UNIVERSITÄT
BERN

Three Essays in Financial Economics

A Ph.D. thesis submitted to the University of Bern

at the faculty of Business, Economics and Social Sciences,

supervised by Prof. Dr. Philip Valta,

in partial fulfilment of the requirements for the degree of

Doctor Rerum Oeconomicarum

by

Jan Pichler

July, 2023

The faculty accepted this thesis on November 2, 2023 at the request of the reviewers Prof. Dr. Philip Valta and Prof. Dr. Michael Rockinger as dissertation, without wishing to comment on the views expressed therein.

Never confuse education with intelligence, you can have a PhD and still be an idiot.

— Richard P. Feynman

Acknowledgments

First, I would like to acknowledge and give my warmest thanks to my supervisor Prof. Dr. Philip Valta. He offered me the opportunity to write this thesis, gave me the academic freedom to choose my projects independently, and never exerted any pressure despite my doubts and the difficulties I faced. I would also like to thank my secondary advisor Prof. Dr. Julien Cujean, for the many discussions on both research and other more trivial matters.

My gratitude also goes to my independent examiner Prof. Dr. Michael Rockinger, for taking the time to evaluate this thesis and especially for giving the course *Empirical Asset Pricing* at the Swiss Finance Institute (SFI) in Lausanne, which inspired me for the first essay of this thesis.

Special thanks go to my working colleagues and fellow Ph.D. students at the Institute for Financial Management at the University of Bern - Sascha, Marc, Christian, Samuel, and Karin - for their helpful and constructive thoughts, witty banters, and moral support.

I also feel grateful to my Alma Mater, the University of Bern, for the terrific twelve years I could spend here: It was truly a privilege to be taught, challenged, and promoted here. While I learned a lot, both in general and field-specific, the biggest gift of all was the inspiring people I met here who enriched my life and continue to do so.

Furthermore, I would like to express my deepest gratitude to my family, friends, and colleagues for not bearing any grudges because I did not spend as much time with them as they deserved and for their forgiveness when my mind was distracted.

Finally, and from the bottom of my heart: Thank you, Dominique, Lara, Mum, and Dad, for what words cannot describe...

Jan Pichler

Preface

In 2013, the Royal Swedish Academy of Sciences decided to award The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel to Eugene F. Fama, Lars Peter Hansen, and Robert J. Shiller *"for their empirical analysis of asset prices,"* and summarized their findings as follows:

> *There is no way to predict the price of stocks and bonds over the next few days or weeks. But it is quite possible to foresee the broad course of these prices over longer periods, such as the next three to five years. These findings, which might seem both surprising and contradictory, were made and analyzed by this year's Laureates, Eugene Fama, Lars Peter Hansen, and Robert Shiller.*

The first two sentences are astonishingly close to describing the content of this thesis, which consists of three single-authored essays. Below I will elaborate more closely on them, but one conclusion of the first essay is that new information is anticipated by financial markets, quickly priced in, and does not predict returns in the subsequent days and weeks. The second essay investigates this at a much lower frequency (up to seconds), showing that pricing in new information happens almost immediately and that patterns around news publication are consistent with rational pricing. The third essay is in line with the second sentence by providing evidence that the level of profitability can predict returns over at least the subsequent twelve months.

In what follows, I briefly summarize the three essays' contents and contextualize their aggregate findings within the context of this thesis to provide a collective conclusion.

Essay 1: Is the Stock Market's Reaction to News Predictable?

The first essay investigates return patterns around news events by analyzing the largest news dataset studied in finance so far: 4.4 million news headlines between January 1996 and December 2019 on firms listed in North America. I use a finance-specific sentiment dictionary to classify these news headlines into positive and negative ones and contrast the results of this approach to classifications based on supervised learning models trained on the market reaction to the news. These supervised learning models include the multinomial Naïve Bayes method and several rudimentary neural networks.

This paper contributes to the literature by showing that supervised learning models outperform the sentiment dictionary approach traditionally used in finance. Furthermore, it provides evidence that financial markets anticipate news in the weeks ahead, that the new information is quickly priced in, and that there is no drift afterward.

Essay 2: Microanalysis of the Stock Market's Reaction to News

The second essay is a follow-up on the first: It merges the same news dataset with the data on the trades at the New York Stock Exchange (NYSE) between January 2011 and December 2019. This seemingly trivial process is technically challenging because the NYSE Trades and Quotes (TAQ) data contains every trade with a time stamp precision of nanoseconds, i.e., hundreds of millions of trades per trading day, and is several terabytes of raw data. The resulting dataset contains 2.3 million observations and covers the eight hours before and after news publication at the second frequency. I use the models from the first essay to classify the news into positive and negative, show return patterns for the two types of news, and test trading strategies that react instantly to the new information.

This study contributes to the literature by providing multiple pieces of evidence that support the efficient market hypothesis at high frequencies. First, I confirm the finding of the first essay that financial markets anticipate new information and show that they react instantly and price in the new information within minutes. Second, the tested trading strategies yield surprisingly low returns. Finally, the average return and volatility patterns around all news are highly consistent with rational pricing: The elevated volatility before news expresses the uncertainty about the content of the news (markets usually know that information is coming because the firms often schedule a news release, but they do not know the content) and I show that holding stocks in the 6.5 hours (one trading day) before news yields an average excess return of 0.1%.

Essay 3: A Reevaluation of Profitability and its Trend

The third essay tests multiple measures of profitability and their trend regarding their ability to predict stock returns. Because such predictors challenge the Efficient Market Hypothesis (EMH), they are called anomalies. Reviewing these anomalies is necessary because they sometimes change over time and many disappear when analyzing a different period (especially post-publication) or, even worse, when just applying proper asset pricing tests. I cover the period from June 1980 to December 2021 and mainly analyze six different profitability measures concerning their level, trend, and level relative to the industry's mean.

This paper makes multiple contributions to the literature. First, I show that the trend-of-profitability effect described in Akbas et al. (2017) is mainly driven by the period 2000 to 2006 and has been reversing since then. This finding is also robust against slight changes in their methodology. Second, I confirm that the cleaning of Compustat's Selling, General and Administrative (SG&A) cost variable by re-adding Research and Development (R&D) expenses, as described in Ball et al. (2015), improves not only their profitability measure but also the one used in Fama and French (2015). Third, I show that the difference to the industry's mean yields strong results in Fama-MacBeth regressions; however, it does not translate into high value-weighted portfolio returns and therefore lags the absolute level of profitability as a predictor for future returns. Fourth, I propose to use a different measure of value compared to the popular book equity-to-market equity ratio, namely Gross Profit (GP) minus SG&A divided by market equity because the first seems to have lost its predictive power while the latter has not and was also a stronger predictor before. This measure of value is an ideal complement to measures of profitability.

Collectively, my thesis contributes to the debate on market efficiency and stock return predictability. The first essay finds that financial markets already anticipate news in the weeks ahead, price them in quickly, and that individual news is not a medium- or long-term return predictor. The second essay shows that this pattern can also be found at the intraday level, that most of the new information is priced in immediately, and that the average drift afterward is minimal and only lasts a few minutes. The third essay contrasts the evidence of market efficiency from the first two with long-term return predictability based on profitability measures.

While the third essay is not proof against market efficiency because there are theoretical risk-related explanations for excess returns of highly profitable stocks, I consider them at least questionable. Furthermore, I would also like to highlight that two of the risk factors of Fama and French, namely size and value, have had negative returns for over a decade. Should this persist, it raises the question of whether markets were efficient and are not anymore, or if they were not and are now. From a behavioral perspective, we as humans may be capable of correctly assessing the impact of individual new information but are ignorant about focusing on what is truly relevant in this ever-expanding sea of information.

References

Akbas, F., Jiang, C., Koch, P. D., 2017. The trend in firm profitability and the cross-section of stock returns. The Accounting Review 92, 1–32.

Ball, R., Gerakos, J., Linnainmaa, J. T., Nikolaev, V. V., 2015. Deflating profitability. Journal of Financial Economics 117, 225–248.

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Contents

List of Tables

List of Figures

Acronyms

ANN Artificial Neural Network. 32, 33

ATO Asset Turnover. 124, 127

BERT Bidirectional Encoder Representations from Transformers. 32, 42, 54

CAPM Capital Asset Pricing Model. 28

CBOW Continuous Bag-Of-Words. 42

CNN Convolutional Neural Network. 26, 43, 44, 47, 48, 73, 95

CPU Central Processing Unit. 33

CRSP Center for Research in Security Prices. 29, 34, 35, 38, 73, 74, 132

CTO Capital Turnover. 124, 127

CUSIP Committee on Uniform Security Identification Procedures. 34, 74, 80, 85, 87, 89, 92, 102, 103, 110–116

DNN Deep Neural Network. 32, 33, 53

EMH Efficient Market Hypothesis. 8, 25, 27, 28, 52, 65, 67, 72

EPFL École Polytechnique Fédérale de Lausanne. 5

ESG Environmental, Social, and Governance. 28

FNN Feedforward Neural Network. 33

GAAP United States Generally Accepted Accounting Principles. 35

GAN Generative Adversarial Network. 43

GDP Gross Domestic Product. 28

GLM Generalized Linear Model. 53

GP Gross Profit. 124, 129, 130

GPU Graphic (or General) Processing Unit. 33

GRU Gated Recurrent Unit. 26, 43–45, 47, 48, 73, 79, 96

HHI Herfindahl–Hirschman Index. 126, 132, 162, 163, 197–199

IDC International Data Corporation. 33

LASSO Least Absolute Shrinkage and Selection Operator. 53

LDA Latent Dirichlet Analysis. 31

LSA Latent Semantic Analysis. 31

LSI Latent Semantic Indexing. 31

LSTM Long- Short-Term Memory. 26, 43–45, 47, 48, 73, 95

MAE Mean Absolute Error. 46

MAP Maximum A Posteriori. 42

MD&A Management Discussion and Analysis. 31

MLP Multilinear Perceptron. 33

MNIST Modified National Institute of Standards and Technology. 33

MSE Mean Squared Error. 46

NASDAQ National Association of Securities Dealers Automated Quotations. 25, 34, 65, 66, 68, 73

NLP Natural Language Processing. 25, 32, 36, 42, 43, 54, 65–67, 72, 73, 91

NYSE New York Stock Exchange. 8, 25, 34, 65, 66, 68, 73, 75, 137, 145, 163, 166, 183, 185, 187

OLS Ordinary Least Squares. 53

PCR Principal Component Regression. 53

PLS Partial Least Squares. 53

pLSA probabilistic Latent Semantic Analysis. 31

PM Profit Margin. 124, 127

R&D Research and Development. 124, 125, 127, 129, 130, 136, 145, 149, 150, 153, 163

RBF Radial Basis Function. 33

RIC Reuters Instrument Code. 34, 35

RNA Return on Net Operating Assets. 124, 127

RNN Recurrent Neural Network. 26, 33, 42–45, 47, 48, 73, 79, 95, 96

ROA Return on Assets. 124, 127

ROE Return On Equity. 124, 127

SEC United States Securities and Exchange Commission. 30

SG&A Selling, General and Administrative. 124, 125, 127, 129, 130, 145, 147, 148, 163

SIC Standard Industrial Classification. 126, 129, 131, 132, 143, 163

SVM Support Vector Machine. 31, 33, 53

TAQ Trades and Quotes. 8, 66, 68, 73

TF-IDF Term Frequency - Inverse Document Frequency. 30, 41

TNIC Text-based Network Industrial Classification. 126, 131, 132, 143, 162, 163, 180, 181, 199

VIF Variance Inflation Factor. 145, 146, 148, 149, 182–187

Essay I:

Is the Stock Market's Reaction to News Predictable?

Jan Pichler*

ABSTRACT

This paper investigates the return predictability based on financial news by studying the largest dataset investigated so far in the literature: Over 4.4 million news headlines on firms listed in North America. I show that already the words of the Loughran and McDonald (2011) sentiment dictionary have some predictive power and that the Naïve Bayes classifier machine learning algorithm, which considers the most frequent words of the corpus it is fed, rather than predefined words of a dictionary, has higher predictive power for subsequent returns. Neural networks (CNN, RNN-LSTM, and RNN-GRU), which can also model relationships between words, even reach slightly higher accuracies. Although I show that a slow trader, only starting to trade on news the day after it is published, could potentially still make some profits (less than 0.10% per trade before transaction costs), the cumulative mean returns of the 21 trading days (one month) before and 63 trading days (three months) after the publication of news provide strong evidence for rational behavior and the market efficiency hypothesis: Markets start to anticipate the news about one month before and have priced in part of the impact upon publication. The news is then quickly digested (less than two days), and there is no clear trend afterward.

Keywords: News, Information, Natural Language Processing, Machine Learning, Supervised Learning

JEL Classification Numbers: C38, C45, G12, G14

*University of Bern, Faculty of Business, Economics and Social Sciences. Engehaldenstrasse 4, 3012 Bern, Switzerland. Email: jan.pichler@unibe.ch

I Introduction

This paper investigates the transmission of information in financial markets, a topic generally also referred to as the economics of information. Ever since Fama proposed the Efficient Market Hypothesis (EMH) in 1970, its validity has been a central discussion in the academic literature. Most importantly, Grossman and Stiglitz (1980) show theoretically that if information is costly, prices cannot fully reflect the information. Here, I provide empirical evidence on whether the market prices in new information efficiently. Furthermore, I focus on various technical aspects of modern Natural Language Processing (NLP) and show how these tools can be used to investigate the EMH based on textual data.

News is the latest information the market gets. In order to be priced in, it had to be read, interpreted, and traded accordingly by humans up to very recently. While the interpretation of numerical data goes back to the first days of modern finance, text analysis has a much shorter history in finance. Loughran and McDonald (2016) provide an overview of the scientific literature on textual analysis in accounting and finance and show that this research area has gone from purely human interpretation to automated analysis with word lists and, more recently, with the application of supervised machine learning techniques.

This paper compares the Loughran and McDonald (2011) sentiment dictionary's and several supervised learning algorithms' ability to predict the stock market's reaction to firm-specific news. I use a dataset of over 4.4 million news headlines from *Reuters North America* between January 1996 and December 2019 on firms traded at New York Stock Exchange (NYSE), NYSE Arca, and National Association of Securities Dealers Automated Quotations (NASDAQ). It is almost five times as large as the second largest dataset ever studied on the individual stock level in this research area: Heston and Sinha (2017), who use about 900'000 *Reuters* articles between 2003 and 2010.

Sentiment plays a central role in this study, and it is essential to distinguish two different concepts: *text sentiment* and *market* or *investor sentiment*. While the former refers to the "tone" of the text and is also used in many fields outside of finance, the latter refers to the general "mood" of the investors. The groundbreaking work of Baker and Wurgler (2006) launched a discussion in the finance literature surrounding the role of *investor sentiment* in the stock market. On the other hand, the pioneering studies of Tetlock (2007) and Tetlock et al. (2008) investigate the effect of *text sentiment* in the media on the aggregated stock market using the Harvard psychological dictionary. Garcia (2013) extends the studies of Tetlock by using the finance-specific sentiment dictionary of Loughran and McDonald (2011) and showing that the *text sentiment* in the media is a stronger predictor of the aggregated stock market during recessions. Similarly, my application of the Loughran and McDonald (2011) sentiment dictionary shows that the *text sentiment* of corporate news has predictive power for individual stock prices. In contrast to that, the supervised learning algorithms map directly to the stock market reaction, i.e., returns. While this approach drops the concept of *text sentiment*, the returns affiliated with a particular news headline can be interpreted as a measure of the *investor sentiment* caused by this headline. This allows the supervised learning models to capture predictive elements that have nothing to do with *text sentiment* but comes at the cost of losing some interpretability.

There are some further important differences between the applied models. The Loughran and McDonald (2011) sentiment dictionary contains 2'335 words perceived either as positive or negative. As headlines are relatively short (here, 10.5 words on average after preprocessing), about two-thirds cannot be classified because they do not contain any of the words from the dictionary. In contrast, all supervised learning algorithms use a vocabulary with the 17'000 most frequent words (98.5% of all words) and can therefore make a prediction for all observations. While a smaller vocabulary size makes the prediction task harder, not requiring a prediction on all samples may make the task easier.

The applied supervised learning algorithms are the Naïve Bayes classifier and three different neural networks (Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) with Long- Short-Term Memory (LSTM) units, and RNN with Gated Recurrent Units (GRUs)), each implemented both as a classification as well as a regression model. The Naïve Bayes classifier is a *bag-of-words*[1] model, meaning it assumes independence between words, which is also why it is called naïve. Neural networks overcome this limitation and are designed to model the relationship between words.

The cumulative excess returns of the 21 trading days (one month) prior and 63 trading days (three months) after the publication of the news classified by the sentiment dictionary shows that markets react within a very short period to the news (less than two days). Therefore, I use the cumulative return of the trading day after the news is published and the trading day before that day as the label for training the supervised learning models. Hence, news headlines are classified as positive if the return affiliated with them is positive and as negative otherwise.[2] The Loughran and McDonald (2011) sentiment dictionary has an out-of-sample accuracy (fraction of correct predictions) of 53.4%, the Naïve Bayes classifier of 54.7%, and the neural networks range between 54.8%, and 55.4% when using all predictions. As mentioned above, the Loughran and McDonald (2011) sentiment dictionary fails to classify about 65% of the headlines because they do not contain any words from the dictionary. Comparing its predictions with the same number of the most certain predictions from the other models leads to an outperformance (accuracy) on the negative predictions of 2.6 to 4.1 percentage points on the negative predictions and 6.2 to 14.9 percentage points on the positive predictions.

Investigating the above-described time frame for the supervised learning models confirms that markets, on average, follow a rational behavior: Markets start to anticipate the news about one month before and have priced in about half of the impact upon publication. The news is then quickly digested (less than two days), and there is no clear trend afterward. On average, the difference in the cumulative returns across two days between positive and

[1] It is called *bag-of-words* because you treat text like a bag filled with different pieces, each representing a word. One may shake this bag, but models following this approach will yield the same results, meaning that the order of the words is completely ignored in these models.

[2] Because news is published anytime while markets are closed most of the time, using this definition when working with daily returns ensures that the price impact is captured and keeps the time window as short as possible. Using a longer time frame would primarily increase the noise in the data.

negative news is around two percentage points. There are no realizable returns on the day after news publication (i.e., for a slow trader) for the predictions of the Loughran and McDonald (2011) sentiment index; however, there are minimal realizable returns left for the predictions of the supervised learning models (0.1% to 0.2% before transaction costs). This contradicts the sentiment theory, which would predict an overreaction and then reversal afterward.

I only find a tiny difference between the Naïve Bayes classifier and the neural networks because most of the learning is spent on creating good word representations in the embedding layer of the networks, and there seems to be only minimal learning of relationships between words. Using general pre-trained word embeddings or training them on a finance-specific task would be a possible solution for this issue. There is a trade-off when choosing between formulating the problem as a classification or as a regression task because the transformation drops a lot of information but, at the same time, gets rid of any potential outlier problem. Although I do not find any significant difference between using the neural networks as classification and as regression models, I would generally still advocate using them as regression models in such a setting because the point estimates are more useful for economic interpretation, and there are other methods to get rid of the outlier problem.

This paper makes multiple contributions to the existing literature. First, I show that the return pattern around news is highly consistent with the EMH: Financial markets anticipate the news and start pricing them in before publication, then quickly digest them upon revelation, and there is no drift afterward. Second, I show that machine learning algorithms can outperform the in finance traditionally used Loughran and McDonald (2011) sentiment dictionary. Furthermore, the detailed description of the text processing can serve as a "handbook" for other researchers.

This paper proceeds as follows: Section II covers the related literature on return predictability and text analysis in accounting and finance. Additionally, there is a small history of deep neural networks for the interested reader. Section III describes the data and the text preprocessing, while Section IV describes the different models. The empirical results are shown in Section V and discussed in Section VI before I conclude in Section VII.

II Related Literature

The first subsection covers the literature on return predictability and the closest related papers. Next, I give an overview of text analysis in accounting and finance; however, for a more thorough overview of textual analysis in accounting and finance, I refer the reader to Loughran and McDonald (2016). Methodologically, this paper is related to the fast-growing literature on what is nowadays known as deep learning. Hence, I provide the interested reader with a small history of deep neural networks.

II-A Return Predictability

Predicting stock returns has been an ongoing task among both practitioners and academics since the early days of stock markets. The predictive power of countless variables has been tested over the years. They could be grouped into stock related (e.g., past prices, returns, volume), firm characteristics (e.g., industry, variables from financial statements and derived ratios, Environmental, Social, and Governance (ESG) scores), macro variables (e.g., interest rates, inflation rates, Gross Domestic Product (GDP) growth, industrial production, unemployment) and many alternative data sources (e.g., Facebook, Twitter, Amazon reviews). It is important to note that (excess) return predictability is not necessarily evidence against the EMH but could also be found due to time-varying expected returns. Although it is possible to test the efficiency hypothesis based on some equilibrium asset pricing model, there is a joint-hypothesis problem, as already Fama (1970) noted - One tests the asset pricing model and the anomaly at the same time, and it is impossible to split the evidence between market inefficiency and a poor asset pricing model.

Nevertheless, countless studies have been published reporting indicators that yield anomalous excess returns compared to some asset pricing models like the Capital Asset Pricing Model (CAPM) (Sharpe, 1964; Lintner, 1965; Mossin, 1966) or the Fama-French three-factor model (Fama and French, 1993). While some interpret this as evidence against market efficiency, others claim to have found a priced factor that should be included in the equilibrium model. This led to the “factor zoo” we have today and is also, to a large extent, the result of excessive data mining.[3]

This paper investigates the possibility of relating news texts to the market’s reaction, i.e., predicting positive and negative (classification) stock reactions or even providing point estimations (regression) of the market’s reaction to the publication of a particular news headline. Being able to do so is by no means evidence against market efficiency because markets have to react to news in order to be efficient. It would only be evidence against efficiency if a medium- or long-term signal could be extracted. Obviously, being able to analyze the latest news faster than other market participants and with high accuracy would yield profitable trading strategies.

In his pioneering work, Tetlock (2007) shows that the sentiment (measured with the Harvard IV-4 psychosociological dictionary) of the Wall Street Journal column “Abreast of the Market” has predictive power for aggregated stock market prices (Dow Jones Industrial Average). Tetlock et al. (2008) extend this work by providing evidence that words also have predictive power for the market’s reaction and earnings on a firm-specific level. This gave rise to a growing literature focusing on the relationship between text data and market prices. Loughran and McDonald (2011) create their finance-specific sentiment dictionary

[3]Researchers and publishers induce many biases. By not publishing non-significant results, publishers give researchers a strong incentive to only report the significant results they found on their particular dataset and only provide limited insight into how many variables (or even datasets) they initially checked. Having a validation and/or test dataset is not standard in this field, but it would at least partially solve the problem of dataset-specific results (some bias would remain as hardly anyone ever checks the performance on the final dataset).

and show that the text sentiment of 10-K filings has predictive power for the market's reaction. All these studies face the issue that they heavily rely on the negative words and that the measures from just the positive words do not yield significant results. This issue is solved by Jegadeesh and Wu (2013), who use both the Harvard IV-4 and the Loughran and McDonald (2011) sentiment dictionary and compute the word weights by regressing on the market reaction to the publication of 10-K documents. Garcia (2013) extends this literature by showing that the predictive power of sentiment in media texts is more pronounced during recessions. The two largest media datasets studied in finance so far are Uhl (2014) and Hillert et al. (2014). Uhl (2014) analyzes the relationship between *Reuters* proprietary sentiment classification of 3.6 million *Reuters* news and returns. Hillert et al. (2014) use 2.2 million articles from 45 U.S. newspapers and show that firms with higher media coverage exhibit stronger momentum and that it depends on article tone, i.e., sentiment.

Apart from Uhl (2014), the aforementioned studies either use media articles or 10-K filings. It is unclear to what extent both document types contain new information. While media articles can cover very recent events, they may also interpret older events, and 10-Ks cover the whole past year of a company. It is highly likely that the *Reuters North America* news covers more recent events to a much larger extent and therefore contains more new information. Due to the similarity of the dataset, Heston and Sinha (2017) is probably the closest related paper to this one:[4] they analyze 900'754 articles from *Thomson Reuters* between 2003 and 2010. A proprietary *Thomson Reuters* neural network classifies news stories according to their sentiment. They then form portfolios based on the average news sentiment in a one-week formation period and find that the top quintile outperforms the bottom quintile by about 1.2% in the previous seven days, 2% on the publication day, 0.17% on the subsequent day, and about 0.2% in the following seven days.

There have been a few attempts to predict the stock market reaction based on financial news in the computer science community. However, they all have questionable true generalizability as they all suffer from at least one of the following issues: small news samples, very short sample period, only tested for a few large firms, or feature engineering decisions that could drive the results. E.g., Peng and Jiang (2015) use a dataset containing 106'521 news articles from *Thomson Reuters* and 447'145 from Bloomberg between October 2006 and December 2013. However, after merging with stock market data from the Center for Research in Security Prices (CRSP), they are left with a training set of 65'646 samples, a validation set of 10'941, and a test set of 9'911 samples. Taking information from the past five days' price movements into account, they reach accuracies between 52% and almost 57% at predicting an up or down stock price movement based on its next day's closing price. Note that apart from also considering past prices, their approach differs in many other aspects compared to this study.

[4]Like Heston and Sinha (2017), Uhl (2014) uses the sentiment classification of *Thomson Reuters* proprietary neural network; however, he investigates the relationship with the aggregated market and not the individual stocks.

II-B Text Analysis in Accounting and Finance

Although research in text analysis has a long history in accounting and finance (see Jones and Shoemaker (1994)), it was a niche corner throughout the 20th century. Data collection was a costly process before documents were available digitally, and the fact that text analysis had to be done manually by humans only made it worse. Therefore, essentially all early studies suffered from small samples, e.g., Frazier et al. (1984) looked at the 1978 annual reports of 74 firms, which was a comparatively large sample at the time. The digitalization wave of the 1990s at least partly solved the problem of costly data collection, e.g., the United States Securities and Exchange Commission (SEC) required all firms to submit their filings digitally starting in 1996 and made the files publicly available. However, depending on the task, there are still no datasets easily available, or if they are, the desired labels are usually missing. Moving from human text analysis to automatic text analysis by machines is a very active research field in computer science nowadays, and despite all the progress, still only partially solved.

As text has a particular hierarchical structure (letters form words, which then form sentences, from which paragraphs are built), early statisticians naturally focused on the level of the words as they undoubtedly contain most of the information. One branch of text analysis research in accounting and finance investigates the readability of texts. Based on the average number of words per sentence and the percentage of complex words, Li (2008) proposed his *Fog Index* as a measure of readability and shows that firms with lower reported earnings tend to have annual reports that are harder to read. A high *Fog Index* (meaning the annual report is difficult to read) indicates less investment efficiency (Biddle et al., 2009), less small investor trading and holding (Miller, 2010; Lawrence, 2013), and a larger analyst forecast dispersion (Lehavy et al., 2011). However, most researchers are more interested in extracting the content of texts rather than some characteristic like readability. Word count models usually only take a subset of words into account. This subset is also called a dictionary because every word belongs to one or multiple categories. There are general dictionaries like the General Inquirer dictionaries (merged from the Harvard-IV-4 and the Lasswell dictionaries) but also field-specific dictionaries like the Loughran and McDonald (2011) dictionary for accounting and finance. The need for field-specific dictionaries comes from the fact that the meaning of some words depends on the field. E.g., *vice, liability*, and *depreciation* are negative words according to the Harvard dictionary, but it is unclear whether these words really indicate a pessimistic tone in a financial text. Apart from the raw word counts, where all the words have the same weights, different weighting schemes have been proposed. Term Frequency - Inverse Document Frequency (TF-IDF) developed by Jones (1972) is the most prominent of them. If the amount of unique words is heavily imbalanced between the different categories and they are expected to be distributed equally, another popular choice is weights that offset this imbalance. Dictionary-based methods are one example of the *bag-of-words* approach. Despite the many uncaptured aspects of language, these word count-based measures usually already produce decent results.

With the advances in computing power over the last decades, researchers started taking another approach by using supervised learning algorithms. Naïve Bayes is one of the oldest and most established methods and was first used in finance by Antweiler and Frank (2004). They look at 1.5 million stock message postings on *Yahoo Finance* and *Raging Bull* but only find a weak relationship with stock returns. This may be explained by the fact that they only use 1'000 examples to calibrate their model. Others use larger training sets (e.g., Li (2010) uses 30'000 sentences from forward-looking statements in the Management Discussion and Analysis (MD&A) section of 10-K filings), but a limited number of training examples is a potential problem for all supervised learning algorithms. Hence, there is usually a trade-off between the marginal improvement and the marginal cost of an additional training example.[5] Some studies overcome this issue by using readily available labels. E.g., Jegadeesh and Wu (2013) use a Naïve Bayes framework to determine the weights of positive or negative words in 10-K filings based on the market reaction. They find that both positive and negative words are significantly related to the market reaction. As 10-K filings are publicly available and have a particular structure, they are the most popular financial texts for analyzing. Probably one of the most exciting datasets, the analyst reports of the *Investext* database, was studied by Huang et al. (2014).[6] They classify over 27 million sentences into positive, neutral, and negative sentiment based on general word lists (accuracy around 50%), finance-specific word lists (accuracy above 60%) and a Naïve Bayes classifier trained on 10'000 examples that reaches an accuracy of above 75%.

Latent Semantic Analysis (LSA) (also called Latent Semantic Indexing (LSI)) (Deerwester et al., 1990) is a *bag-of-words* model which aims to identify common themes in documents or a corpus of documents. Hofmann (2001) improves LSA (based on singular value decomposition) by introducing probabilistic Latent Semantic Analysis (pLSA), which defines a generative model for the words and uses maximum likelihood estimation. The concept was further enhanced with the introduction of Latent Dirichlet Analysis (LDA) (Blei et al., 2003) by also defining a generative model for the mixing proportions of the topics on the document level. So far, only very few studies in accounting and finance apply LDA. Dyer et al. (2017) use it to show that the length increase of 10-K filings between 1996 and 2013 is mainly related to three of 150 topics: fair value, internal controls, and risk factor disclosures. Huang et al. (2018) compare the topics extracted by LDA in management conference calls with those in subsequent analyst reports. They find that analysts pick up the topics from conference calls but also discuss further topics. Furthermore, they show that investors also react to analysts simply confirming management's conference call discussions.

Purda and Skillicorn (2015) use a Random Forest (Breiman, 2001) and Support Vector Machine (SVM) combination to determine which words are best at predicting fraud. Apart from words expected to have high predictive power (e.g., *legal* or *settlement*), there were surprisingly many small and very common words (e.g., *or, at, on, as, is, its, may, it, no*) with high predictive power. These words (usually referred to as *Stop Words*) are often re-

[5] Researchers may introduce a bias here by adding samples up to the point where they are satisfied with the results.

[6] Unfortunately, most institutions providing analyst reports nowadays have contracts with Refinitiv preventing them from providing access to the database at a large scale.

moved in natural language processing because it is perceived that they do not contain much of information and are more noise than anything else. However, the results of Purda and Skillicorn (2015) indicate, that people writing fraudulently heavily tend to use more *Stop Words*; hence, they do contain information in this case. They further show that the change in their truth measure from one report to the next of the same firm adds very substantial predictive power, highlighting the potential benefit of controlling for the text source.

Wang et al. (2014) investigate the return predictability of the text sentiment of articles on *Seeking Alpha* and messages on *StockTwits*, and Mishev et al. (2020) compare the sentiment extraction ability of numerous models in the field of finance.[7] They start with sentiment dictionaries like the aforementioned of Loughran and McDonald (2011), continue with Artificial Neural Networks (ANNs) as word and sentence encoders, and end with the latest available transformers models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018).

This is also symbolic for the whole field of NLP, which moved towards using ANNs because they are well suited for high-dimensional problems. Thus, the next subsection provides a short history of the development of these deep neural networks.

II-C A Small History of Deep Neural Networks

Today, Deep Neural Networks (DNNs) are state-of-the-art models for automatic language processing. Deep learning algorithms have had exceptional success in pattern recognition and labeling problems in recent years. However, this success wave, which started around 2006, and led to the term deep learning, is only the latest development wave. E.g., one of the most crucial concepts, backpropagation, is based on the chain rule, which was developed at the end of the seventeenth century by Leibniz (1684) and Newton.[8] Although calculus and the algebra developed by the two were extensively used to solve optimization problems, it was not until the middle of the nineteenth century that gradient descent was introduced as a technique to iteratively approximate the solution of an optimization problem (Cauchy, 1847). In the 1940s, a field known as *cybernetics* emerged in which a perceptron with linear functions was introduced. As the field developed, it moved to non-linear functions and multilayer perceptrons (see Figure 1) in the 1960s and 1970s. Werbos (1981) develops the idea of using backpropagation to train these first ANNs, and Rumelhart et al. (1986b) present some of the first results from applying backpropagation. The field then became known as

[7] They use the *Financial Phrase Bank* from Malo et al. (2014) with 4'846 English sentences labeled by financial experts with positive, neutral, and negative, and the *SemEval 2017 TASK 5* from Atzeni et al. (2017) with 2'510 news headlines also labeled by financial experts with a sentiment score between minus one and one.

[8] There was a dispute between Leibniz and Newton about who invented calculus. Newton claims to have started working on it in 1666 at the age of 23 but did not publish anything until decades later, and even then did so only as a minor annotation in one of his publications. The main publications of Newton on the topic were in 1687, 1693, and 1704. On the other hand, Leibniz started around 1674 and published his work ten years later. L'Hôpital (1696) recognized that the work of the two was mostly about the same calculus. The consensus nowadays is that although each of the two knew the work of the other, they invented calculus independently (mainly because they took different paths to develop the idea of calculus).

connectionism in the 1980s and the early 1990s because of the importance of the connections between the "neurons." Interest then shifted in the 1990s towards other machine learning techniques than ANNs until their renaissance in 2006, when Hinton et al. (2006) demonstrated that a neural network could outperform the Radial Basis Function (RBF) kernel SVM on the Modified National Institute of Standards and Technology (MNIST)[9] benchmark.

The success of ANNs in recent years was supported greatly by the improved availability of large datasets as well as the continuing growth of their size.[10] As computing power increased exponentially over the past century and its price decreased similarly,[11] it is possible nowadays to train simple ANNs with reasonable training times on any personal desktop or laptop computer which has multiple Central Processing Unit (CPU) cores. Using a high-end workstation with multiple powerful Graphic (or General) Processing Units (GPUs), one can train computationally complex ANNs with many layers on large training datasets within days or even hours.[12]

ANNs were originally developed as mathematical models with an architecture similar to a biological brain (Rosenblatt, 1962; Rumelhart et al., 1986a; McCulloch and Pitts, 1988). Although similar in architecture, ANNs are very different from biological brains. Mainly because they are limited to the physical properties of current computing machines which work on a binary electrical system. Communication between neurons in biological neural networks happens mostly via chemical neurotransmitters, rather than electronic signals (these are mostly used for communication within a neuron).[13] An ANN consists of nodes and weighted connections between them. The nodes represent the neurons, and the connections represent the synapses of a biological brain. The network is activated by providing some input to one or several nodes, and the signal then spreads through the network and results in some output. One can distinguish between ANNs with connections that form cycles (called feedback or recurrent neural network) and the ones that do not. ANNs with no cycles are called Feedforward Neural Networks (FNNs), e.g., a perceptron (Rosenblatt, 1958). Figure 1 shows a Multilinear Perceptron (MLP) with an input layer, which is processed by two hidden layers that produce an output layer.

Nearly all deep learning models can be divided into the following components: the dataset, some cost function, an optimization algorithm, and the model architecture. Theoretically, one can replace any component independently from the others, but some combinations are more common than others because they are easier to optimize. For a deeper insight into

[9] The MNIST database contains images of handwritten digits.

[10] Hilbert and López (2011) estimate that in 1986 around 1% of the five exabytes (one exabyte equals one million terabytes) of data worldwide was digitally stored, and in 2007 it was around 93% of the 500 exabytes. The International Data Corporation (IDC) estimates in their 2018 white paper that the current 33 zettabytes (one zettabyte equals 1'000 exabytes) will increase to 175 zettabytes by 2025. This corresponds to a growth of 25% p.a. between 1986 and 2007, then 46% p.a. till 2018, and an expected growth of 27% p.a. for the coming years.

[11] See Roser and Ritchie (2019) for an overview of the development of computing power.

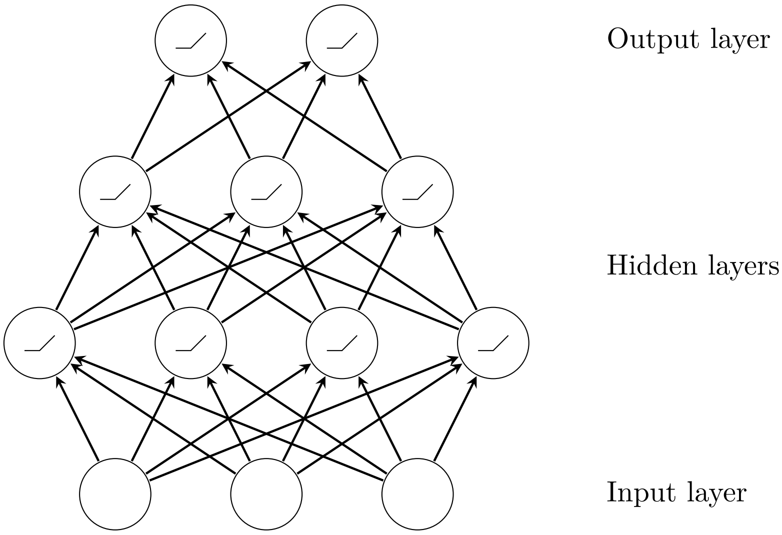
[12] GPUs are very efficient at processing the most frequent mathematical operations involved in training an ANN like matrix operations.

[13] Although most connections between two neurons (synapses) are chemical, there is a small fraction of electrical synapses (gap junctions).

DNNs, I refer the reader to Graves (2012) and Goodfellow et al. (2016). Graves (2012) focuses on RNNs, while Goodfellow et al. (2016) try to cover the whole field of deep learning and also include the basic calculus and algebra behind the algorithms.

Figure 1: Multilayer Perceptron

This figure shows a multilayer perceptron with two hidden layers. The ⟋ stand for the application of a ReLU function, a popular activation function because the input can have no relationship to the output or a linear relationship.



III Data

The used dataset contains all *Reuters North America* news from the beginning of January 1996 till the end of December 2019. There are 38.10 million news headlines in total; however, most of them get dropped due to various reasons: 24.72 million articles are on a financial instrument identified by a Reuters Instrument Code (RIC), of which only 7.96 million have a Committee on Uniform Security Identification Procedures (CUSIP) code.[14] Next, I drop all observations with a CUSIP 8 code which is not included in CRSP/*Compustat* merged database (common stocks of firms traded at NYSE, NASDAQ, or NYSE Arca, including firms incorporated outside the U.S.). The dataset then consists of 6.31 million observations. Table 1 shows the distribution of the number of subjects (i.e, firms) affiliated with each news article. Because the applied models cannot distinguish between which text part is related to which firm, I drop all observations on multiple firms. Including them could be done in one of the following two ways: One could either use the average return of the firms as the label, which may result in averaging out any valuable information and adding a lot of noise, or have the same news for each firm as a separate observation, which would put more

[14] The most frequent news headlines which do not have a CUSIP are on currencies, indices, or firms not listed in North America.

(most likely too much)[15] weight on these observations. Inspection of a sample of articles with many affiliated firms shows that these articles are mostly statements that these firms will or have published earnings results or that analysts have updated their ratings on these firms.

The observations with only one affiliated firm contain 280'804 exact duplicates.[16] Furthermore, I drop another 106'367 observations, which are considered duplicates. These observations are identified by computing a similarity measure for each headline with all the headlines of the same firm on the same date.[17] The dataset containing 5'002'242 observations is then merged with the CRSP/*Compustat* merged database, which has the corresponding return data for 91% of the observations. Hence, the dataset before any text preprocessing consists of 4'542'855 observations.

Table 1: Number of Subjects (i.e., firms) per News Article

This table shows the distribution of the number of subjects (i.e., firms) affiliated with each news article. The vast majority of articles are on just one firm (85.3%). Inspection of a sample of headlines on multiple firms shows that they are just statements that these firms will or have published earnings results or that analysts have updated their ratings on these firms.

Number of firms	Observations	% of total
1	5'389'413	85.3%
2	510'341	8.1%
3	180'520	2.9%
4	89'209	1.4%
5	50'397	0.8%
6-10	64'480	1.0%
11-20	16'620	0.3%
21-30	11'479	0.2%
31-40	2'430	0.0%
41+	0	0.0%

III-A News Headlines and Preprocessing

Text cleaning is an important issue when working with natural language. All characters are converted to lowercase, and the company RICs (embedded in the text as follows:<RIC>) are stripped from the headlines.[18] Furthermore, all punctuation, special characters, and numbers are removed.[19] A more difficult problem is contractions in the English language. While it is an easy task to contract since two contractable words can only be contracted to

[15] The weight of the multiple firm news would increase from 14.7% to 37.5%.

[16] Exact duplicates are observations with the same date, subject, and headline.

[17] I us the *SequenceMatcher* function of the Python *difflib* package to compute the similarity between two headlines. Values above 0.97 are considered duplicates (only the last version is kept). It makes sense to use a high value like 0.97 because there are a lot of very similar headlines, e.g., the only difference being United States Generally Accepted Accounting Principles (GAAP) earnings vs. non-GAAP earnings.

[18] This was already done for the computation of the similarity scores.

[19] The removed punctuation and special characters are the following: !"#$%&()*+,-./:;<=>?@[\]_^'{|}~\t \n'. \n is the encoding of a new line and \t of a tab.

exactly one possible contraction, expansion is a rather difficult task because for the same contraction multiple possible expansions exist and the right one has to be determined from the context. I use the Python module *pycontractions*' precise method to expand the contractions.

So far, all text cleaning steps are standard practice in NLP for most applications. A more controversial discussion surrounds the removal of *Stop Words*.[20] Neither is there a general rule on when to remove *Stop Words* nor a universally used list of them. The effect of the removal of *Stop Words* has been studied for several applications (e.g., see Munková et al. (2014) or Ghag and Shah (2015)), but they yielded no clear evidence on when to remove them and when not. From a logical point of view, it makes sense not to remove anything if the model should develop a profound understanding of language. These would generally be huge models trained on massive datasets (e.g., the whole Wikipedia corpus). For tasks that only require a shallow understanding of language and for which the available training datasets are small, it is much harder to trade off the advantage of removing the noise of *Stop Words* with the disadvantage of losing information. I do not remove any *Stop Words* because the semantic meaning can entirely change due to *Stop Words* like *no*, *not*, etc. However, I remove any occurrence of an apostrophe followed by an *s* (*'s*), i.e., the possessive *s* since the contractions have been expanded before. As described before, many duplicates were identified based on the unprocessed headlines. I use the same procedure to check for exact duplicates in the processed headlines and find a surprisingly large number of 92'605 which I also remove. This indicates how cautious it was to only consider unprocessed headlines with a similarity measure above 0.97 as duplicates. The dataset now consists of 4'450'250 observations.

