). Princeton University Press.

Peng, L. (2005). Learning with information capacity constraints. *The Journal of Financial and Quantitative Analysis*, *40*(2), 307–329.

Peng, L., & Xiong, W. (2006). Investor attention, overconfidence and category learning. *Journal of Financial Economics*, *80*, 563–602.

Rapach, D. E., Ringgenberg, M., & Zhou, G. (2016). Short interest and aggregate stock returns. *Journal of Financial Economics*, *121*(1), 46–65.

Roondiwala, M., Patel, H., & Varma, S. (2017). Predicting stock prices using lstm. *International Journal of Science and Research (IJSR)*, *6*.

Ryans, J. (2017). Using the edgar log file data set. *Research Methods & Methodology in Accounting eJournal*.

Sicherman, N., Loewenstein, G., Seppi, D. J., & Utkus, S. P. (2015). Financial Attention. *The Review of Financial Studies*, *29*(4), 863–897.

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*, 1455–1508.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, 391–420.

Yadav, A., Jha, C. K., & Sharan, A. (2020). Optimizing lstm for time series prediction in indian stock market [International Conference on Computational Intelligence and Data Science]. *Procedia Computer Science*, *167*, 2091–2100.

Yuan, Y. (2015). Market-wide attention, trading, and stock returns. *Journal of Financial Economics*, *116*(3), 548–564.

Zhang, Y., Chu, G., & Shen, D. (2021). The role of investor attention in predicting stock prices: The long short-term memory networks perspective. *Finance Research Letters*, *38*(100).

Appendix I

Table A.1: Correlation Plot

	A^{AVol}	A^{ERet}	A^{PRet}	A^{52wH}	A^{HisH}	$A^{\#AC}$	A^{CAD}	A^{Inflow}	$A^{Outflow}$	A^{Media}	A^{Google}	A^{EDGAR}
A^{AVol}	1	0.01	0.19	0.11	0.14	-0.14	0.07	0.31	0.27	0.13	0.33	-0.04
A^{ERet}		1	0.06	0.17	0.15	-0.01	0.15	0.09	0.06	0.05	0.04	0.03
A^{PRet}			1	0.7	0.29	0.12	0.61	-0.36	-0.4	0.24	-0.24	0
A^{52wH}				1	0.77	0.39	0.66	-0.41	-0.43	0.39	-0.13	-0.05
A^{HisH}					1	0.43	0.54	-0.26	-0.25	0.39	0.15	-0.12
$A^{\#AC}$						1	0.24	-0.63	-0.52	0.46	-0.43	0.46
A^{CAD}							1	-0.39	-0.43	0.28	0	-0.1
A^{Inflow}								1	0.96	-0.34	0.39	-0.11
$A^{Outflow}$									1	-0.33	0.34	-0.03
A^{Media}										1	-0.12	0.12
A^{Google}											1	-0.35
A^{EDGAR}												1

Table A.1 shows the cross-correlation between the 12 investor attention proxies A^{AVol}, A^{ERet}, A^{PRet}, A^{52wH},A^{HisH}, $A^{\#AC}$, A^{CAD}, A^{Inflow}, $A^{Outflow}$, A^{Media}, $A^{Google}$and A^{EDGAR} .

Table A.2: Summary Statistics Jan 1980 - Dec 2017

Name	1st Quartile	Median	3rd Quartile	Skewness	$p(1)$
A^{AVol}	-0.72	-0.10	0.52	0.98	0.48
A^{ERet}	-0.49	0.10	0.60	-0.63	0.18
A^{PRet}	-0.64	0.08	0.53	0.02	0.93
A^{52wH}	-0.18	0.41	0.64	-2.40	0.88
A^{HisH}	-0.41	0.39	0.80	-1.63	0.94
$A^{\#AC}$	-0.84	0.03	0.76	-0.11	0.99
A^{CAD}	-0.45	-0.05	0.61	0.88	0.93
A^{Inflow}	-0.64	-0.32	-0.31	2.08	0.80
$A^{Outflow}$	-0.60	-0.35	-0.27	2.31	0.77
A^{Media}	-0.44	-0.12	0.22	1.71	0.61
A^{Google}	-0.63	-0.20	0.33	0.91	0.90
A^{EDGAR}	-0.56	0.14	0.68	-0.43	0.82

Table A.2 reports the 1st- and 3rd quartiles, median, skewness and first-order autocorrelation of the twelve attention proxies A^{AVol}, A^{ERet}, A^{PRet}, A^{52wH},A^{HisH}, $A^{\#AC}$, A^{CAD}, A^{Inflow}, $A^{Outflow}$, A^{Media},
A^{Google} and A^{EDGAR} for the time period Jan 1980 to June 2017 (when all measures are complete). All variables are standardized to normal. This time frame is the same as the one studied by Chen et al. (2022).

Appendix II: A Brief Introduction to Neural Networks

Imagine a neural network containing one input layer, one hidden layer with two neurons and an output layer. From the input layer, input values are sent to each neuron inside the hidden layer. Inside each neuron, weights are assigned to the inputs and a bias is added.

$$x_1 \rightarrow (x_1 \cdot w_1) + x_2 \rightarrow (x_2 \cdot w2) + bias \tag{1}$$

At this point, what we have is equal to a multivariate linear regression:

$$(x_1 \cdot w_1) + (x_2 \cdot w_2) + b \tag{2}$$

Before the value is fed to the output layer however, the hidden layer applies a nonlinear activation function $(\cdot)$ to the sum of (1):

$$y = (x_1 w_1 + x_2 w_2) + b \tag{3}$$

Normally, the activation function "squashes" the sum of (1) so that its output takes a value where y is

$$1 \leq y \leq 1, 0 \leq y \leq 1 \ or \ 0 \leq y \tag{4}$$

Lastly, each neuron in the hidden layer feeds its output to the output layer where the same process is repeated in an output neuron, though not necessarily using the same activation function. Some of the most common nonlinear activation functions ReLu, Tanh, Sigmoid and Softplus. Neural Networks are able to model high-dimensional relationships between variables and can become highly complex as more hidden layers are added to the network, thus they are called Deep Neural Networks (DNNs) when more hidden layers are added.

Albeit being highly effective for many nonlinear prediction situations, the neural network we have studied above suffers when the prediction problem is of a time series nature. In a time series, the prediction of yt+1 at timet is potentially dependent on earlier information from time t , $t-1$ and $t-2$ and so on. As the timesteps relevant to predict y_{t+1} increases, the neural network, more specifically called a feed-forward neural network, runs into problems since it has no notion of order in time. The information inside a feed-forward neural network can not run in "cycles" and thus it has no "memory" of earlier timesteps. Recurrent Neural Networks (RNNs) mitigate this problem by containing loops that allow the information to cycle through the network. The RNN can be thought of as several copies of the same neural network for each time step that each can feed information to the network next to it such that NN_{t-1} feeds to NN_t that feeds to NN_{t+1} and so on. In effect, information can be stored across time steps as memory for predictions. Thus, RNNs are especially effective when combating sequential data such as natural language processing, speech recognition and time series prediction.

A traditional RNN often becomes futile when one is dealing with predictions dependent on information many time steps apart. Hochreiter and Schmidhuber (1997) introduced the Long-Short Term Memory Network (LSTM), a type of RNN that is able to store both long- and short term memory.