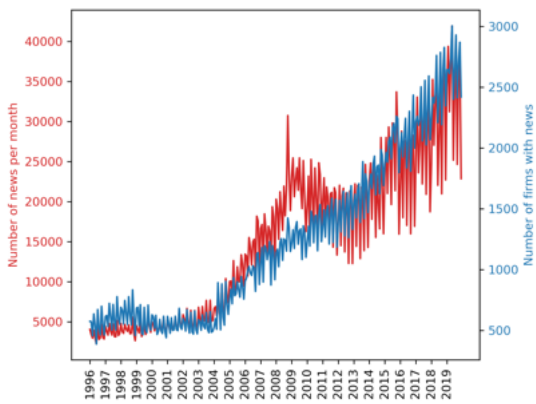
Figure 2 (a) shows that the number of observations is not equally distributed over time. Overall, the almost sevenfold increase in monthly news, which started in 2004, is mainly driven by the increase in covered firms. However, Figure 2 (b) shows that also the intensity of news coverage changed over time. While it was relatively low in the 1990s, it started to increase after 1999 before peaking during the financial crisis. The intensity then dropped and has been steady at around twelve news per firm per month for the last eight years of the sample. Headlines have an average length of 10.50, and Figure 3 (a) shows the distribution of the headline length. I do not impose a minimum or a maximum headline length and thus neither trim the headlines nor drop any observations.

There are 159'078 unique words in all the headlines. The frequency of the words differs vastly and is shown in the log-scaled Figure 3 (b). Apart from above mentioned stop-words removal, there is another controversial preprocessing step: word stemming/lemmatizing. Reducing inflected or derived words to their word stem by removing their ending is called stemming, and lemmatizing is the more sophisticated process of changing words to their base or root form. Stemming reduces the number of unique words in the dataset to 130'956

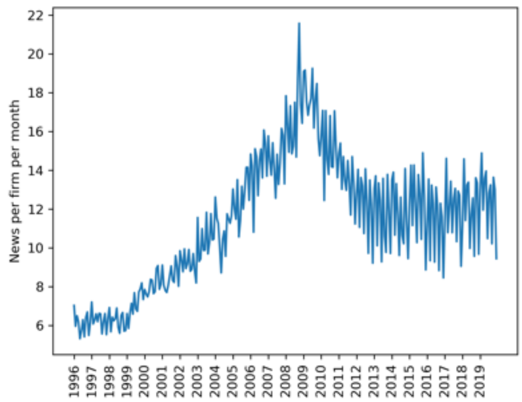
[20] The term *Stop Words* refers to any words that are removed in the preprocessing of texts in NLP. Usually, these are words that are frequent but contain little to no relevant information (e.g. is, to, where, who, etc.).

Figure 2: Observations per Month and News per Firm

The red line in Figure (a) shows the number of observations per month (left scale). While it was almost steady at around 5'000 from 1996 to 2004, it then started to increase to about 30'000 in 2009. After the financial crisis, there was a drop back to around 20'000 in 2013 before there was again an increase to around 35'000. The blue line shows the number of different companies about which at least one news headline was published (right scale). It was at 600 between 1996 and 2004 before it steadily increased to almost 3'000 at the end. In total, there is news on 3'950 distinct firms in the dataset. Figure (b) shows each month's average number of news per firm. There was a steady increase between 1999 and the financial crisis in 2009. The development afterward indicates that the increase in the observations per month after 2013, shown in Figure (a), is driven by an increase in the number of firms rather than by more intense news coverage of the same firms.



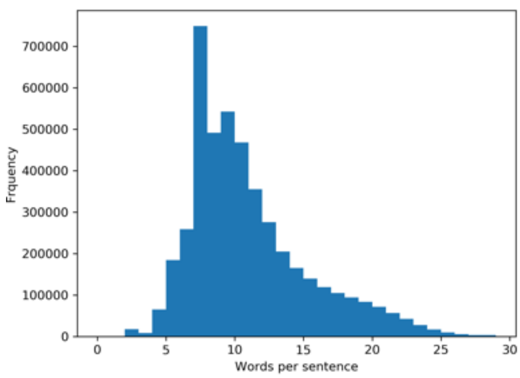
(a) Observations per month



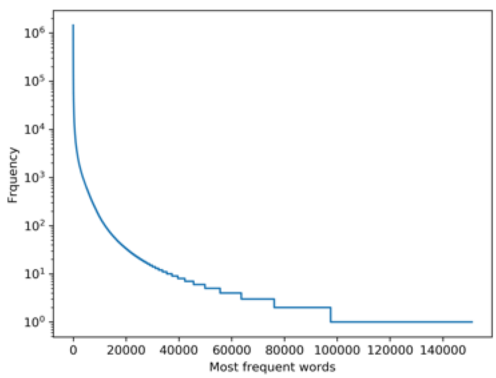
(b) News per firm per month

Figure 3: Words per Headline and Word Frequencies

Figure (a) shows the distribution of the number of words per headline. The average length is 10.50, and the distribution is heavily right-skewed. 46 observations have a length of zero words, meaning that all the content of the headline was dropped in the preprocessing, and the maximum length is 66 words. Figure (b) shows the frequencies of all words in the dataset. Table A.1 in the appendix shows the 30 most frequent words and their frequency. While there are 151'037 unique words in the dataset, the models in this paper only take the 17'000 most frequent words into account, which make up 98.5% of the total number of words. Over 53'000 words appear only once and roughly another 21'000 only twice.



(a) Words per headline



(b) Word frequencies

and lemmatizing to 151'037.[21]Unfortunately, the literature does not provide clear guidance on when to use stemming/lemmatizing, and its influence on results is application dependant. I use the lemmatized version since it reduces the number of unique words and because an inspection of some samples shows that stemming changes the meaning of some words. In undisclosed results, I compare the performance between stemming, lemmatizing, and doing neither for a couple of models, and I find that the impact is neglectable. There are over 53'000 words that only appear once, and the 100 most frequent words make up 50.3% of the total number of words (46'712'364).[22] For all models except for the sentiment dictionary, where the vocabulary size is given (2'689), one has to choose a vocabulary size.[23] A too-small vocabulary size means dropping potentially valuable information, while a too-large value results in overfitting. In undisclosed results, I compare the performance for different values for some models and find that results are robust for values between 3'000 and 30'000. I choose to use a vocabulary size of 17'000 because it is estimated that a native English-speaking adult has a vocabulary of about 17'000 words (Goulden et al., 1990). The 17'000 most frequent words make up 98.5% of the total number of words in my dataset. Table A.2 in the appendix shows an extract of the dataset, and Table A.3 in the appendix shows a random sample of 20 headlines before the preprocessing and afterward in order for the reader to get an idea of the applied changes in the preprocessing.

III-B Return Data

As mentioned above, the return data is from CRSP/*Compustat* merged database. Since the merge is conducted such that day 1 is the trading day after the news is published, the market could either react on day 0 (if it was a trading day and the news was published before market closing) or on day 1. To prevent large outliers from driving the results, I exclude all observations with a cumulative return above 100% over days 0 and 1,[24] which is the case for 8'643 observations (0.2%). The first column of Table 2 shows the descriptive statistics of the cumulative returns of day 0 and day 1 after removing the observations with returns above 100%. Figure 4 shows some descriptive statistics for each trading day, starting 21 trading days (one month) before the news publication and ending 63 trading days (three months) afterward. For readability, the minimum and maximum are not included as they generally range between -73% to -93% and +265% to +1'025%, respectively. 90% of the daily returns are between -4% and +4%, except for around the news publication days, where this spread increases by about 80%. The daily volatility is generally slightly above 3% and increases to about 7% on news publication days. The slight increase in volatility after three months indicates that the news cluster a bit around earnings announcements which usually occur more or less exactly three months later. I show the same timeframe for the

[21]I use the *PorterStemmer* from the python *nltk* package for the word stemming and *WordNetLemmatizer* for the lemmatizing.

[22]Table A.1 in the appendix shows the 30 most frequent words.

[23]Vocabulary size is the number of most frequent words taken into account.

[24]The models are trained on the cumulative return of days 0 and 1 (except for the sentiment dictionary).

predictions of the models in Figure 5 in order to investigate how the market digests news and if any medium-term signal can be extracted. However, there is the caveat that I only analyze daily returns and not intraday data, and due to this caveat, I cannot estimate any potential gain a fast trader, who reacts instantly to the news, could make. Nevertheless, I provide an estimate of the returns a slow trader, who reacts on the trading day after the news is published, could achieve. Obviously, any returns before the publication day would not be realizable for such a trader. Since the returns are computed as the change in the closing prices, the return on the publication day would also not be realizable. But even the return of the day after might not be realizable since the news could be published after market closing, and the trader could not trade at the closing price. Therefore, I compute the realizable returns as the change between day 1's opening and closing price. Descriptive statistics for these returns are shown in the second column of Table 2.

Figure 4: Descriptive Statistics for News Timeframe

This figure shows the means, standard deviations, 5%, and 95% quantiles of the returns for each trading day, starting 21 trading days (one month) before the news publication and ending 63 trading days (three months) afterward. 90% of the daily returns are between -4% and +4%, except for around the news publication days, where this spread increases by about 80%. The daily volatility is generally slightly above 3% and increases to about 7% on news publication days. Returns are computed as the change in closing prices.

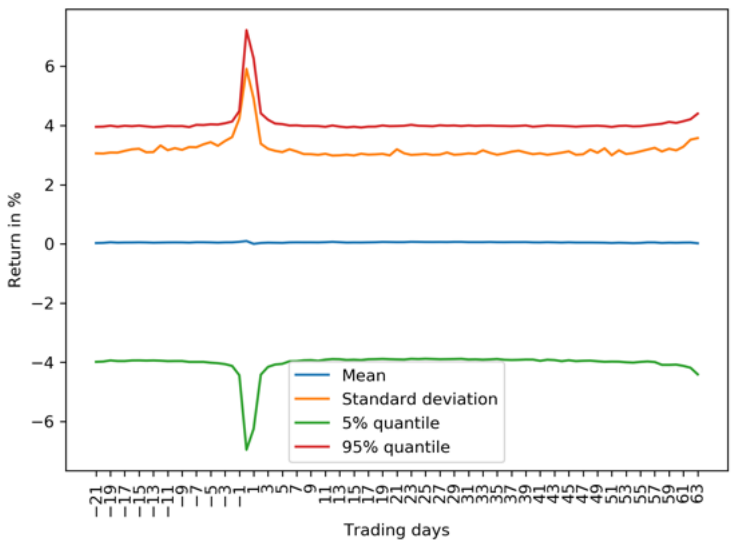


Table 2: Descriptive Statistics of the Cumulative Return of Days 0 and 1 and the Realizable Returns

This table shows the descriptive statistics of the cumulative return of day 0 and day 1 (trading day after news publication) after removing the cumulative returns above 100% in the first column and of the realizable returns (computed as the difference between day 1's opening and closing price) in the second column. The number of observations is 4'441'607.

	cumulative returns days 0 and 1	realizable returns
Maximum	99.99%	121.89%
95% percentile	10.38%	4.98%
Mean	0.10%	−0.03%
5% percentile	−10.30%	−5.12%
Minimum	−92.21%	−75.52%
Standard deviation	7.57%	3.68%

IV Methodology

This paper aims to map financial news headlines to future returns. Traditionally, one would define some measure of sentiment and compute it for each particular piece of text, then check if this measure of sentiment has predictive power for future returns. Using a *bag-of-words* model, where the words are categorized according to some dictionary, is the most common approach and has become popular in many research areas. Another *bag-of-words* approach is the multinomial Naïve Bayes method. This supervised learning algorithm takes the most frequent words into account and computes the best weight of each word based on some training sample. Neural networks are a text-processing approach that became very popular in the computer science community but has not been used frequently in the field of finance so far. Since neural networks can model many aspects of language, which the *bag-of-words* models ignore, they could provide superior predictions.

There are several options on how to approach the task at hand here. The first is to also train a model for sentiment classification and then see if stocks outperform after positively classified news and underperform after negative news. The problem with this approach is that it requires having lots of labeled training samples. While the Naïve Bayes classifier may reach decent results on about 10'000 training samples (e.g., see Huang et al. (2014), who classify analyst reports), neural networks require many more training samples. Therefore, I take a far less labor-intensive approach than human classification by categorizing the news according to the cumulative return of days 0 and 1 (positive if above zero, negative otherwise). As day 1 is the trading day after the news publication and day 0 the trading day before, using the cumulative return of the two ensures that the market's reaction is included. Another option is to make it a regression problem and directly map the texts to the cumulative returns of days 0 and 1. As the label is now the discrete rather than the categorized discrete variable, the network may be able to use this additionally provided information to produce superior predictions.

In this paper, I compare the abovementioned approaches to predict returns based on financial news. Table 3 provides an overview of the applied models, which are explained in more detail in the subsequent subsections. Note that I do not lay out all the concepts and mathematics of the neural networks and refer the reader to Graves (2012) and Goodfellow et al. (2016) in case any concept of neural networks is unknown.

Table 3: Model Overview

This table provides an overview of the different text classification/regression methods applied in this paper. Classified returns refer to the cumulative returns of days 0 and 1 being categorized as positive if above zero and as negative otherwise.

Model	Maps to
Loughran and McDonald (2011) sentiment dictionary	text sentiment
Multinomial Naïve Bayes	classified returns
CNN	classified returns / returns
RNN-LSTM	classified returns / returns
RNN-GRU	classified returns / returns

IV-A Sentiment Dictionary

Loughran and McDonald (2011) provide the most popular semantic dictionary for financial texts. They gathered words from 10-K filings between 1994 and 2008, and their dictionary consists of 2'335 negative and 354 positive words. The fact that the negative words outnumber the positive words by a factor of about 6.6 is a potential problem. If the positive words do not occur about 6.6 times more often (which is highly unlikely and not the case in the used dataset), the sentiment measured with this dictionary tends to be strongly negative on average. Therefore, it cannot be concluded that news tends to be negative on average, even if this measure suggests it.
Term weighting schemes have a long history in information retrieval from texts and at least go back to the 1950s (e.g., see Luhn (1957)). TF-IDF is one of the (if not the) most popular weighting scheme; however, many others have been proposed (e.g., see Aizawa (2003)). Words are weighted by their frequency relative to their occurrence in all documents, and the reasoning why the TF-IDF measure could be more appropriate than the raw counts comes from the underlying assumption that the informational content of words is inversely related to their frequency. Intuitively, this seems to be true to a certain extent, and empirical evidence indicates that it improves results in many cases. However, its benefits would be extremely limited in this case. First of all, it would not systematically mitigate the issue of the imbalanced dictionary. Secondly, its effect on short texts is minimal since most headlines only contain one word from the dictionary, weighting would not change the classification. Hence, I use the raw word counts to compute the sentiment measure.

IV-B Multinomial Naïve Bayes

The Naïve Bayes method, also called simple Bayes or independence Bayes, assumes that the inputs (here: words) are independent of each other. This assumption is obviously strongly violated in the context of natural language. Nevertheless, this simple model already performs well on many tasks and even outperforms more advanced models in situations where only very few training examples are available. Naïve Bayes classifier for text classification is trained on labeled text samples, where each sample is a vector with the length of the vocabulary V and the word counts as its entries. As for the neural networks, I choose to only include the 17'000 most frequent words. The Naïve Bayes classifier maximizes the Maximum A Posteriori (MAP) class c_{map} according to:

$$c_{\text{map}} = \underset{c \in C}{\arg\max} \left[\log \hat{P}(c) + \sum_{1 \leq w \leq n_d} \log \hat{P}(w_k|c) \right], \tag{1}$$

where $\log \hat{P}(c)$ is the prior of class c, depending on its relative frequency, and $\log \hat{P}(w_k|c)$ is a weight that indicates how strong the word w is associated with class c. For a more detailed description of the Naïve Bayes classifier, I refer to chapter 13 of Schütze et al. (2008).

IV-C NLP with Neural Networks

In NLP with ANNs, the first step is to set up the vocabulary of all words present in the data.[25] Each word is then encoded as a vector with the length of this vocabulary, which has all zeros but a one on the position of the particular word (1-of-V coding). From these word encodings, a sentence is then encoded as a matrix with the total vocabulary length on the first dimension and the sentence length on the second dimension. This sentence representation can then be fed to a neural network with an embedding layer as the first layer. The embedding layer will map each word to a point in a high-dimensional vector space (usually three to 300 dimensional), i.e., each word is defined by a vector with the length of the dimensionality of the particular vector space. Alternatively, these vectors can be learned in a pre-training process, and the sentences are directly encoded with the pre-trained vectors (as a consequence, there would be no embedding layer in the network). Mikolov et al. (2013a) introduce *Word2Vec*, where the vector representations are learned with a one-layer FNN. The network takes N 1-of-V encoded words as inputs, maps them to projection layer P with a $N \times D$ projection matrix, and has one hidden layer.[26] The network's output is a probability distribution over words which differs for the two models they propose: the Continuous Bag-Of-Words (CBOW) model and the skip-gram model. In the CBOW model, a word has to be predicted from the n previous and the n following words. The skip-gram model is the opposite: one word is used to predict the n previous and n following words.[27] More recently, Devlin et al. (2018) from *Google AI Language* propose BERT, which is currently the best peer-reviewed model in many NLP tasks. The main advantage of pre-training vector representation is that they can be learned on large unlabeled text corpora, which do not

[25] After choosing the vocabulary size (here: 17'000), and dropping the excluded words.

[26] They use between 500 and 1'000 units in the hidden layer.

[27] Earlier in 2013, Mikolov et al. (2013b) had already proposed an RNN for learning the vector representation of words, but the CBOW and the skip-gram model significantly outperformed this approach on various tasks. For details see Mikolov et al. (2013a).

need to be task specific. To this day, there is no general rule in which cases pre-learned representations outperform the ones learned task specifically. However, they are beneficial when only a few labeled examples are available or if the whole text corpus is small.
Choosing a reasonable vocabulary size (number of most frequent words considered) and the vector embedding dimensionality is crucial to prevent overfitting. The number of parameters of the embedding layer is the product of the two, usually making them responsible for most of the model parameters. The two most widely used architectures for NLP tasks are CNNs and RNNs. RNNs use different memory cells, LSTM and GRU being the most popular. However, many other architectures and models have been proposed recently (e.g., attention-based models or Generative Adversarial Networks (GANs)), as this is a very active area of research.
The models applied here use dropouts and regularization to improve the learning process, but it is optional to implement them in this case. As the model capacities are far too large for the limited number of samples, early stopping is essential to prevent overfitting in this case. I stop training when the loss function does not improve for three epochs,[28] and all networks are implemented in *Tensorflow*.

IV-D Convolutional Neural Networks (CNN)

CNNs are very popular in computer vision, where one is processing images or videos (e.g. see Krizhevsky et al. (2012) or Simonyan and Zisserman (2014)). Convolution is done across one or multiple dimensions with meaningful order (e.g. two dimensions for images and an additional time dimension for videos), which means it can also be applied to NLP tasks as the sequence of words has a meaningful order (e.g. see Kalchbrenner et al. (2014)). Despite the fact that CNNs also achieve decent results in NLP tasks, they are not the quantum leap they were in computer vision. Other models and network architectures achieve similar performance like the RNNs from the next section. I use a very basic CNN setup and implement dropouts and regularization which help to improve the learning process. Table 4 provides a detailed overview of the applied CNN's architecture and the used parameters.

IV-E Recurrent Neural Networks (RNN)

RNNs are the most popular neural network type for sequence data. Unlike CNNs, which were not intentionally developed for sequence data, RNNs were designed for situations where past observations or inputs matter for the current output. These one-directional RNNs were extended to bidirectional RNNs for sequences where one observation not only depends on its predecessors but also on its successors, which makes them more adequate for language tasks. The first RNNs suffered from numerical instability, an issue fixed with the introduction of memory cells like LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Chung

[28] An epoch is when the model used the whole dataset once to update its parameters.

Table 4: CNN Architecture

This table shows the architecture of the applied CNN. The max_tokens hyperparameter is set to 17'000 to drop all words occurring less than roughly 100 times in the dataset. This prevents the model from overfitting on less frequent words and reduces the model size significantly. The dimensionality of the vector representation is set to four. The kernel size of the first convolutional layers determines how many words can be reached per filter, and by stacking two convolutional layers, the maximum of possible word dependencies is equal to the product of the two kernels (here: $3 \cdot 3 = 9$). The global average pooling layer removes one dimension by averaging across the second dimension (sentence length). Alternatively, one may also use another pooling procedure, like max pooling. One dense layer is then added with 20 units and some L2 regularization to prevent overfitting too fast. The last layer is determined by the labels: the number of units equals the number of labels per example (1), and the activation depends on the type of labels (*sigmoid* for binary classes, *softmax* for multiple classes, and *linear* for numerical). The network is trained by minimizing the binary cross-entropy (classification) or the mean squared error (regression) using the *Adam* optimizer (learning rate=0.001, beta 1=0.9, beta 2=0.999, epsilon=1e-7, and amsgrad=False) and the mini-batch size is 4'096.

Layer	Parameters	Output dimension	Number of parameters
TextVectorization	max_tokens=17'000, output_dim=4	None, None, 4	68'000
Conv1D	filters=24, kernel_size=3, padding='same', activation='relu'	None, None, 24	312
Conv1D	filters=12, kernel_size=3, padding='same', activation='relu'	None, None, 12	876
GlobalAveragePooling1D	-	None, 12	0
Dense	units=20, activation='relu'	None, 20	220
Dropout	rate=0.1	None, 20	0
Dense	units=1, activation='linear' / activation='sigmoid'	None, 1	21
		Total number of parameters	69'469

et al., 2014).[29] Table 5 provides a detailed overview of the applied RNN with LSTM units and Table 6 of the applied RNN with GRU units.

Table 5: RNN-LSTM Architecture

This table shows the architecture of the applied RNN with LSTM units. The max_tokens hyperparameter is set to 17'000 to drop all words occurring less than roughly 100 times in the dataset. This prevents the model from overfitting on less frequent words and reduces the model size significantly. The dimensionality of the vector representation is set to four. The model uses ten LSTM units, a dense layer with 20 units, and some L2 regularization to prevent overfitting too fast. The last layer is determined by the labels: the number of units equals the number of labels per example (1), and the activation depends on the type of labels (*sigmoid* for binary classes, *softmax* for multiple classes, and *linear* for numerical). The network is trained by minimizing binary cross-entropy (classification) or the mean squared error (regression) using the *Adam* optimizer (learning rate=0.001, beta 1=0.9, beta 2=0.999, epsilon=1e-7, and amsgrad=False) and the mini-batch size is 4'096.

Layer	Parameters	Output dimension	Number of parameters
TextVectorization	max_tokens=17'000, output_dim=4	None, None, 4	68'000
Bidirectional	LSTM(units=10)	None, 20	1'200
Dense	units=20, activation='relu', kernel_regularizer=(L2=0.001)	None, 20	420
Dropout	rate=0.1	None, 20	0
Dense	units=1, activation='linear' / activation='sigmoid'	None, 1	21
		Total number of parameters:	69'641

IV-F Evaluation

I divide the dataset into a training (70% (3'109'125 examples), 01.01.1996-31.03.2016), a validation (15% (666'241 examples), 31.03.2016-26.04.2018), and a test dataset (15% (666'241

[29] Other memory cells have been proposed, but to this day, LSTM and GRU are the most popular choices.

Table 6: RNN-GRU Architecture

This table shows the architecture of the applied RNN with GRU units. The max_tokens hyperparameter is set to 17'000 to drop all words occurring less than roughly 100 times in the dataset. This prevents the model from overfitting on less frequent words and reduces the model size significantly. The dimensionality of the vector representation is set to four. The model uses ten GRU units, a dense layer with 20 units, and some L2 regularization to prevent overfitting too fast. The last layer is determined by the labels: the number of units equals the number of labels per example (1), and the activation depends on the type of labels (*sigmoid* for binary classes, *softmax* for multiple classes, and *linear* for numerical). The network is trained by minimizing binary cross-entropy (classification) or the mean squared error (regression) using the *Adam* optimizer (learning rate=0.001, beta 1=0.9, beta 2=0.999, epsilon=1e-7, and amsgrad=False) and the mini-batch size is 4'096.

Layer	Parameters	Output dimension	Number of parameters
TextVectorization	max_tokens=17'000, output_dim=4	None, None, 4	68'000
Bidirectional	GRU(units=10)	None, 20	960
Dense	units=20, activation='relu', kernel_regularizer=(L2=0.001)	None, 20	420
Dropout	rate=0.1	None, 10	0
Dense	units=1, activation='linear' / activation='sigmoid'	None, 1	21
		Total number of parameters:	69'401

examples), 26.04.2018-31.12.2020). The training dataset is used to train the models, the validation dataset to estimate the generalization ability when deciding on model architecture and for hyperparameter optimization, and the test dataset to estimate the true generalization. Splitting the dataset without shuffling has its advantages and drawbacks. The main reason for doing so is to prevent the claim that there was predictability in the earlier time period of the dataset and that it has vanished nowadays because markets have become more efficient. Another reason is that there may still be some highly similar headlines in the dataset, which should be considered duplicates. This is only a minor issue if one does not shuffle the dataset before splitting it because the duplicates will be in the same dataset split. If this were not the case, there would be information leakage between the training, validation, and test dataset, resulting in too-optimistic error estimates. However, it also results in the fact that the training, validation, and test dataset differ in many aspects. As shown previously, the total news coverage (Figure 2 (a)) and the coverage per firm (Figure 2 (b)) changed over time. With the burst of the dot-com bubble and the financial crisis, the training dataset covers two major market downturns; In contrast, the validation dataset covers a relatively calm period of increasing stock prices. Although stock market indices increased over the time of the test dataset, it was a much more volatile period (e.g., U.S.-China tariffs, interest rate hikes), resulting in negative cumulative returns around news publications. Table 7 provides the descriptive statistics for the training, validation, and test dataset.

I report the accuracy of the classification models and for the regression models also their Mean Squared Error (MSE) and Mean Absolute Error (MAE). The accuracy is computed according to Equation 2:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{2}$$

Table 7: Descriptive Statistics of Training, Validation, and Test Dataset

This table shows the descriptive statistics of the training, validation, and test dataset. The training dataset consists of 3'109'125 observations (70%) between 01.01.1996 and 31.03.2016, the validation dataset of 666'241 observations (15%) between 31.03.2016 and 26.04.2018, and the test dataset of 666'241 observations (15%) between 26.04.2018 and 31.12.2019. Although it may seem redundant to show standard deviation and variance (of the cumulative returns of days 0 and 1), it is relevant to distinguish the two when comparing the model results. The cumulative return of days 0 and 1 is the label (either directly in the regression models or classified to zero and one for the classification models) the models are trained to predict. The realizable return is the return of the trading day after the news publication (day 1) computed as the difference between the opening and the closing price.

	Training dataset	Validation dataset	Test dataset
Standard deviation	7.09%	7.91%	9.27%
Variance	50.27%	62.63%	85.87%
Mean cumulative return day 0 and 1	0.13%	0.06%	−0.07%
Mean realizable return day 1	−0.00%	−0.09%	−0.11%
Positive returns (> 0)	1'602'479 (51.54%)	345'332 (51.83%)	342'941 (51.47%)
Negative returns (≤ 0)	1'506'646 (48.56%)	320'909 (48.12%)	323'300 (48.53%)
Number of observations	3'109'125	666'241	666'241

where TP/TN is the number of true positive/negative predictions, and FP/FN is the number of false positive/negative predictions. Therefore, accuracy is simply the fraction of correct predictions. Cumulative mean returns of the trading day after the news publication and the previous trading day (days 0 and 1) of all positive and negative predictions are also provided for all models. Furthermore, I estimate the realizable returns if one reacted on the trading day after the news is published (difference between the opening and closing price of day 1). To check for any anticipation and post-news drift or reversal, I investigate the cumulative returns for the previous 21 (one month) and the next 63 trading days (three months).

V Results

Table 8 shows the results of all models. The classification according to the Loughran and McDonald (2011) sentiment dictionary, has an accuracy between 53.2% and 54.1% among the three datasets. Because there are no parameters to be optimized for this model, generalization is no issue, and the best estimation for the true accuracy would be across the whole dataset (53.4%). As there are only 2'689 words in the dictionary, only 33.4% of all the headlines can be classified, and due to the imbalance of the sentiment dictionary (2'335 negative vs. 354 positive words), only 5.7% are classified as positive vs. 27.6% as negative. The returns of the classified news are clearly in line with the measured sentiment, but the realizable returns on the next trading are not. It is worth noting that while the fractions of positive news and words are relatively stable among the three datasets, the validation dataset, which does not contain any major crisis, has significantly fewer negative words than the other two datasets.

In contrast to the sentiment dictionary classification, the Naïve Bayes classifier and all neural networks must classify all news. While this makes the task much harder, they can use the 17'000 most frequent words compared to predefined 2'689 words. The Naïve Bayes classifier reaches an accuracy of 54.7% on the test dataset and tends towards classifying positive news because returns are positive on average (51.5% in the training dataset). The returns are according to their classification, and the model generalizes well to the validation and test dataset. While for the sentiment dictionary classification, the realizable returns of the trading day after the news publication are not according to the predictions, they are for the Naïve Bayes classifier with a return difference between positive and negative predictions of 0.15% on the test dataset.

All neural networks have a slightly higher accuracy than the Naïve Bayes classifier, ranging between +0.12% and +0.54% on the test dataset. They all generalize well because of early stopping while training them (otherwise, the networks all start to overfit very soon as the fitting capacity of the models is too large for the used training dataset). The similar results of the different neural networks, and the fact that they are similar to the Naïve Bayes classifier, indicate that almost all predictive power comes from the words. Therefore, the networks learn reasonable embeddings for predicting returns but only minimal relationships between the words. It is therefore also not relevant whether a CNN, an RNN with LSTM, or with GRU units is used. One key observation is that the proportion of positive and negative predictions can vary significantly without impacting the accuracy. Different training runs with the same model specifications of all neural networks result in the fraction of positive predictions varying between around 45% and around 75%. This leads to the conclusion that the models can learn to classify about 25% as negative news with relatively decent certainty and that a fraction of positive news is also well-classifiable. However, at least about 30% of the news headlines are more or less randomly classified.

Figure 5 shows the cumulative sum of excess returns to the predictions of the different models for the 21 trading days (one month) prior to the news publication and the 63 trading days (three months) afterward. There is clear evidence, that markets start to steadily anticipate the news at least one month before publication, and by the time the news is published, part of the impact has already been priced in. Markets then react within less than two trading days, and there is only minimal change afterward. The missing anticipation and slight reversal of positive news of the sentiment dictionary's predictions is most likely due to the small sample size (40'429 observations). As stated above, no parameter optimization is involved in the application of the sentiment dictionary; hence, the best estimation is across the whole dataset. Figure 6 confirms that there is anticipation of good news and indeed only very minimal reversal, if any, across the whole dataset.

Although it would be possible to also distinguish the "certainty" for the sentiment dictionary by considering that the sum of positive minus the sum of negative words is larger than one or smaller than minus one for some headlines, it is very rarely the case. Therefore, regarding the sentiment dictionary, I only report how the returns of the two classes develop. For all other models, I distinguish between five different levels of certainty (classification) / thresholds (regression) for each class. The dynamics are the same for all subsets (e.g., no under- or overreaction to the best news). However, the subsets with the highest certainty of being negative according to the Naïve Bayes classifier, the RNN-LSTM, and the RNN-GRU networks are a bit odd in the sense that the anticipation does not seem to be in line with the other models.

The certainty scores (classification) or point estimates (regression) also allow for a more adequate comparison between the sentiment dictionary and the machine learning algorithms. Table 9 shows the performance of each model on the 40'429 most positive and the 192'563 most negative predictions. While all machine learning models also reach a higher accuracy on the negative predictions by 2.6 to 4.1 percentage points, they crush the sentiment dictionary's accuracy on the positive predictions by 6.2 to 14.9 percentage points. The Naïve Bayes classifier's performance is close to the neural networks' on the positive predictions; however, it is significantly below the neural networks on the most positive predictions. The accuracies of the different levels in Figure 6 also indicate that the neural networks compute superior certainty scores or point estimates. E.g., the 58'682 positive predictions with the highest certainty of the Naïve Bayes classifier have an accuracy of 61.9%, while the CNN classification has a higher accuracy (64.4%) on more predictions (85'249). The same seems to be the case for the negative predictions: the 60'238 most negative predictions of the Naïve Bayes classifier have an accuracy of 63.8%, and the CNN classification has an accuracy of 65.8% on the 60'498 most negative predictions. However, an exact comparison is difficult because either the level/threshold or the number of predictions of such a comparison is chosen arbitrarily.

[30]The accuracy of the positive predictions is also called precision or positive predictive value, and the accuracy of the negative predictions is also called negative predictive value.

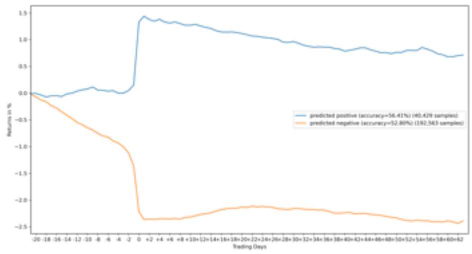
Table 8: Model Results

This table summarizes the performance of the different models. For the regression models, the Mean Absolute Error (MAE) is compared to the standard deviation of the dataset, while the Mean Squared Error (MSE) is compared to the variance of the dataset. n positive and n negative are the number of positive/negative predictions. Return positive/return negative are the mean cumulative returns of days 0 and 1 (on which the models were trained) of the positive/negative predictions, and the next column shows their difference. Real. ret. positive and real. ret. negative are the mean returns of day 1 (computed based on the difference between the opening and closing price of that day) of the positive/negative predictions, and the next column shows their difference. Accuracy is measured based on the cumulative returns of days 0 and 1.

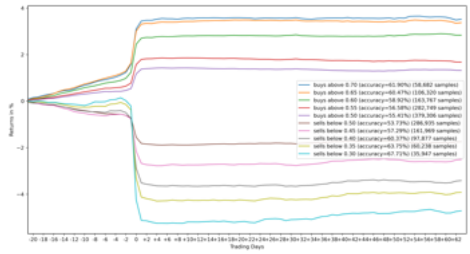
Model	Dataset	MAE dataset std.	MSE dataset var.	n positive	n negative	Return positive	Return negative	Return Δ	Real. ret. positive	Real. ret. negative	Real. ret. Δ	Accuracy
Sentiment dictionary (classification)	Train:	-	-	171'590 (5.52%)	875'074 (28.15%)	1.08%	−0.57%	1.65%	−0.01%	−0.00%	−0.00%	53.21%
	Validation:	-	-	42'051 (6.31%)	159'575 (23.95%)	1.50%	−0.96%	2.46%	−0.17%	−0.08%	−0.09%	54.06%
	Test:	-	-	40'429 (6.07%)	192'563 (28.90%)	1.17%	−1.08%	2.25%	−0.21%	−0.11%	−0.10%	53.43%
Naïve Bayes (classification)	Train:	-	-	1'736'473 (55.85%)	1'372'652 (44.15%)	0.94%	−0.88%	1.82%	0.10%	−0.14%	0.24%	56.92%
	Validation:	-	-	389'196 (58.42%)	277'045 (41.58%)	0.74%	−0.89%	1.63%	−0.05%	−0.14%	0.09%	54.89%
	Test:	-	-	379'306 (56.93%)	286'935 (43.07%)	0.77%	−1.18%	1.95%	−0.05%	−0.20%	0.15%	54.69%
CNN (classification)	Train:	-	-	1'750'025 (56.29%)	1'359'100 (43.71%)	0.94%	−0.90%	1.84%	0.09%	−0.12%	0.21%	56.81%
	Validation:	-	-	378'015 (56.74%)	288'226 (43.26%)	0.91%	−1.04%	1.95%	−0.05%	−0.13%	0.08%	55.41%
	Test:	-	-	360'304 (54.08%)	305'937 (45.92%)	0.90%	−1.22%	2.12%	−0.04%	−0.19%	0.15%	54.96%
RNN-LSTM (classification)	Train:	-	-	1'922'891 (61.85%)	1'186'234 (38.15%)	0.89%	−1.09%	1.98%	0.09%	−0.16%	0.25%	57.31%
	Validation:	-	-	408'446 (61.21%)	257'795 (38.69%)	0.81%	−1.11%	1.92%	−0.05%	−0.13%	0.08%	55.28%
	Test:	-	-	396'256 (59.48%)	269'985 (40.24%)	0.77%	−1.31%	2.08%	−0.05%	−0.20%	0.15%	54.83%
RNN-GRU (classification)	Train:	-	-	2'084'649 (67.05%)	1'024'436 (32.95%)	0.79%	−1.18%	1.97%	0.08%	−0.17%	0.25%	56.86%
	Validation:	-	-	459'402 (68.95%)	206'839 (31.05%)	0.75%	−1.46%	2.31%	−0.06%	−0.15%	0.09%	55.94%
	Test:	-	-	455'508 (68.37%)	210'733 (31.63%)	0.71%	−1.77%	2.48%	−0.06%	−0.23%	0.17%	55.43%
CNN (regression)	Train:	4.19 (7.09)	47.56 (50.27)	1'989'906 (64.00%)	1'119'219 (36.00%)	0.92%	−1.26%	2.18%	0.08%	−0.16%	0.24%	56.14%
	Validation:	4.55 (7.91)	55.27 (62.63)	420'971 (63.19%)	245'270 (36.81%)	0.79%	−1.19%	1.98%	−0.08%	−0.10%	0.02%	55.27%
	Test:	5.22 (9.27)	81.83 (85.87)	407'573 (61.18%)	258'668 (38.82%)	0.78%	−1.41%	2.19%	−0.04%	−0.23%	0.19%	54.81%
RNN-LSTM (regression)	Train:	4.12 (7.09)	47.14 (50.27)	1'897'323 (61.02%)	1'211'802 (38.98%)	0.94%	−1.12%	2.06%	0.08%	−0.14%	0.22%	55.86%
	Validation:	4.45 (7.91)	58.24 (62.63)	428'432 (64.31%)	237'809 (35.69%)	0.83%	−1.23%	2.05%	−0.07%	−0.10%	0.03%	55.49%
	Test:	5.14 (9.27)	80.66 (85.87)	409'569 (61.47%)	256'672 (38.53%)	0.57%	−1.91%	2.48%	−0.07%	−0.19%	0.12%	54.98%
RNN-GRU (regression)	Train:	4.15 (7.09)	47.27 (50.27)	1'788'792 (57.53%)	1'320'333 (42.47%)	0.98%	−1.01%	1.99%	0.09%	−0.13%	0.22%	55.69%
	Validation:	4.49 (7.91)	58.70 (62.63)	433'966 (65.14%)	232'275 (34.86%)	0.80%	−1.30%	2.10%	−0.08%	−0.10%	0.02%	55.30%
	Test:	5.17 (9.27)	85.87 (85.87)	414'144 (62.16%)	252'097 (37.84%)	0.78%	−1.47%	2.25%	−0.07%	−0.19%	0.12%	54.81%

Figure 5: Cumulative Mean Excess returns on Positive/Negative News

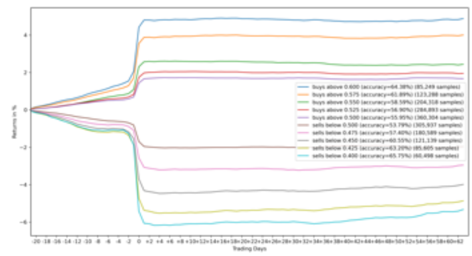
This figure shows the cumulative sum of the mean excess returns (in excess of the average daily returns) of positive and negative news according to the indicated model for the 21 trading days (one month) prior to the news publication and the 63 trading days (three months) afterward of the test dataset. There is a clear spike on the publication day, where the stock price changes according to the news. There is still a clearly different reaction the day after the news is published because markets may not be able to react on the news publication day due to markets already being closed. Markets start anticipating the news at least one month before, and by the time the news is published, part of the impact has already been priced in. Markets then react within less than one trading day, and there is no change afterward.



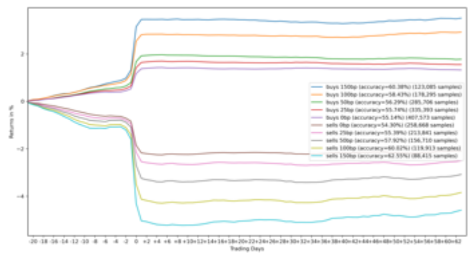
(a) Sentiment dictionary classification



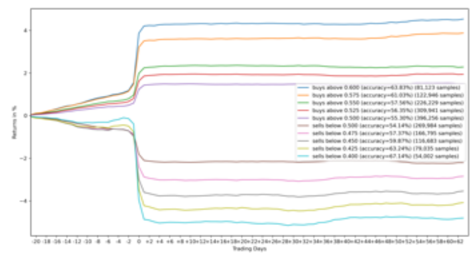
(b) Naïve Bayes classification



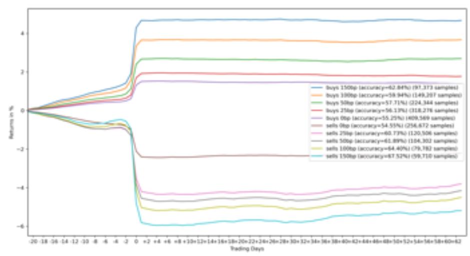
(c) CNN classification



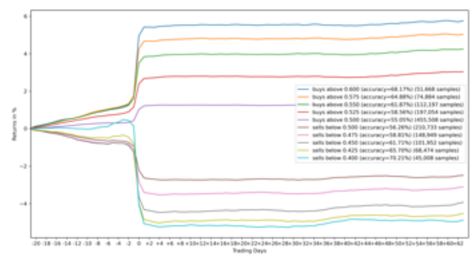
(d) CNN regression



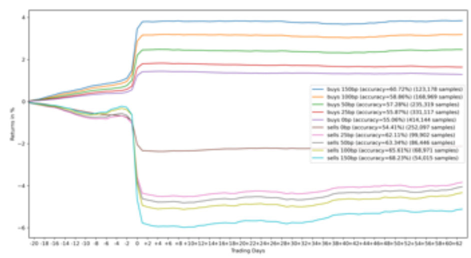
(e) RNN-LSTM classification



(f) RNN-LSTM regression



(g) RNN-GRU classification



(h) RNN-GRU regression

Figure 6: Sentiment Dictionary Cumulative Mean Sum of Excess Returns on Positive/Negative News

This figure shows the cumulative sum of the mean excess returns (in excess of the average daily returns) of positive and negative news according to the sentiment dictionary for the 21 trading days (one month) prior to the news publication and the 63 trading days (three months) afterward of the whole dataset. There is a clear spike on the publication day, where the stock price changes according to the news. There is still a clearly different reaction the day after the news is published because markets may not be able to react on the news publication day due to markets already being closed. Markets start anticipating the news at least one month before, and by the time the news is published, part of the impact has already been priced in. Markets then react within less than one trading day, and there is no change afterward.

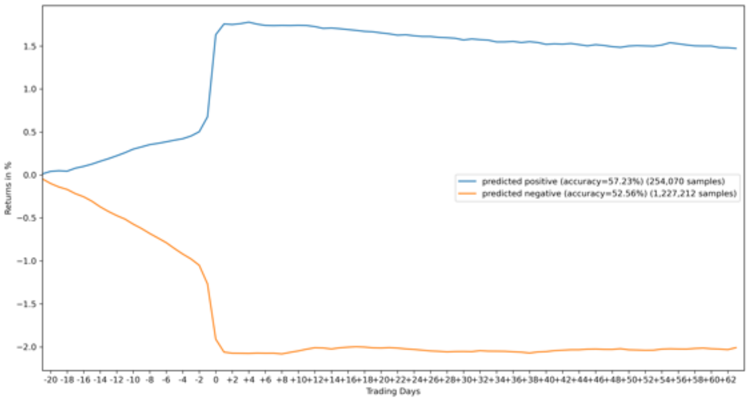


Table 9: Model Comparison to Sentiment Dictionary

This table summarizes the performance of the different models on the 40'429 most positive and the 192'563 most negative predictions, which allows for a more adequate comparison between the sentiment dictionary and the supervised learning algorithms. Return positive/return negative are the mean cumulative returns of days 0 and 1 (on which the models were trained) of the positive/negative predictions, and the next column shows their difference. Real. ret. positive and real. ret. negative are the mean returns of day 1 (computed based on the difference between the opening and closing price of that day) of the positive/negative predictions, and the next column shows their difference. Accuracy is measured based on the cumulative returns of days 0 and 1.

Model	Return positive	Return negative	Return Δ	Real. ret. positive	Real. ret. negative	Real. ret. Δ	Accuracy positive[30]	Accuracy negative[30]
Sentiment dictionary (classification)	1.17%	−1.08%	2.25%	−0.21%	−0.11%	−0.10%	56.41%	52.80%
Naïve Bayes (classification)	1.87%	−1.74%	3.61%	0.06%	−0.26%	0.32%	62.66%	56.29%
CNN (classification)	3.95%	−1.95%	5.90%	−0.04%	−0.26%	0.22%	70.10%	56.95%
RNN-LSTM (classification)	3.97%	−1.81%	5.78%	−0.07%	−0.24%	0.17%	69.65%	56.51%
RNN-GRU (classification)	4.11%	−1.92%	6.03%	−0.06%	−0.25%	0.19%	70.39%	56.88%
CNN (regression)	4.20%	−1.84%	6.04%	−0.15%	−0.27%	0.09%	70.09%	54.54%
RNN-LSTM (regression)	4.67%	−1.94%	6.61%	−0.16%	−0.22%	0.06%	71.33%	54.70%
RNN-GRU (regression)	4.61%	−1.89%	6.50%	−0.16%	−0.21%	0.05%	71.15%	54.39%

VI Discussion

As Tetlock (2007) points out, sentiment theory predicts that short-term returns will be (partially) reversed in the long run (since they were an overreaction), whereas information theory predicts that they will persist indefinitely (because they represent the rational price impact). This paper strongly supports the information theory's view and the EMH by showing that financial markets react quickly (less than two days) to news and that there is no clear trend in stock returns afterward. Furthermore, markets seem to start anticipating news at least one month before publication and have priced in part of the impact upon publication. There are multiple potential explanations for this anticipation. News on accounting measures like earnings might be known in advance because companies tend to release information on some key figures before publishing the full financial report. Another source could be that there might be some rumors on which some people trade before they are confirmed or the use of insider information. Brunnermeier (2005) provides a theoretical framework where early informed market participants buy (sell) on positive (negative) rumors and sell (buy) when the news becomes publicly available. He claims that the informed trader can profit twice: first, when he receives the signal early, and second when the news is published because markets overreact. While the first is certainly true, my results indicate that, on a daily basis, the second does not seem to be the case.

Despite this strong evidence for market efficiency, the empirical results presented here do not deny the existence of a profitable trading strategy based on news. While the results of this paper show that a slow trader, who reacts on the trading day after the news is published, could only make a minimal profit per trade (less than 0.10% per trade before transaction costs), they also indicate that the first to react upon publication could make significant profits since most of the price impact is left when the news is published. In order to seriously elaborate on the feasibility of such a strategy, one would have to work with intraday data, which is beyond the scope of this paper. Nevertheless, it is worth discussing it in a more hypothetical manner. Automatically analyzing the news is obviously faster than any human can, and having a reasonable point estimate of the price impact from a regression model would help to take transaction costs as well as the price impact of trades into account. Such a strategy would most likely work for news published during trading hours; however, it may not for news published during off-trading hours because speed is no advantage anymore. Furthermore, any thorough investigation should consider the potentially very limited capacity of such a strategy due to the short time frame and limited liquidity of small firms (although, as Heston and Sinha (2017) (who use a very similar dataset) show, most news is on large firms with the largest 20% accounting for more than half of the news). The fact that the realizable returns of the validation set are significantly lower than the ones of the test dataset should also raise some concerns. As Garcia (2013) points out, daily return predictability based on news is concentrated in recessions, which favors the test dataset because it has negative returns on average. Therefore, these returns may only exist in market downturns because bad news travels slowly (see Hong et al. (2000) and Frazzini (2006)).

While one should appreciate the decent performance, simplicity, and interpretability of the classification with the Loughran and McDonald (2011) sentiment dictionary, there is little room for improvement besides extending the dictionary and using a different weighting scheme. Hence, it is adequate in cases where one really needs to measure sentiment, but supervised learning methods are better suited when many labeled examples are available. Although the Naïve Bayes classifier makes the naïve assumption of words being independent, it should be the first choice for any quick analysis because of its simplicity (the only hyperparameter to choose is the number of words to take into account) and the fact that it already performs decently on most text classification tasks. Besides the limitation that the underlying assumption imposes, the lack of a straightforward regression version of Naïve Bayes, which works well for text, is another disadvantage. Ordinary Least Squares (OLS) and other standard techniques fail in most language applications because of the high dimensionality of the word counts. Gentzkow et al. (2019) discuss several penalized linear possibilities like Least Absolute Shrinkage and Selection Operator (LASSO) (L_1 penalty, see Tibshirani (1996)), ridge (L_2 penalty, see Hoerl and Kennard (1970)), the "elastic net" (a mixture of L_1 and L_2 penalty, see Zou and Hastie (2005)), and the log penalty (see Candes et al. (2008)) which all can be interpreted as posterior maximization under some prior.[31] They further elaborate on other approaches like dimensionality reduction (e.g., Principal Component Regression (PCR) and Partial Least Squares (PLS)) and non-linear models (e.g., Generalized Linear Models (GLMs), SVMs, regression trees, and DNNs). One of the many advantages of DNNs is the fact that the same network can be applied to binary classification, multi-classification as well as regression problems by simply changing the activation function of the output layer (*sigmoid* for binary classification, *softmax* for multi-classification, and *linear* for regression). Nowadays, the almost limitless ways to design DNNs truly make them universal function approximators, and well-suited for high-dimensional problems. Nevertheless, this design freedom is also one of their weaknesses because it can be very time-consuming to find a reasonable architecture, and comparison is more difficult, which is why I applied very basic architectures. This leaves room for improvement, and one may pursue a similar project with new exciting ideas from the deep learning community. While the difference between the Naïve Bayes classifier and the neural networks may look neglectable according to some of the empirical results presented here, they also provide evidence that as long as one does not let the overfitting problem get out of hand, the Naïve Bayes classifier's results are like a lower bound for the neural networks.

Having a third (neutral) category has been proposed to me on several occasions, but it is not sensible from my point of view. First, it would require an arbitrary choice for the return threshold between these three categories instead of the logical choice of zero for separating positive from negative. Then, one would have to decide whether to train the models to predict three classes or to drop the neutral category and predict the two remaining classes. While the former may sound appealing at first, it does not make sense because the mod-

[31] E.g., ridge regression assumes independent Gaussian priors on each coefficient while LASSO assumes a Laplacian prior (see Park and Casella (2008) and Hans (2009)).

els would have to learn a third category without knowing that all categories are actually aligned across the return dimension, thus making the problem harder than it already is. To make this a bit clearer, consider how the Naïve Bayes classifier works: In the binary case, it computes a weight for each word that indicates if this word is evidence for the positive class (weight >0.5) or the negative class (weight <0.5). With three categories, it would have to compute three weights for each word instead of one (which increases the number of parameters dramatically and makes overfitting a potential threat), meaning that each weight is only the evidence for each class and not simultaneously evidence against any other class like in the binary case. On the other hand, dropping the neutral category would mean losing training samples (potentially a lot, depending on the threshold choice), which is only beneficial if it substantially decreases the noise in the data; however, I doubt that this would be the case.

Whilst the implementation of a third category in the training process does not make much sense due to the aforementioned reasons, the thresholds implemented in Figure 6 effectively create a third (neutral) category and are highly recommended if one were to trade based on such a model. Because trading means facing a choice between three options (buying, selling, or doing nothing), one should have a third category based on the certainty or point estimate because they allow taking transaction costs and risk aversion into account.

In my opinion, the two most promising approaches for enhancing the neural networks' performance would be to use pre-trained embeddings (e.g., BERT) or the implementation of auxiliary loss(es). The idea of auxiliary loss was introduced with *GoogleLeNet* (Szegedy et al., 2014) and means requiring one or multiple layers of a network to predict the same as the output layer and adding the loss metrics of these predictions, scaled by some factor less than one, to the loss of the final prediction. Initially, the intention was to improve the optimization process by increasing the gradient signal; however, this issue is nowadays usually solved by adding batch normalization layers. Nevertheless, the idea of using auxiliary losses to improve the feature extraction of hidden layers stuck, but hidden layers are now usually asked to predict something different from the final layer, e.g., in NLP, the word before and after (skip-gram model) to train the embeddings. The difference to pre-training is that the learning of the features/embeddings takes both the auxiliary loss and the final loss into account, while pre-trained embeddings would be learned independently of the final task.

Because new data is generated every day, the training dataset could double within less than eight years at the current speed of 35'000 news per month (see Figure 2 (a)). Discussing the effect of a larger training sample is therefore essential. While the sentiment dictionary's predictions would not profit, the Naïve Bayes classifier could close its generalization gap.[32] However, to improve the models' performance on the training dataset, the additional observations would have to be more predictable than the ones used before, which is unlikely.[33]

[32] The generalization gap is the difference between the models' performance on the training and the test dataset.

[33] This is the case because the Naïve Bayes classifier always computes the optimal weights on the training dataset and now has to fit a larger number of observations with the same number of parameters.

On the other hand, neural networks would benefit a lot from a larger training sample. Unlike for the Naïve Bayes classifier, the generalization gap would most likely be similar since it is mainly controlled by the early stopping (which prevents the optimization algorithm from reaching a local minimum). Since a larger training set would prevent the model from overfitting so soon, it would lead to better performance on the training set while keeping the generalization gap more or less constant. Hence, neural networks are the superior models as more and more data becomes available.

VII Conclusion

This paper shows that the stock market's reaction to new information is predictable. The words of the Loughran and McDonald (2011) sentiment dictionary already have some predictive power, but using this dictionary approach has several drawbacks. First, for short texts, it fails to classify a large fraction because they do not contain any of the dictionary's words. Second, its performance is limited, there is no room for improvement, and it cannot capitalize on more data. The Naïve Bayes classifier machine learning algorithm overcomes the first problem because it considers the most frequent words of the corpus it is fed, rather than predefined words of a dictionary, and can therefore classify all texts. It slightly outperforms the Loughran and McDonald (2011) sentiment dictionary already when comparing the accuracy of all its predictions and does so substantially when comparing the same number of predictions. However, there is also limited room for improvement, and additional data would not lead to substantial improvements. Neural networks, which additionally can model the relationship between words, even reach slightly higher accuracies than the Naïve Bayes classifier. Nevertheless, the very basic implementations of this paper could be improved substantially, and these models would benefit the most from additional data.

Although I show that a slow trader only starting to trade on the *Reuters* news the day after it is published could potentially still make some profits (less than 0.10% per trade before transaction costs), the cumulative mean returns of the 21 trading days (one month) before and 63 trading days (three months) after the publication of news provide strong evidence for rational behavior and the market efficiency hypothesis: Markets start to anticipate the news about one month before and have priced in part of the impact upon publication. The news is then quickly digested (less than two days), and there is no clear trend afterward. This contradicts the sentiment theory, which would predict overreaction and then reversal afterward.

Appendix Essay 1

Table A.1: 30 Most Frequent Words

This table shows the 30 most frequent words and their frequency.

Word	Frequency
to	1'498'475
inc	1'105'562
of	933'363
q	780'691
on	681'764
in	635'637
nyse	598'224
imbalance	583'041
shares	577'752
says	573'011
mln	494'075
from	492'698
corp	476'050
shr	439'966
for	417'868
order	411'535
buy	400'707
and	398'389
side	391'139
brief	390'626
price	359'793
target	350'768
the	306'472
s	284'365
sell	280'766
view	255'381
co	253'078
raises	250'985
us	233'075
pct	231'359

Table A.2: Extract of Dataset

This table illustrates to the reader what the dataset looks like.

Index	Headline	Date	CUSIP 8	Next trading day	Return next trading day
0	indian phone privatisation plan hits fresh snag	1996-01-01	00206R10	1996-01-02	0.019651
1	"toy story" back on top of hollywood box office	1996-01-01	25468710	1996-01-02	0.033970
2	gecc <ge.n> sets $300 million five-year eurobond	1996-01-02	36960410	1996-01-03	0.001701
3	cigarette makers claim fda exceeds authority - wsj	1996-01-02	02209S10	1996-01-03	0.001361
4	gecc <ge.n> prices $300 million five-year eurobond	1996-01-02	36960410	1996-01-03	0.001701
5	qatar to name partner for major petchem venture	1996-01-02	67459910	1996-01-03	−0.017341
6	gilead science<gild.o>says phase i topical ophthalmic cidofovir study started	1996-01-02	37555810	1996-01-03	0.042146
7	goldinger portfolio losses seen at $100 mln -paper	1996-01-02	72027950	1996-01-03	0.053763
8	at&t <t.n> says to take after-tax charge of about $4 bln in q4	1996-01-02	00206R10	1996-01-03	0.014989
9	at&t <t.n> sees about 30,000 of workforce cuts being involuntary	1996-01-02	00206R10	1996-01-03	0.014989
10	at&t <t.n> says new jersey to see reductions of 6,000-7,000 jobs	1996-01-02	00206R10	1996-01-03	0.014989
⋮	⋮	⋮	⋮	⋮	⋮
4542971	s&w seed co - new credit facility replaces an existing facility	2019-12-30	78513510	2019-12-31	0.055276
4542972	brief-glacier bancorp declares special dividend	2019-12-30	37637Q10	2019-12-31	0.006126
4542973	s&w seed co - entered into 3-yr extension with rooster capital on $9.3 mln real...	2019-12-30	78513510	2019-12-31	0.055276
4542974	s&w seed co - new maturity date for note is nov 30, 2022	2019-12-30	78513510	2019-12-31	0.055276
4542975	brief-nw natural files oregon general rate case	2019-12-30	66765N10	2019-12-31	0.008756
4542976	brief-s&w secures new $35 million working capital facility with cibc	2019-12-30	78513510	2019-12-31	0.055276
4542977	update 2-lockheed martin hits 2019 f-35 delivery target of 131 jets	2019-12-30	53983010	2019-12-31	−0.006430
4542978	brief-lockheed martin delivers 134 f-35 aircrafts in 2019	2019-12-30	53983010	2019-12-31	−0.006430
4542979	microsoft says north korea-linked hackers stole sensitive information	2019-12-30	59491810	2019-12-31	0.000698
4542980	huawei's 2019 revenue to jump 18%, forecasts "difficult" 2020	2019-12-30	02079K30	2019-12-31	−0.000239

Table A.3: Headlines Before and After Preprocessing

This table shows a random sample of 20 headlines before and after preprocessing to illustrate how much changes.

Index	Before preprocessing	After preprocessing
200214	newfield exploration <nfx.n> says has dry hole offshore australia	newfield exploration says has dry hole offshore australia
290235	norfolk southern <nsc.n> q3 revs $1.51 bln vs $1.54 bln	norfolk southern revs bln vs bln
411856	ford unit plans minimum 1.0 bln euro 2009 bond at swaps +200 bps area wed -banker	ford unit plans minimum bln euro bond at swaps bps area wed banker
552068	update 1-uk's sainsbury q4 sales up, confident on targets	update uk sainsbury sales up confident on targets
568464	u.s. stocks seen flat with focus back on earnings	stocks seen flat with focus back on earnings
876133	loral space sued by hedge fund over sale of preferred stock	loral space sued by hedge fund over sale of preferred stock
1132978	nyse order imbalance <duk.n> 241600 shares on sell side	nyse order imbalance shares on sell side
1199295	nyse order imbalance <nwl.n> 128700 shares on buy side	nyse order imbalance shares on buy side
1299345	lsb industries inc <lxu.n> sees q4 pre-tax income, although positive, will be lower	lsb industries inc sees pre tax income although positive will be lower
1364525	s&p equity research cuts conocophillips <cop.n> price target by $4 to $58; rating strong buy	equity research cuts conocophillips price target by to rating strong buy
1890003	auto alert - strattec security corp <strt.o> q3 shr $0.02	auto alert strattec security corp shr
1973201	brief-research alert-jefferies cuts american axle price target	brief research alert jefferies cuts american axle price target
2019934	update 1-cognex q3 profit beats, sees q4 rev below street view	update cognex profit beats sees rev below street view
2072962	covenant transportation group inc <cvti.o> q4 revenue $162 mln	covenant transportation group inc revenue mln
2328036	research alert-microstrategy: fbr raises to outperform - theflyonthewall.com	research alert microstrategy fbr raises to outperform theflyonthewall com
2363103	research alert-delta air lines: jp morgan raises price target	research alert delta air lines jp morgan raises price target
2485972	resolute forest posts quarterly loss due to one-time tax charge	resolute forest posts quarterly loss due to one time tax charge
2560588	research alert-cenovus energy : firstenergy capital cuts target price	research alert cenovus energy firstenergy capital cuts target price
3168685	cohen & steers announces preliminary assets under management february 29, 2016	cohen steers announces preliminary assets under management february
3557877	brief-travelzoo says holger bartel left co's board of directors	brief travelzoo says holger bartel left co board of directors

References

Aizawa, A., 2003. An information-theoretic perspective of tf–idf measures. Information Processing & Management 39, 45 – 65.

Antweiler, W., Frank, M. Z., 2004. Is all that talk just noise? the information content of internet stock message boards. The Journal of Finance 59, 1259–1294.

Atzeni, M., Dridi, A., Reforgiato Recupero, D., 2017. Fine-grained sentiment analysis on financial microblogs and news headlines. In: *Semantic Web Challenges: 4th SemWebEval Challenge at ESWC 2017, Portoroz, Slovenia, May 28-June 1, 2017, Revised Selected Papers*, Springer, pp. 124–128.

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. Journal of Finance 61, 1645–1680.

Biddle, G. C., Hilary, G., Verdi, R. S., 2009. How does financial reporting quality relate to investment efficiency? Journal of Accounting and Economics 48, 112–131.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Brunnermeier, M. K., 2005. Information leakage and market efficiency. The Review of Financial Studies 18, 417–457.

Candes, E. J., Wakin, M. B., Boyd, S. P., 2008. Enhancing sparsity by reweighted l 1 minimization. Journal of Fourier Analysis and Applications 14, 877–905.

Cauchy, A., 1847. Méthode générale pour la résolution des systemes d'équations simultanées. Comp. Rend. Sci. Paris 25, 536–538.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 .

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R., 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41, 391–407.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805.

Dyer, T., Lang, M., Stice-Lawrence, L., 2017. The evolution of 10-k textual disclosure: Evidence from latent dirichlet allocation. Journal of Accounting and Economics 64, 221–245.

Fama, E. F., 1970. Efficient capital markets: A review of theory and empirical work. The Journal of Finance 25, 383–417.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.

Frazier, K. B., Ingram, R. W., Tennyson, B. M., 1984. A methodology for the analysis of narrative accounting disclosures. Journal of Accounting Research pp. 318–331.

Frazzini, A., 2006. The disposition effect and underreaction to news. The Journal of Finance 61, 2017–2046.

Garcia, D., 2013. Sentiment during recessions. The Journal of Finance 68, 1267–1300.

Gentzkow, M., Kelly, B., Taddy, M., 2019. Text as data. Journal of Economic Literature 57, 535–74.

Ghag, K. V., Shah, K., 2015. Comparative analysis of effect of stopwords removal on sentiment classification. In: *2015 International Conference on Computer, Communication and Control (IC4)*, pp. 1–6.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, http://www.deeplearningbook.org.

Goulden, R., Nation, P., Read, J., 1990. How Large Can a Receptive Vocabulary Be? Applied Linguistics 11, 341–363.

Graves, A., 2012. Supervised sequence labelling. In: *Supervised sequence labelling with recurrent neural networks*, Springer, pp. 5–13.

Grossman, S. J., Stiglitz, J. E., 1980. On the impossibility of informationally efficient markets. The American Economic Review 70, 393–408.

Hans, C., 2009. Bayesian lasso regression. Biometrika 96, 835–845.

Heston, S. L., Sinha, N. R., 2017. News vs. sentiment: Predicting stock returns from news stories. Financial Analysts Journal 73, 67–83.

Hilbert, M., López, P., 2011. The world's technological capacity to store, communicate, and compute information. Science 332, 60–65.

Hillert, A., Jacobs, H., Müller, S., 2014. Media makes momentum. The Review of Financial Studies 27, 3467–3501.

Hinton, G. E., Osindero, S., Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural computation 18, 1527–1554.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Hoerl, A. E., Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42, 177–196.

Hong, H., Lim, T., Stein, J. C., 2000. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. The Journal of Finance 55, 265–295.

Huang, A. H., Lehavy, R., Zang, A. Y., Zheng, R., 2018. Analyst information discovery and interpretation roles: A topic modeling approach. Management Science 64, 2833–2855.

Huang, A. H., Zang, A. Y., Zheng, R., 2014. Evidence on the information content of text in analyst reports. The Accounting Review 89, 2151–2180.

Jegadeesh, N., Wu, D., 2013. Word power: A new approach for content analysis. Journal of Financial Economics 110, 712–729.

Jones, K. S., 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation .

Jones, M. J., Shoemaker, P. A., 1994. Accounting narratives: A review of empirical studies of content and readability. Journal of Accounting Literature 13, 142.

Kalchbrenner, N., Grefenstette, E., Blunsom, P., 2014. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188 .

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105.

Lawrence, A., 2013. Individual investors and financial disclosure. Journal of Accounting and Economics 56, 130–147.

Lehavy, R., Li, F., Merkley, K., 2011. The effect of annual report readability on analyst following and the properties of their earnings forecasts. The Accounting Review 86, 1087–1115.

Leibniz, G. W., 1684. Nova methodus pro maximis et minimis, itemque tangentibus, quae nec fractas, nec irrationales quantitates moratur, et singulare pro illis calculi genus. Acta eruditorum pp. 467–473.

L'Hôpital, G. F. A., 1696. Analyse des infiniment petits pour l'intelligence des lignes courbes. n/a.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics 45, 221–247.

Li, F., 2010. The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. Journal of Accounting Research 48, 1049–1102.

Lintner, J., 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. The Review of Economics and Statistics 47, 13–37.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance 66, 35–65.

Loughran, T., McDonald, B., 2016. Textual analysis in accounting and finance: A survey. Journal of Accounting Research 54, 1187–1230.

Luhn, H. P., 1957. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development 1, 309–317.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., Takala, P., 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology 65, 782–796.

McCulloch, W. S., Pitts, W., 1988. Neurocomputing: Foundations of research. ch. A Logical Calculus of the Ideas Immanent in Nervous Activity pp. 15–27.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .

Mikolov, T., Yih, W.-t., Zweig, G., 2013b. Linguistic regularities in continuous space word representations. In: *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751.

Miller, B. P., 2010. The effects of reporting complexity on small and large investor trading. The Accounting Review 85, 2107–2143.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., Trajanov, D., 2020. Evaluation of sentiment analysis in finance: from lexicons to transformers. IEEE Access 8, 131662–131682.

Mossin, J., 1966. Equilibrium in a capital asset market. Econometrica 34, 768–783.

Munková, D., Munk, M., Vozár, M., 2014. Influence of stop-words removal on sequence patterns identification within comparable corpora. In: Trajkovik, V., Anastas, M. (eds.), *ICT Innovations 2013*, Springer International Publishing, Heidelberg, pp. 67–76.

Park, T., Casella, G., 2008. The bayesian lasso. Journal of the American Statistical Association 103, 681–686.

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

Purda, L., Skillicorn, D., 2015. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. Contemporary Accounting Research 32, 1193–1223.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review 65, 386.

Rosenblatt, F., 1962. Principles of neurodynamics (spartan, new york). Principles of Neurodynamics .

Roser, M., Ritchie, H., 2019. Technological progress. Our World in Data Https://ourworldindata.org/technological-progress.

Rumelhart, D. E., Hinton, G. E., McClelland, J. L., et al., 1986b. A general framework for parallel distributed processing. Parallel Distributed Processing: Explorations in the Microstructure of Cognition 1, 26.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., 1986a. Learning internal representations by error propagation. Nature 323, 533–536.

Schütze, H., Manning, C. D., Raghavan, P., 2008. Introduction to information retrieval, vol. 39. Cambridge University Press Cambridge.

Sharpe, W. F., 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. Journal of Finance 19, 425–442.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2014. Going deeper with convolutions. CoRR abs/1409.4842.

Tetlock, P. C., 2007. Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance 62, 1139–1168.

Tetlock, P. C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. The Journal of Finance 63, 1437–1467.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 267–288.

Uhl, M. W., 2014. Reuters sentiment and stock returns. Journal of Behavioral Finance 15, 287–298.

Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H., Zhao, B. Y., 2014. Crowds on wall street: Extracting value from social investing platforms. CoRR abs/1406.1137.

Werbos, P., 1981. Applications of advances in nonlinear sensitivity analysis, systems modeling and optimization. Proceedings of the 70th IFIP .

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.

Essay II:

Microanalysis of the Stock Market's Reaction to News

Jan Pichler*

ABSTRACT

This paper studies the market's micro reaction (second frequency) to more than 2.3 million news events for stocks traded in the U.S. I use the Natural Language Processing (NLP) models from Pichler (2023) to classify the news into positive/negative and show that the market nowadays prices in new information within two to three minutes on average. A trading strategy that starts buying (selling) one second after good (bad) news is released for one minute, then holds the position for one minute before liquidating it generates statistically significant profits. The strategy yields slightly more than one basis point, which is economically highly significant considering the short investment period. However, due to this short period, profits are limited and estimated to be USD 50 to 85 million for the sample period of nine years. Furthermore, I show that volatility is elevated around news and that investors, on average, get a risk premium of 0.05% for holding stock in the previous 6.5 hours and another 0.05% when news is released. Hence, the resolution of uncertainty concerning the information content is compensated with a risk premium, i.e., decreases to zero at the resolution; however, there is no compensation on average for the uncertainty regarding how to interpret the information.

Keywords: News, Information, Natural Language Processing, Machine Learning, Supervised Learning, Market Micro Reaction

JEL Classification Numbers: C38, C45, G12, G14

*University of Bern, Faculty of Business, Economics and Social Sciences. Engehaldenstrasse 4, 3012 Bern, Switzerland. Email: jan.pichler@unibe.ch

I Introduction

Ever since Fama proposed the Efficient Market Hypothesis (EMH) in 1970 (Fama, 1970), its validity has been an open debate among academics. On the one hand, empirical researchers investigate whether public information[1] is correctly reflected in current prices, i.e., has no predictive power for future returns. This gave rise to the anomaly and factor zoo we have today (e.g., see Cochrane (2011); Harvey et al. (2015); Feng et al. (2020)).[2] On the other hand, event studies of the return pattern around the revelation of information are conducted to examine how it gets incorporated into asset prices and whether it is compatible with the EMH.

For most of the research in accounting and finance's history, numerical information like balance sheets, income statements, and cash flow statements were the only data sources for empirical studies. Hence, analyzing the return pattern around earnings announcements was the most popular choice of the event study approach (e.g., see Ball and Brown (1968); Bernard and Thomas (1989, 1990)). The progress in computing and Natural Language Processing (NLP) of the last decades opened the route to also investigate text data sources. Nowadays, NLP is one of the most active research areas in computer science, and progress is mainly driven by large private companies like *Google*, *Microsoft*, *Apple*, and *Meta Platforms*, who each already invest billions per year and are even increasing their spending rapidly. However, this just seems to be the rational thing to do for these companies because the progress in this area of science is just astonishing, and so is the market potential of these technologies. Other fields outside computer science are sailing in the fairway of their progress and are slowly adapting their methods and technologies.

This paper extends the event study line of research by using the NLP models from Pichler (2023) to investigate the micro reaction of financial markets to new information at the second frequency. I use 2.3 million *Reuters* news from between January 2011 and December 2019, each concerning only a single one of 3'552 firms listed at the New York Stock Exchange (NYSE) or the National Association of Securities Dealers Automated Quotations (NASDAQ). Hence, it is the largest dataset ever studied at intraday frequency in the literature and provides crucial empirical evidence on the transmission of information. The only even larger news dataset (almost 4.5 million observations) is the one used in Pichler (2023), which is the same dataset but goes back to 1996 and only considers daily returns.[3] I use the NLP models from Pichler (2023) to classify the news into positive and negative, and the goal here is to investigate how financial markets digest new information and if profitable trading strategies exist after it has been published. In the following, I briefly summarize

[1] Information available to at least most major market participants can be referred to as public information even though not everyone has access to it.

[2] An anomaly refers to an empirical finding inconsistent with the EMH, i.e., a variable with predictive power for future excess returns (in excess of the ones proposed by the assumed asset pricing model). Similarly, a factor also refers to a variable with predictive power for future excess returns. However, they are explained by the additional risk, and the variable should therefore be part of the asset pricing model. Hence, many variables started as anomalies and later became factors when they were related to risk.

[3] Uhl (2014) analyzes 3.6 million news event;however, he investigates the relationship with the aggregated market and not the individual stocks.

the essential concepts and conclusions of Pichler (2023) before elaborating on the contents of this paper.

Sentiment is a key term that refers to two different concepts which, although closely related, are distinct. In NLP and finance, we have to distinguish between *text sentiment* and *market* or *investor sentiment*. While the former refers to the "tone" of the text and is also used in many fields outside of finance, the latter refers to the general "mood" of the investors. The groundbreaking work of Baker and Wurgler (2006) launched a discussion in the finance literature surrounding the role of *investor sentiment* in the stock market. On the other hand, the pioneering studies of Tetlock (2007) and Tetlock et al. (2008) investigate the effect of *text sentiment* in the media on the aggregated stock market using the Harvard psychological dictionary. Garcia (2013) extends the studies of Tetlock by using the finance-specific sentiment dictionary of Loughran and McDonald (2011) and shows that the *text sentiment* in the media is a stronger predictor for the aggregated stock market during recessions.

Although it can also work the other way around, the rationale is generally that *text sentiment* causes (or at least is a good proxy for) *investor sentiment*. Pichler (2023) compares different NLP models' abilities to predict the market reaction to news headlines. While the Loughran and McDonald (2011) sentiment dictionary aims to measure *text sentiment*, it is not unambiguous if the supervised learning models, trained to predict the market reaction, measure *investor sentiment* towards some news or *text sentiment*. Markets also react positively towards news with bad *text sentiment*, which has been expected to be even worse and vice versa. While this is an argument that they measure *investor sentiment*, one has to keep in mind that the models use text as input, hence rely on measuring some form of *text sentiment* which is related to *investor sentiment*. Independent of this "chicken or egg" discussion, Pichler (2023) shows that supervised learning models like the Naïve Bayes classifier or neural networks outperform the sentiment dictionary of Loughran and McDonald (2011). The neural networks' performance in Pichler (2023) is very limited because they are only trained on the news dataset, which is too small to model language appropriately. One should use pre-trained models to increase their performance and fine-tune them to the specific task. While Pichler (2023) also heavily focuses on the technical aspects of the NLP models, this paper is purely dedicated to the financial market aspects. Therefore, it is sufficient for understanding this paper to just see them as a way of classifying news without having any more profound understanding of how they work.

Pichler (2023) provides evidence that the market already anticipates news in the weeks before and then quickly prices in the new information. However, due to the limitation of only using market data at the daily frequency, Pichler (2023) only indicates that some profits are left on the table after new information is published. In this paper, I use intraday Trades and Quotes (TAQ) data from the NYSE and NASDAQ to show the market reaction up to the second frequency and estimate potential profits of trading strategies that react instantly to the news. For comparability, I use the classifications from Pichler (2023) based on the models trained on the daily returns (i.e., without retraining them). This paper's results confirm that the main finding of Pichler (2023) also holds at the intraday level: markets

already anticipate news and price them in quickly. The latest sample period suggests that it nowadays takes markets, on average, two to three minutes to price in new information.
Furthermore, I show that a simple trading strategy, which acquires a position for one minute and holds it for another minute before liquidating it, delivers statistically significant returns. While their economic magnitude is significant considering the investment period, it is limited due to the relatively short investment period of several minutes. Total trading profits for the whole nine years sample period are estimated to be USD 50 to 85 million. Additionally, I indicate that returns before the news also have predictive power for the market's reaction. Measured over a long enough period (i.e., eight hours), a trading strategy based on them is similarly profitable as the NLP models. Results are robust against changes in the parameters of the trading strategies (length of buying or holding period as well as maximum trading amount).
Furthermore, I show that volatility is increased around news and gets compensated with 0.05% in the 6.5 hours before news publication and another 0.05% at news publication. However, the elevated volatility after news publication does not get compensated. One may interpret this as follows: The resolution of uncertainty concerning the information content is compensated with a risk premium, but there is no compensation on average for the uncertainty regarding how to interpret the information.
This paper adds several novel aspects to the existing literature. To my knowledge, it is the first paper that analyzes the stock market's reaction to firm-specific news at the second frequency. Moreover, it does so at large scale with 2.3 million observations. The results contribute to the discussion of the EMH's validity by showing that markets react very efficiently to new information. The observed return pattern is highly consistent with market efficiency for the following reasons: First, markets anticipate the type of news (positive/negative), and prices already move accordingly in the hours before the new information arrives. Second, incorporating the news in prices is estimated to only take two to three minutes on average. And finally, the return pattern of all news, i.e., independent of any NLP model, indicates that there is a risk premium for holding stocks in the hours before new information is expected. The tested trading strategies show that already classifications with relatively low accuracies result in trading profits. Hence, more sophisticated NLP models paired with recent advances of the field would likely yield much higher profits.
This paper proceeds as follows: Section II covers the related literature on the transmission of information, while Section III describes the methodology and data. The empirical results are shown in Section IV, and their robustness is elaborated in Section V. Section VI discusses the broader implications of the results before I conclude in Section VII.

II Related Literature

This paper is also related to the literature on NLP in finance, but since this is already described in Pichler (2023) and more extensively in Loughran and McDonald (2016), I refer the interested reader to these two papers. Here, I first give a short overview of market microstructure, which has become increasingly important due to the rise of high-frequency trading, and then focus more extensively on the transmission of information, also referred to as the economics of information.

II-A Market Microstructure

The speed of trading experienced a dramatic acceleration with the introduction of electronic trading and has since been way beyond human trading speed. However, the increase in speed did not stop there and has since increased alongside technological progress. This can be illustrated by the accuracy of the stock exchanges' timestamps. While it was seconds in 1993, it became milliseconds in 2003, microseconds in 2015, and nanoseconds in 2016/2017.[4] Although trading is not at the nanosecond frequency, but in the order of milliseconds or microseconds, it is almost at the physical boundary.[5] Since the application in this paper is not about gathering and processing market information faster than anyone else, the news trading strategies described in this paper would be considered algorithmic trading, but not high-frequency trading.

In general, algorithmic trading has increased the liquidity and informativeness of quotes; namely, it narrowed spreads, reduced adverse selection, and reduced trade-related price discovery (Hendershott et al., 2011). Therefore, one might think that the rise of high-frequency trading would have led to a consolidation of the exchanges. But instead, we observe higher fragmentation of equity markets in the U.S. and Europe and new exchanges entering the market (O'hara, 2015). E.g., in the U.S., there are currently 22 exchanges, of which more than half were founded in the 21st century.[6] In this paper, I use the trades from five major exchanges in the U.S.[7] For a more detailed description of high-frequency trading and today's market microstructure, see O'hara (2015).

II-B Rational Agents

Due to the lack of means to measure information quantitatively, the transmission of information has been studied theoretically rather than empirically for a long time. Hayek (1945) establishes the role of the price system as a way of aggregating dispersed information, and he points to its economic efficiency, which would be hard, if not impossible, to

[4] See the description of NYSE TAQ data: https://www.nyse.com/publicdocs/nyse/data/Daily_TAQ_Client_Spec_v3.3b.pdf. Note that the error in the time measurement can be up to 100 microseconds.

[5] High-frequency traders usually rely on a signal from the exchange. Hence, information has to travel from the exchange to the trader's server, where it is processed, and the order signal has to travel back to the exchange (within exchange trading strategies, where the traders' servers are close to the exchanges' server). Alternatively, they rely on signals from two different exchanges. Therefore, they need both signals to arrive at their server, process them, and return the orders (price difference arbitrage). Today's computers rely on electrons that travel at 50-99% of the speed of light, the fastest speed known to humanity, at about 30 cm per nanosecond. Hence, every additional 15 cm from the exchanges server adds at least one nanosecond, and the signal processing always requires several computation steps, hence adding additional time.

[6] See https://www.sec.gov/divisions/marketreg/mrexchanges.shtml

[7] NYSE, NYSE Arca, NYSE American, NYSE National and NASDAQ.

overcome by a central planner. In that sense, Fama (1970) formulates the Efficient Market Hypothesis (EMH): assuming fully informed rational agents, prices represent these rational traders' beliefs. It is worth noting that the idea of trading leading to prices that fully reflect all available information goes at least back to Bachelier (1900). Tirole (1982) and Milgrom and Stokey (1982) point out that no one would be incentivized to trade in such a market, and since we observe trading, these models do not accurately represent reality. Following this logic and assuming that traders face some cost to become informed, Grossman and Stiglitz (1980) show theoretically that markets cannot be in equilibrium, i.e., fully arbitraged. Traders would only pay the cost of becoming informed if the potential benefits cover the costs; hence, there must be some arbitrage opportunity. Furthermore, it is neither an equilibrium in their model if no trader is informed nor if all traders are informed, a finding also shared by more recent works (e.g., see Foucault et al. (2013)).

Bagehot (1971) brings up the idea that random investors would lose to informed traders in an efficient market. While models of the type of Grossman and Stiglitz (1980) have two agents (informed and uninformed traders), he additionally has a market maker as the third agent. The market maker will lose on average to the informed traders and win to the uninformed traders, leaving the uninformed traders as the losers of this game. Based on Bagehot's idea, Glosten and Milgrom (1985) study the effect on the bid-ask spread, showing that the presence of informed traders leads to a positive bid-ask spread which is negatively related to trade size and is positively related to information quality. As Krishnan (1992) points out, the binary version of Kyle (1985)'s model is essentially equivalent to Glosten and Milgrom (1985), and the pricing rule of Kyle (1985) is the bid-ask quotes of Glosten and Milgrom (1985).

The aforementioned papers and models assume that there are differently informed traders; however, they all assume homogeneity in the traders' priors. Aumann (1976) shows that rational agents with the same priors must come to the same posterior if their posterior for a given event is common knowledge, i.e., they cannot agree to disagree even if they reach their posterior using different information. Miller (1977) combines the idea of dispersion in beliefs, i.e., different priors, with short-selling constraints. His model predicts that stocks with larger dispersion in beliefs and more significant short-selling constraints are more overpriced. Furthermore, it can explain market phenomena like the closed-end fund discount (Lee et al., 1991), volatility anomaly (high idiosyncratic volatility stocks have low returns and vice versa) (Ang et al., 2006, 2009), betting against beta anomaly (high systematic volatility stocks have lower returns than they should) (Frazzini and Pedersen, 2014), and the issuance anomaly (low returns post-issuance) (Daniel and Titman, 2006). More recent models like He and Wang (1995) and Tetlock (2010) picked up the idea of heterogeneity among investors' beliefs, and they show that the release of information resolves this uncertainty or information asymmetry, leading to convergence in investor beliefs.

II-C Behavioral Finance

After relaxing the assumption of homogeneously informed agents, research turned to another crucial element: the processing of information by the agent (or, in the words of Shiller (2003): From efficient markets theory to behavioral finance). The book of Shiller (2015)[8] addresses many aspects of human behavior[9] like herding (note that aggregate irrational behavior can even be caused by perfectly rational individuals), overconfidence (overconfidence in general and caused by feedback loops, i.e., positive returns in the past must indicate positive returns in the future), and new era thinking, witch can all lead to asset bubbles. Another important aspect of human decision-making pointed out by Shiller (2015) is the attention mechanism. While psychologists had studied it long before economists got interested, artificial intelligence scientists recently got outright enthusiastic about it (e.g., see Bahdanau et al. (2014)). Economists tend to think of attention as a curse that makes humans prone to systematic misjudgment; however, it also seems to be a blessing and a fundamental driver of human intelligence because it enables us to be very efficient at task-specific separation of relevant information from noise with a very limited dataset (a process called feature selection in machine learning).

The media take up a central role in disseminating and interpreting information, both from an efficiency and a behavioral point of view. Empirical research in finance has shown that more prominently posted news (the front page of *The New York Times*) has a greater impact than the same news in a less prominent spot (Klibanoff et al., 1998), high media coverage predicts positive returns and vice-versa (Fang and Peress, 2009), local media predicts local trading (Engelberg and Parsons, 2011) and that the absence of media (e.g., during strikes) decreases the trading volume and increases volatility (Peress, 2014). Although the media generally seems to increase efficiency, Solomon (2012) shows that specialized investor relations firms tend to "spin" their clients' news such that they reach higher and more positive media coverage. Hence, these firms experience higher news announcement returns, which are reversed at earnings announcements, when the news cannot be spun anymore.

Whether economic agents under- or overreact to new information is a key question in information economics. The best-known investor underreaction is the post-earnings-announcement drift[10] (Ball and Brown, 1968; Bernard and Thomas, 1989, 1990), famously called the "granddaddy of underreaction events" by Fama (1998). Chan et al. (1996) point out that both the post-earnings-announcement drift as well as price momentum (Jegadeesh and Titman, 1993) are evidence for investor underreaction to new information (if current prices would reflect all currently available information, past prices should have no predictive power). Hong and Stein (1999) develop a model where they have two groups with different information sets: the news watchers and the price momentum traders, which leads to short-term underreaction and long-term overreaction. Hong et al. (2000) test the model empirically and find that stocks with low analyst coverage experience higher price momentum

[8] The first edition was published in 2000, and the third in 2015.

[9] Countless experiments show that humans are prone to many cognitive errors, e.g., see Kahneman et al. (1982).

[10] For an overview of the literature on the post-earnings-announcement drift, see Fink (2021).

profits and that the effect is larger among past losers. Short-selling constraints sometimes rationalize this asymmetry between the long and the short leg of momentum. However, there is also a behavioral explanation: The disposition effect (Shefrin and Statman, 1985), which is caused by a combination of prospect theory[11] and mental accounting (Thaler, 1985) and describes the tendency to sell securities that have increased in value rather than those that have decreased since they were purchased. Frazzini (2006) provides evidence for this effect by showing that the post-announcement drift of mutual funds is more pronounced when the news has the same sign as recent capital gains. More recent research provides evidence that the post-earnings-announcement drift is at least weaker when applying the current standards for testing anomalies (Hou et al., 2020) and that it may even have vanished since around 2010 (Martineau, 2021). Nevertheless, the persistence of price momentum indicates that there is still an underreaction to some other information.

In the spirit of the post-earnings announcement drift, Odean (1998) argues that investor overconfidence leads to an underreaction to abstract, statistical, and highly relevant information (e.g., earnings announcements) and overreaction to salient, anecdotal, and less relevant information (e.g., some media story) and hence to higher expected volumes. While the research on the first part of this statement is discussed in the previous paragraph, Tetlock (2007) investigates the second part by showing that media sentiment has predictive power for stock returns in the same direction (i.e., positive sentiment predicts positive returns and vice versa) and trading volume depends on the absolute magnitude of sentiment in either direction. Because he does not find a relationship between text sentiment and fundamental risk measures, the irrational price reaction must be attributed to investor sentiment caused by the media text sentiment.[12] However, Tetlock et al. (2008) reach a different conclusion when using news on S&P 500 firms instead of the general market: Negative words predict negative fundamentals, i.e., the news contains fundamental information, and there is a short-term underreaction to the news stories.

II-D Modern Tools for the Transmission of Information

More recent research investigates search engines (e.g., *Google Search*) and social media platforms (e.g., *Facebook*, *Twitter*), which are just modern tools of information transmission. Hence, the research questions that were applied to the classical media also correspond to them. *Google Search* volume of individual stocks is positively related to their short-term returns, which revert within one year (Da et al., 2011), and the negative words predict short-term reversal of the general market (Da et al., 2015). Bartov et al. (2018) use three different dictionaries[13] and a Naïve Bayes classifier to categorize Twitter messages into positive, negative, and neutral and find that these messages can predict future firm fundamentals. This

[11] Humans tend to draw more negative utility from negative events than positive utility from equally positive events (Kahneman et al., 1982).

[12] While this measure of investor sentiment from text sentiment follows a “bottom-up” approach, Baker and Wurgler (2006) measure the investor sentiment “top-down” from macroeconomic variables.

[13] Loughran and McDonald (2011), Harvard HV-4, and Hu and Liu (2004).

effect is more pronounced for messages focusing on fundamentals, which is precisely what Tetlock et al. (2008) show for the negative words in classical media articles. A general issue of most empirical papers on the transmission of information is the relatively small sample size, usually ranging from a few hundred observations to a couple of thousand (e.g., 3'709 samples in Tetlock (2007)). Papers looking at the individual stock level instead of the aggregated market[14]) often just analyze the data from a few selected stocks, which puts their generalizability into question. Two other papers also overcome these limitations and have similar sample sizes: Tetlock (2010) covers 2.2 million news stories between 1979 and 2007, and Heston and Sinha (2017) use a similar sample from *Reuters News* with over 900'000 news stories between 2003 and 2010. I also overcome both limitations by covering more than 2.3 million news headlines regarding all stocks traded at the major exchanges in the U.S. (3'552 firms), but unlike the two previously mentioned papers, which work with daily market data, I provide evidence at the second precision. This eliminates the problem of fuzzy merging of news with their corresponding returns, which naturally arises when using daily data (also an issue in Pichler (2023)), but still leaves the caveat that market participants have more time to digest news published outside of trading hours than news published during trading hours.

This paper introduces several novel aspects to the existing literature. To my knowledge, it is the first paper that analyzes the stock market's reaction to firm-specific news at the second frequency. Moreover, it does so at large scale with 2.3 million observations. The results contribute to the discussion of the EMH's validity by showing that markets react very efficiently to new information. The observed return pattern is highly consistent with market efficiency for the following reasons: First, markets anticipate the type of news (positive/negative), and prices already move accordingly in the hours before the new information arrives. Second, incorporating the news in prices is estimated to only take two to three minutes on average. And finally, the return pattern of all news, i.e., independent of any NLP model, indicates that investors get a risk premium for holding stocks in the hours before new information is expected.

III Methodology & Data

This paper inherits the NLP models and their predictions from Pichler (2023). Conceptually, the dataset of Pichler (2023) looks as follows: On the one hand, it has the text of the *Reuters* news headline, and on the other hand, the market's reaction to the news, i.e., the price change from before to after the news. The NLP models are put in between as tools to predict the market's reaction from the news headline's text. The NLP models are described in more detail in Pichler (2023), and I just provide a short overview in the next paragraph.

[14] Uhl (2014) uses 3.6 million *Reuters* news and analyzes the relationship between aggregated market returns and *Reuters*' proprietary sentiment classification of the news.

The first two models are so-called *bag-of-words*[15] models, i.e., the order of the words is ignored, namely the Loughran and McDonald (2011) sentiment dictionary and the Multinomial Naïve Bayes classifier. While the sentiment dictionary assigns a minus one to the negative words and a plus one to the positive words of the dictionary (zero to all other words) and classifies the text based on the sum of the assigned values, the Multinomial Naïve Bayes classifier computes to optimal weight for each word based on a labeled training sample. Due to this parameter learning, the Multinomial Naïve Bayes classifier belongs to the family of supervised learning models. The three applied types of neural networks (Convolutional Neural Network (CNN), Recurrent Neural Network (RNN)-Long- Short-Term Memory (LSTM), and RNN-Gated Recurrent Unit (GRU)) also belong to this model family. While the Naïve Bayes classifier assumes independence between the words, neural networks are able to model dependencies between the words. Each neural network is once trained to provide a binary label (positive/negative, i.e., classification) and once to predict the price change caused by the news (regression). The binary labels are just the transformation to plus one for all the positive returns and zero for all the negative returns. Tables A.1, A.2, and A.3 in the appendix describe the models' architecture in more detail. The dataset is split into training, validation, and test data because all models, apart from the sentiment dictionary, optimize their parameters based on the training data, and the validation dataset is needed for the early stopping criterion of the neural network training. Hence, the test dataset was not involved in the learning process and therefore provides out-of-sample results.

Pichler (2023) mainly shows that the more sophisticated NLP models also yield superior performance for tasks in finance and are able to map directly to returns instead of text sentiment. However, potential trading profits after news publication are only investigated superficially because the restriction of daily data prevents a deeper analysis. To fill this gap, this paper merges the *Reuters* news dataset from Pichler (2023) with the trades from the NYSE TAQ database, but I only use the data of the nine years period between January 2011 and December 2019 due to the availability of intraday data.[16] The news dataset and all text preprocessing steps are described in detail in Pichler (2023). Note that the number of news per firm per month has been around twelve throughout this sample period, but the number of firms with news increased from around 1'500 in 2011 to around 2'500 at the end of 2019, hence increasing the number of observations per month from about 18'000 to roughly 30'000.[17]

I only use trades from the NYSE, NYSE Arca, NYSE American, NYSE National, and NASDAQ (exchange = A, C, N, P, T or Q), marked as regular trades (sale condition = blank), automatic execution trades (sale condition = E), or intermarket sweep orders (sale condition = F) that were not corrected, changed, or signified as cancel or error (correction indicator = 00). For 89.2% of the news data, the NYSE TAQ data has the corresponding intraday

[15] It is called *bag-of-words* because you treat text like a bag filled with different pieces, each representing a word. One may shake this bag, but models following this approach will yield the same results, meaning that the order of the words is completely ignored in these models.

[16] NYSE also offers data further back, but I only have access to the data from 2011 onwards.

[17] See Figure 3-2 in Pichler (2023).

price and volume data. While the data is split 70%-15%-15% into training, validation, and test dataset in Pichler (2023), it is about a 50%-25%-25% split here due to the available NYSE TAQ data not covering the training sample period from the beginning of 1996 to the end of 2010.

Returns are computed by adjusting prices according to the Center for Research in Security Prices (CRSP) variable *FACPR* (stock splits, reverse splits) and taking the CRSP variable *DIVAMT* (cash dividends, liquidation, spin-offs) into account at the start of the corresponding trading days.[18]

One second after positive (negative) news, the tested trading strategies would take the buy (sell) side for one minute, then hold the stock for one minute before starting to liquidate the position till no more shares are left. The relatively short trading periods are chosen in order to be able to liquidate the acquired position and because the return pattern of Figure 3 suggests that the market's reaction time is relatively short. For the results not to be purely driven by the largest trades, a maximum buying (selling) amount per news is specified. However, this maximum is set relatively high at USD five million, which limits less than 2% of the observations, and the effect of changes in this parameter (as well as of the above described) is further discussed in Section V. At each second, the number of currently held shares is computed by adjusting the number of shares held in the previous second with the CRSP variable *FACSHR*[18] (adjustment in the first second the trading day) plus/minus the traded shares. The trading profit corresponding to a particular news headline is then computed as the dot product of the price and trading volume (positive in the buying period and negative in the selling period) vectors plus the dot product of the shares held and the dividend vector.

It is important to note that for the trading strategies, the samples with the same Committee on Uniform Security Identification Procedures (CUSIP) and next trading second are mean aggregated to reasonably estimate the potential trading profits. This reduces the samples from 2.33 million to 1.05 million. However, this procedure does not account for the fact that news about the same company could be released in two subsequent seconds. But this is only a minor issue since the probability of such observations is low due to the next second being the first trading second of a trading day for most observations and the relatively short trading period.

Table 1 describes the different datasets and shows that the training and validation sets have a very similar return distribution while the returns are, on average, lower and a bit more dispersed in the test dataset. However, this difference seems reasonable and should not affect any of my conclusions. Table 2 indicates that the next trading second for about

[18] For *FACPR*, *FACSHR*, and *DIVAMT*, only first-level distribution codes 1 (ordinary dividend), 2 (liquidation dividend), 3 (exchanges and reorganizations), 4 (subscription rights), and 5 (splits and stock dividends) are considered. Since exchanges and reorganizations (distribution code 3) are both included in *FACPR* and *DIVAMT*, I treat them like cash dividends and set *FACPR* such that there is no adjustment. The correction with the CRSP dataset is not perfect, resulting in a few unlikely returns. But since their influence is neglectable and identification is tricky because most returns around news are highly unlikely observations, I leave them in the sample.

Table 1: Descriptive Statistics of Training, Validation, and Test Dataset

This table shows the descriptive statistics of the training, validation, and test dataset after being merged with the NYSE data for the time between 01.01.2011 and 31.12.2019. Note that the dataset of Pichler (2023) started in January 1991, was sorted by time, and split 70%-15%-15%. Therefore, the size of the training dataset here is only about 40% of the original size in Pichler (2023), while it is around 90% for the validation and test dataset. The training dataset covers the period between 01.01.2011 and 31.03.2016, the validation between 31.03.2016 and 26.04.2018, and the test dataset between 26.04.2018 and 31.12.2019. All values are in basis points except for the last three rows, which are the number of observations. Note that the mean return and volatility of the time window [-1h,0h) is smaller than that of the other time windows because it does not include an opening hour return for most observations, while the others do.

	Training dataset	Validation dataset	Test dataset
Mean return [-1h, 0h)	3.96	4.48	2.32
Mean return [-1h, +1h)	11.83	10.29	−1.55
Mean return [0h, +1h)	8.01	6.36	3.45
Mean return [0h, +2h)	9.45	6.23	−4.49
Standard deviation [-1h, 0h)	162.97	230.32	292.16
Standard deviation [-1h, +1h)	518.97	618.70	760.80
Standard deviation [0h, +1h)	495.68	586.16	711.88
Standard deviation [0h, +2h)	510.13	597.03	729.55
Positive returns [-1h, +1h) (> 0)	605’598 (51.70%)	311’689 (51.88%)	314’531 (50.96%)
Negative returns [1h, +1h) (≤ 0)	565’883 (48.30%)	280’117 (48.12%)	302’628 (49.04%)
Number of observations	1’171’481	600’806	617’159

Table 2: Distribution of Observations Within Trading Hours

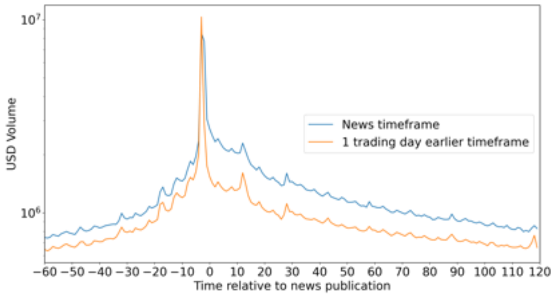
This table shows the distribution of the next trading seconds after news publication over the trading times. A regular trading day in New York starts at 09:30 and ends at 16:00, i.e. lasts for 6.5 hours. Because most news is published outside trading hours, 09:30:00 is the next trading second for most news, skewing the distribution heavily to the start of the trading sessions. The other news concentration is towards the end of the trading day. Note that because I do not account for trading halts; hence, the next trading second could be a second when the stock market is open, but the particular stock’s trading is halted.

	Training dataset	Validation dataset	Test dataset
09:30-10:00	74.94%	77.86%	74.70%
10:00-11:00	3.64%	3.06%	2.88%
1100:-12:00	2.92%	2.65%	2.29%
12:00-13:00	2.64%	2.17%	2.08%
13:00-14:00	2.32%	1.92%	1.77%
14:00-15:00	2.10%	1.66%	1.57%
15:00-16:00	11.44%	10.68%	14.71%

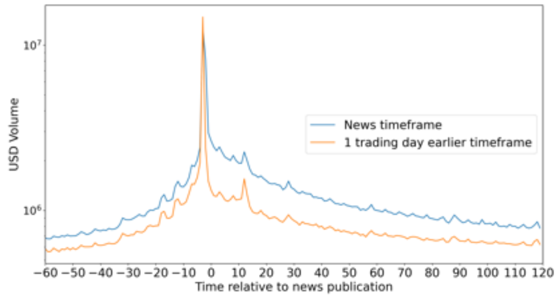
three-quarters of the observations is the first second of a trading session, which is primarily due to news being published around the clock and trading only taking place 6.5 hours per workday. Interestingly, there is also a lot of news in the last hour of the trading sessions. A potential problem would arise if this spike in the last trading hour was driven by companies halting trading of their stock(s) for some news release which is then republished by *Reuters*. Since I do not account for trading halts, I would observe this as a delayed market reaction, but my results suggest otherwise. However, it slightly affects the trading strategies since these observations yield zero profit due to no trading taking place. Although there is no trading for slightly less than 20% of the observations, it is unclear whether this is due to trading halts or simply no trades happening due to illiquidity.

Figure 1: USD Volumes Around News Publication

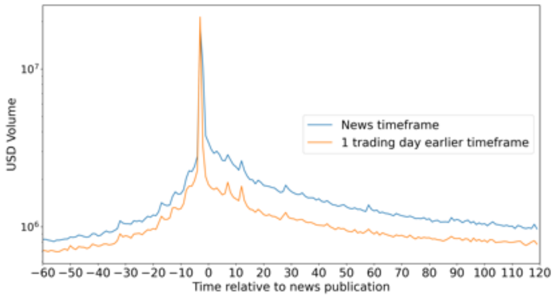
This figure shows the average USD volume per minute around the publication of news (blue line) and for the same time window one trading day earlier (orange line). Because most news is published outside trading hours, zero on the time axis coincides with the beginning of a trading day, and therefore the volume is naturally elevated around it. To visually mitigate the spike at time zero, the y-axis is log scaled.



(a) USD volume training dataset



(b) USD volume validation dataset



(c) USD volume test dataset

Figure 1 shows the average USD volume per minute around the publication of news (blue line) and for the same time window one trading day earlier (orange line). Because most news is published outside of trading hours, zero on the time axis coincides for roughly 75% of the observations with the beginning of a trading day, and therefore the volume is naturally elevated around it. In order to estimate the additional volume caused by the news, it is compared to the same timeframe one trading day earlier. Table 3 summarizes the data shown in Figure 1 numerically. On average, the trading volume increases due to news by around 20% in the hour before publication, more than 50% in the first hour, and around 30% in the second hour after publication. The higher volume level of the test dataset compared to the other two datasets is most likely caused by the overall trend of increasing volumes (note that the y-axes of Subfigures (a) - (c) have different log scales).

Table 3: USD Volumes Around News Publications

This table summarizes the average trading volumes in million USD per minute around news publications. Because most news is published outside trading hours, zero on the time axis coincides with the beginning of a trading day, and therefore the volume is naturally elevated around it. It is therefore compared to the same time timeframe one trading day earlier.

	Training dataset	Validation dataset	Test dataset
Volume news timeframe [-1h, 0h)	1.27	1.28	1.59
Volume 24h before [-1h, 0h)	1.06	1.04	1.33
Volume change [-1h, 0h)	+19.83%	+23.50%	+19.82%
Volume news timeframe [0h, +1h)	1.57	1.50	1.87
Volume 24h before [0h, +1h)	1.05	0.92	1.22
Volume change [0h, +1h)	+50.09%	+62.10%	+52.94%
Volume news timeframe [+1h, +2h)	0.92	0.87	1.10
Volume 24h before [+1h, +1h)	0.72	0.65	0.83
Volume change [+1h, +2h)	+27.97%	33.07%	+31.69%

IV Results

First, I check the accuracies of the models' forecasts concerning the returns, shown in Table 4. Accuracy is defined as the fraction of correct predictions out of all predictions. E.g., the first entry of the first column states that the return in the 6.5 hours before the news (including the first second after the news) is according to the sentiment dictionary's predictions (positive/negative) in 54.86% out of all training dataset predictions. No time window would yield an exact comparison to the numbers in Table 8 of Pichler (2023); however, the first two columns should be roughly similar[19] since they include the time before the news and the minute after the news is published. The subsequent rows indicate that the models have some predictive power for the post-publication returns and that this relationship gets weaker the longer the time lag between news and returns becomes. This already indicates that instant trading on the news should still be profitable, but also suggests that the time window is relatively short. All models have very similar relationships to returns, except for the sentiment dictionary, which is about 2 percentage points lower. Because the sentiment dictionary does not use any optimized parameters, the comparison between the training, validation, and test dataset already provides some evidence that markets have become more efficient over time at anticipating and digesting the news.

The graphs in Figure 2 are the counterparts to Figure 5 in Pichler (2023) for the eight trading hours before and after news publication and show the same pattern as for the daily returns. The news is anticipated before its publication and then quickly priced in. Figure 3 zooms in on the time frame and displays 30, 10, and one minute around news publication. Most of the reaction is instantly at the next trading second, but there is still a drift afterward, which takes around two to three minutes in the test dataset. Since the test dataset

[19] One could argue that they should be slightly lower because the models were trained on the daily returns used in Pichler (2023), but on the other hand, the relationship between the news and the returns may be stronger due to the higher precision of matching returns to news (daily vs. second).

Table 4: Accuracies of Model Predictions

This table displays the accuracies of the different models with respect to the returns around news publications. The training dataset covers the time period between 01.01.2011 and 31.03.2016, the validation between 31.03.2016 and 26.04.2018, and the test dataset between 26.04.2018 and 31.12.2019. Descriptive statistics for the three datasets can be found in Table 1. Note that the time windows here always contain the impact of the news, unlike the [-1,0) window in the descriptive statistics in Table 1.

Time window in minutes:		(-390,0]	(-60,0]	[0,60)	[1,61)	[5,65)
Sentiment dictionary	Train:	54.86%	54.40%	54.59%	51.35%	50.96%
	Validation:	54.46%	54.67%	53.55%	50.77%	50.18%
	Test:	53.88%	54.52%	52.88%	50.40%	49.80%
Naïve Bayes	Train:	56.45%	57.54%	56.43%	53.11%	52.35%
	Validation:	55.51%	56.78%	55.44%	52.56%	54.89%
	Test:	55.18%	56.27%	54.54%	52.18%	51.83%
CNN (classification)	Train:	56.56%	57.56%	56.49%	53.06%	52.36%
	Validation:	56.22%	57.41%	55.80%	52.71%	52.01%
	Test:	55.54%	56.59%	54.81%	52.30%	51.83%
RNN-LSTM (classification)	Train:	56.80%	57.86%	56.74%	53.23%	52.36%
	Validation:	56.18%	57.31%	55.56%	52.48%	51.77%
	Test:	55.39%	56.37%	54.61%	52.20%	51.82%
RNN-GRU (classification)	Train:	56.94%	56.97%	56.75%	53.18%	52.36%
	Validation:	56.74%	56.95%	56.10%	52.86%	52.18%
	Test:	55.95%	55.26%	55.35%	52.71%	52.27%
CNN (regression)	Train:	56.56%	57.74%	56.34%	52.86%	52.11%
	Validation:	56.16%	57.45%	55.61%	52.53%	51.61%
	Test:	55.43%	56.22%	54.67%	52.18%	51.79%
RNN-LSTM (regression)	Train:	56.36%	57.58%	56.17%	52.79%	52.04%
	Validation:	56.31%	57.63%	55.81%	52.57%	51.80%
	Test:	55.63%	56.45%	54.88%	52.26%	51.81%
RNN-GRU (regression)	Train:	56.22%	57.47%	56.10%	52.79%	52.06%
	Validation:	56.09%	57.51%	55.74%	52.56%	51.81%
	Test:	55.48%	56.30%	54.75%	52.19%	51.75%

covers the most recent time period, a comparison with Figures A.3 (training dataset) and A.4 (validation dataset) in the appendix indicates that this drift has become slightly shorter over time.

Table 5 shows the results for the test dataset of the trading strategy that starts buying (selling) one second after the news is published for one minute, then holds the stock (short position) for one minute before starting to liquidate the position. The trading profit of the negative news is the profit from shorting, and the values in brackets below are the corresponding t-test values against the mean profit for positive and negative and against zero for <Profit all>. Note that because the t-value of <Profit negative> tests the mean of a long position in the negative news observations against the mean, a negative t-value is desired. Across all models, buying following positive news generates profits above the mean profit, and following negative news profits below it. These results are almost always statistically significant. Buying and shorting taken together also generate statistically significant profits above zero for all models. Since the average buying trade generates a positive return (e.g., USD 36.60 in the test dataset (used in Table 5) according to the descriptive statistics of the different trading strategies shown in Table A.4 in the appendix) and all models apart from the sentiment dictionary predict more positive news, the test against zero has to be viewed with some skepticism. However, the separate results for the positive and negative predictions provide clear evidence, that also a test against a weighted average (i.e., n_{pos} times mean profit minus n_{neg} times mean profit) would be statistically significant. Increasing the threshold for classifying the news tends to increase the profits per trade and their significance but also decreases the number of trades. An estimate for the overall trading profit can be computed by multiplying the columns <Profit all> and <n all>, and this number decreases when increasing the thresholds.

Since the average trading amount of the test dataset is USD 797'239.91 (see Table A.4 in the appendix) USD 79.72 is one basis point, meaning that the trading profits of the test dataset range from 0.91 basis points to 3.95 basis points. This may sound very small, but one must remember that the investment period is also short. While the buying (selling) and holding period is fixed at one minute each,liquidating the position on average takes longer due to the volume pattern shown in Figure 1. Since the economic significance of one basis point at this frequency is not straightforward for most people, the following annualization assuming an average holding period of five minutes, 6.5 hours of trading, and 252 trading days illustrates it: $1.0001^{6.5 \cdot 60 \cdot 252/5} - 1 = 613.85\%$ p.a. Table A.14 in the appendix shows this computation for one to six basis points and also for assuming a financial year with 24 hours and 360 days instead of a trading year. These computations just aim to give the reader some perspective of what a tiny return means at these frequencies. However, the annualized returns are nothing realizable in the setting discussed here since the investing time is heavily centered around the first trading minutes of a day, and it is neither possible to always be fully invested nor with large amounts.

Probably the best estimation of the realizable return is provided in Table A.7 in the appendix, which shows the results for the whole dataset. It indicates that profits would be around USD 60 per trade (ignoring the sentiment dictionary and the RNN-GRU)[20] on 840'811 trades in the nine years window, i.e., total profits of around USD 50 million. The three last columns in Table 5 show the fractions of profitable trades and below in brackets of losing trades. Note that they do not sum to one since up to 2% of the trades yield a zero profit. The fact that most trading strategies lose on more trades than they gain on in case of positive news is concerning but only found in the test dataset.[21] Since the Naïve Bayes classifier does not rely on any hyperparameters, the validation set is like another test set for this model. Hence results of the validation dataset shown in Table A.6 in the appendix indicate that this seems to be related to the dataset. Generally, the fraction of profitable trades is slightly higher for the negative news than for the positive news, which might be because the models were trained on daily returns and have a bias toward predicting a positive market

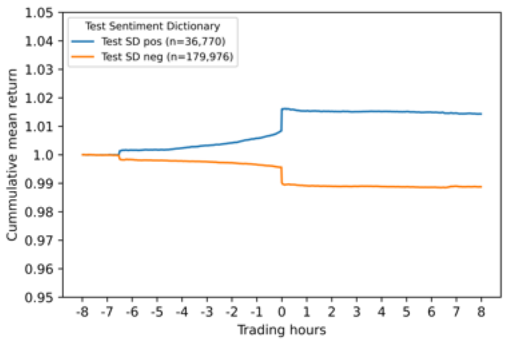
[20] The two models are ignored here because the sentiment dictionary fails to classify most news which heavily limits the trading capacity, and for the RNN-GRU a weak parameter solution was chosen deliberately to show that these solutions are also possible.

[21] Compare with the results of the training dataset (Table A.5 in the appendix), the validation dataset (Table A.6 in the appendix), or the whole dataset (Table A.6 in the appendix).

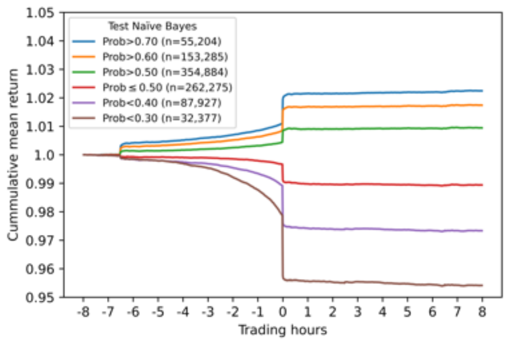
reaction. Short-selling constraints are an alternative explanation for this observation.
Most models have very similar results, but there is a slight tendency of the neural networks to outperform the Naïve Bayes classifier. The RNN-GRU's tendency to lean heavily towards a positive prediction leads to the weakest results among the neural networks. However, this is not a general problem of the RNN-GRU networks but instead of the used parameter solution, which is deliberately chosen in Pichler (2023) to show that also such solutions are possible. The sentiment dictionary provides competitive results with the other models on its predictions. Nevertheless, the overall profits are lower because it cannot classify most headlines.

Figure 2: Cumulative Mean Returns on Positive/Negative News - Test Dataset

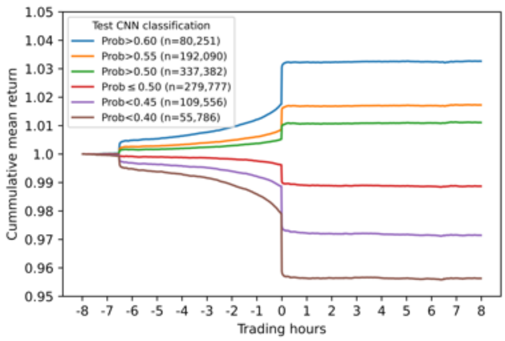
This figure shows the cumulative mean returns of positive and negative news according to the indicated model for the eight hours before and after news publication for the test dataset. The x-axis shows the trading time relative to news publication at zero. Markets start anticipating the news and then digest them quickly, and very soon, there is no trend in the stock price. Figures A.1 and A.2 in the appendix show the training and validation dataset results.



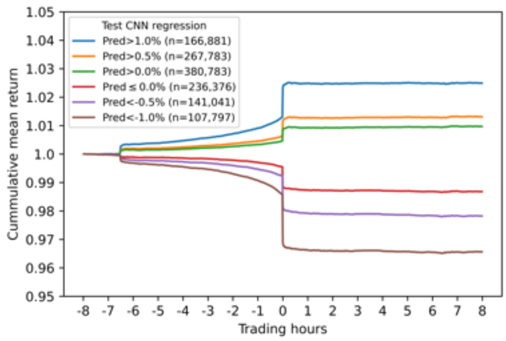
(a) Sentiment Dictionary Classification



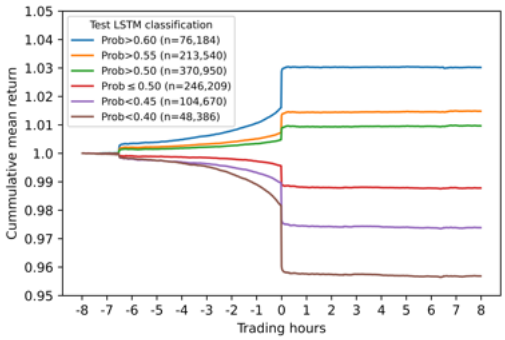
(b) Naïve Bayes Classification



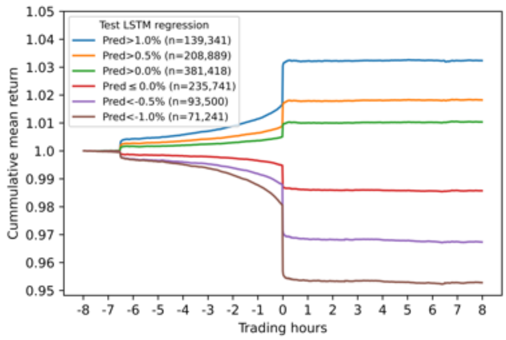
(c) CNN Classification



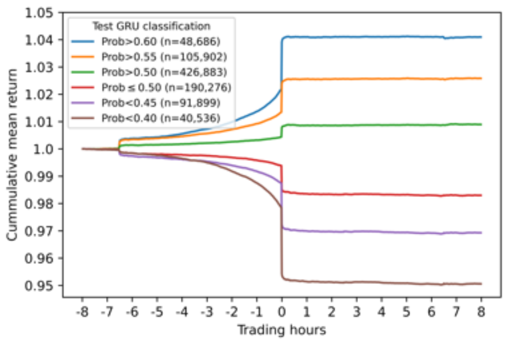
(d) CNN Regression



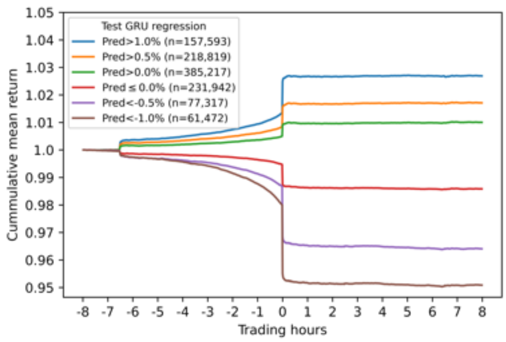
(e) RNN-LSTM Classification



(f) RNN-LSTM Regression



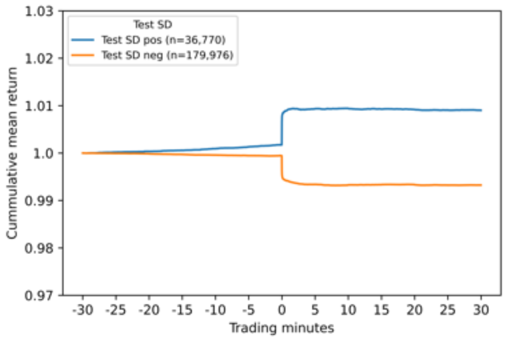
(g) RNN-GRU Classification



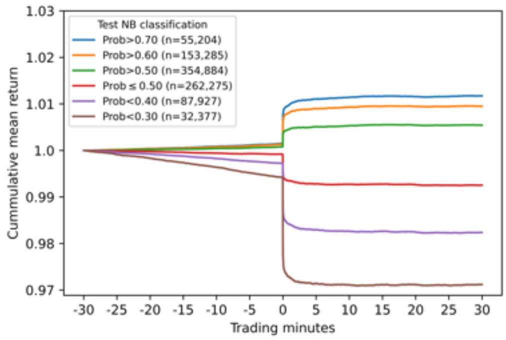
(h) RNN-GRU Regression

Figure 3: Zoom in on Cumulative Mean Returns on Positive/Negative News - Test Dataset SD and NB

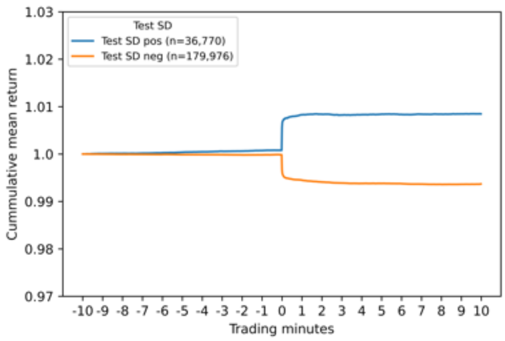
This figure zooms in on the cumulative mean returns of positive and negative news according to the sentiment dictionary and the Naïve Bayes classification for the eight hours before and after news publication for the test dataset. The x-axis shows the trading time relative to news publication at zero. Markets start anticipating the news and then digest them quickly, and very soon, there is no trend in the stock price. Figures A.3 and A.4 in the appendix show the training and validation dataset results.



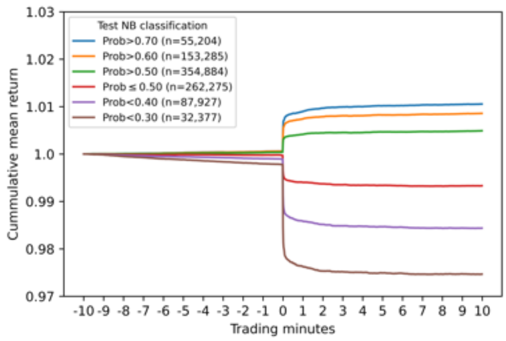
(a) Sentiment Dictionary +/- 30 min.



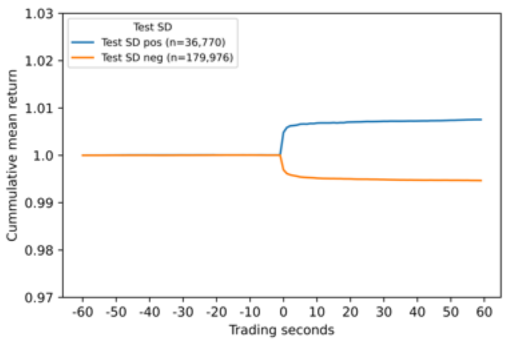
(b) Naïve Bayes +/- 30 min.



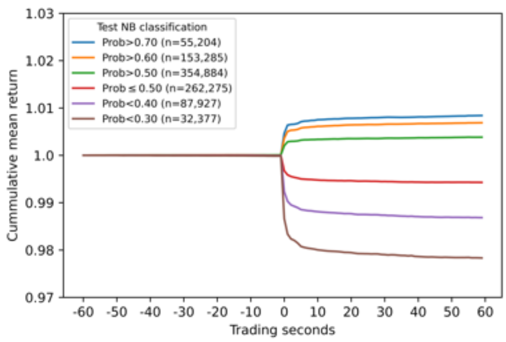
(c) Sentiment Dictionary +/- 10 min.



(d) Naïve Bayes +/- 10 min.



(e) Sentiment Dictionary +/- 1 min.



(f) Naïve Bayes +/- 1min.

Table 5: Model Comparison for the Test Dataset [1 sec., 1 min., 1 min., 5m max.]

The strategies shown here start trading one second after the news is published for one minute, then hold the position for one minute before starting to fully liquidate the position with a maximum of USD five million per trade. Each model predicts a probability (classification) or a return (regression), and the thresholds are used to classify the headlines into positive ($\geq$ positive threshold) and negative ($\leq$ negative threshold). Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 61'681 observations, there is no trading in the buying/selling period, and 74 observations are excluded because not all shares are sold till the end of the eight hours window.

Model	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Fractions positive	Fractions negative	Fractions all
SD classification	[na/na]	331.13	40.21	71.96	14′885	121′484	136′369	49.97%	52.14%	51.90%
		(3.59)	(−2.48)	(3.16)				(49.12%)	(46.62%)	(46.89%)
Naïve Bayes (classification)	[0.50/0.50]	85.84	24.04	58.15	119′919	97′379	217′298	48.23%	50.97%	49.46%
		(1.53)	(−1.63)	(2.92)				(48.35%)	(47.52%)	(48.03%)
	[0.60/0.40]	199.43	196.88	198.52	43′834	24′404	68′238	50.35%	51.02%	50.59%
		(3.40)	(−3.67)	(5.59)				(48.74%)	(47.49%)	(48.29%)
	[0.70/0.30]	190.25	206.15	196.40	15′700	9′906	25′606	50.80%	51.01%	50.88%
		(2.02)	(−2.53)	(3.51)				(48.35%)	(47.52%)	(48.03%)
CNN (classification)	[0.50/0.50]	115.61	55.31	87.73	116′845	100′453	217′298	48.40%	51.16%	49.68%
		(2.38)	(−2.55)	(4.40)				(48.31%)	(47.29%)	(47.88%)
	[0.55/0.45]	154.31	132.42	147.93	74′095	30′458	104′553	48.83%	51.58%	49.63%
		(3.01)	(−3.00)	(5.43)				(50.04%)	(47.14%)	(49.20%)
	[0.60/0.40]	288.71	134.55	224.21	26′754	19′250	46′004	50.82%	51.49%	51.10%
		(4.06)	(−2.46)	(4.55)				(48.31%)	(47.29%)	(47.88%)
RNN-LSTM (classification)	[0.50/0.50]	93.65	38.92	70.10	123′784	93′514	217′298	48.30%	51.05%	49.49%
		(1.75)	(−2.04)	(3.52)				(48.67%)	(47.41%)	(48.19%)
	[0.55/0.45]	123.67	172.47	135.52	86′771	27′832	114′603	47.85%	51.44%	48.72%
		(2.37)	(−3.50)	(5.07)				(51.03%)	(47.18%)	(50.10%)
	[0.60/0.40]	278.57	139.62	225.07	23′581	14′762	38′343	50.53%	51.20%	50.79%
		(3.62)	(−2.21)	(3.62)				(48.67%)	(47.41%)	(48.19%)
RNN-GRU (classification)	[0.50/0.50]	72.98	78.88	74.40	165′239	52′059	217′298	48.12%	51.10%	48.83%
		(1.30)	(−2.32)	(3.73)				(48.34%)	(47.73%)	(48.08%)
	[0.55/0.45]	314.94	161.34	247.47	33′354	26′129	59′483	50.52%	51.42%	50.92%
		(4.82)	(−3.21)	(5.11)				(48.73%)	(47.40%)	(48.15%)
	[0.60/0.40]	246.85	138.76	200.12	17′137	13′051	30′188	50.84%	50.99%	50.90%
		(2.75)	(−2.09)	(2.97)				(48.34%)	(47.73%)	(48.08%)
CNN (regression)	[0.0%/ 0.0%]	83.34	26.28	59.01	124′645	92′653	217′298	48.31%	51.11%	49.50%
		(1.44)	(−1.69)	(2.96)				(48.71%)	(47.11%)	(48.06%)
	[0.5%/-0.5%]	107.46	105.20	106.80	92′612	37′932	130′544	48.18%	51.36%	49.11%
		(1.93)	(−2.50)	(3.49)				(50.63%)	(47.45%)	(49.71%)
	[1.0%/-1.0%]	224.48	157.38	197.10	43′202	29′775	72′977	50.30%	51.66%	50.85%
		(3.58)	(−3.10)	(3.88)				(48.71%)	(47.11%)	(48.06%)
RNN-LSTM (regression)	[0.0%/ 0.0%]	94.68	38.42	70.13	122′475	94′823	217′298	48.26%	51.07%	49.49%
		(1.78)	(−2.03)	(3.52)				(48.84%)	(46.91%)	(48.15%)
	[0.5%/-0.5%]	226.71	190.67	214.52	50′072	25′595	75′667	50.20%	51.86%	50.76%
		(3.83)	(−3.34)	(4.19)				(48.80%)	(46.84%)	(48.14%)
	[1.0%/-1.0%]	292.15	230.70	270.15	35′019	19′537	54′556	50.20%	51.77%	50.76%
		(4.42)	(−3.67)	(4.57)				(48.84%)	(46.91%)	(48.15%)
RNN-GRU (regression)	[0.0%/ 0.0%]	99.84	44.12	75.36	121′844	95′454	217′298	48.26%	51.07%	49.50%
		(1.93)	(−2.21)	(3.78)				(48.79%)	(46.56%)	(48.10%)
	[0.5%/-0.5%]	212.61	222.74	215.56	51′504	21′202	72′706	50.17%	51.89%	50.67%
		(3.58)	(−3.70)	(4.49)				(48.84%)	(46.83%)	(48.26%)
	[1.0%/-1.0%]	277.08	259.56	271.64	37′483	16′851	54′334	50.24%	52.04%	50.80%
		(4.25)	(−3.82)	(4.53)				(48.79%)	(46.56%)	(48.10%)

V Robustness

Since this paper aims to investigate the market's micro reaction to new information, using all eight models from Pichler (2023) is in itself a significant robustness test. Nevertheless, I investigate the robustness of my results as follows: Subsection V-A displays the return patterns around news which arrive within trading hours. The remaining subsections show how the trading strategies are affected by the parameter choices of the maximum trading amount (Subsection V-B), holding time (Subsection V-C), and buying/selling time (Subsection V-D). Since there is no material difference between the models, all tables and figures in this section only contain results for the classifications by the sentiment dictionary and the Naïve Bayes classifier. The results for all models can be found in the appendix and are indicated in the description of each table or figure.

V-A Within Trading Hours News

I define "within trading hours news" as news published at least 30 minutes after the market opening and 30 minutes before the market closing. This definition only leaves about 13% of all observations since most news is published outside of trading hours. Note that since firms are required to publish relevant information outside of trading hours or halt trading of their stock, this subset is not just a random selection and may be biased towards being less relevant. Figure 4 displays the cumulative mean return of the eight hours (left side, Subfigures (a) and (c)) and the 30 minutes (right side, Subfigures (b) and (d)) before and after news publication of the classifications according to the sentiment dictionary and the Naïve Bayes classifier.[22]

It shows that there is also anticipation for this subset of news and that there is no drift after two to three minutes. Subfigure (b)'s shape is a bit surprising since the orange curve (negative news) is flat with a small increase before publication. Note that the red curve in Subfigure (b), which displays the cumulative mean return of all the negative news, has the same shape, but this pattern disappears when applying a higher threshold, i.e., only considering more negative news. Together with the results of the training and validation dataset displayed in Figures A.9 and A.10 in the appendix, I would argue that this pattern is sample-specific rather than systematic. However, the most important observation is the disappearance of the jump at time zero and that almost the whole price impact is before and not after the news.

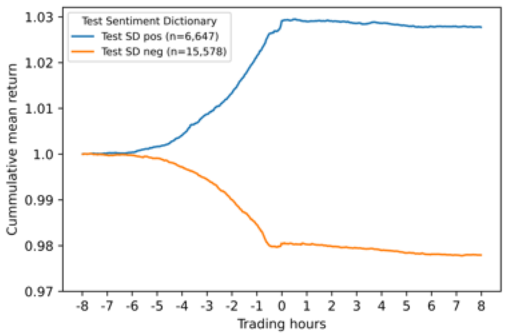
There are multiple possible explanations for this pattern. First, these news observations could be interpretations of or additions to earlier news, with no incremental new information for stock prices. Second, the speed of information transmission matters for news within trading hours while it does not for news outside of trading hours, and *Reuters* may simply be slower than other news providers like *CNBC*, *Bloomberg*, etc. Alternatively, it would be

[22] Figures A.5, A.6, and A.7 in the appendix show the graphs for all models for the training, validation, and test dataset (eight hours before and after). Figures A.8, A.9, and A.10 in the appendix show the graphs for all models for the training, validation, and test dataset (30 minutes before and after).

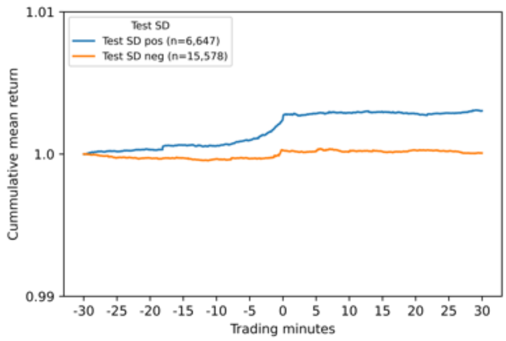
consistent with an intraday momentum pattern. Since all machine learning algorithms are trained to predict returns, it could very well be the case that they pick it up. However, two arguments against this hypothesis: First, all models take the headline text as input and not past returns. Therefore, it seems much more reasonable that they learn the type of news on a positive/negative sentiment or content scale rather than a high/low momentum scale. Second, classification by the sentiment dictionary results in the same pattern.

Figure 4: Cumulative Mean Returns on Positive/Negative News of SD and NB- Test Dataset Within Trading Hours

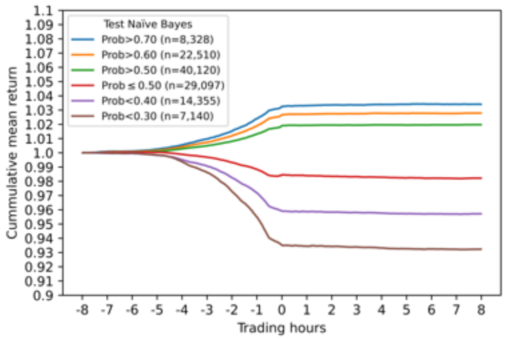
This figure shows the cumulative mean returns of positive and negative news according to the indicated model for the eight hours (left side)/30 minutes (right side) before and after news publication for the test dataset. The x-axis shows the trading time relative to news publication at zero. Markets start anticipating the news and then digest them quickly, and very soon, there is no trend in the stock price. Figures A.5, A.6, and A.7 in the appendix show the graphs for all models for the training, validation, and test dataset (eight hours before and after). Figures A.8, A.9, and A.10 in the appendix show the graphs for all models for the training, validation, and test dataset (30 minutes before and after).



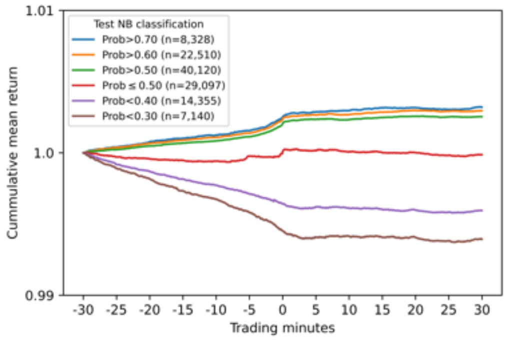
(a) Sentiment Dictionary Classification



(b) Naïve Bayes Classification



(c) Naïve Bayes Classification



(d) CNN Classification

V-B Maximum Trading Amount

The trading strategies contain a maximum amount per trade because the results would otherwise be mainly determined by very few observations. Table 6 compares having a trading maximum of USD one, five, and ten million for the sentiment dictionary and the Naïve Bayes classifier. Note that the results of the other models can be found in Tables A.7, A.8, and A.9 in the appendix, and that all comparisons mentioned below are based on the averages over all models.

Increasing the maximum trading amount only increases the trading profits marginally but significantly reduces the statistical power. On the other hand, reducing the maximum trading amount to USD one million decreases the profits by about 43%, despite increasing the t-values significantly. Over the whole sample, the average trade size in the base case (trading maximum of USD five million) is USD 803'597.60 and increases to USD 999'627.87 when using a maximum of USD ten million. Hence, increasing the maximum per trade decreases the profits in basis points on average by 21.53%, indicating that larger, more liquid stocks' prices adapt even faster to new information. This is supported by the observations that the average trading amount decreases by 51.84% to USD 387'037.64 when applying a maximum per trade of USD one million, therefore increasing the profits in basis points by 18.90% on average. Despite these differences, results are robust against changes in the maximum amount per trade parameter.

V-C Holding Time

The trading strategies have a defined holding period to profit from the difference between the trading average before and after it. Table 7 compares a holding period of 30 seconds, one, and two minutes for the sentiment dictionary and the Naïve Bayes classifier. Note that results for the other models can be found in Tables A.7, A.10, and A.11 in the appendix, and all comparisons mentioned below are based on the averages over all models.
Increasing the holding period from one to two minutes only increases the trading profits by 4.73%, and decreasing it to 30 seconds decreases profits by 6.57% on average. Since the average trading amount is independent of the holding time, profits in basis points change at the same rates. These results indicate that increasing the holding period could increase profits; however, the return patterns suggest that the benefit should be very limited beyond a couple of minutes. Hence, the results also seem to be robust against changes in the holding period parameter.

V-D Buying/Selling Time

The trading strategies have a defined time period in which the position is acquired (buying upon positive news and selling on negative news). Table 8 compares a buying/selling period of 30 seconds, one, and two minutes for the sentiment dictionary and the Naïve Bayes classifier. Note that results for the other models can be found in Tables A.7, A.12, and A.13 in the appendix, and all comparisons mentioned below are based on the averages over all models. Prolonging the buying/selling period from one to two minutes increases the trading profits by 49.70%, and shortening it to 30 seconds decreases profits by 39.67% on average. The t-stats also change accordingly, i.e., a longer buying/selling period results in larger t-stats. However, the length of the buying/selling period also changes the average trading amount: USD 565'442.36 (30 seconds), USD 803'597.60 (one minute), and USD 1'166'171.35 (two minutes). Hence, the profits in basis points only increase by 3.16% when

Table 6: Trading Strategy of SD and NB [1 sec., 1 min., 1 min., 1/5/10m max.] - Whole Dataset

The strategies shown here start trading one second after the news is published for one minute, then hold the position for one minute before starting to fully liquidate the position with a maximum of USD one/five/ten million per trade. Each model predicts a probability (classification) or a return (regression) and the thresholds are used to classify the headlines into positive (≥ positive threshold) and negative (≤ negative threshold). Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 209'667 observations, there is no trading in the buying/selling period, and 265 (one/five/ten million maximum) are excluded because not all shares are sold till the end of the eight hours window. Tables A.9 (one million maximum), A.7 (five million maximum), and A.8 (ten million maximum) in the appendix show the results of all models.

Model	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Accuracy positive	Accuracy negative	Accuracy all
Trading maximum: USD 1 million										
SD classification	[na/na]	75.57 (4.50)	16.92 (−6.02)	25.34 (6.30)	67′724	403′555	471′279	50.52% (48.48%)	51.19% (47.44%)	51.10% (47.59%)
Naïve Bayes (classification)	[0.50/0.50]	46.35 (5.30)	21.55 (−5.90)	35.70 (10.34)	479′897	360′914	840′811	49.86% (47.15%)	50.92% (47.88%)	50.31% (47.41%)
	[0.60/0.40]	87.18 (8.75)	38.01 (−5.09)	70.03 (11.19)	188′920	101′216	290′136	51.21% (47.70%)	50.67% (47.76%)	51.02% (47.72%)
	[0.70/0.30]	93.81 (6.27)	48.05 (−3.92)	77.65 (8.24)	70′870	38′711	109′581	51.82% (47.15%)	50.48% (47.88%)	51.34% (47.41%)
Trading maximum: USD 5 million										
SD classification	[na/na]	182.91 (3.91)	12.12 (−4.70)	36.66 (3.64)	67′724	403′555	471′279	50.52% (48.49%)	51.24% (47.39%)	51.14% (47.55%)
Naïve Bayes (classification)	[0.50/0.50]	99.70 (3.24)	4.18 (−3.55)	58.70 (6.73)	479′897	360′914	840′811	49.83% (47.21%)	50.95% (48.10%)	50.31% (47.53%)
	[0.60/0.40]	169.64 (5.74)	31.70 (−3.17)	121.52 (8.01)	188′920	101′216	290′136	51.17% (47.74%)	50.57% (47.86%)	50.96% (47.78%)
	[0.70/0.30]	204.27 (4.85)	17.48 (−1.73)	138.29 (6.19)	70′870	38′711	109′581	51.75% (47.21%)	50.26% (48.10%)	51.23% (47.53%)
Trading maximum: USD 10 million										
SD classification	[na/na]	248.72 (3.68)	−7.50 (−3.68)	29.32 (2.06)	67′724	403′555	471′279	50.54% (48.46%)	51.23% (47.40%)	51.13% (47.56%)
Naïve Bayes (classification)	[0.50/0.50]	122.29 (2.12)	−26.34 (−2.37)	58.49 (4.80)	479′897	360′914	840′811	49.84% (47.24%)	50.94% (48.13%)	50.31% (47.55%)
	[0.60/0.40]	200.07 (4.26)	−25.95 (−1.45)	121.22 (5.77)	188′920	101′216	290′136	51.18% (47.73%)	50.55% (47.88%)	50.96% (47.78%)
	[0.70/0.30]	236.30 (3.62)	−84.82 (0.06)	122.86 (3.99)	70′870	38′711	109′581	51.73% (47.24%)	50.23% (48.13%)	51.20% (47.55%)

increasing to two minutes and decrease by 14.25% when decreasing to 30 seconds. Note that the buying/selling period parameter strongly affects the number of observations with no trading in the buying/selling period: 284'467 (30 seconds), 209'667 (one minute), and 147'785 (two minutes). But at the same time, prolonging the buying/selling period also increases the number of observations where not all shares are sold till the end of the eight hours window: 161 (30 seconds), 265 (one minute), and 448 (two minutes). Although the profits depend heavily on the length of the buying/selling period, they are qualitatively also robust against changes in this parameter.

However, there is one odd case where results change qualitatively: The profits of a long

Table 7: Trading Strategy of SD and NB [1 sec., 1 min., 30 sec./1min./2 min., 5m max.] - Whole Dataset

The strategies shown here start trading one second after the news is published for one minute, then hold the position for 30 seconds / one minute / two minutes before starting to fully liquidate the position with a maximum of USD five million per trade. Each model predicts a probability (classification) or a return (regression) and the thresholds are used to classify the headlines into positive (≥ positive threshold) and negative (≤ negative threshold). Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 209'667 observations, there is no trading in the buying/selling period. For the different holding periods, the following number of observations are excluded because not all shares are sold till the end of the eight hours window: 255 (30 seconds), 265 (one minute), and 279 (two minutes). Tables A.11 (30 seconds holding period), A.7 (one minute holding period), and A.10 (two minutes holding period) in the appendix show the results of all models.

Model	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Accuracy positive	Accuracy negative	Accuracy all
				Holding period: 30 seconds						
SD classification	[na/na]	167.04	5.63	28.83	67′726	403′557	471′283	50.29%	50.90%	50.81%
		(3.81)	(−4.36)	(3.15)				(48.48%)	(47.52%)	(47.66%)
Naïve Bayes (classification)	[0.50/0.50]	94.84	5.85	56.64	479′902	360′919	840′821	49.76%	50.69%	50.16%
		(3.40)	(−3.71)	(7.02)				(47.28%)	(47.84%)	(47.48%)
	[0.60/0.40]	157.69	33.28	114.29	188′923	101′217	290′140	50.94%	50.45%	50.77%
		(5.77)	(−3.35)	(8.19)				(47.73%)	(47.78%)	(47.75%)
	[0.70/0.30]	182.85	38.71	131.93	70′871	38′711	109′582	51.49%	50.35%	51.09%
		(4.61)	(−2.33)	(6.41)				(47.28%)	(47.84%)	(47.48%)
				Holding period: 1 minute						
SD classification	[na/na]	182.91	12.12	36.66	67′724	403′555	471′279	50.52%	51.24%	51.14%
		(3.91)	(−4.70)	(3.64)				(48.49%)	(47.39%)	(47.55%)
Naïve Bayes (classification)	[0.50/0.50]	99.70	4.18	58.70	479′897	360′914	840′811	49.83%	50.95%	50.31%
		(3.24)	(−3.55)	(6.73)				(47.21%)	(48.10%)	(47.53%)
	[0.60/0.40]	169.64	31.70	121.52	188′920	101′216	290′136	51.17%	50.57%	50.96%
		(5.74)	(−3.17)	(8.01)				(47.74%)	(47.86%)	(47.78%)
	[0.70/0.30]	204.27	17.48	138.29	70′870	38′711	109′581	51.75%	50.26%	51.23%
		(4.85)	(−1.73)	(6.19)				(47.21%)	(48.10%)	(47.53%)
				Holding period: 2 minutes						
SD classification	[na/na]	185.94	23.72	47.03	67′721	403′552	471′273	50.63%	51.54%	51.41%
		(3.77)	(−4.43)	(4.08)				(48.60%)	(47.41%)	(47.58%)
Naïve Bayes (classification)	[0.50/0.50]	101.52	23.70	68.12	479′891	360′906	840′797	49.97%	51.23%	50.51%
		(3.45)	(−3.84)	(6.96)				(47.15%)	(48.37%)	(47.58%)
	[0.60/0.40]	175.12	16.96	119.94	188′915	101′216	290′131	51.48%	50.59%	51.17%
		(5.65)	(−2.10)	(6.79)				(47.65%)	(48.17%)	(47.83%)
	[0.70/0.30]	211.77	19.80	143.95	70′868	38′711	109′579	52.05%	50.36%	51.45%
		(4.74)	(−1.43)	(5.30)				(47.15%)	(48.37%)	(47.58%)

(short) position after negative news for the training dataset displayed in Table A.5 in the appendix are above (below) the mean profit when classifying all news. I am not too concerned with this for the following reasons: First, increasing the threshold and not classifying all news mitigates the problem; hence, it is caused by hard-to-classify observations. Second, I confirm in undisclosed results that it is driven by the large observations and vanishes when reducing the maximum trading amount to USD one million or when increasing the buying/selling period to two minutes.

Table 8: Trading Strategy of SD and NB [1 sec., 30 sec./1 min./2 min., 1 min., 5m max.] - Whole Dataset

The strategies shown here start trading one second after the news is published for 30 seconds / one minute / two minutes, then hold the position for one minute before starting to fully liquidate the position with a maximum of USD five million per trade. Each model predicts a probability (classification) or a return (regression) and the thresholds are used to classify the headlines into positive (≥ positive threshold) and negative (≤ negative threshold). Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 147'785 observations, there is no trading in the buying/selling period. For the different buying/selling period periods, the following number of observations are excluded because not all shares are sold till the end of the eight hours window: 161 (30 seconds), 265 (one minute), and 448 (two minutes). Tables A.13 (30 seconds buying/selling period period), A.7 (one minute buying/selling period), and A.12 (two minutes buying/selling period period) in the appendix show the results of all models.

Model	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Accuracy positive	Accuracy negative	Accuracy all
				Buying/selling period: 30 seconds						
SD classification	[na/na]	154.67	−6.42	15.65	60′227	379′449	439′676	50.09%	50.11%	50.11%
		(3.84)	(−3.59)	(1.98)				(48.48%)	(47.62%)	(47.74%)
Naïve Bayes (classification)	[0.50/0.50]	72.76	−15.73	34.39	433′934	332′181	766′115	49.63%	50.11%	49.84%
		(2.16)	(−2.30)	(4.71)				(47.33%)	(47.88%)	(47.53%)
	[0.60/0.40]	122.16	3.61	80.12	165′426	90′886	256′312	50.80%	50.18%	50.58%
		(4.39)	(−2.24)	(6.31)				(47.64%)	(47.80%)	(47.70%)
	[0.70/0.30]	138.97	−22.04	80.78	61′628	34′880	96′508	51.27%	50.07%	50.84%
		(3.46)	(−0.74)	(4.35)				(47.33%)	(47.88%)	(47.53%)
				Buying/selling period: 1 minute						
SD classification	[na/na]	182.91	12.12	36.66	67′724	403′555	471′279	50.52%	51.24%	51.14%
		(3.91)	(−4.70)	(3.64)				(48.49%)	(47.39%)	(47.55%)
Naïve Bayes (classification)	[0.50/0.50]	99.70	4.18	58.70	479′897	360′914	840′811	49.83%	50.95%	50.31%
		(3.24)	(−3.55)	(6.73)				(47.21%)	(48.10%)	(47.53%)
	[0.60/0.40]	169.64	31.70	121.52	188′920	101′216	290′136	51.17%	50.57%	50.96%
		(5.74)	(−3.17)	(8.01)				(47.74%)	(47.86%)	(47.78%)
	[0.70/0.30]	204.27	17.48	138.29	70′870	38′711	109′581	51.75%	50.26%	51.23%
		(4.85)	(−1.73)	(6.19)				(47.21%)	(48.10%)	(47.53%)
				Buying/selling period: 2 minutes						
SD classification	[na/na]	194.19	47.43	69.46	74′365	421′043	495′408	50.70%	51.58%	51.45%
		(3.45)	(−5.12)	(4.97)				(48.64%)	(47.74%)	(47.87%)
Naïve Bayes (classification)	[0.50/0.50]	128.79	56.32	97.96	518′525	383′985	902′510	50.24%	51.42%	50.74%
		(4.38)	(−4.96)	(8.69)				(47.14%)	(48.25%)	(47.52%)
	[0.60/0.40]	233.17	55.11	172.05	210′147	109′834	319′981	51.59%	50.83%	51.33%
		(7.14)	(−2.98)	(8.57)				(47.69%)	(48.08%)	(47.83%)
	[0.70/0.30]	271.97	103.91	214.10	79′454	41′728	121′182	52.18%	50.60%	51.64%
		(5.67)	(−2.83)	(6.89)				(47.14%)	(48.25%)	(47.52%)

VI Discussion

This paper analyses the largest news dataset ever studied in the financial literature at the second frequency and provides strong support for the EMH. The return pattern around news indicates that financial markets start anticipating the news before publication and pricing the new information in takes about two to three minutes nowadays (see Figure 3). This finding is consistent with Tetlock (2010), who finds four return patterns around news, which indicate that some investors trade before the news and others afterward, concluding that there is asymmetric information among market participants. Furthermore, the time to digest the news has decreased over time (compare Figure 3 with Figures A.1 and A.2 in the appendix). Unsurprisingly, this seems to be a long-term trend since Patell and Wolfson (1984) find that in 1976/1977, it took about five to ten minutes to price in new information. It is also in line with Busse and Green (2002), who show that markets react within one minute to *CNBC TV* analyst calls (they probably find a lower reaction time since they are biased towards larger companies compared to my sample).

Despite this strong support for market efficiency, I show that instantly buying after positive and shorting after negative news for a short time period (results disclosed here go up to two minutes, but also times up to five minutes yield similar results) is a profitable trading strategy that yields significantly positive profits. However, I also show that the capacity of this strategy is very limited, and profits would be USD 50 to 85 million over the whole nine years period (2011-2019). Apart from the limited capacity, implementing such a strategy would face several difficulties. While the pure transaction costs are almost zero nowadays, a trader pursuing such a strategy would create additional demand and have a price impact.[23] Short-selling constraints are another potential issue; however, one could also just implement the long strategy.[24]

This limited profitability may also stem from intraday trading patterns. Holbrook Working already described in the 1950s the prices' tendencies to slightly move in one direction, then revert to the initial price causing a "jiggling" of prices (Working, 1958). This causes subsequent price changes to be negatively correlated. However, the empirical evidence on intraday autocorrelation is not as unambiguous despite a tendency towards negative intraday autocorrelations (e.g., see Zhou (1996); Andersen et al. (2001) and Aït-Sahalia et al. (2011)). This most likely stems from the fact that most empirical papers use returns at some (relatively high) frequency instead of subsequent price changes. I find a statistically highly significant (t-value: -263.48) small negative average one-minute autocorrelation of -2.19% in the eight hours before and after news publication. When only looking at the 30 minutes after news publication, the autocorrelation becomes more negative with a value of -5.71%.

Gao et al. (2018) show that the first 30 minutes of each trading session predict price movements in the last 30 minutes, and they call this effect "intraday momentum." Their reasoning for using the first 30 minutes is that the volume is elevated in this period. They further argue that it is due to the information revealed outside of trading hours and conclude that it takes about 30 minutes to price them in. Two comments on Gao et al. (2018): First, their "intraday momentum" is unrelated to the return patterns I find due to its definition. Second, I would argue that concluding that it takes 30 minutes to price in new information, simply based on the observed trading volume patterns, is inadequate. While I agree that new information elevates the trading volume in the first minutes, multiple explanations exist for the entire first (and last) 30 minutes. Gao et al. (2018) point out themselves that

[23] In this paper, I use the prices of all executed trades and simply assume that one would have taken one side of these trades.

[24] Since the investment periods are only a few minutes long on average and highly concentrated in the first trading minutes, it would not be self-financing anyway.

institutional investors use the liquidity in this period to rebalance their portfolios. Hence, the liquidity level in the first and last 30 minutes could also be a Nash equilibrium[25] of the institutional investors. Day traders are another reason for the volume pattern since they open positions in the morning and close them before the end of a trading session.[26]
The finding that the market anticipates the news also suggests that returns before publication should have some predictive power for the returns after it. Therefore, one could also use past returns as the trading signal. The results for this approach are shown in Table 9. They depend heavily on the chosen time window for measuring the past return and are best at time horizons below five minutes and above two hours when only considering the profits and their statistical significance. However, the accuracies at the short horizon are similar to the ones of the medium term and only increase at the long horizon. Therefore, one should be careful about interpreting too much into the predictive power of short-term returns. Including the last overnight return boosts the predictive power significantly (the case for about 75% of the observations when using the last eight hours) and works as well as the NLP models. One explanation could be that sophisticated and potentially partly informed traders choose to trade a trading day in advance in order to be able to trade larger volumes due to the daily volume distribution. This is not necessarily because there is insufficient volume to acquire the desired position but rather to reduce the price impact. Another driver could be a positive autocorrelation of news: positive past news result in positive past returns, and since good past news are followed by other good news, positive past returns predict a positive reaction to future news. In my sample, the average autocorrelation per firm of the predictions ranges between 7.76% and 14.04%, i.e., confirming that positive news is indeed followed by other positive news. The true autocorrelation of news is probably larger than the values provided by the models because they do not provide the true label of the news but rather a prediction. Since the next trading day's return is used as the label to train the models in Pichler (2023), the autocorrelation of the labels, which is at 22.86%, provides an alternative estimator for the true autocorrelation of news. The positive autocorrelation of news indicates that the predictive power of a "news sentiment index" that aggregates the news over a certain time period seems a promising approach for future research. Heston and Sinha (2017) is one of the few comparable studies which looks at about 900'000 news (also from *Reuters*) between 2003 and 2010 and creates such an index. They show that while one particular news only has predictive power for up to two days, the weekly aggregate can predict returns for up to one quarter.

This paper offers several key insights about the transmission of information in financial markets without suffering from weak statistical power, which is often an issue in the empirical evidence of this literature. First of all, it is essential to highlight some assump-

[25] As institutional investors rely on liquidity because they need to trade large amounts, it is a dominating strategy to trade when also expecting others to trade; hence no one has the incentive to deviate once high volume times are established.

[26] Note that the pattern caused by day traders would be the opposite of what Gao et al. (2018) find; hence, it does not explain their finding regarding the return pattern.

Table 9: Comparison of Past Returns as Predictors [1 sec., 1 min., 1 min., 5m max.]

The strategies shown here start trading one second after the news is published for one minute, then hold the position for one minute before starting to fully liquidate the position with a maximum of USD five million per trade. The direction of trading depends on the past return (not including the jump at news publication) at different horizons (one minute to eight hours), i.e., buying after past returns $\geq 0\%$ and selling otherwise. Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 209' 667 observations, there is no trading in the buying/selling period, and 265 observations are excluded because not all shares are sold till the end of the eight hours window.

Predictor (past return)	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Accuracy positive	Accuracy negative	Accuracy all
1 minute	[0.0%/ 0.0%]	73.13	−33.50	24.63	458′399	382′412	840′811	48.63%	49.89%	49.20%
		(1.21)	(−1.41)	(2.82)				(49.68%)	(49.20%)	(49.46%)
5 minutes	[0.0%/ 0.0%]	79.88	−28.84	27.12	432′734	408′077	840′811	48.62%	49.64%	49.12%
		(1.62)	(−1.76)	(3.11)				(49.90%)	(49.17%)	(49.55%)
10 minutes	[0.0%/ 0.0%]	67.53	−42.27	13.53	427′243	413′568	840′811	48.68%	49.67%	49.17%
		(0.81)	(−0.86)	(1.55)				(49.88%)	(49.10%)	(49.50%)
30 minutes	[0.0%/ 0.0%]	63.39	−46.56	9.29	427′082	413′729	840′811	48.66%	49.57%	49.10%
		(0.55)	(−0.57)	(1.06)				(49.98%)	(49.13%)	(49.56%)
1 hour	[0.0%/ 0.0%]	72.00	−37.53	18.32	428′706	412′105	840′811	48.78%	49.68%	49.22%
		(1.11)	(−1.17)	(2.10)				(49.88%)	(49.00%)	(49.45%)
2 hours	[0.0%/ 0.0%]	85.25	−23.64	31.98	429′511	411′300	840′811	49.01%	49.89%	49.44%
		(1.98)	(−2.10)	(3.67)				(49.67%)	(48.77%)	(49.23%)
4 hours	[0.0%/ 0.0%]	89.07	−18.73	37.02	434′816	405′995	840′811	49.26%	50.15%	49.69%
		(2.24)	(−2.42)	(4.24)				(49.43%)	(48.49%)	(48.98%)
8 hours	[0.0%/ 0.0%]	101.58	8.87	62.57	487′035	353′776	840′811	49.73%	50.91%	50.23%
		(3.31)	(−3.89)	(7.17)				(48.96%)	(47.73%)	(48.44%)

tions regarding the nature of the dataset. One can reasonably assume that for most news events in my sample, market participants know that some information will arrive as well as when;[27] however, they are not sure about the content and its price implications. Therefore, the excess return[28] pattern of all news displayed in Figure 5 can be used to study the impact of resolving uncertainty. It reveals that there is a risk premium for holding stocks where the arrival of new information is expected (red line). The risk premium is divided into about 0.05% for the 6.5 hours before publication and a premium at publication, which is not as straightforward to estimate. Simply taking the jump at time zero would overestimate the premium because it is also an overnight return for roughly three-quarters of the observations. However, the pattern at the next overnight return (at 6.5 hours) indicates only a minimal impact of the overnight return; hence, an estimation of another 0.05% seems reasonable. Another interesting observation is the shape of the risk premium: it is growing exponentially the closer the point in time gets, where uncertainty is resolved. However, the strong day jump the day before (at -6.5h), which is then reversed in the next 90 minutes, is

[27]Companies schedule the release of earnings information in advance and tend to schedule the release of other relevant information in advance.

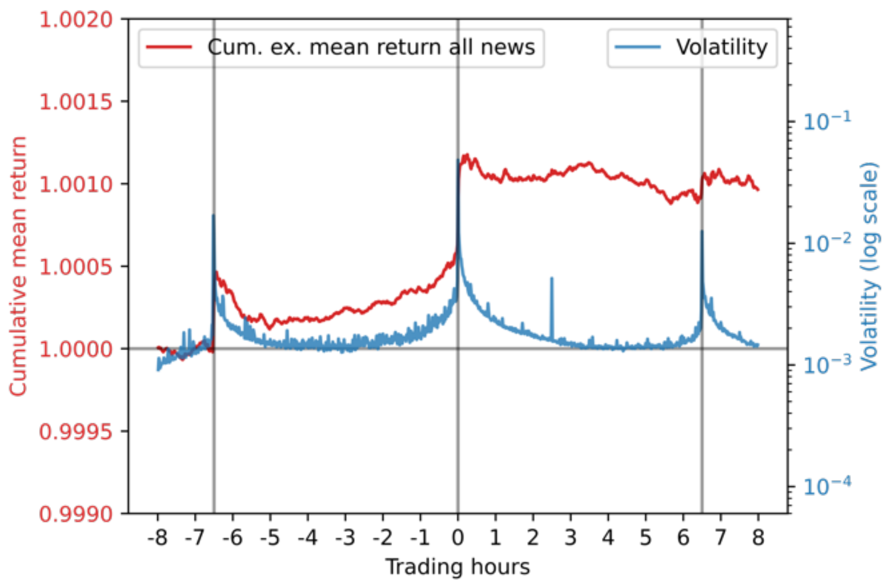
[28]In excess of the average one-minute S&P 500 return of the sample period.

puzzling.

When interpreting the volatility pattern of Figure 5, one must again consider that some of it is due to intraday patterns. The U-shape of intraday volatility is well documented in the literature (e.g., see Figure 1 of Chan et al. (1991) or Figure 1 of Engle and Sokalska (2012)) and according to Ederington and Lee (1993) related to macroeconomic news.[29] However, the increase associated with firm-specific news is clearly above the normal intraday pattern (see pattern at -6.5h and +6.5h). Apart from the volatility of overnight returns, which is compensated by a positive average excess return, the usual volatility increase at the beginning and toward the end of a trading session is not associated with a compensating return. But the volatility increase before the news is compensated by a higher return, whereas the elevated volatility after publication is not. One may interpret this as follows: The resolution of uncertainty concerning the information content is compensated with a risk premium, i.e., the risk premium decreases to zero at the resolution, but there is no compensation on average for the uncertainty regarding how to interpret the information.

Figure 5: Cumulative Mean Excess Return and Volatility of all News

This figure shows the cumulative mean return in excess of the average one-minute return of the S&P 500 and its volatility during the sample period for all news. The increase of the mean return on the left side is the risk premium for holding stocks where the market expects new information to arrive. Volatility is log scaled (right side) and set such that the "normal" level is at the 0% cumulative return. The vertical grey lines mark the points where about three-quarters of the observations have the overnight return; hence the pattern between the two of them is the typical U-shaped pattern of intraday volatility.



[29] Ederington and Lee (1993) use bond and forex data, which does not necessarily fully translate to stock volatility patterns.

VII Conclusion

This paper investigates the market micro reaction to new information and shows that it nowadays takes two to three minutes for financial markets to price it in. Furthermore, I provide evidence that this timeframe has gotten shorter over time. I show that a trading strategy that buys after good news and sells after bad for a short time period (up to a few minutes) yields statistically significant returns. However, profits are limited due to the short trading time and are estimated to be about USD 50 to 85 million for the nine years sample period. Since the market anticipates the news, past returns work similarly well as predictors for the trading profits as the used news classification models, as long as the return is measured over a long enough period (e.g., the previous eight trading hours). Additionally, I show that volatility increases before the news and investors receive a risk premium for holding stocks before news of about 0.05% in the previous 6.5 hours and another 0.05% when the news is published. While volatility remains elevated after the news is published, there is no return compensation for it. One may interpret this as follows: The resolution of uncertainty concerning the information content is compensated with a risk premium, i.e., the risk premium decreases to zero at the resolution, but there is no compensation on average for the uncertainty regarding how to interpret the information.

Appendix Essay 2

Table A.1: CNN Architecture

This table shows the architecture of the applied CNN. The max_tokens hyperparameter is set to 17'000 to drop all words occurring less than roughly 100 times in the dataset. This prevents the model from overfitting on less frequent words and reduces the model size significantly. The dimensionality of the vector representation is set to four. The kernel size of the first convolutional layers determines how many words can be reached per filter, and by stacking two convolutional layers, the maximum of possible word dependencies is equal to the product of the two kernels (here: $3 \cdot 3 = 9$). The global average pooling layer removes one dimension by averaging across the second dimension (sentence length). Alternatively, one may also use another pooling procedure, like max pooling. One dense layer is then added with 20 units and some L2 regularization to prevent overfitting too fast. The last layer is determined by the labels: the number of units equals the number of labels per example (1), and the activation depends on the type of labels (*sigmoid* for binary classes, *softmax* for multiple classes, and *linear* for numerical). The network is trained by minimizing the binary cross-entropy (classification) or the mean squared error (regression) using the *Adam* optimizer (learning rate=0.001, beta 1=0.9, beta 2=0.999, epsilon=1e-7, and amsgrad=False) and the mini-batch size is 4'096.

Layer	Parameters	Output dimension	Number of parameters
TextVectorization	max_tokens=17'000, output_dim=4	None, None, 4	68'000
Conv1D	filters=24, kernel_size=3, padding='same', activation='relu'	None, None, 24	312
Conv1D	filters=12, kernel_size=3, padding='same', activation='relu'	None, None, 12	876
GlobalAveragePooling1D	-	None, 12	0
Dense	units=20, activation='relu'	None, 20	220
Dropout	rage=0.1	None, 20	0
Dense	units=1, activation='linear' / activation='sigmoid'	None, 1	21
		Total number of parameters	69'469

Table A.2: RNN-LSTM Architecture

This table shows the architecture of the applied RNN with LSTM units. The max_tokens hyperparameter is set to 17'000 to drop all words occurring less than roughly 100 times in the dataset. This prevents the model from overfitting on less frequent words and reduces the model size significantly. The dimensionality of the vector representation is set to four. The model uses ten LSTM units, a dense layer with 20 units, and some L2 regularization to prevent overfitting too fast. The last layer is determined by the labels: the number of units equals the number of labels per example (1), and the activation depends on the type of labels (*sigmoid* for binary classes, *softmax* for multiple classes, and *linear* for numerical). The network is trained by minimizing binary cross-entropy (classification) or the mean squared error (regression) using the *Adam* optimizer (learning rate=0.001, beta 1=0.9, beta 2=0.999, epsilon=1e-7, and amsgrad=False) and the mini-batch size is 4'096.

Layer	Parameters	Output dimension	Number of parameters
TextVectorization	max_tokens=17'000, output_dim=4	None, None, 4	68'000
Bidirectional	LSTM(units=10)	None, 20	1'200
Dense	units=20, activation='relu', kernel_regularizer=(L2=0.001)	None, 20	420
Dropout	rate=0.1	None, 20	0
Dense	units=1, activation='linear' / activation='sigmoid'	None, 1	21
		Total number of parameters:	69'641

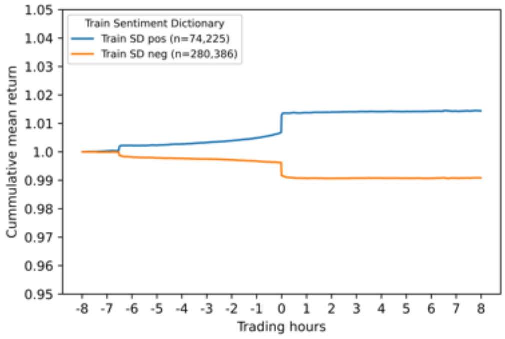
Table A.3: RNN-GRU Architecture

This table shows the architecture of the applied RNN with GRU units. The max_tokens hyperparameter is set to 17'000 to drop all words occurring less than roughly 100 times in the dataset. This prevents the model from overfitting on less frequent words and reduces the model size significantly. The dimensionality of the vector representation is set to four. The model uses ten GRU units, a dense layer with 20 units, and some L2 regularization to prevent overfitting too fast. The last layer is determined by the labels: the number of units equals the number of labels per example (1), and the activation depends on the type of labels (*sigmoid* for binary classes, *softmax* for multiple classes, and *linear* for numerical). The network is trained by minimizing binary cross-entropy (classification) or the mean squared error (regression) using the *Adam* optimizer (learning rate=0.001, beta 1=0.9, beta 2=0.999, epsilon=1e-7, and amsgrad=False) and the mini-batch size is 4'096.

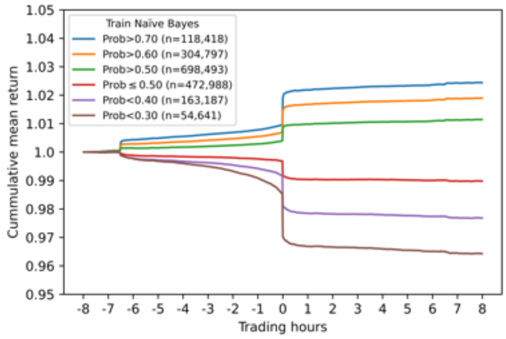
Layer	Parameters	Output dimension	Number of parameters
TextVectorization	max_tokens=17'000, output_dim=4	None, None, 4	68'000
Bidirectional	GRU(units=10)	None, 20	960
Dense	units=20, activation='relu', kernel_regularizer=(L2=0.001)	None, 20	420
Dropout	rate=0.1	None, 10	0
Dense	units=1, activation='linear' / activation='sigmoid'	None, 1	21
		Total number of parameters:	69'401

Figure A.1: Cumulative Mean Returns on Positive/Negative News - Training Dataset

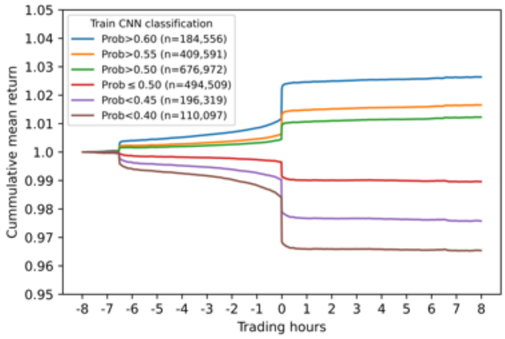
This figure shows the cumulative mean returns of positive and negative news according to the indicated model for the eight hours before and after news publication for the training dataset. The x-axis shows the trading time relative to news publication at zero. Markets start anticipating the news and then digest them quickly, and very soon, there is no trend in the stock price.



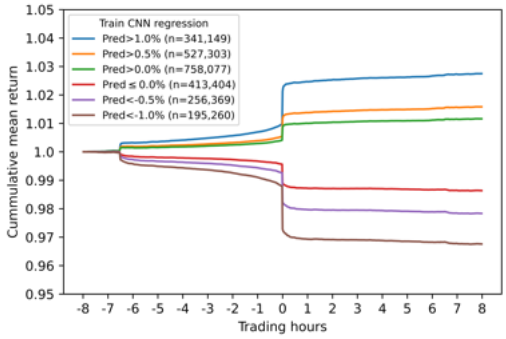
(a) Sentiment Dictionary Classification



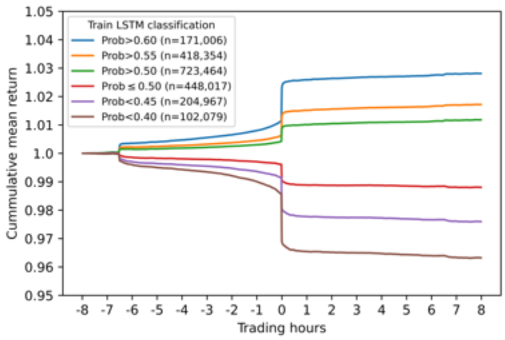
(b) Naïve Bayes Classification



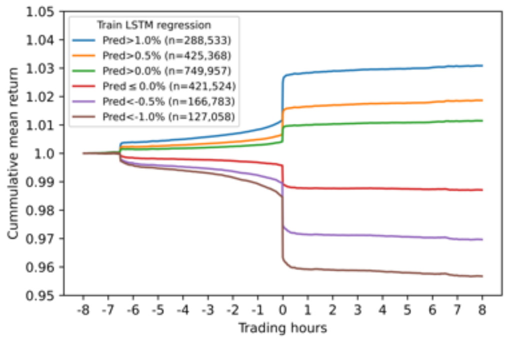
(c) CNN Classification



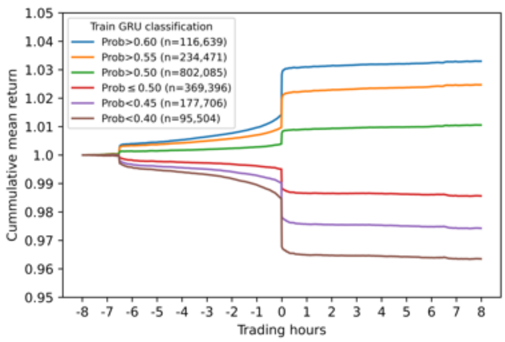
(d) CNN Regression



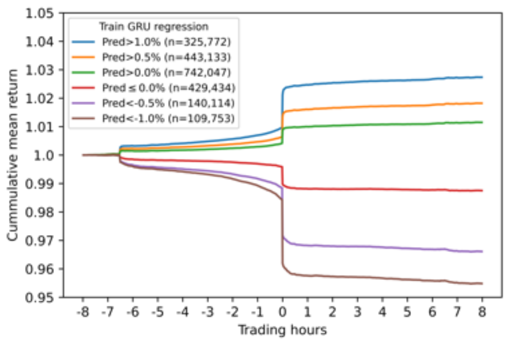
(e) RNN-LSTM Classification



(f) RNN-LSTM Regression



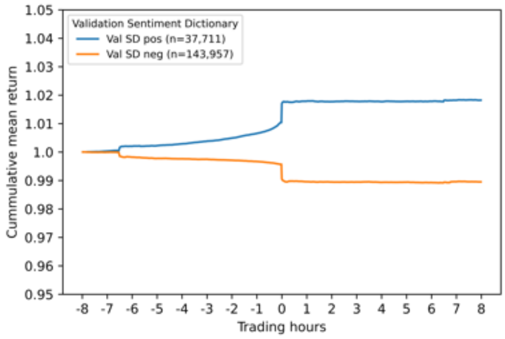
(g) RNN-GRU Classification



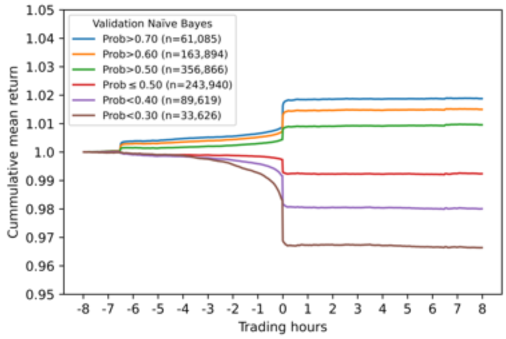
(h) RNN-GRU Regression

Figure A.2: Cumulative Mean Returns on Positive/Negative News - Validation Dataset

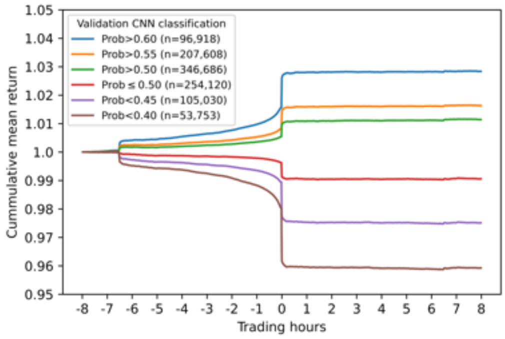
This figure shows the cumulative mean returns of positive and negative news according to the indicated model for the eight hours before and after news publication for the validation dataset. The x-axis shows the trading time relative to news publication at zero. Markets start anticipating the news and then digest them quickly, and very soon, there is no trend in the stock price.



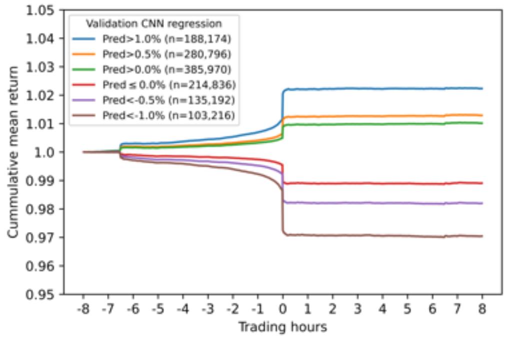
(a) Sentiment Dictionary Classification



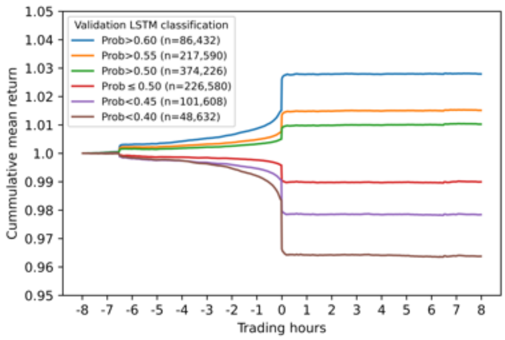
(b) Naïve Bayes Classification



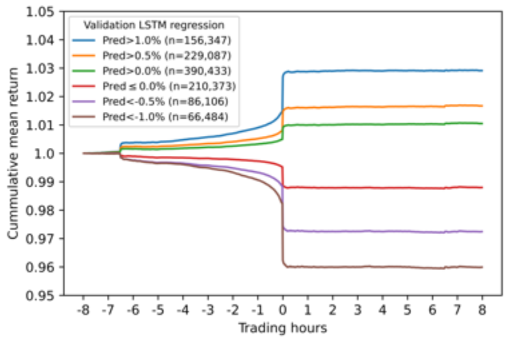
(c) CNN Classification



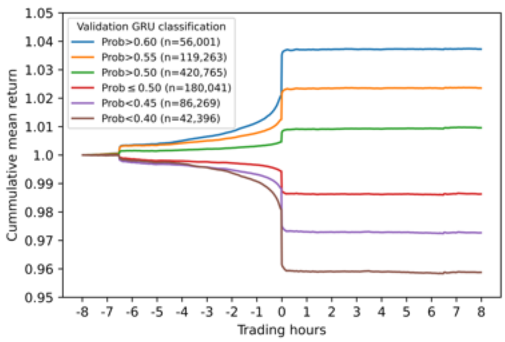
(d) CNN Regression



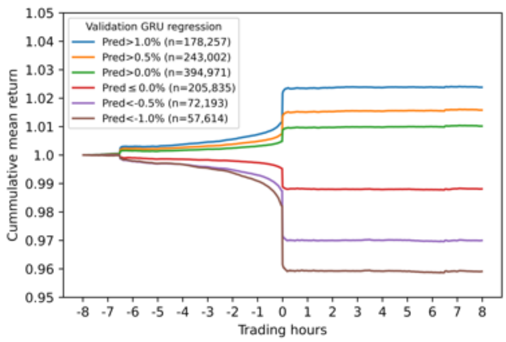
(e) RNN-LSTM Classification



(f) RNN-LSTM Regression



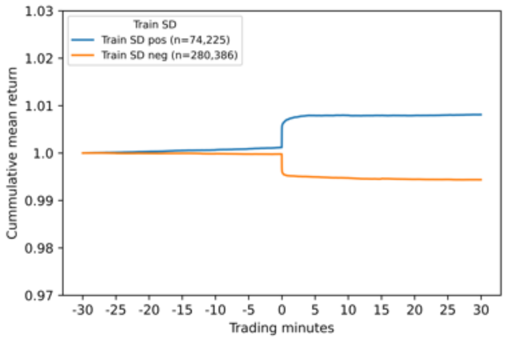
(g) RNN-GRU Classification



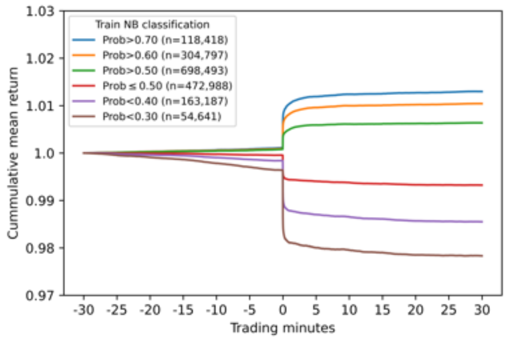
(h) RNN-GRU Regression

Figure A.3: Zoom in on Cumulative Mean Returns on Positive/Negative News - Training Dataset SD and NB

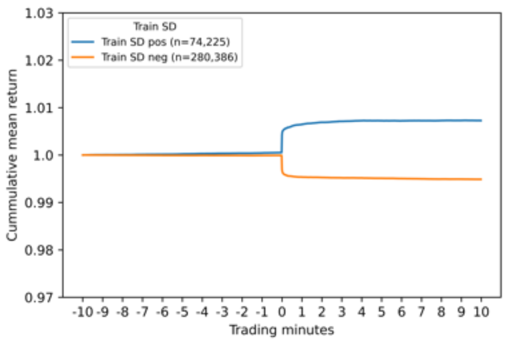
This figure zooms in on the cumulative mean returns of positive and negative news according to the sentiment dictionary and the Naïve Bayes classification for the eight hours before and after news publication for the training dataset. The x-axis shows the trading time relative to news publication at zero. Markets start anticipating the news and then digest them quickly, and very soon, there is no trend in the stock price. Figure 3 shows the test dataset results, and A.4 in the appendix shows the validation dataset results.



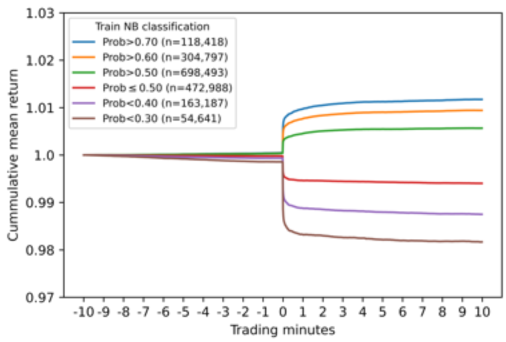
(a) Sentiment Dictionary +/- 30 min.



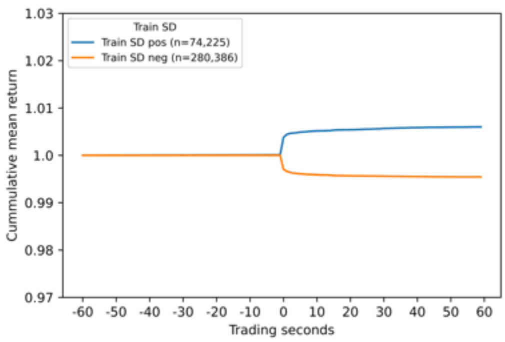
(b) Naïve Bayes +/- 30 min.



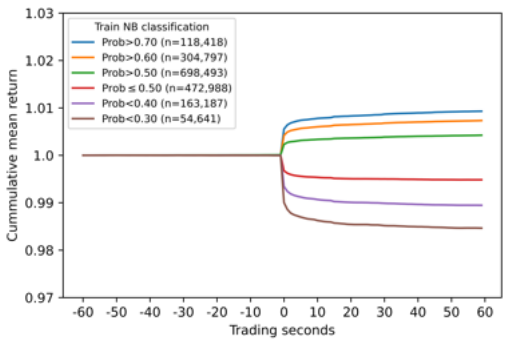
(c) Sentiment Dictionary +/- 10 min.



(d) Naïve Bayes +/- 10 min.



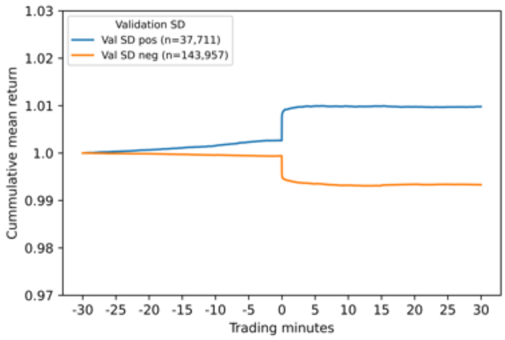
(e) Sentiment Dictionary +/- 1 min.



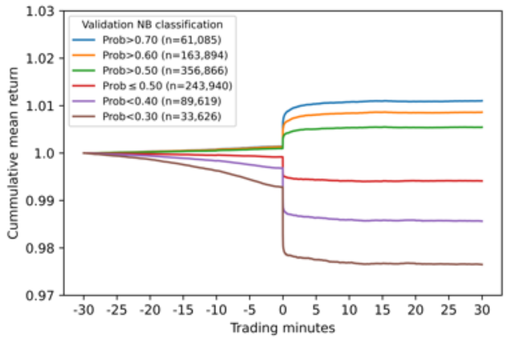
(f) Naïve Bayes +/- 1min.

Figure A.4: Zoom in on Cumulative Mean Returns on Positive/Negative News - Validation Dataset SD and NB

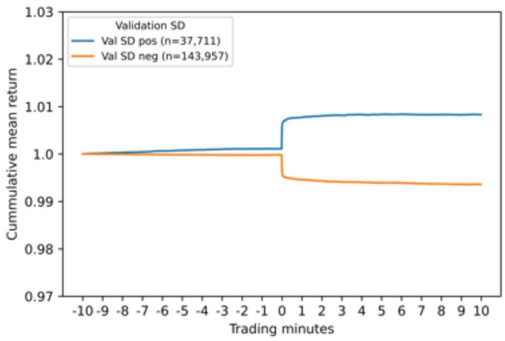
This figure zooms in on the cumulative mean returns of positive and negative news according to the sentiment dictionary and the Naïve Bayes classification for the eight hours before and after news publication for the validation dataset. The x-axis shows the trading time relative to news publication at zero. Markets start anticipating the news and then digest them quickly, and very soon, there is no trend in the stock price. Figure 3 shows the test dataset results, and A.3 in the appendix shows the training dataset results.



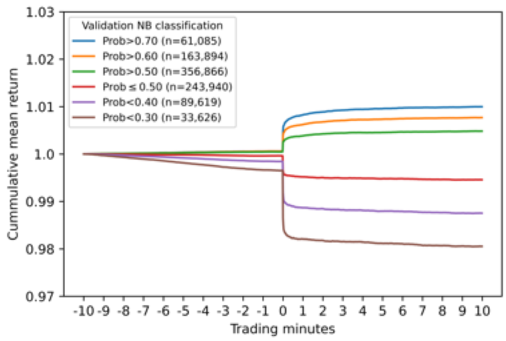
(a) Sentiment Dictionary +/- 30 min.



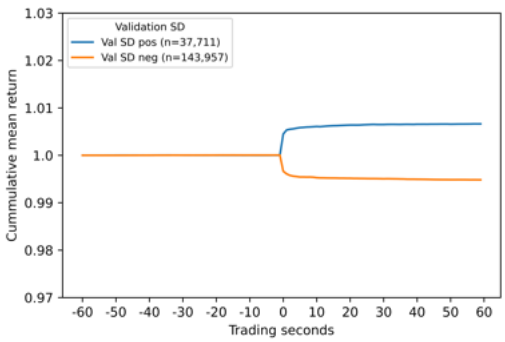
(b) Naïve Bayes +/- 30 min.



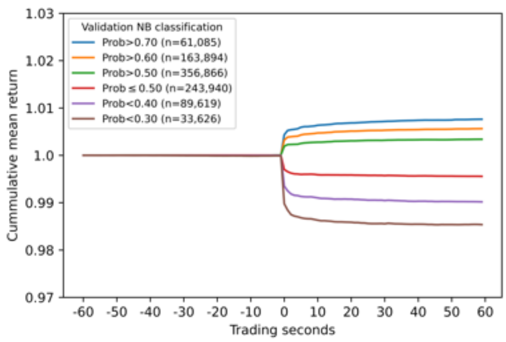
(c) Sentiment Dictionary +/- 10 min.



(d) Naïve Bayes +/- 10 min.



(e) Sentiment Dictionary +/- 1 min.



(f) Naïve Bayes +/- 1min.

Table A.4: Descriptive Statistics of the Trading Strategies

This table shows the descriptive statistics for trading profit and amount of the different trading strategies and datasets. The first parameter of each strategy is the waiting period between news publication and trading start, the next is the trading time (buy/sell), the third is the holding period, and the fourth is the maximum trading amount per sample.

	Trading profit				Trading amount			
	Train	Validation	Test	Full sample	Train	Validation	Test	Full sample
Trading strategy: [1sec., 1 min., 1 min., 5m max.]								
Mean	74.39	34.27	36.60	55.11	632′715.96	551′651.93	548′290.04	591′793.37
Std.	7′629.15	7′203.97	9′290.29	7′998.26	1′037′261.59	980′799.04	995′041.21	1′014′148.44
Min.	−378′519.86	−271′678.04	−693′077.42	−693′077.42	39.50	30.00	11.90	11.90
25%-quantile	−158.96	−133.85	−140.74	−147.88	47′918.52	38′480.95	37′985.00	42′787.25
Median	0.00	0.00	−1.00	0.00	220′724.30	176′521.75	172′706.50	196′274.30
75%-quantile	156.93	132.22	120.50	140.82	707′653.67	567′163.92	546′355.49	630′296.27
Max.	1′749′576.86	567′158.70	1′299′850.28	1′749′576.86	5′000′000.00	4′999′999.99	4′999′999.99	5′000′000.00
Observations	426′017.00	197′570.00	217′298.00	840′811.00	426′017.00	197′570.00	217′298.00	840′811.00
Trading strategy: [1sec., 1 min., 1 min., 10m max.]								
Mean	101.53	53.90	66.67	81.11	705′775.51	618′419.20	625′393.58	664′395.83
Std.	10′362.87	9′803.36	13′574.81	11′162.06	1′419′730.90	1′352′536.14	1′413′108.09	1′402′956.30
Min.	−719′288.82	−593′071.68	−1′194′164.85	−1′194′164.85	39.50	30.00	11.90	11.90
25%-quantile	−158.84	−133.85	−141.00	−147.87	47′918.52	38′480.95	37′985.00	42′787.25
Median	0.00	0.00	−1.00	0.00	220′724.30	176′521.75	172′706.50	196′276.24
75%-quantile	157.24	132.32	120.50	141.00	707′653.67	567′163.92	546′355.49	630′299.35
Max.	2′120′033.37	1′117′329.25	1′678′749.02	2′120′033.37	9′999′999.99	10′000′000.00	9′999′999.98	10′000′000.00
Observations	426′017.00	197′570.00	217′298.00	840′811.00	426′017.00	197′570.00	217′298.00	840′811.00
Trading strategy: [1sec., 1 min., 1 min., 1m max.]								
Mean	29.54	9.88	−0.25	17.20	378′050.61	336′261.82	330′314.67	355′881.43
Std.	2′992.63	3′029.47	3′592.32	3′166.49	374′790.95	358′319.13	355′556.74	366′745.49
Min.	−245′442.72	−113′707.21	−227′674.42	−245′442.72	39.50	30.00	11.90	11.90
25%-quantile	−152.22	−130.42	−136.64	−142.67	47′918.52	38′480.95	37′985.00	42′787.25
Median	0.00	0.00	−0.94	0.00	220′721.60	176′513.85	172′700.05	196′264.30
75%-quantile	150.23	129.06	118.25	136.26	707′144.59	567′013.63	546′257.78	630′033.52
Max.	713′176.27	178′582.98	352′077.28	713′176.27	1′000′000.00	1′000′000.00	1′000′000.00	1′000′000.00
Observations	426′017.00	197′570.00	217′298.00	840′811.00	426′017.00	197′570.00	217′298.00	840′811.00
Trading strategy: [1sec., 1 min., 2 min., 5m max.]								
Mean	72.47	28.10	17.85	47.77	632′727.48	551′656.01	548′296.50	591′801.75
Std.	8′565.76	8′142.44	10′364.86	8′971.99	1′037′267.92	980′805.69	995′046.47	1′014′154.64
Min.	−542′603.72	−444′379.04	−601′207.43	−601′207.43	39.50	30.00	11.90	11.90
25%-quantile	−194.89	−165.88	−171.81	−181.50	47′922.84	38′481.90	37′986.00	42′790.18
Median	0.00	0.00	−1.00	0.00	220′733.00	176′516.56	172′709.00	196′281.38
75%-quantile	189.99	162.69	148.47	172.00	707′656.42	567′179.67	546′372.72	630′316.37
Max.	1′779′859.77	664′550.70	1′299′850.28	1′779′859.77	5′000′000.00	4′999′999.99	4′999′999.99	5′000′000.00
Observations	426′009.00	197′567.00	217′295.00	840′797.00	426′009.00	197′567.00	217′295.00	840′797.00
Trading strategy: [1sec., 1 min., 30 sec., 5m max.]								
Mean	62.37	38.89	42.00	51.62	632′707.64	551′646.46	548′285.21	591′786.72
Std.	7′097.69	6′593.89	8′570.19	7′396.35	1′037′256.68	980′795.58	995′037.90	1′014′144.25
Min.	−452′287.46	−270′726.55	−895′830.03	−895′830.03	39.50	30.00	11.90	11.90
25%-quantile	−135.00	−116.00	−121.79	−126.68	47′916.40	38′478.75	37′983.36	42′785.00
Median	0.00	0.00	−0.50	0.00	220′720.17	176′515.88	172′702.05	196′264.37
75%-quantile	135.70	115.20	104.43	122.00	707′634.77	567′161.11	546′335.48	630′285.06
Max.	1′803′577.29	495′022.43	1′299′850.28	1′803′577.29	5′000′000.00	4′999′999.99	4′999′999.99	5′000′000.00
Observations	426′023.00	197′572.00	217′300.00	840′821.00	426′023.00	197′572.00	217′300.00	840′821.00
Trading strategy: [1sec., 2 min., 1 min., 5m max.]								
Mean	82.77	27.30	8.16	50.03	1′010′578.92	854′345.81	854′514.33	932′776.67
Std.	10′268.83	10′023.19	12′066.00	10′708.18	1′368′611.24	1′272′528.65	1′280′403.87	1′325′842.54
Min.	−542′603.72	−542′914.68	−601′207.43	−601′207.43	22.00	28.78	11.50	11.50
25%-quantile	−337.94	−272.27	−287.93	−307.74	87′507.38	65′232.00	65′328.05	75′331.94
Median	0.00	0.00	−1.00	−0.02	415′084.68	313′195.93	311′702.80	359′913.76
75%-quantile	324.85	270.70	259.35	293.59	1′306′916.73	1′009′105.03	992′445.54	1′151′443.64
Max.	1′779′859.77	707′417.27	1′299′850.28	1′779′859.77	5′000′000.00	5′000′000.00	5′000′000.00	5′000′000.00
Observations	453′300.00	214′914.00	234′378.00	902′510.00	453′300.00	214′914.00	234′378.00	902′510.00
Trading strategy: [1sec., 30 sec., 1 min., 5m max.]								
Mean	49.45	44.88	47.92	48.03	404′717.82	359′467.32	358′028.48	382′065.63
Std.	6′019.11	5′403.65	7′761.39	6′387.73	787′055.51	753′099.83	777′634.66	777′082.26
Min.	−364′360.86	−270′726.55	−895′830.03	−895′830.03	13.12	37.00	11.90	11.90
25%-quantile	−82.68	−72.16	−74.00	−77.83	28′451.75	23′905.00	23′318.75	25′917.87
Median	0.00	0.00	−0.05	0.00	124′637.95	104′334.25	100′023.40	112′613.70
75%-quantile	84.02	74.77	67.00	77.18	404′847.97	332′722.16	315′112.59	363′453.37
Max.	1′541′613.33	495′022.43	1′299′850.28	1′541′613.33	5′000′000.00	4′999′999.98	4′999′999.96	5′000′000.00
Observations	389′820.00	177′965.00	198′395.00	766′115.00	389′820.00	177′965.00	198′395.00	766′115.00

Table A.5: Trading Strategy of all Models [1 sec., 1 min., 1 min., 5m max.] - Training Dataset

The strategies shown here start trading one second after the news is published for one minute, then hold the position for one minute before starting to fully liquidate the position with a maximum of USD five million per trade. Each model predicts a probability (classification) or a return (regression), and the thresholds are used to classify the headlines into positive ($\geq$ positive threshold) and negative ($\leq$ negative threshold). Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 87'309 observations, there is no trading in the buying/selling period, and 125 observations are excluded because not all shares are sold till the end of the eight hours window.

Model	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Accuracy positive	Accuracy negative	Accuracy all
SD classification	[na/na]	184.40	−13.10	18.63	36′417	190′202	226′619	50.87%	50.88%	50.88%
		(2.61)	(−3.17)	(1.46)				(48.14%)	(48.00%)	(48.02%)
Naïve Bayes (classification)	[0.50/0.50]	108.20	−28.10	50.68	246′215	179′802	426′017	50.39%	50.75%	50.54%
		(1.86)	(−2.03)	(4.34)				(46.65%)	(48.99%)	(47.41%)
	[0.60/0.40]	164.00	−70.08	83.15	96′918	51′133	148′051	51.53%	50.46%	51.16%
		(3.39)	(−0.12)	(4.06)				(47.33%)	(48.06%)	(47.58%)
	[0.70/0.30]	237.85	−110.45	125.19	37′335	17′850	55′185	52.31%	49.42%	51.37%
		(4.05)	(0.62)	(4.57)				(46.65%)	(48.99%)	(47.41%)
CNN (classification)	[0.50/0.50]	99.69	−40.47	39.83	244′053	181′964	426′017	50.36%	50.71%	50.51%
		(1.39)	(−1.50)	(3.41)				(47.29%)	(48.02%)	(47.57%)
	[0.55/0.45]	113.22	−56.69	66.62	161′913	61′173	223′086	50.43%	50.46%	50.44%
		(1.83)	(−0.51)	(4.23)				(48.43%)	(48.18%)	(48.36%)
	[0.60/0.40]	165.60	−75.77	74.52	62′353	37′789	100′142	51.54%	50.63%	51.19%
		(2.81)	(0.03)	(2.69)				(47.29%)	(48.02%)	(47.57%)
RNN-LSTM (classification)	[0.50/0.50]	97.80	−38.10	44.52	259′011	167′006	426′017	50.34%	50.90%	50.56%
		(1.31)	(−1.54)	(3.81)				(47.17%)	(47.94%)	(47.45%)
	[0.55/0.45]	119.13	−18.95	82.07	163′895	60′113	224′008	50.44%	50.73%	50.52%
		(2.12)	(−1.58)	(5.12)				(48.47%)	(47.76%)	(48.28%)
	[0.60/0.40]	183.20	−108.85	75.60	55′846	32′580	88′426	51.72%	50.68%	51.33%
		(3.15)	(0.75)	(2.34)				(47.17%)	(47.94%)	(47.45%)
RNN-GRU (classification)	[0.50/0.50]	66.45	−95.55	22.21	309′677	116′340	426′017	49.65%	50.08%	49.77%
		(−0.49)	(0.75)	(1.90)				(47.07%)	(48.26%)	(47.60%)
	[0.55/0.45]	208.54	−74.66	89.34	77′104	56′042	133′146	51.72%	50.58%	51.24%
		(4.48)	(0.01)	(3.55)				(47.18%)	(48.05%)	(47.55%)
	[0.60/0.40]	197.54	−89.46	70.56	40′646	32′255	72′901	51.84%	50.44%	51.22%
		(3.09)	(0.33)	(1.91)				(47.07%)	(48.26%)	(47.60%)
CNN (regression)	[0.0%/ 0.0%]	95.54	−39.98	43.97	263′894	162′123	426′017	50.21%	50.79%	50.43%
		(1.19)	(−1.45)	(3.76)				(47.54%)	(48.31%)	(47.84%)
	[0.5%/-0.5%]	127.41	−125.89	52.08	183′496	77′664	261′160	50.34%	50.14%	50.28%
		(2.58)	(1.54)	(2.99)				(48.48%)	(48.56%)	(48.50%)
	[1.0%/-1.0%]	201.86	−132.48	72.23	92′126	58′339	150′465	51.27%	50.33%	50.90%
		(4.52)	(1.54)	(2.64)				(47.54%)	(48.31%)	(47.84%)
RNN-LSTM (regression)	[0.0%/ 0.0%]	98.21	−38.97	43.06	254′738	171′279	426′017	50.25%	50.70%	50.43%
		(1.32)	(−1.54)	(3.68)				(47.42%)	(48.04%)	(47.62%)
	[0.5%/-0.5%]	192.67	−185.05	78.83	111′320	48′024	159′344	51.28%	50.40%	51.02%
		(4.56)	(2.65)	(2.95)				(47.52%)	(48.28%)	(47.75%)
	[1.0%/-1.0%]	219.05	−231.18	73.95	76′585	36′419	113′004	51.43%	50.57%	51.15%
		(4.75)	(3.30)	(2.11)				(47.42%)	(48.04%)	(47.62%)
RNN-GRU (regression)	[0.0%/ 0.0%]	106.71	−30.09	49.01	246′326	179′691	426′017	50.21%	50.54%	50.35%
		(1.75)	(−1.97)	(4.19)				(47.48%)	(48.14%)	(47.66%)
	[0.5%/-0.5%]	196.92	−213.51	89.22	113′657	40′436	154′093	51.29%	50.37%	51.05%
		(4.74)	(3.07)	(3.22)				(47.48%)	(48.27%)	(47.68%)
	[1.0%/-1.0%]	218.45	−183.70	109.29	81′880	30′507	112′387	51.39%	50.40%	51.12%
		(4.83)	(2.16)	(3.19)				(47.48%)	(48.14%)	(47.66%)

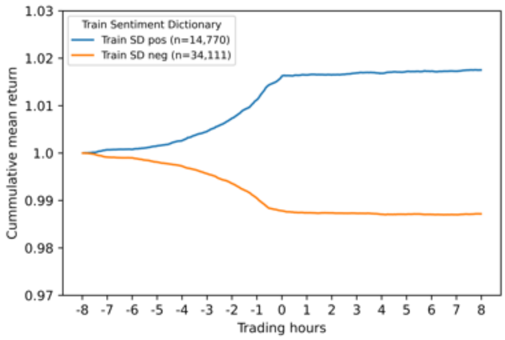
Table A.6: Trading Strategy of all Models [1 sec., 1 min., 1 min., 5m max.] - Validation Dataset

The strategies shown here start trading one second after the news is published for one minute, then hold the position for one minute before starting to fully liquidate the position with a maximum of USD five million per trade. Each model predicts a probability (classification) or a return (regression), and the thresholds are used to classify the headlines into positive (≥ positive threshold) and negative (≤ negative threshold). Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 60'695 observations, there is no trading in the buying/selling period, and 66 observations are excluded because not all shares are sold till the end of the eight hours window.

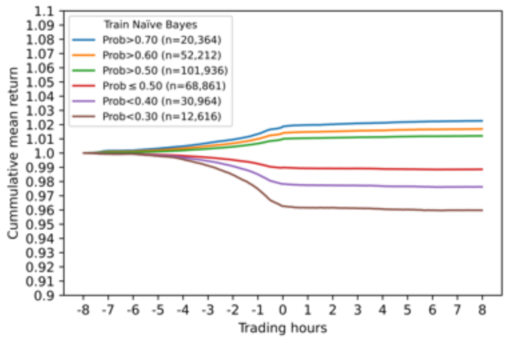
Model	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Accuracy positive	Accuracy negative	Accuracy all
SD classification	[na/na]	49.20	27.64	30.91	16′431	91′888	108′319	50.21%	50.80%	50.71%
		(0.25)	(−2.26)	(1.57)				(48.68%)	(47.18%)	(47.41%)
Naïve Bayes (classification)	[0.50/0.50]	96.25	49.93	76.61	113′807	83′763	197′570	50.31%	51.35%	50.75%
		(2.39)	(−2.74)	(4.73)				(47.40%)	(47.18%)	(47.32%)
	[0.60/0.40]	156.37	76.83	128.72	48′198	25′690	73′888	51.18%	50.33%	50.89%
		(3.38)	(−2.25)	(4.58)				(47.65%)	(47.84%)	(47.72%)
	[0.70/0.30]	150.66	55.22	114.36	17′851	10′957	28′808	51.41%	50.94%	51.24%
		(2.11)	(−1.23)	(2.55)				(47.40%)	(47.18%)	(47.32%)
CNN (classification)	[0.50/0.50]	102.09	56.92	82.82	113′303	84′267	197′570	50.32%	51.36%	50.76%
		(2.58)	(−3.01)	(5.11)				(47.67%)	(47.41%)	(47.57%)
	[0.55/0.45]	108.33	129.88	114.39	76′092	29′781	105′873	50.34%	51.02%	50.53%
		(2.45)	(−3.57)	(5.10)				(48.05%)	(47.41%)	(47.87%)
	[0.60/0.40]	188.76	192.43	190.08	33′843	19′056	52′899	51.16%	51.10%	51.14%
		(3.57)	(−4.07)	(5.21)				(47.67%)	(47.41%)	(47.57%)
RNN-LSTM (classification)	[0.50/0.50]	97.33	62.41	83.54	119′582	77′988	197′570	50.31%	51.59%	50.82%
		(2.45)	(−3.09)	(5.15)				(48.15%)	(47.81%)	(48.03%)
	[0.55/0.45]	100.40	149.40	112.87	80′767	27′556	108′323	50.40%	50.60%	50.45%
		(2.24)	(−3.84)	(5.05)				(47.88%)	(47.66%)	(47.83%)
	[0.60/0.40]	209.22	167.28	194.32	28′668	15′802	44′470	50.65%	50.52%	50.60%
		(3.70)	(−3.29)	(4.39)				(48.15%)	(47.81%)	(48.03%)
RNN-GRU (classification)	[0.50/0.50]	60.41	40.13	55.14	146′207	51′363	197′570	49.09%	50.31%	49.41%
		(1.13)	(−1.92)	(3.40)				(48.26%)	(47.66%)	(48.01%)
	[0.55/0.45]	177.87	104.93	148.76	38′915	25′852	64′767	50.89%	51.01%	50.93%
		(3.45)	(−2.83)	(4.23)				(47.98%)	(47.49%)	(47.79%)
	[0.60/0.40]	213.86	119.83	175.09	21′230	14′893	36′123	50.53%	50.80%	50.64%
		(3.31)	(−2.46)	(3.45)				(48.26%)	(47.66%)	(48.01%)
CNN (regression)	[0.0%/ 0.0%]	93.29	58.68	79.85	120′846	76′724	197′570	50.30%	51.69%	50.84%
		(2.29)	(−2.96)	(4.93)				(48.11%)	(47.26%)	(47.80%)
	[0.5%/-0.5%]	93.71	79.97	89.78	91′651	36′714	128′365	50.20%	51.00%	50.43%
		(2.06)	(−2.57)	(3.85)				(48.07%)	(47.51%)	(47.91%)
	[1.0%/-1.0%]	131.67	105.56	122.20	50′432	28′721	79′153	50.63%	51.19%	50.84%
		(2.56)	(−2.85)	(3.52)				(48.11%)	(47.26%)	(47.80%)
RNN-LSTM (regression)	[0.0%/ 0.0%]	100.73	66.56	87.15	119′086	78′484	197′570	50.32%	51.66%	50.85%
		(2.55)	(−3.27)	(5.38)				(48.11%)	(47.27%)	(47.85%)
	[0.5%/-0.5%]	129.76	110.23	123.98	56′830	23′882	80′712	50.70%	51.21%	50.85%
		(2.63)	(−2.68)	(3.53)				(47.94%)	(47.24%)	(47.73%)
	[1.0%/-1.0%]	127.11	154.84	135.86	40′979	18′885	59′864	50.57%	51.09%	50.74%
		(2.25)	(−3.17)	(3.15)				(48.11%)	(47.27%)	(47.85%)
RNN-GRU (regression)	[0.0%/ 0.0%]	94.46	54.86	78.50	117′934	79′636	197′570	50.23%	51.53%	50.76%
		(2.29)	(−2.92)	(4.84)				(48.22%)	(47.15%)	(47.93%)
	[0.5%/-0.5%]	136.32	125.52	133.55	58′866	20′318	79′184	50.61%	51.16%	50.75%
		(2.83)	(−2.76)	(3.72)				(47.98%)	(47.22%)	(47.78%)
	[1.0%/-1.0%]	120.38	129.58	122.88	43′553	16′224	59′777	50.46%	51.07%	50.62%
		(2.10)	(−2.58)	(2.81)				(48.22%)	(47.15%)	(47.93%)

Figure A.5: Cumulative Mean Returns on Positive/Negative News - Training Dataset Within Trading Hours

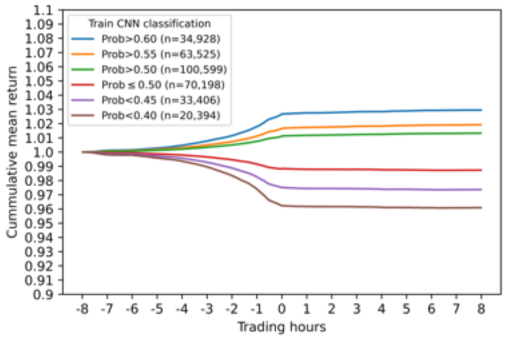
This figure shows the cumulative mean returns of positive and negative news according to the indicated model for the eight hours before and after news publication for the training dataset. The x-axis shows the trading time relative to news publication at zero. Markets start anticipating the news and then digest them quickly, and very soon, there is no trend in the stock price. Figures A.6 and A.7 in the appendix show the validation and test dataset results.



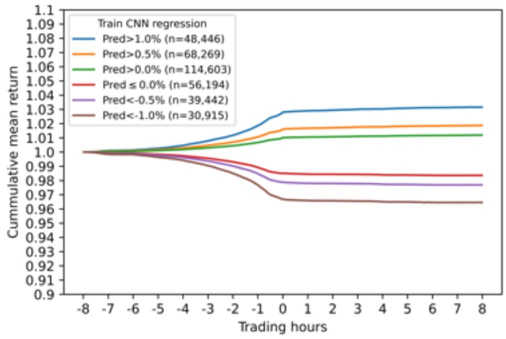
(a) Sentiment Dictionary Classification



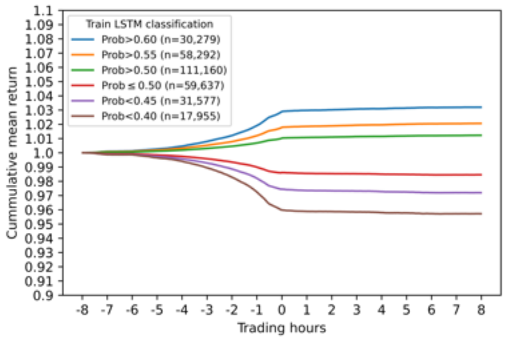
(b) Naïve Bayes Classification



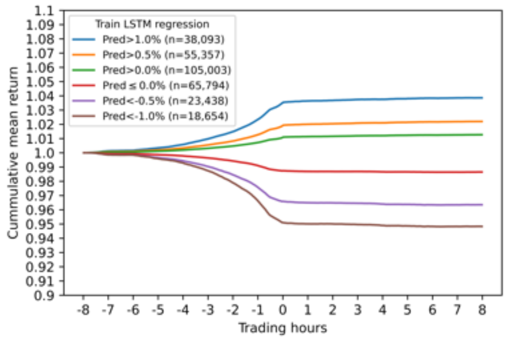
(c) CNN Classification



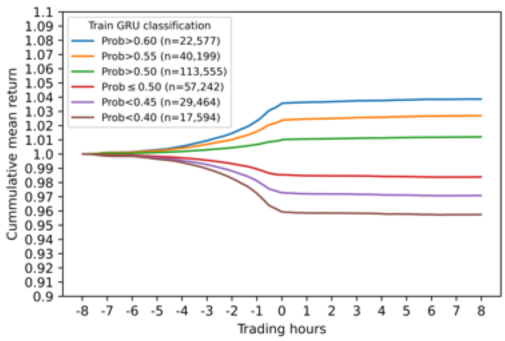
(d) CNN Regression



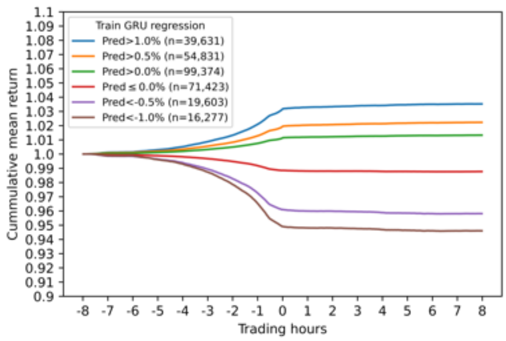
(e) RNN-LSTM Classification



(f) RNN-LSTM Regression



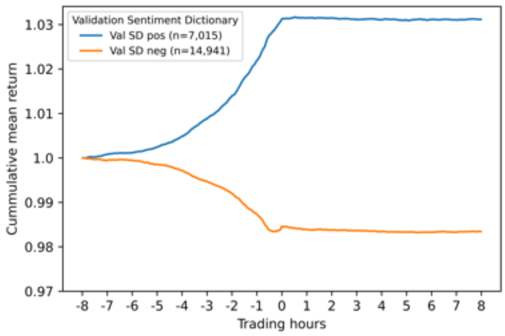
(g) RNN-GRU Classification



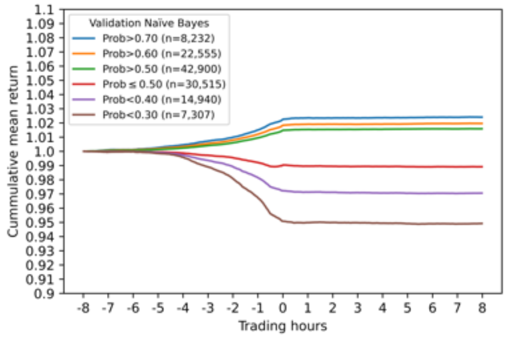
(h) RNN-GRU Regression

Figure A.6: Cumulative Mean Returns on Positive/Negative News - Validation Dataset Within Trading Hours

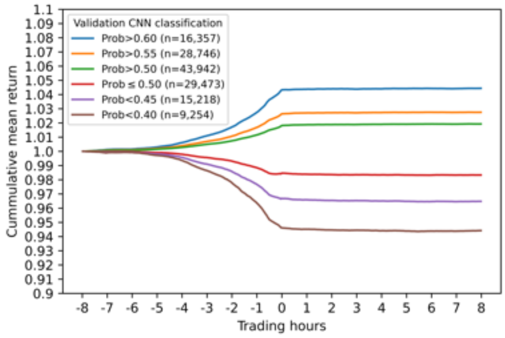
This figure shows the cumulative mean returns of positive and negative news according to the indicated model for the eight hours before and after news publication for the validation dataset. The x-axis shows the trading time relative to news publication at zero. Markets start anticipating the news and then digest them quickly, and very soon, there is no trend in the stock price. Figures A.5 and A.7 in the appendix show the training and test dataset results.



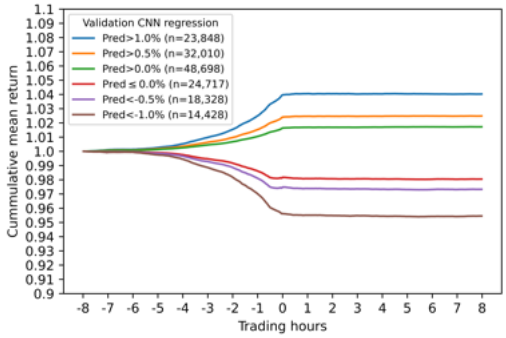
(a) Sentiment Dictionary Classification



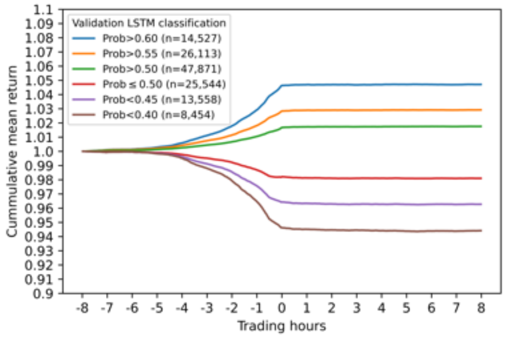
(b) Naïve Bayes Classification



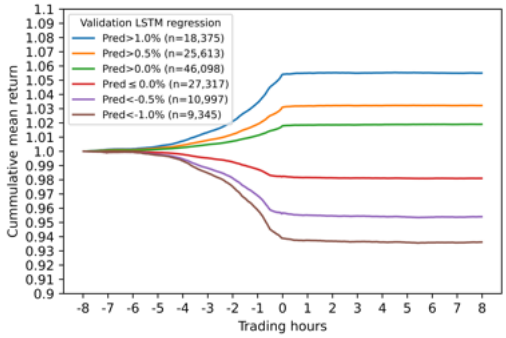
(c) CNN Classification



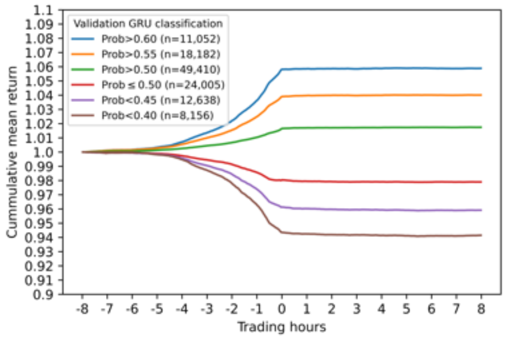
(d) CNN Regression



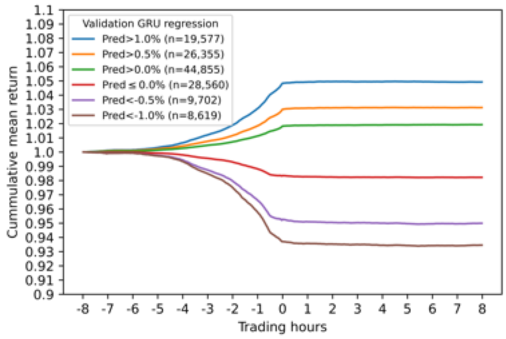
(e) RNN-LSTM Classification



(f) RNN-LSTM Regression



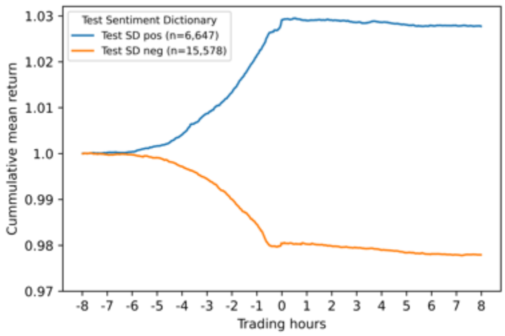
(g) RNN-GRU Rlassification



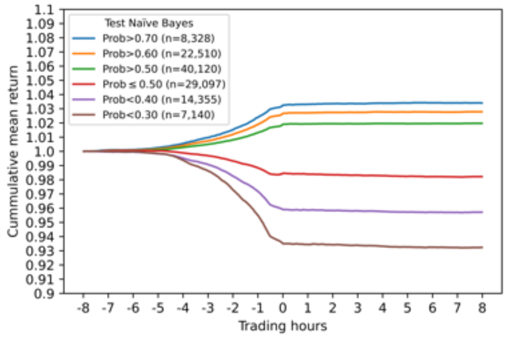
(h) RNN-GRU Regression

Figure A.7: Cumulative Mean Returns on Positive/Negative News - Test Dataset Within Trading Hours

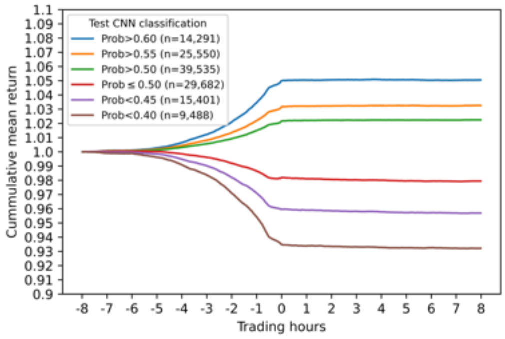
This figure shows the cumulative mean returns of positive and negative news according to the indicated model for the eight hours before and after news publication for the test dataset. The x-axis shows the trading time relative to news publication at zero. Markets start anticipating the news and then digest them quickly, and very soon, there is no trend in the stock price. Figures A.5 and A.6 in the appendix show the training and validation dataset results.



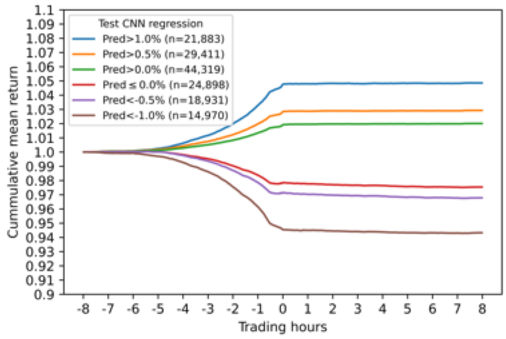
(a) Sentiment Dictionary Classification



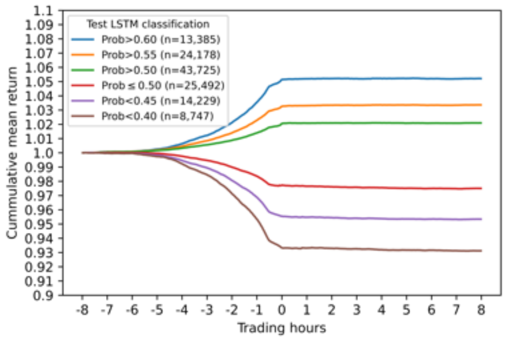
(b) Naïve Bayes Classification



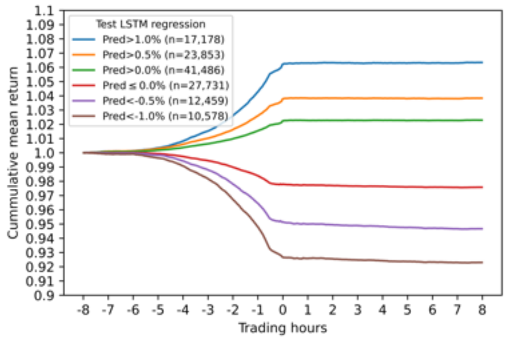
(c) CNN Classification



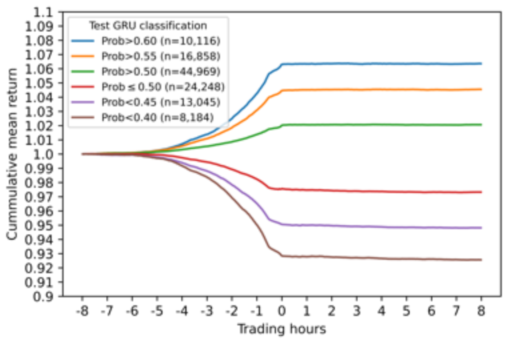
(d) CNN Regression



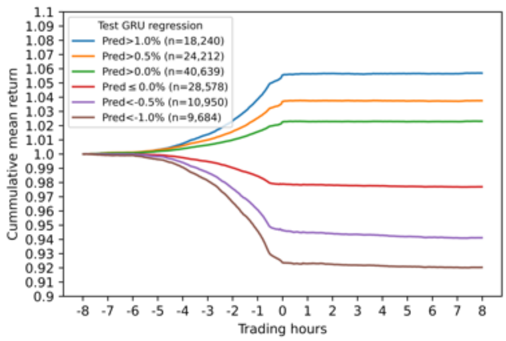
(e) RNN-LSTM Classification



(f) RNN-LSTM Regression



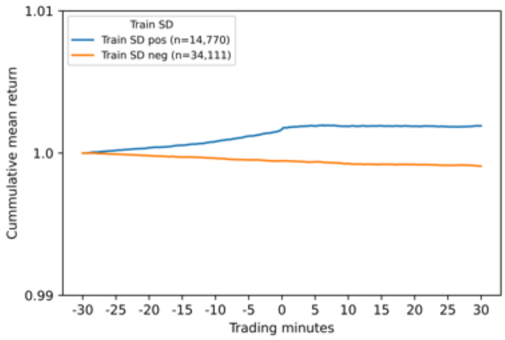
(g) RNN-GRU Classification



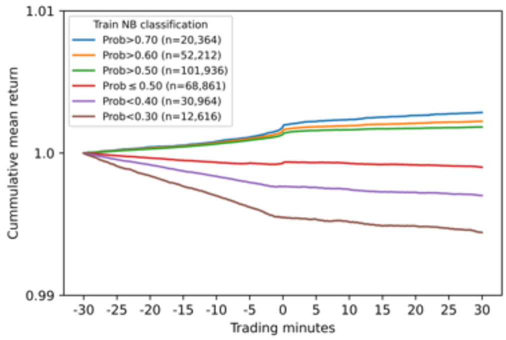
(h) RNN-GRU Regression

Figure A.8: Cumulative Mean Returns on Positive/Negative News - Training Dataset Within Trading Hours - 30 min.

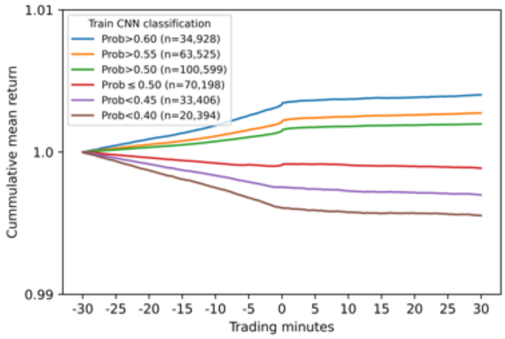
This figure shows the cumulative mean returns of positive and negative news according to the indicated model for the 30 minutes before and after news publication for the training dataset. The x-axis shows the trading time relative to news publication at zero. Markets start anticipating the news and then digest them quickly, and very soon, there is no trend in the stock price. Figures A.9 and A.10 in the appendix show the validation and test dataset results.



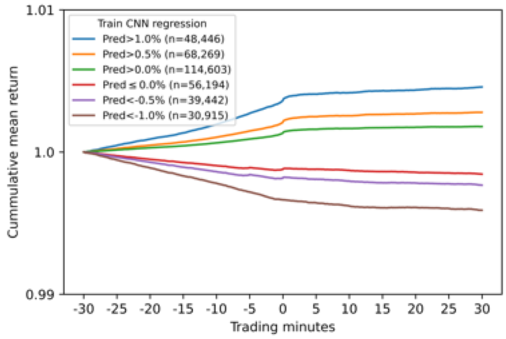
(a) Sentiment Dictionary Classification



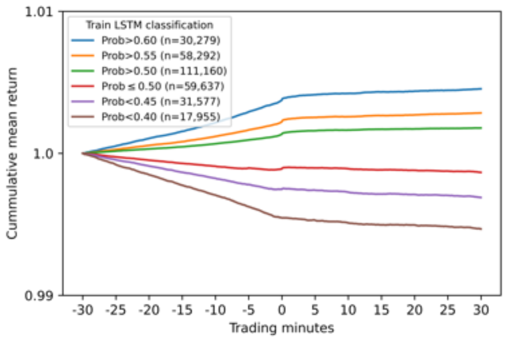
(b) Naïve Bayes Classification



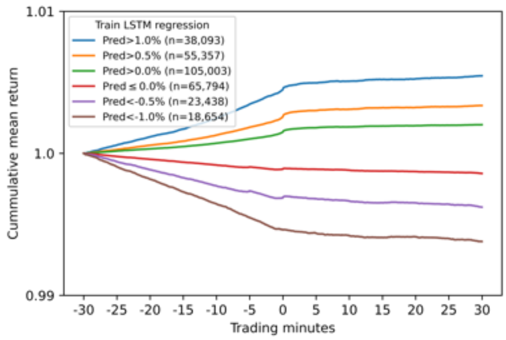
(c) CNN Classification



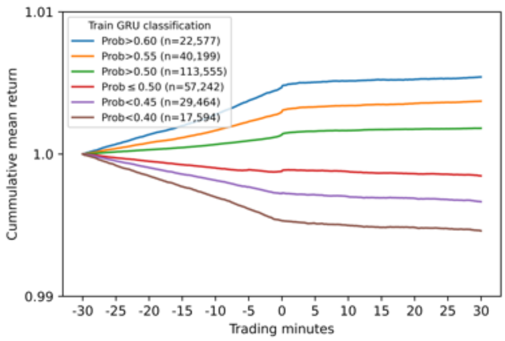
(d) CNN Regression



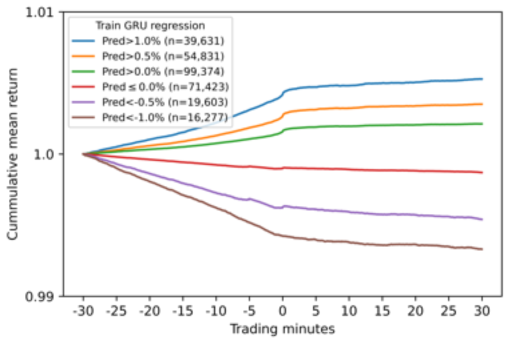
(e) RNN-LSTM Classification



(f) RNN-LSTM Regression



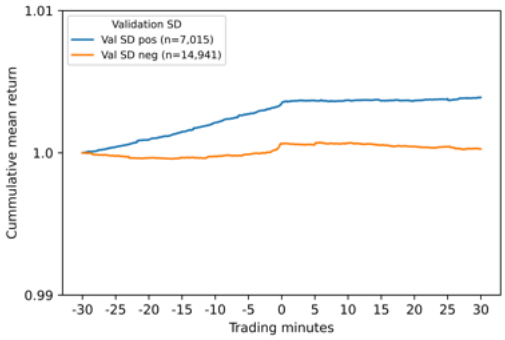
(g) RNN-GRU Classification



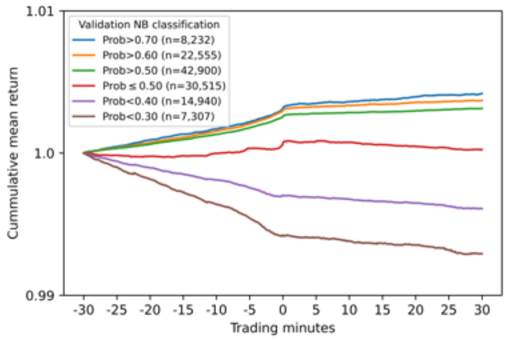
(h) RNN-GRU Regression

Figure A.9: Cumulative Mean Returns on Positive/Negative News - Validation Dataset Within Trading Hours - 30 min.

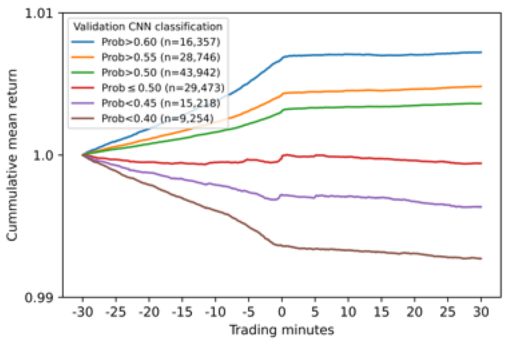
This figure shows the cumulative mean returns of positive and negative news according to the indicated model for the 30 minutes before and after news publication for the validation dataset. The x-axis shows the trading time relative to news publication at zero. Markets start anticipating the news and then digest them quickly, and very soon, there is no trend in the stock price. Figures A.8 and A.10 in the appendix show the training and test dataset results.



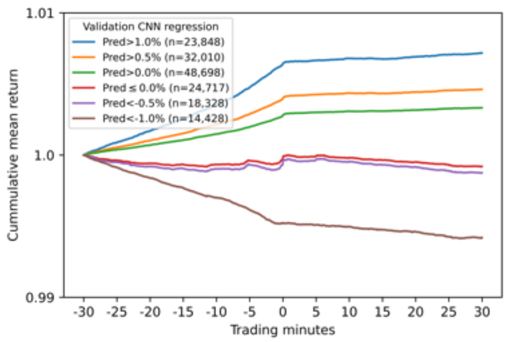
(a) Sentiment Dictionary Classification



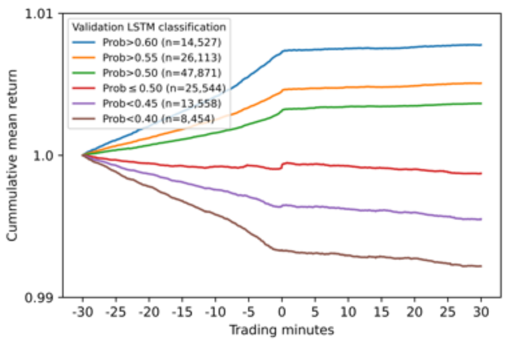
(b) Naïve Bayes Classification



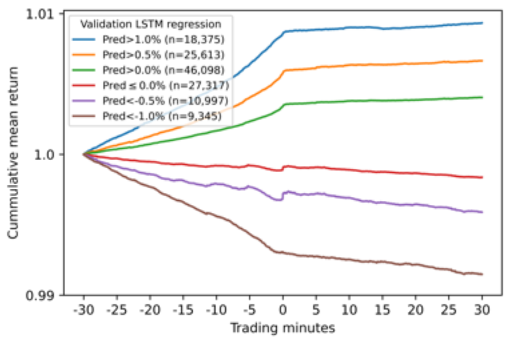
(c) CNN Classification



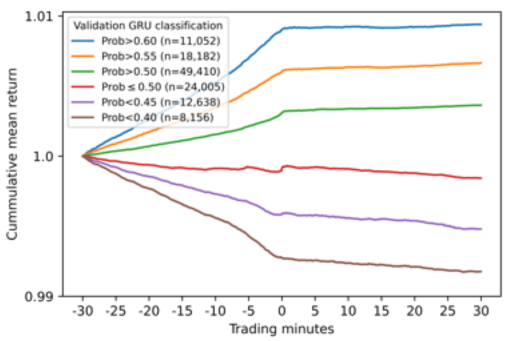
(d) CNN Regression



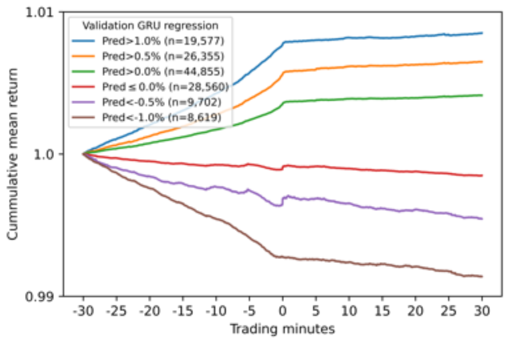
(e) RNN-LSTM Classification



(f) RNN-LSTM Regression



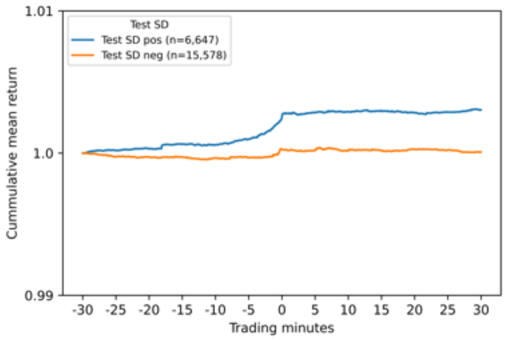
(g) RNN-GRU Classification



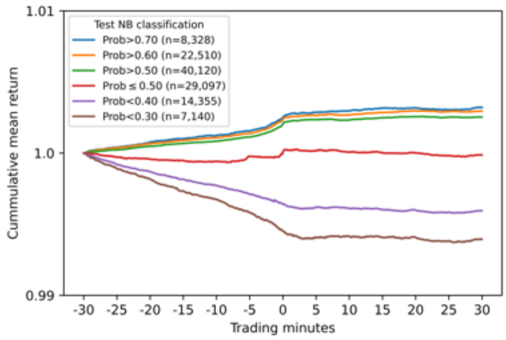
(h) RNN-GRU Regression

Figure A.10: Cumulative Mean Returns on Positive/Negative News - Test Dataset Within Trading Hours - 30 min.

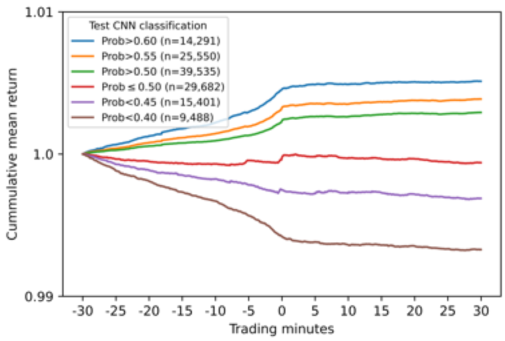
This figure shows the cumulative mean returns of positive and negative news according to the indicated model for the 30 minutes before and after news publication for the test dataset. The x-axis shows the trading time relative to news publication at zero. Markets start anticipating the news and then digest them quickly, and very soon, there is no trend in the stock price. Figures A.8 and A.9 in the appendix show the training and validation dataset results.



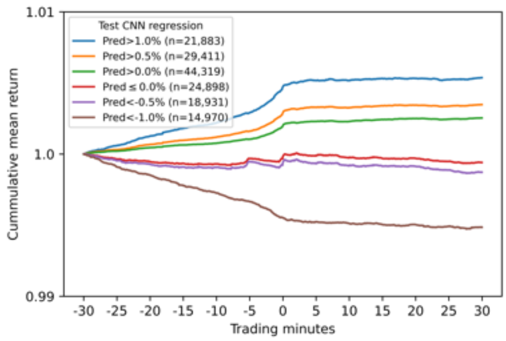
(a) Sentiment Dictionary Classification



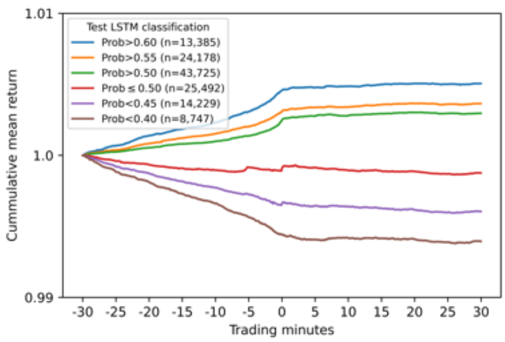
(b) Naïve Bayes Classification



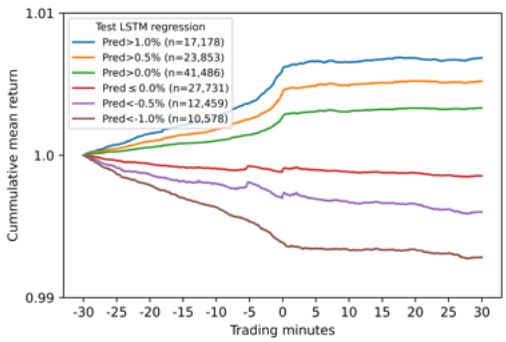
(c) CNN Classification



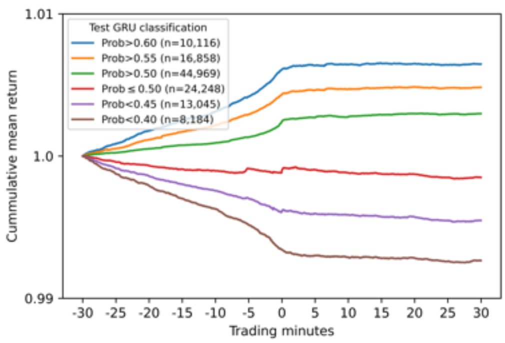
(d) CNN Regression



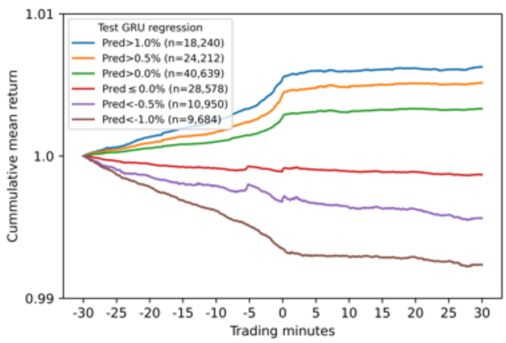
(e) RNN-LSTM Classification



(f) RNN-LSTM Regression



(g) RNN-GRU Classification



(h) RNN-GRU Regression

Table A.7: Trading Strategy of all Models [1 sec., 1 min., 1 min., 5m max.] - Whole Dataset

The strategies shown here start trading one second after the news is published for one minute, then hold the position for one minute before starting to fully liquidate the position with a maximum of USD five million per trade. Each model predicts a probability (classification) or a return (regression), and the thresholds are used to classify the headlines into positive ($\geq$ positive threshold) and negative ($\leq$ negative threshold). Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 209’667 observations, there is no trading in the buying/selling period, and 265 observations are excluded because not all shares are sold till the end of the eight hours window.

Model	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Accuracy positive	Accuracy negative	Accuracy all
SD classification	[na/na]	182.91	12.12	36.66	67′724	403′555	471′279	50.52%	51.24%	51.14%
		(3.91)	(−4.70)	(3.64)				(48.49%)	(47.39%)	(47.55%)
Naïve Bayes (classification)	[0.50/0.50]	99.70	4.18	58.70	479′897	360′914	840′811	49.83%	50.95%	50.31%
		(3.24)	(−3.55)	(6.73)				(47.21%)	(48.10%)	(47.53%)
	[0.60/0.40]	169.64	31.70	121.52	188′920	101′216	290′136	51.17%	50.57%	50.96%
		(5.74)	(−3.17)	(8.01)				(47.74%)	(47.86%)	(47.78%)
	[0.70/0.30]	204.27	17.48	138.29	70′870	38′711	109′581	51.75%	50.26%	51.23%
		(4.85)	(−1.73)	(6.19)				(47.21%)	(48.10%)	(47.53%)
CNN (classification)	[0.50/0.50]	104.00	8.11	62.18	474′153	366′658	840′811	49.87%	50.98%	50.35%
		(3.49)	(−3.85)	(7.13)				(47.62%)	(47.68%)	(47.64%)
	[0.55/0.45]	121.53	36.52	97.72	312′063	121′394	433′457	50.03%	50.88%	50.27%
		(4.08)	(−3.65)	(8.29)				(48.72%)	(47.73%)	(48.44%)
	[0.60/0.40]	198.03	44.74	139.43	122′922	76′081	199′003	51.28%	50.97%	51.16%
		(5.79)	(−3.23)	(6.82)				(47.62%)	(47.68%)	(47.64%)
RNN-LSTM (classification)	[0.50/0.50]	96.62	6.49	60.33	502′326	338′485	840′811	49.83%	51.10%	50.34%
		(3.03)	(−3.62)	(6.92)				(47.75%)	(47.78%)	(47.76%)
	[0.55/0.45]	115.47	67.46	103.07	331′399	115′487	446′886	49.75%	50.87%	50.04%
		(3.80)	(−4.72)	(8.69)				(49.00%)	(47.59%)	(48.63%)
	[0.60/0.40]	211.10	18.80	140.19	108′077	63′134	171′211	51.18%	50.76%	51.02%
		(5.88)	(−2.17)	(5.70)				(47.75%)	(47.78%)	(47.76%)
RNN-GRU (classification)	[0.50/0.50]	66.66	−22.45	43.37	621′069	219′742	840′811	49.11%	50.38%	49.44%
		(0.94)	(−1.54)	(4.97)				(47.66%)	(47.99%)	(47.81%)
	[0.55/0.45]	223.99	25.55	140.71	149′348	108′004	257′352	51.23%	50.89%	51.09%
		(7.31)	(−3.01)	(7.28)				(47.74%)	(47.76%)	(47.74%)
	[0.60/0.40]	212.15	11.99	125.59	78′993	60′192	139′185	51.27%	50.65%	51.00%
		(5.16)	(−1.94)	(4.55)				(47.66%)	(47.99%)	(47.81%)
CNN (regression)	[0.0%/ 0.0%]	91.85	1.35	56.17	509′339	331′472	840′811	49.76%	51.09%	50.29%
		(2.70)	(−3.30)	(6.44)				(47.97%)	(47.75%)	(47.88%)
	[0.5%/-0.5%]	113.89	−18.69	75.06	367′718	152′288	520′006	49.76%	50.65%	50.02%
		(3.77)	(−1.48)	(5.79)				(48.92%)	(48.03%)	(48.66%)
	[1.0%/-1.0%]	187.54	−0.01	115.12	185′723	116′820	302′543	50.87%	50.88%	50.87%
		(6.19)	(−2.00)	(5.63)				(47.97%)	(47.75%)	(47.88%)
RNN-LSTM (regression)	[0.0%/ 0.0%]	97.79	6.37	60.33	496′252	344′559	840′811	49.77%	51.02%	50.28%
		(3.09)	(−3.66)	(6.92)				(47.93%)	(47.55%)	(47.81%)
	[0.5%/-0.5%]	183.65	−14.07	122.59	218′187	97′489	315′676	50.88%	50.99%	50.91%
		(6.42)	(−1.36)	(6.03)				(47.92%)	(47.64%)	(47.84%)
	[1.0%/-1.0%]	210.78	−12.66	137.24	152′558	74′830	227′388	50.92%	51.02%	50.95%
		(6.68)	(−1.27)	(5.45)				(47.93%)	(47.55%)	(47.81%)
RNN-GRU (regression)	[0.0%/ 0.0%]	101.88	8.97	62.68	486′053	354′758	840′811	49.73%	50.91%	50.23%
		(3.34)	(−3.89)	(7.19)				(47.98%)	(47.47%)	(47.84%)
	[0.5%/-0.5%]	184.21	−16.40	130.48	223′989	81′944	305′933	50.86%	50.96%	50.88%
		(6.49)	(−1.20)	(6.43)				(47.92%)	(47.63%)	(47.84%)
	[1.0%/-1.0%]	204.95	14.17	151.39	162′886	63′573	226′459	50.87%	51.01%	50.91%
		(6.56)	(−1.94)	(6.04)				(47.98%)	(47.47%)	(47.84%)

Table A.8: Trading Strategy of all Models [1 sec., 1 min., 1 min., 10m max.] - Whole Dataset

The strategies shown here start trading one second after the news is published for one minute, then hold the position for one minute before starting to fully liquidate the position with a maximum of USD ten million per trade. Each model predicts a probability (classification) or a return (regression), and the thresholds are used to classify the headlines into positive (≥ positive threshold) and negative (≤ negative threshold). Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 209'667 observations, there is no trading in the buying/selling period, and 265 observations are excluded because not all shares are sold till the end of the eight hours window.

Model	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Accuracy positive	Accuracy negative	Accuracy all
SD classification	[na/na]	248.72	−7.50	29.32	67′724	403′555	471′279	50.54%	51.23%	51.13%
		(3.68)	(−3.68)	(2.06)				(48.46%)	(47.40%)	(47.56%)
Naïve Bayes (classification)	[0.50/0.50]	122.29	−26.34	58.49	479′897	360′914	840′811	49.84%	50.94%	50.31%
		(2.12)	(−2.37)	(4.80)				(47.24%)	(48.13%)	(47.55%)
	[0.60/0.40]	200.07	−25.95	121.22	188′920	101′216	290′136	51.18%	50.55%	50.96%
		(4.26)	(−1.45)	(5.77)				(47.73%)	(47.88%)	(47.78%)
	[0.70/0.30]	236.30	−84.82	122.86	70′870	38′711	109′581	51.73%	50.23%	51.20%
		(3.62)	(0.06)	(3.99)				(47.24%)	(48.13%)	(47.55%)
CNN (classification)	[0.50/0.50]	129.32	−18.76	64.75	474′153	366′658	840′811	49.88%	50.98%	50.36%
		(2.45)	(−2.74)	(5.32)				(47.64%)	(47.70%)	(47.67%)
	[0.55/0.45]	146.31	−8.67	102.90	312′063	121′394	433′457	50.04%	50.87%	50.27%
		(2.85)	(−2.08)	(6.19)				(48.71%)	(47.74%)	(48.44%)
	[0.60/0.40]	238.02	−9.42	143.42	122′922	76′081	199′003	51.25%	50.94%	51.13%
		(4.51)	(−1.67)	(4.90)				(47.64%)	(47.70%)	(47.67%)
RNN-LSTM (classification)	[0.50/0.50]	120.66	−22.41	63.06	502′326	338′485	840′811	49.84%	51.10%	50.35%
		(2.05)	(−2.50)	(5.18)				(47.78%)	(47.81%)	(47.79%)
	[0.55/0.45]	139.41	24.44	109.70	331′399	115′487	446′886	49.75%	50.86%	50.04%
		(2.60)	(−2.93)	(6.54)				(49.00%)	(47.61%)	(48.64%)
	[0.60/0.40]	241.96	−68.72	127.40	108′077	63′134	171′211	51.15%	50.73%	51.00%
		(4.30)	(−0.26)	(3.64)				(47.78%)	(47.81%)	(47.79%)
RNN-GRU (classification)	[0.50/0.50]	84.59	−71.25	43.87	621′069	219′742	840′811	49.11%	50.36%	49.44%
		(0.20)	(−0.34)	(3.60)				(47.71%)	(48.02%)	(47.84%)
	[0.55/0.45]	256.76	−38.44	132.87	149′348	108′004	257′352	51.20%	50.88%	51.07%
		(5.40)	(−1.14)	(4.87)				(47.77%)	(47.77%)	(47.77%)
	[0.60/0.40]	242.24	−72.35	106.19	78′993	60′192	139′185	51.23%	50.62%	50.97%
		(3.78)	(−0.18)	(2.76)				(47.71%)	(48.02%)	(47.84%)
CNN (regression)	[0.0%/ 0.0%]	112.98	−32.12	55.78	509′339	331′472	840′811	49.77%	51.09%	50.29%
		(1.66)	(−2.07)	(4.58)				(47.97%)	(47.76%)	(47.89%)
	[0.5%/-0.5%]	135.02	−75.47	73.38	367′718	152′288	520′006	49.76%	50.64%	50.02%
		(2.46)	(−0.17)	(4.06)				(48.92%)	(48.03%)	(48.66%)
	[1.0%/-1.0%]	214.20	−51.06	111.77	185′723	116′820	302′543	50.87%	50.86%	50.87%
		(4.41)	(−0.79)	(3.92)				(47.97%)	(47.76%)	(47.89%)
RNN-LSTM (regression)	[0.0%/ 0.0%]	121.15	−23.43	61.91	496′252	344′559	840′811	49.78%	51.01%	50.28%
		(2.06)	(−2.48)	(5.09)				(47.95%)	(47.57%)	(47.82%)
	[0.5%/-0.5%]	217.16	−91.72	121.77	218′187	97′489	315′676	50.88%	50.97%	50.91%
		(4.82)	(0.25)	(4.31)				(47.92%)	(47.66%)	(47.84%)
	[1.0%/-1.0%]	238.63	−106.25	125.14	152′558	74′830	227′388	50.90%	50.99%	50.93%
		(4.82)	(0.54)	(3.58)				(47.95%)	(47.57%)	(47.82%)
RNN-GRU (regression)	[0.0%/ 0.0%]	125.10	−20.83	63.53	486′053	354′758	840′811	49.74%	50.90%	50.23%
		(2.23)	(−2.64)	(5.22)				(47.99%)	(47.50%)	(47.86%)
	[0.5%/-0.5%]	219.49	−113.93	130.18	223′989	81′944	305′933	50.86%	50.94%	50.88%
		(4.94)	(0.74)	(4.58)				(47.92%)	(47.66%)	(47.85%)
	[1.0%/-1.0%]	235.02	−81.18	146.26	162′886	63′573	226′459	50.86%	50.98%	50.89%
		(4.77)	(0.00)	(4.15)				(47.99%)	(47.50%)	(47.86%)

Table A.9: Trading Strategy of all Models [1 sec., 1 min., 1 min., 1m max.] - Whole Dataset

The strategies shown here start trading one second after the news is published for one minute, then hold the position for one minute before starting to fully liquidate the position with a maximum of USD one million per trade. Each model predicts a probability (classification) or a return (regression), and the thresholds are used to classify the headlines into positive (≥ positive threshold) and negative (≤ negative threshold). Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 209’667 observations, there is no trading in the buying/selling period, and 265 observations are excluded because not all shares are sold till the end of the eight hours window.

Model	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Accuracy positive	Accuracy negative	Accuracy all
SD classification	[na/na]	75.57	16.92	25.34	67′724	403′555	471′279	50.52%	51.19%	51.10%
		(4.50)	(−6.02)	(6.30)				(48.48%)	(47.44%)	(47.59%)
Naïve Bayes (classification)	[0.50/0.50]	46.35	21.55	35.70	479′897	360′914	840′811	49.86%	50.92%	50.31%
		(5.30)	(−5.90)	(10.34)				(47.15%)	(47.88%)	(47.41%)
	[0.60/0.40]	87.18	38.01	70.03	188′920	101′216	290′136	51.21%	50.67%	51.02%
		(8.75)	(−5.09)	(11.19)				(47.70%)	(47.76%)	(47.72%)
	[0.70/0.30]	93.81	48.05	77.65	70′870	38′711	109′581	51.82%	50.48%	51.34%
		(6.27)	(−3.92)	(8.24)				(47.15%)	(47.88%)	(47.41%)
CNN (classification)	[0.50/0.50]	47.91	22.51	36.83	474′153	366′658	840′811	49.89%	50.95%	50.35%
		(5.51)	(−6.14)	(10.67)				(47.56%)	(47.55%)	(47.55%)
	[0.55/0.45]	56.09	44.06	52.72	312′063	121′394	433′457	50.06%	50.99%	50.32%
		(6.03)	(−6.16)	(11.23)				(48.69%)	(47.62%)	(48.39%)
	[0.60/0.40]	84.27	49.42	70.95	122′922	76′081	199′003	51.34%	51.09%	51.24%
		(6.91)	(−5.46)	(9.13)				(47.56%)	(47.55%)	(47.55%)
RNN-LSTM (classification)	[0.50/0.50]	43.85	22.34	35.19	502′326	338′485	840′811	49.85%	51.06%	50.34%
		(4.89)	(−5.91)	(10.19)				(47.74%)	(47.66%)	(47.71%)
	[0.55/0.45]	52.40	52.07	52.31	331′399	115′487	446′886	49.78%	50.98%	50.09%
		(5.57)	(−6.70)	(10.83)				(48.97%)	(47.49%)	(48.59%)
	[0.60/0.40]	88.52	42.66	71.61	108′077	63′134	171′211	51.18%	50.88%	51.07%
		(6.79)	(−4.45)	(7.36)				(47.74%)	(47.66%)	(47.71%)
RNN-GRU (classification)	[0.50/0.50]	28.51	14.75	24.91	621′069	219′742	840′811	49.16%	50.42%	49.49%
		(2.31)	(−3.84)	(7.21)				(47.64%)	(47.86%)	(47.73%)
	[0.55/0.45]	101.98	40.58	76.21	149′348	108′004	257′352	51.25%	50.99%	51.14%
		(9.25)	(−5.43)	(9.86)				(47.72%)	(47.65%)	(47.69%)
	[0.60/0.40]	86.15	35.30	64.16	78′993	60′192	139′185	51.30%	50.79%	51.08%
		(5.76)	(−3.84)	(6.11)				(47.64%)	(47.86%)	(47.73%)
CNN (regression)	[0.0%/ 0.0%]	42.49	21.65	34.27	509′339	331′472	840′811	49.79%	51.06%	50.29%
		(4.67)	(−5.75)	(9.93)				(47.93%)	(47.66%)	(47.82%)
	[0.5%/-0.5%]	52.62	23.65	44.13	367′718	152′288	520′006	49.81%	50.72%	50.08%
		(5.71)	(−4.19)	(8.56)				(48.87%)	(47.96%)	(48.60%)
	[1.0%/-1.0%]	89.20	32.83	67.43	185′723	116′820	302′543	50.91%	50.97%	50.93%
		(8.48)	(−4.62)	(8.40)				(47.93%)	(47.66%)	(47.82%)
RNN-LSTM (regression)	[0.0%/ 0.0%]	44.85	22.61	35.74	496′252	344′559	840′811	49.80%	50.99%	50.29%
		(5.02)	(−6.03)	(10.35)				(47.92%)	(47.45%)	(47.76%)
	[0.5%/-0.5%]	89.23	38.48	73.55	218′187	97′489	315′676	50.90%	51.08%	50.96%
		(9.01)	(−4.67)	(9.12)				(47.90%)	(47.54%)	(47.79%)
	[1.0%/-1.0%]	99.91	42.29	80.95	152′558	74′830	227′388	50.93%	51.12%	50.99%
		(8.91)	(−4.47)	(7.96)				(47.92%)	(47.45%)	(47.76%)
RNN-GRU (regression)	[0.0%/ 0.0%]	46.12	22.42	36.12	486′053	354′758	840′811	49.75%	50.87%	50.22%
		(5.17)	(−6.12)	(10.46)				(47.95%)	(47.37%)	(47.79%)
	[0.5%/-0.5%]	89.14	37.87	75.41	223′989	81′944	305′933	50.88%	51.08%	50.93%
		(9.06)	(−4.31)	(9.21)				(47.90%)	(47.52%)	(47.80%)
	[1.0%/-1.0%]	100.13	52.11	86.65	162′886	63′573	226′459	50.90%	51.11%	50.96%
		(9.11)	(−4.88)	(8.56)				(47.95%)	(47.37%)	(47.79%)

Table A.10: Trading Strategy of all Models [1 sec., 1 min., 2 min., 5m max.] - Whole Dataset

The strategies shown here start trading one second after the news is published for one minute, then hold the position for two minutes before starting to fully liquidate the position with a maximum of USD five million per trade. Each model predicts a probability (classification) or a return (regression), and the thresholds are used to classify the headlines into positive ($\geq$ positive threshold) and negative ($\leq$ negative threshold). Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 209'667 observations, there is no trading in the buying/selling period, and 279 observations are excluded because not all shares are sold till the end of the eight hours window.

Model	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Accuracy positive	Accuracy negative	Accuracy all
SD classification	[na/na]	185.94	23.72	47.03	67′721	403′552	471′273	50.63%	51.54%	51.41%
		(3.77)	(−4.43)	(4.08)				(48.60%)	(47.41%)	(47.58%)
Naïve Bayes (classification)	[0.50/0.50]	101.52	23.70	68.12	479′891	360′906	840′797	49.97%	51.23%	50.51%
		(3.45)	(−3.84)	(6.96)				(47.15%)	(48.37%)	(47.58%)
	[0.60/0.40]	175.12	16.96	119.94	188′915	101′216	290′131	51.48%	50.59%	51.17%
		(5.65)	(−2.10)	(6.79)				(47.65%)	(48.17%)	(47.83%)
	[0.70/0.30]	211.77	19.80	143.95	70′868	38′711	109′579	52.05%	50.36%	51.45%
		(4.74)	(−1.43)	(5.30)				(47.15%)	(48.37%)	(47.58%)
CNN (classification)	[0.50/0.50]	103.63	24.46	69.11	474′148	366′649	840′797	50.00%	51.24%	50.54%
		(3.53)	(−3.94)	(7.06)				(47.65%)	(47.85%)	(47.73%)
	[0.55/0.45]	126.52	19.15	96.45	312′059	121′393	433′452	50.24%	51.00%	50.45%
		(4.30)	(−2.36)	(7.10)				(48.79%)	(47.93%)	(48.55%)
	[0.60/0.40]	204.57	37.69	140.77	122′919	76′081	199′000	51.51%	51.13%	51.37%
		(5.64)	(−2.45)	(5.96)				(47.65%)	(47.85%)	(47.73%)
RNN-LSTM (classification)	[0.50/0.50]	95.82	23.53	66.72	502′322	338′475	840′797	49.98%	51.40%	50.55%
		(3.11)	(−3.76)	(6.82)				(47.90%)	(47.94%)	(47.92%)
	[0.55/0.45]	115.31	60.07	101.03	331′396	115′484	446′880	49.89%	50.98%	50.17%
		(3.77)	(−3.68)	(7.38)				(49.14%)	(47.81%)	(48.80%)
	[0.60/0.40]	205.68	12.04	134.28	108′075	63′134	171′209	51.29%	50.91%	51.15%
		(5.30)	(−1.56)	(4.79)				(47.90%)	(47.94%)	(47.92%)
RNN-GRU (classification)	[0.50/0.50]	62.35	−6.55	44.34	621′064	219′733	840′797	49.21%	50.57%	49.56%
		(1.05)	(−1.75)	(4.53)				(47.88%)	(48.20%)	(48.02%)
	[0.55/0.45]	225.85	10.19	135.34	149′345	108′001	257′346	51.42%	50.97%	51.23%
		(6.86)	(−1.91)	(6.10)				(47.79%)	(47.99%)	(47.87%)
	[0.60/0.40]	217.65	−3.88	121.84	78′992	60′192	139′184	51.31%	50.73%	51.06%
		(4.98)	(−1.13)	(3.92)				(47.88%)	(48.20%)	(48.02%)
CNN (regression)	[0.0%/ 0.0%]	89.96	17.05	61.22	509′332	331′465	840′797	49.87%	51.32%	50.44%
		(2.74)	(−3.40)	(6.26)				(48.02%)	(47.96%)	(48.00%)
	[0.5%/-0.5%]	109.60	−1.84	76.97	367′711	152′282	519′993	49.91%	50.80%	50.17%
		(3.52)	(−1.67)	(5.30)				(49.06%)	(48.16%)	(48.80%)
	[1.0%/-1.0%]	183.74	20.77	120.81	185′718	116′817	302′535	51.10%	50.95%	51.04%
		(5.63)	(−2.24)	(5.29)				(48.02%)	(47.96%)	(48.00%)
RNN-LSTM (regression)	[0.0%/ 0.0%]	96.96	23.08	66.69	496′245	344′552	840′797	49.89%	51.26%	50.45%
		(3.15)	(−3.79)	(6.82)				(47.98%)	(47.77%)	(47.91%)
	[0.5%/-0.5%]	184.77	17.33	133.06	218′182	97′484	315′666	51.10%	51.06%	51.09%
		(6.06)	(−1.93)	(5.87)				(47.98%)	(47.88%)	(47.95%)
	[1.0%/-1.0%]	201.47	16.33	140.54	152′553	74′828	227′381	51.14%	51.11%	51.13%
		(5.86)	(−1.71)	(4.99)				(47.98%)	(47.77%)	(47.91%)
RNN-GRU (regression)	[0.0%/ 0.0%]	100.71	24.76	68.67	486′046	354′751	840′797	49.84%	51.14%	50.39%
		(3.34)	(−3.95)	(7.02)				(48.05%)	(47.82%)	(47.99%)
	[0.5%/-0.5%]	178.51	15.87	134.95	223′983	81′939	305′922	51.02%	50.98%	51.01%
		(5.82)	(−1.77)	(5.90)				(48.04%)	(47.92%)	(48.01%)
	[1.0%/-1.0%]	192.86	20.36	144.44	162′881	63′570	226′451	51.07%	50.98%	51.05%
		(5.63)	(−1.71)	(5.10)				(48.05%)	(47.82%)	(47.99%)

Table A.11: Trading Strategy of all Models [1 sec., 1 min., 30 sec., 5m max.] - Whole Dataset

The strategies shown here start trading one second after the news is published for one minute, then hold the position for 30 seconds before starting to fully liquidate the position with a maximum of USD five million per trade. Each model predicts a probability (classification) or a return (regression), and the thresholds are used to classify the headlines into positive (≥ positive threshold) and negative (≤negative threshold). Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 209'667 observations, there is no trading in the buying/selling period, and 255 observations are excluded because not all shares are sold till the end of the eight hours window.

Model	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Accuracy positive	Accuracy negative	Accuracy all
SD classification	[na/na]	167.04	5.63	28.83	67′726	403′557	471′283	50.29%	50.90%	50.81%
		(3.81)	(−4.36)	(3.15)				(48.48%)	(47.52%)	(47.66%)
Naïve Bayes (classification)	[0.50/0.50]	94.84	5.85	56.64	479′902	360′919	840′821	49.76%	50.69%	50.16%
		(3.40)	(−3.71)	(7.02)				(47.28%)	(47.84%)	(47.48%)
	[0.60/0.40]	157.69	33.28	114.29	188′923	101′217	290′140	50.94%	50.45%	50.77%
		(5.77)	(−3.35)	(8.19)				(47.73%)	(47.78%)	(47.75%)
	[0.70/0.30]	182.85	38.71	131.93	70′871	38′711	109′582	51.49%	50.35%	51.09%
		(4.61)	(−2.33)	(6.41)				(47.28%)	(47.84%)	(47.48%)
CNN (classification)	[0.50/0.50]	97.78	8.07	58.66	474′159	366′662	840′821	49.79%	50.72%	50.20%
		(3.58)	(−3.91)	(7.27)				(47.63%)	(47.71%)	(47.66%)
	[0.55/0.45]	111.01	36.22	90.06	312′066	121′395	433′461	49.98%	50.66%	50.17%
		(3.96)	(−3.79)	(8.32)				(48.55%)	(47.73%)	(48.32%)
	[0.60/0.40]	177.58	47.29	127.77	122′923	76′081	199′004	51.04%	50.71%	50.91%
		(5.53)	(−3.47)	(6.85)				(47.63%)	(47.71%)	(47.66%)
RNN-LSTM (classification)	[0.50/0.50]	89.75	4.97	55.62	502′332	338′489	840′821	49.74%	50.79%	50.16%
		(3.02)	(−3.59)	(6.90)				(47.76%)	(47.67%)	(47.73%)
	[0.55/0.45]	106.86	64.30	95.86	331′402	115′490	446′892	49.72%	50.71%	49.98%
		(3.77)	(−4.83)	(8.76)				(48.83%)	(47.54%)	(48.50%)
	[0.60/0.40]	188.36	28.43	129.38	108′078	63′136	171′214	50.96%	50.66%	50.85%
		(5.58)	(−2.54)	(5.68)				(47.76%)	(47.67%)	(47.73%)
RNN-GRU (classification)	[0.50/0.50]	63.90	−16.92	42.77	621′075	219′746	840′821	49.08%	50.21%	49.38%
		(1.08)	(−1.77)	(5.30)				(47.69%)	(47.91%)	(47.78%)
	[0.55/0.45]	204.82	31.47	132.07	149′349	108′007	257′356	51.02%	50.65%	50.86%
		(7.19)	(−3.35)	(7.45)				(47.74%)	(47.78%)	(47.75%)
	[0.60/0.40]	181.68	23.78	113.39	78′993	60′193	139′186	51.05%	50.51%	50.82%
		(4.62)	(−2.35)	(4.42)				(47.69%)	(47.91%)	(47.78%)
CNN (regression)	[0.0%/ 0.0%]	85.59	0.58	52.07	509′346	331′475	840′821	49.69%	50.82%	50.14%
		(2.71)	(−3.28)	(6.46)				(47.98%)	(47.75%)	(47.89%)
	[0.5%/-0.5%]	107.83	−17.72	71.06	367′724	152′291	520′015	49.71%	50.42%	49.92%
		(3.91)	(−1.48)	(5.90)				(48.76%)	(48.03%)	(48.54%)
	[1.0%/-1.0%]	173.09	−4.36	104.58	185′727	116′821	302′548	50.65%	50.65%	50.65%
		(6.16)	(−1.85)	(5.49)				(47.98%)	(47.75%)	(47.89%)
RNN-LSTM (regression)	[0.0%/ 0.0%]	91.93	6.44	56.90	496′259	344′562	840′821	49.72%	50.76%	50.15%
		(3.17)	(−3.72)	(7.05)				(47.94%)	(47.53%)	(47.80%)
	[0.5%/-0.5%]	174.51	−11.17	117.17	218′193	97′492	315′685	50.67%	50.83%	50.72%
		(6.66)	(−1.43)	(6.19)				(47.93%)	(47.59%)	(47.82%)
	[1.0%/-1.0%]	198.22	−7.41	130.55	152′563	74′832	227′395	50.71%	50.83%	50.75%
		(6.82)	(−1.42)	(5.55)				(47.94%)	(47.53%)	(47.80%)
RNN-GRU (regression)	[0.0%/ 0.0%]	96.61	10.03	60.08	486′060	354′761	840′821	49.66%	50.64%	50.07%
		(3.49)	(−4.03)	(7.45)				(47.95%)	(47.37%)	(47.78%)
	[0.5%/-0.5%]	175.40	−8.66	126.10	223′994	81′947	305′941	50.64%	50.81%	50.69%
		(6.75)	(−1.43)	(6.70)				(47.95%)	(47.59%)	(47.85%)
	[1.0%/-1.0%]	196.17	18.80	146.38	162′891	63′575	226′466	50.71%	50.91%	50.77%
		(6.85)	(−2.12)	(6.25)				(47.95%)	(47.37%)	(47.78%)

Table A.12: Trading Strategy of all Models [1 sec., 2 min., 1 min., 5m max.] - Whole Dataset

The strategies shown here start trading one second after the news is published for two minutes, then hold the position for one minute before starting to fully liquidate the position with a maximum of USD five million per trade. Each model predicts a probability (classification) or a return (regression), and the thresholds are used to classify the headlines into positive ($\geq$ positive threshold) and negative ($\leq$ negative threshold). Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 147'785 observations, there is no trading in the buying/selling period, and 448 observations are excluded because not all shares are sold till the end of the eight hours window.

Model	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Accuracy positive	Accuracy negative	Accuracy all
SD classification	[na/na]	194.19	47.43	69.46	74′365	421′043	495′408	50.70%	51.58%	51.45%
		(3.45)	(−5.12)	(4.97)				(48.64%)	(47.74%)	(47.87%)
Naïve Bayes (classification)	[0.50/0.50]	128.79	56.32	97.96	518′525	383′985	902′510	50.24%	51.42%	50.74%
		(4.38)	(−4.96)	(8.69)				(47.14%)	(48.25%)	(47.52%)
	[0.60/0.40]	233.17	55.11	172.05	210′147	109′834	319′981	51.59%	50.83%	51.33%
		(7.14)	(−2.98)	(8.57)				(47.69%)	(48.08%)	(47.83%)
	[0.70/0.30]	271.97	103.91	214.10	79′454	41′728	121′182	52.18%	50.60%	51.64%
		(5.67)	(−2.83)	(6.89)				(47.14%)	(48.25%)	(47.52%)
CNN (classification)	[0.50/0.50]	131.40	56.22	98.80	511′102	391′408	902′510	50.27%	51.42%	50.77%
		(4.45)	(−5.05)	(8.77)				(47.66%)	(47.73%)	(47.68%)
	[0.55/0.45]	164.88	71.69	138.45	335′775	132′887	468′662	50.57%	51.32%	50.78%
		(5.43)	(−3.77)	(8.85)				(48.72%)	(47.73%)	(48.44%)
	[0.60/0.40]	259.95	111.28	203.56	136′072	83′166	219′238	51.63%	51.39%	51.54%
		(6.65)	(−4.07)	(7.78)				(47.66%)	(47.73%)	(47.68%)
RNN-LSTM (classification)	[0.50/0.50]	120.34	55.48	94.40	541′581	360′929	902′510	50.22%	51.54%	50.75%
		(3.94)	(−4.84)	(8.38)				(47.85%)	(47.63%)	(47.77%)
	[0.55/0.45]	149.02	99.98	136.12	354′611	126′545	481′156	50.31%	51.16%	50.53%
		(4.77)	(−4.50)	(8.62)				(49.01%)	(47.76%)	(48.68%)
	[0.60/0.40]	255.07	74.85	189.22	119′135	68′596	187′731	51.47%	51.33%	51.42%
		(6.05)	(−2.86)	(6.08)				(47.85%)	(47.63%)	(47.77%)
RNN-GRU (classification)	[0.50/0.50]	78.32	27.65	64.79	661′564	240′946	902′510	49.51%	50.83%	49.86%
		(1.75)	(−2.91)	(5.75)				(47.83%)	(47.91%)	(47.86%)
	[0.55/0.45]	278.42	65.79	189.76	164′752	117′835	282′587	51.61%	51.25%	51.46%
		(7.72)	(−3.36)	(7.60)				(47.74%)	(47.82%)	(47.77%)
	[0.60/0.40]	275.29	71.62	188.07	86′676	64′919	151′595	51.47%	51.18%	51.35%
		(5.79)	(−2.73)	(5.39)				(47.83%)	(47.91%)	(47.86%)
CNN (regression)	[0.0%/ 0.0%]	116.79	54.01	92.25	549′751	352′759	902′510	50.08%	51.45%	50.62%
		(3.75)	(−4.75)	(8.18)				(48.00%)	(47.87%)	(47.95%)
	[0.5%/-0.5%]	135.05	49.75	109.81	396′887	166′745	563′632	50.25%	51.03%	50.48%
		(4.20)	(−3.21)	(6.70)				(48.97%)	(48.05%)	(48.70%)
	[1.0%/-1.0%]	225.64	91.59	174.41	206′757	127′921	334′678	51.20%	51.19%	51.20%
		(6.42)	(−4.11)	(6.99)				(48.00%)	(47.87%)	(47.95%)
RNN-LSTM (regression)	[0.0%/ 0.0%]	123.44	57.92	96.92	537′187	365′323	902′510	50.14%	51.45%	50.67%
		(4.06)	(−5.03)	(8.60)				(47.91%)	(47.52%)	(47.78%)
	[0.5%/-0.5%]	218.59	94.93	180.87	243′876	107′025	350′901	51.19%	51.37%	51.25%
		(6.58)	(−3.83)	(7.27)				(47.99%)	(47.66%)	(47.89%)
	[1.0%/-1.0%]	236.62	110.52	195.71	170′715	81′974	252′689	51.30%	51.46%	51.35%
		(6.28)	(−3.80)	(6.32)				(47.91%)	(47.52%)	(47.78%)
RNN-GRU (regression)	[0.0%/ 0.0%]	124.47	54.28	95.24	526′654	375′856	902′510	50.10%	51.35%	50.62%
		(4.07)	(−4.93)	(8.45)				(48.06%)	(47.63%)	(47.94%)
	[0.5%/-0.5%]	212.23	105.32	184.03	250′719	89′839	340′558	51.07%	51.33%	51.14%
		(6.37)	(−3.82)	(7.28)				(48.09%)	(47.67%)	(47.98%)
	[1.0%/-1.0%]	229.95	114.63	198.14	182′736	69′610	252′346	51.14%	51.28%	51.17%
		(6.17)	(−3.65)	(6.39)				(48.06%)	(47.63%)	(47.94%)

Table A.13: Trading Strategy of all Models [1 sec., 30 sec., 1 min., 5m max.] - Whole Dataset

The strategies shown here start trading one second after the news is published for 30 seconds, then hold the position for one minute before starting to fully liquidate the position with a maximum of USD five million per trade. Each model predicts a probability (classification) or a return (regression), and the thresholds are used to classify the headlines into positive (≥ positive threshold) and negative (≤ negative threshold). Profits are per sample in USD, i.e., the total profit could be computed by multiplying <Profit all> with <n all>. <Profit positive> (<Profit negative>) is the profit from buying (selling) after positive (negative) news, and <Profit all> from buying after positive and shorting after negative news. The values in brackets below the profits are the t-tests against the mean profit of buying, holding, and selling after all news for the positive and negative profits and against zero for buying the positive and shorting the negative news (<Profit all>). The last three columns show the fractions of profitable trades and, in brackets below, the number of unprofitable trades (they do not sum to one due to some trades yielding a zero return). Because observations with identical CUSIP and next trading second are mean aggregated, n differs from the previous analysis. For 284'467 observations, there is no trading in the buying/selling period, and 161 observations are excluded because not all shares are sold till the end of the eight hours window.

Model	Threshold [pos/neg]	Profit positive	Profit negative	Profit all	n positive	n negative	n all	Accuracy positive	Accuracy negative	Accuracy all
SD classification	[na/na]	154.67	−6.42	15.65	60′227	379′449	439′676	50.09%	50.11%	50.11%
		(3.84)	(−3.59)	(1.98)				(48.48%)	(47.62%)	(47.74%)
Naïve Bayes (classification)	[0.50/0.50]	72.76	−15.73	34.39	433′934	332′181	766′115	49.63%	50.11%	49.84%
		(2.16)	(−2.30)	(4.71)				(47.33%)	(47.88%)	(47.53%)
	[0.60/0.40]	122.16	3.61	80.12	165′426	90′886	256′312	50.80%	50.18%	50.58%
		(4.39)	(−2.24)	(6.31)				(47.64%)	(47.80%)	(47.70%)
	[0.70/0.30]	138.97	−22.04	80.78	61′628	34′880	96′508	51.27%	50.07%	50.84%
		(3.46)	(−0.74)	(4.35)				(47.33%)	(47.88%)	(47.53%)
CNN (classification)	[0.50/0.50]	73.62	−15.33	34.57	429′789	336′326	766′115	49.65%	50.12%	49.85%
		(2.20)	(−2.36)	(4.74)				(47.61%)	(47.84%)	(47.70%)
	[0.55/0.45]	83.83	1.79	61.17	283′414	108′145	391′559	49.80%	50.38%	49.96%
		(2.64)	(−2.36)	(6.33)				(48.28%)	(47.80%)	(48.15%)
	[0.60/0.40]	140.23	13.66	91.48	108′082	67′708	175′790	50.90%	50.42%	50.71%
		(4.40)	(−2.37)	(5.36)				(47.61%)	(47.84%)	(47.70%)
RNN-LSTM (classification)	[0.50/0.50]	67.33	−19.82	31.93	454′912	311′203	766′115	49.59%	50.17%	49.83%
		(1.69)	(−1.98)	(4.38)				(47.77%)	(47.76%)	(47.76%)
	[0.55/0.45]	79.95	29.72	67.21	302′829	102′814	405′643	49.53%	50.48%	49.77%
		(2.42)	(−3.56)	(6.91)				(48.49%)	(47.58%)	(48.26%)
	[0.60/0.40]	145.85	−14.08	86.33	95′261	56′470	151′731	50.80%	50.39%	50.65%
		(4.34)	(−1.19)	(4.15)				(47.77%)	(47.76%)	(47.76%)
RNN-GRU (classification)	[0.50/0.50]	50.48	−40.90	27.12	570′251	195′864	766′115	49.00%	49.95%	49.25%
		(0.24)	(−0.40)	(3.72)				(47.71%)	(47.98%)	(47.83%)
	[0.55/0.45]	159.98	−0.84	91.91	131′628	96′598	228′226	50.95%	50.34%	50.69%
		(5.73)	(−2.10)	(5.73)				(47.65%)	(47.85%)	(47.73%)
	[0.60/0.40]	139.25	−15.14	71.83	69′768	54′087	123′855	50.91%	50.23%	50.61%
		(3.53)	(−1.13)	(3.09)				(47.71%)	(47.98%)	(47.83%)
CNN (regression)	[0.0%/ 0.0%]	65.68	−21.38	31.00	460′925	305′190	766′115	49.54%	50.17%	49.79%
		(1.56)	(−1.85)	(4.25)				(47.96%)	(47.90%)	(47.94%)
	[0.5%/-0.5%]	83.09	−50.08	44.52	333′456	135′979	469′435	49.62%	50.06%	49.75%
		(2.71)	(0.10)	(4.03)				(48.40%)	(48.17%)	(48.34%)
	[1.0%/-1.0%]	135.08	−43.31	65.45	162′779	104′220	266′999	50.52%	50.28%	50.43%
		(4.80)	(−0.20)	(3.65)				(47.96%)	(47.90%)	(47.94%)
RNN-LSTM (regression)	[0.0%/ 0.0%]	71.00	−15.64	35.05	448′268	317′847	766′115	49.57%	50.11%	49.80%
		(2.00)	(−2.29)	(4.80)				(47.95%)	(47.77%)	(47.89%)
	[0.5%/-0.5%]	139.79	−54.02	79.09	190′491	86′855	277′346	50.55%	50.48%	50.53%
		(5.41)	(0.23)	(4.43)				(47.91%)	(47.75%)	(47.86%)
	[1.0%/-1.0%]	162.92	−59.91	88.50	133′080	66′726	199′806	50.56%	50.43%	50.52%
		(5.80)	(0.42)	(4.05)				(47.95%)	(47.77%)	(47.89%)
RNN-GRU (regression)	[0.0%/ 0.0%]	74.45	−12.59	37.28	438′924	327′191	766′115	49.55%	50.01%	49.74%
		(2.27)	(−2.56)	(5.11)				(47.96%)	(47.64%)	(47.87%)
	[0.5%/-0.5%]	140.56	−60.32	85.90	195′474	73′067	268′541	50.54%	50.45%	50.52%
		(5.49)	(0.45)	(4.93)				(47.89%)	(47.82%)	(47.87%)
	[1.0%/-1.0%]	159.57	−31.48	105.06	141′910	56′650	198′560	50.58%	50.51%	50.56%
		(5.74)	(−0.54)	(4.82)				(47.96%)	(47.64%)	(47.87%)

Table A.14: Annualizing Examples for Different 5 or 10 min. Returns

This table shows the annualized returns for one to six basis points per five or ten minutes. The two columns on the left assume a trading hour year (6.5 hours and 252 days), and the two on the right side a financial year (24 hours and 360 days).

Holding time / Basis points	Assuming 6.5h and 252 days		Assuming 24h and 360 days	
	5 Min.	10 Min.	5 Min.	10 Min.
1	6.14e2%	1.67e2%	1.42e5%	3.67e3%
2	4.99e3%	6.14e2%	2.01e8%	1.42e5%
3	3.63e4%	1.81e3%	2.85e11%	5.34e6%
4	2.59e5%	4.99e3%	4.03e14%	2.01e8%
5	1.85e6%	1.35e4%	5.70e17%	7.55e9%
6	1.32e7%	3.62e4%	8.05e20%	2.84e11%

References

Aït-Sahalia, Y., Mykland, P. A., Zhang, L., 2011. Ultra high frequency volatility estimation with dependent microstructure noise. Journal of Econometrics 160, 160–175.

Andersen, T. G., Bollerslev, T., Das, A., 2001. Variance-ratio statistics and high-frequency data: Testing for changes in intraday volatility patterns. The Journal of Finance 56, 305–327.

Ang, A., Hodrick, R. J., Xing, Y., Zhang, X., 2006. The cross-section of volatility and expected returns. The Journal of Finance 61, 259–299.

Ang, A., Hodrick, R. J., Xing, Y., Zhang, X., 2009. High idiosyncratic volatility and low returns: International and further us evidence. Journal of Financial Economics 91, 1–23.

Aumann, R. J., 1976. Agreeing to Disagree. The Annals of Statistics 4, 1236–1239.

Bachelier, L., 1900. Théorie de la spéculation. In: *Annales scientifiques de l'École normale supérieure*, vol. 17, pp. 21–86.

Bagehot, W., 1971. The only game in town. Financial Analysts Journal 27, 12–14.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 .

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. Journal of Finance 61, 1645–1680.

Ball, R., Brown, P., 1968. An empirical evaluation of accounting income numbers. Journal of Accounting Research pp. 159–178.

Bartov, E., Faurel, L., Mohanram, P. S., 2018. Can twitter help predict firm-level earnings and stock returns? The Accounting Review 93, 25–57.

Bernard, V. L., Thomas, J. K., 1989. Post-earnings-announcement drift: delayed price response or risk premium? Journal of Accounting Research 27, 1–36.

Bernard, V. L., Thomas, J. K., 1990. Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. Journal of Accounting and Economics 13, 305–340.

Busse, J. A., Green, T. C., 2002. Market efficiency in real time. Journal of Financial Economics 65, 415–437.

Chan, K., Chan, K. C., Karolyi, G. A., 1991. Intraday volatility in the stock index and stock index futures markets. The Review of Financial Studies 4, 657–684.

Chan, L. K., Jegadeesh, N., Lakonishok, J., 1996. Momentum strategies. The Journal of Finance 51, 1681–1713.

Cochrane, J. H., 2011. Presidential address: Discount rates. The Journal of finance 66, 1047–1108.

Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. The Journal of Finance 66, 1461–1499.

Da, Z., Engelberg, J., Gao, P., 2015. The sum of all fears: investor sentiment and asset prices. The Review of Financial Studies 28, 1–32.

Daniel, K., Titman, S., 2006. Market reactions to tangible and intangible information. Journal of Finance 61, 1605–1643.

Ederington, L. H., Lee, J. H., 1993. How markets process information: News releases and volatility. The Journal of Finance 48, 1161–1191.

Engelberg, J. E., Parsons, C. A., 2011. The causal impact of media in financial markets. the Journal of Finance 66, 67–97.

Engle, R. F., Sokalska, M. E., 2012. Forecasting intraday volatility in the us equity market. multiplicative component garch. Journal of Financial Econometrics 10, 54–83.

Fama, E. F., 1970. Efficient capital markets: A review of theory and empirical work. The Journal of Finance 25, 383–417.

Fama, E. F., 1998. Market efficiency, long-term returns, and behavioral finance. Journal of Financial Economics 49, 283–306.

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

Feng, G., Giglio, S., Xiu, D., 2020. Taming the factor zoo: A test of new factors. The Journal of Finance 75, 1327–1370.

Fink, J., 2021. A review of the post-earnings-announcement drift. Journal of Behavioral and Experimental Finance 29, 100446.

Foucault, T., Pagano, M., Roell, A., Röell, A., 2013. Market liquidity: theory, evidence, and policy. Oxford University Press.

Frazzini, A., 2006. The disposition effect and underreaction to news. The Journal of Finance 61, 2017–2046.

Frazzini, A., Pedersen, L. H., 2014. Betting against beta. Journal of Financial Economics 111, 1–25.

Gao, L., Han, Y., Li, S. Z., Zhou, G., 2018. Market intraday momentum. Journal of Financial Economics 129, 394–414.

Garcia, D., 2013. Sentiment during recessions. The Journal of Finance 68, 1267–1300.

Glosten, L. R., Milgrom, P. R., 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. Journal of Financial Economics 14, 71–100.

Grossman, S. J., Stiglitz, J. E., 1980. On the impossibility of informationally efficient markets. The American Economic Review 70, 393–408.

Harvey, C. R., Liu, Y., Zhu, H., 2015. ...and the cross-section of expected returns. The Review of Financial Studies 29, 5–68.

Hayek, F., 1945. American economic association. The American Economic Review 35, 519–530.

He, H., Wang, J., 1995. Differential information and dynamic behavior of stock trading volume. The Review of Financial Studies 8, 919–972.

Hendershott, T., Jones, C. M., Menkveld, A. J., 2011. Does algorithmic trading improve liquidity? The Journal of Finance 66, 1–33.

Heston, S. L., Sinha, N. R., 2017. News vs. sentiment: Predicting stock returns from news stories. Financial Analysts Journal 73, 67–83.

Hong, H., Lim, T., Stein, J. C., 2000. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. The Journal of Finance 55, 265–295.

Hong, H., Stein, J. C., 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. Journal of Finance 54, 2143–2184.

Hou, K., Xue, C., Zhang, L., 2020. Replicating anomalies. The Review of Financial Studies 33, 2019–2133.

Hu, M., Liu, B., 2004. Mining opinion features in customer reviews. In: *AAAI*, vol. 4, pp. 755–760.

Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: implications for stock market efficiency. The Journal of Finance 48, 65–91.

Kahneman, D., Slovic, S. P., Slovic, P., Tversky, A., 1982. Judgment under uncertainty: Heuristics and biases. Cambridge university press.

Klibanoff, P., Lamont, O., Wizman, T. A., 1998. Investor reaction to salient news in closed-end country funds. The Journal of Finance 53, 673–699.

Krishnan, M., 1992. An equivalence between the kyle (1985) and the glosten—milgrom (1985) models. Economics Letters 40, 333–338.

Kyle, A. S., 1985. Continuous auctions and insider trading. Econometrica: Journal of the Econometric Society pp. 1315–1335.

Lee, C. M., Shleifer, A., Thaler, R. H., 1991. Investor sentiment and the closed-end fund puzzle. The Journal of Finance 46, 75–109.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of Finance 66, 35–65.

Loughran, T., McDonald, B., 2016. Textual analysis in accounting and finance: A survey. Journal of Accounting Research 54, 1187–1230.

Martineau, C., 2021. Rest in peace post-earnings announcement drift. Critical Finance Review, Forthcoming .

Milgrom, P., Stokey, N., 1982. Information, trade and common knowledge. Journal of Economic Theory 26, 17–27.

Miller, E. M., 1977. Risk, uncertainty, and divergence of opinion. The Journal of Finance 32, 1151–1168.

Odean, T., 1998. Volume, volatility, price, and profit when all traders are above average. The journal of finance 53, 1887–1934.

O'hara, M., 2015. High frequency market microstructure. Journal of Financial Economics 116, 257–270.

Patell, J. M., Wolfson, M. A., 1984. The intraday speed of adjustment of stock prices to earnings and dividend announcements. Journal of Financial Economics 13, 223–252.

Peress, J., 2014. The media and the diffusion of information in financial markets: Evidence from newspaper strikes. The Journal of Finance 69, 2007–2043.

Pichler, J., 2023. Is the stock market's reaction to news predictable? Doctoral Thesis .

Shefrin, H., Statman, M., 1985. The disposition to sell winners too early and ride losers too long: Theory and evidence. The Journal of Finance 40, 777–790.

Shiller, R. J., 2003. From efficient markets theory to behavioral finance. Journal of Economic Perspectives 17, 83–104.

Shiller, R. J., 2015. Irrational exuberance. In: *Irrational exuberance*, Princeton university press.

Solomon, D. H., 2012. Selective publicity and stock prices. The Journal of Finance 67, 599–638.

Tetlock, P. C., 2007. Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance 62, 1139–1168.

Tetlock, P. C., 2010. Does public financial news resolve asymmetric information? The Review of Financial Studies 23, 3520–3557.

Tetlock, P. C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. The Journal of Finance 63, 1437–1467.

Thaler, R., 1985. Mental accounting and consumer choice. Marketing science 4, 199–214.

Tirole, J., 1982. On the possibility of speculation under rational expectations. Econometrica: Journal of the Econometric Society pp. 1163–1181.

Uhl, M. W., 2014. Reuters sentiment and stock returns. Journal of Behavioral Finance 15, 287–298.

Working, H., 1958. A theory of anticipatory prices. The American Economic Review 48, 188–199.

Zhou, B., 1996. High-frequency data and volatility in foreign-exchange rates. Journal of Business & Economic Statistics 14, 45–52.

Essay III:

A Reevaluation of Profitability and its Trend

Jan Pichler*

ABSTRACT

This paper shows that the level of profitability has been a consistently strong predictor for future returns over the past 40 years. However, the trend in profitability described by Akbas et al. (2017) is not, and its positive predictive power primarily stems from the period between 2000 and 2006 and has been reversing since then. The strongest predictors are gross profit-to-assets (Novy-Marx, 2013), gross profits minus SG&A-to-assets (Ball et al., 2015), and gross profits minus SG&A minus interest-to-book equity (Fama and French, 2015) (only similarly strong as the other two when *Compustat*'s SG&A-is also "corrected" according to Ball et al. (2015), i.e., R&D expenses are readded). While I agree with Novy-Marx (2013) that profitability is an excellent complement to value, I propose to measure value as gross profit minus SG&A–to-market equity instead of book-to-market since book-to-market seems to have lost its predictive power over time. While the difference in profitability to the industry's mean has some predictive power for future returns, it is below the one of the absolute level of profitability, and industry concentration is, at most, of marginal relevance. Nevertheless, profitability can separate strong from weak future performers independently of industry concentrations.

Keywords: Anomaly, Profitability, Trend, Industry Concentration
JEL Classification Numbers: C58, G12, G14

*University of Bern, Faculty of Business, Economics and Social Sciences. Engehaldenstrasse 4, 3012 Bern, Switzerland. Email: jan.pichler@unibe.ch

I Introduction

This paper expands the existing literature on the profitability and anomaly by reevaluating the most important effects described in the literature with more recent data. I provide updated empirical evidence which covers the 40.5 years between June 1980 and December 2021. Such reviews are necessary because, as Linnainmaa and Roberts (2018) point out, many anomalies described in the literature stem from data snooping by only reporting statistically significant results or choosing time frames, data selection, cleaning procedures (e.g., winsorizing), or weighting schemes that yield the desired results. A common flaw many papers share is that they just show tables with average results over the whole sample period and do not bother to show how their described effects evolve. Hence, it is not very surprising that many anomalies cannot be found in other time frames (especially post-publication), when the proper state-of-the-art asset pricing tests are applied (many anomalies heavily rely on hard-to-arbitrage micro-caps), or they can be explained by some common asset pricing factor (see Fama and French (1996)).

The profitability anomaly (more profitable firms yield higher future returns on average than less profitable firms) is well documented in the literature and remains puzzling from a risk perspective. While variation in profitability may identify variation in investors' required rates of returns, the underlying risk story seems questionable. I would argue that firms with low profitability usually tried to improve it and failed; hence, they represent the riskiest stocks and are potentially threatened by bankruptcy, while highly profitable firms did not put their last effort into improving their profitability and are generally far from going bankrupt.

A variety of different profitability measures have been proposed so far, most prominent among them are: Return On Equity (ROE) (Haugen and Baker, 1996), Return on Assets (ROA) (Balakrishnan et al., 2010), Return on Net Operating Assets (RNA) (Soliman, 2008), Profit Margin (PM) (Soliman, 2008), Asset Turnover (ATO) (Soliman, 2008), Capital Turnover (CTO) (Haugen and Baker, 1996), Gross Profit (GP)-to-assets (GP/A) (Novy-Marx, 2013) and GP minus Selling, General and Administrative (SG&A)-to-assets (GPSGA/A) (Ball et al., 2015). Furthermore, Fama and French (2015) use GP minus SG&A and minus interest payments scaled by book equity (GPSGAI/BE) to construct their profitability factor in the Fama-French five-factor model. While all of the aforementioned papers use the current level of profitability, Akbas et al. (2017) show for gross profits-to-assets that the trend in profitability is also not fully priced into today's prices.

In this paper, I focus on GP/A (Novy-Marx, 2013), GPSGA/A (Ball et al., 2015), and GPSGAI/BE Fama and French (2015). Furthermore, I also show evidence on earnings-to-book equity (E/BE), also known as ROE (Haugen and Baker, 1996). Ball et al. (2015) point out that *Compustat*'s SG&A variable also contains Research and Development (R&D) expenses, and they argue that one should add them back to get a cleaner measure of profitability. Hence, I show results for both a "corrected" and an "uncorrected" version of GPSGA/A and GPSGAI/BE. Apart from studying the level effect of the resulting six variables, I apply the methodology of Akbas et al. (2017) to measure the trend in these variables. I highlight that when using quarterly data, the coefficients of a two-year regression with quarterly dummy variables, as in Akbas et al. (2017), yield a measure equal to the average year-over-year growth of the previous four quarters. As a consequence, it ignores any growth between the quarters. Therefore, I also show results using two-year regressions without quarterly dummies, but the results do not change significantly. Fama-MacBeth regressions (Fama and MacBeth, 1973) and value-weighted portfolios are used to assess the importance of the profitability measures as return predictors.

This paper makes multiple contributions to the literature on profitability. First, I show that the level of profitability is a persistent predictor for future returns, while the trend effect is mainly driven by a small time period (2000 to 2006) and cannot be found in the later sample period. Moreover, the portfolio returns based on the different trend measures of profitability are low and statistically insignificant over the whole sample period. The empirical evidence suggests that the level of GP/A, GPSGA/A, and GPSGAI/BE are the strongest predictors (as long as the latter two are "corrected" by re-adding R&D expenses to *Compustat*'s SG&A variable). However, they seem to catch different corners of the spectrum: GP/A excels at finding strong future performers, especially among smaller stocks, while low GPSGA/A or GPSGAI/BE identifies weak future performers.

Second, I confirm Novy-Marx (2013)'s finding that GP/A subsumes the information of asset turnover and gross margin (gross profit-to-sales) well and show that the GPSGA-to-sales margin is also subsumed by GPSGA/A. The results of my double sorts on size and value as the book-to-market (BE/ME) ratio (both parts of the Fama-French three-factor model (Fama and French, 1996)) indicate that these two factors do not predict positive future returns anymore. While I agree with Novy-Marx (2013) that value is an excellent complement to profitability, I propose to measure value (not profitability) as GPSGA (or GPSGAI) scaled by market equity (instead of total assets or book equity for profitability) to overcome the vanishing predictive power of the book-to-market ratio. My results show that this combination leads to portfolios with higher average returns both before and especially after 2010 (sample split at the end of 2010 due to Novy-Marx (2013)'s sample ending there). Over the whole sample period, the combined long-short strategies of a double quintile sort (first on GPSGA/ME, then on profitability) yield raw monthly returns of 1.06% (GP/A), 1.08% (GPSGA/A) and 0.89% (GPSGAI/BE). Note that the portfolios are only rebalanced annually in June and that the Fama-French three-factor alphas are always even higher than the raw returns due to the strategies loading significantly negative on the market (MKT) and the size (SMB) factor, indicating that they are contrarian strategies with more exposure towards larger stocks.

Finally, I investigate the role of industry and industry concentration with respect to profitability. Novy-Marx (2013) shows Fama-MacBeth regressions on profitability measures demeaned by industry (Fama-French 49 industries), i.e., the delta to the industry mean. My results indicate that these deltas can also be predictors for future returns, but unlike suggested by the strong results of their Fama-MacBeth regressions, value-weighted portfolios based on them lag the ones based on the absolute level of profitability significantly and are less consistent over time. Industry concentration measured by the Herfindahl–Hirschman Index (HHI) is at most marginally important for predicting future returns; however, profitability separates poor from strong future performers independently of industry concentration. The same results concerning industry or its concentration can also be found when applying Fama-French 12, 17, or 30 industries, first-level Standard Industrial Classification (SIC), or the Text-based Network Industrial Classification (TNIC) from Hoberg and Phillips (2010, 2016).

This paper proceeds as follows: Section II summarizes literature on profitability, while Section III lays out the methodology and describes the data. Section IV reports the main results on the level, trend, and delta of the level to the industry mean, and the fourth subsection decomposes GPSGA/A in different ways into its components. Section V explores the moderating variables size, value as BE/ME, value as GPSGA/ME, and industry concentration before Section VI concludes.

II Profitability and Related Literature

While value investors aim to buy inexpensive and sell expensive assets, profitability strategies aim to buy highly productive and sell unproductive assets. The superior returns gained from value strategies are arguably compensation for the higher risks of the inexpensive assets and therefore consistent with rational pricing. Novy-Marx (2013) argues that similar to the higher required return by investors of high book-to-market firms, highly productive firms are also expected to yield higher returns. This is illustrated by the mechanical analysis of the dividend discount model in Fama and French (2015), who show that higher expected earnings (which are well proxied by current profitability) imply higher expected returns. While I follow the argument that if variation in productivity identifies the variation in investors' required (or expected) rates of returns, more productive firms should yield higher returns, I still consider the underlying risk story questionable to some degree. One may argue that because the productivity of an already highly productive asset can hardly be increased, it has more downside than upside potential, i.e., it is riskier. However, in the context of entire firms, it seems unlikely that already struggling firms can increase their prices or cut their costs (they are usually in a difficult competitive situation and have tried and failed before); hence, the firms with lower profitabilities are riskier which makes their lower future returns puzzling from a risk perspective. While the general risk story is debatable, it is essential to note that some measures of profitability are clearly related to risk components. E.g., all profitability measures which are scaled by book equity depend on financial leverage and are therefore risk related. In the following subsections, I discuss the most important findings on the absolute level of profitability, its trend, and the relative level of profitability (relative to the industry mean).

II-A Level of Profitability

Profitability and its relation to future returns has been debated for decades in the literature with sometimes mixed results. Most prominently, Fama and French (2006) show that portfolios sorted on earnings-to-book equity yield positive abnormal returns and then raise doubt about it in Fama and French (2008) by also testing the predictor with Fama and MacBeth (1973) regressions. Novy-Marx (2013) argues that earnings-to-book equity (E/BE) is a polluted measure of profitability, and one should instead use gross profits-to-assets (GP/A) to measure it. He provides such clear empirical evidence that Fama and French (2015) include profitability as an additional factor when expanding their three-factor model (Fama and French, 1993) to the five-factor model. However, instead of GP/A, they use gross profits minus SG&A expenses minus interest payments scaled by book equity (GPSGAI/BE). The choice for this measure may be driven by the fact that it is related to the firm's leverage and yields good explanatory results in combination with the value factor BE/ME.

Ball et al. (2015) question Novy-Marx (2013)'s argument that GP/A is truly the cleanest measure for economic profitability and propose to further deduct SG&A from gross profits (i.e., GPSGA/A) since these expenses are also needed for the production of a firm's output. They highlight that the weaker performance of GPSGA/A in many studies is due to the fact that the SG&A variable in *Compustat* (*XSGA* (annual) or *XSGAQ* (quarterly)) is polluted since it also contains R&D expenses. These expenses are not needed to produce a firm's current output but rather spent in order to increase a firm's future earnings and cash flows. I would therefore argue that in a Fama-French five-factor pricing framework, R&D expenses would be more related to the investment factor (Fama and French (2015) compute the investment factor simply as the change in a firm's total assets) rather than the profitability factor. This idea is closely related to the literature on intangible assets and how to measure them (e.g., Chan et al. (2001) propose to compute a stock of R&D capital based on the previous five years' R&D expenses).

Apart from the abovementioned, numerous other indicators have been proposed as measures for a firm's profitability. Most prominent among them are: Return On Equity (ROE) (Haugen and Baker, 1996), Return on Assets (ROA) (Balakrishnan et al., 2010), Return on Net Operating Assets (RNA) (Soliman, 2008), Profit Margin (PM) (Soliman, 2008), Asset Turnover (ATO) (Soliman, 2008), and Capital Turnover (CTO) (Haugen and Baker, 1996). Individual indicators can also be assembled to profitability or failure/bankruptcy scores (they may also include other things like financial leverage, but usually also include some profitability measure(s)): Z-score (Altman, 1968), O-score (Ohlson, 1980), Piotroski's F-score (Piotroski, 2000) or Campbell et al. (2008)'s failure probability. A comparison of these different measures and scores together with many other anomalies is provided by Hou et al. (2015) and for many of them also at different lags and holding periods by Hou et al. (2020).

II-B Trend of Profitability

It may not seem straightforward at first, but the trend in profitability is closely related to earnings surprises, which has a long history in the finance and accounting literature (e.g., Foster et al. (1984)). While earnings and profitability are not the same, they are obviously connected; hence, many findings for one also apply to the other. But how do the trend and surprises fit together?

Earnings surprises are the difference between the actual earnings and the ones that were expected. From a theoretical perspective, one would argue that earnings stem from a stochastic process where the simplest form is something similar to:

$$E_t = E_{t-1} + T_E + \epsilon_t, \tag{1}$$

where the current earnings E_t are the sum of the previous earnings E_{t-1}, some trend component T_E and the error term ϵ_t. Since E_{t-1} is observed, the empirical problem lies in finding a good measure for the trend component. Due to the lack of good proxies for the trend, one could naïvely neglect it and use the previous earnings as expected earnings, hence measuring the latest earnings surprise as the difference between E_t and E_{t-1}. Because this naïve solution is somewhat unsatisfactory, finding other good proxies for expected earnings is an important research question. The most widely used measure for expected earnings are analyst consensus estimates (e.g., Doyle et al. (2006)), but they are systematically biased (Richardson et al., 2004; Chan et al., 2007) and do not reflect market participants' expectations (Malmendier and Shanthikumar, 2007; Mikhail et al., 2007).[1]

As mentioned above, it is difficult to find a good empirical measure for the trend because there are many ways to compute the trend of a time series. Although also different measures are possible for the level (e.g., only taking the last quarter or the previous two years instead of the past four quarters), most research focuses on measures that cover the previous year, four quarters, or twelve months to account for any seasonality effects. Unfortunately, no such "gold standard" exists for the trend. One possibility to measure the trend would be to simply compute the difference between the current and the previous earnings, which is exactly the same as the earnings surprise measure under the naïve assumption of no trend component described above. Alternatively, the trend is part of the expected earnings and, therefore, also related to earnings surprises.

For most of the indicators mentioned in Section II-A, the predictive power of their level has been studied; however, for some variables also the change from the previous to the current level (e.g., Barth et al. (1999) for earnings, Hou et al. (2020) for sales) has been investigated. Akbas et al. (2017) use linear regressions with quarterly dummies over the past eight quarters to measure the trend in GP/A. In this paper, I follow the approach of Akbas et al. (2017) but also investigate other profitability measures besides GP/A. Furthermore,

[1] Chiang et al. (2019) propose measuring the earnings surprise as the difference in the number of estimates that were above and below the actual earnings, scaled by the total number of analyst estimates, as it is a more robust measure than the difference to the consensus estimate. While yielding a more robust earnings surprise measure, this procedure does not use a proxy for expected earnings.

for reasons elaborated in Section III, I also show results for trend measures stemming from regressions without quarterly dummies.

II-C Profitability Relative to Peers

The main reason for investigating a firm's profitability relative to its peers is that profitability can be highly industry-dependent, and the outperformance of more profitable firms may stem from their competitive advantage against their peers. Hence, the relative measure could be more informative than the absolute level. Industry refers to economic actors involved in the same (or at least similar) type of business, and in general, the SIC system is the most widely used for grouping firms. But deciding which firm is a peer (and to what extent) and which is not is notoriously difficult because no two firms are identical, and larger companies often operate in multiple markets. If the definition of peers is too broad, the companies are not all in competition with each other. However, if it is very narrow, there may only be very few peers, and any means estimates are not robust. E.g., the first level SIC could be too broad since, by far, not all companies in the manufacturing group (codes 2000 to 3999) are competing with each other.
Alternative industry groupings (12, 17, 30, or 49 industries) are provided by Eugene F. Fama and Kenneth R. French on the latter's website, and they are often more appropriate in the context of finance. Furthermore, Hoberg and Phillips (2010) propose a pairwise similarity score between firms based on their product descriptions in their 10-K filings.
Novy-Marx (2013) shows in his Table 1 that GP/A demeaned by industry (Fama-French 49 industries) is an even stronger predictor for future returns than the absolute level in Fama-MacBeth regressions. Although this "industry adjustment," as it is also referred to, is used for constructing the factors displayed in Table 10 in Novy-Marx (2013), a direct comparison of value-weighted portfolios with the absolute level is missing.

III Methodology & Data

III-A Methodology

My analysis mainly focuses on the most prominent measures of profitability: GP scaled by total assets (GP/A, Novy-Marx (2013)), GP minus SG&A and interest expenses, scaled by the book value of equity (GPSGAI/BE, Fama and French (2015)) and return on equity (E/BE, Haugen and Baker (1996)). Since Ball et al. (2015) point out that R&D is polluting *Compustat*'s SG&A variable (i.e., R&D expenses are included in *Compustat*'s SG&A), I also investigate GP minus SG&A scaled by total assets: once as GPSGARD/A (*Compustat* items for nominator: *REVTQ-COGSQ-XSGAQ*) and once as GPSGA/A (*Compustat* items for nominator: *REVTQ-COGSQ-XSGAQ+XRDQ*). Furthermore, I apply the same distinction to GPSGAI/BE by distinguishing GPSGARDI/BE (*Compustat* items for nominator: *REVTQ-COGSQ-XSGAQ-XINTQ*) and GPSGAI/BE (*Compustat* items for nominator: *REVTQ-COGSQ-XSGAQ+XRDQ-XINTQ*). Hence, GPSGARDI/BE is the variable used in Fama and French (2015)'s five-factor model and not GPSGAI/BE.

Depreciation is usually not considered in these profitability measures because it is rather the result of a firm's depreciation policy than the assets' true economic decline in value. For tax reasons, firms have an incentive to depreciate their assets faster, which leads to lower profitability after depreciation of younger firms, whereas it increases the profitability of older firms that still create goods and services with fully depreciated equipment or properties. On the other hand, managers with variable payments that depend on measures after depreciation are incentivized to depreciate their assets at a slower pace to boost their compensation. In undisclosed results, I confirm that depreciation is a polluting factor; however, for the sake of readability, no results regarding depreciation and amortization are displayed in this paper.

While the computation of the profitability measures' nominators is straightforward and total assets as a denominator can be found directly in *Compustat*, the computation of the book value of equity is more complicated. In this paper, book equity (BE) is computed according to Novy-Marx (2013):[2] Shareholder equity, plus deferred taxes, minus preferred stock, when available. Stockholders equity is *Compustat*'s *SEQ* variable if available, or else common equity plus the carrying value of the preferred stock (*CEQQ+PSTXQ*) if available, or else total assets minus total liabilities (*ATQ-LTQ*). This is consistent with the definition of Fama and French (1993) for their definition of the HML factor. Deferred taxes are deferred taxes and investment tax credits (*TXDITCQ*). Preferred stock is its redemption value (*PSTRQ*) if available, or else its carrying value (*PSTKQ*).

This paper investigates three different effects of profitability: the absolute level, the trend, and the level relative to the industry's mean. The level effect simply takes the absolute value of the measure and states that firms with a higher value have higher average returns in the subsequent month than firms with a lower value. This is common in the financial literature not only for profitability, but many other anomalies or factors. E.g., Fama and French (1992) use the level of book equity divided by market equity to sort their portfolios for the construction of their value factor in the famous Fama-French three-factor model (and similarly the level of market equity to construct the size factor). Note that I use annual levels despite having quarterly data, i.e., I aggregate the previous four quarters to account for any seasonality in a firm's accounting measures.

The trend effect, on the other hand, uses some measure of the change in the level variable to sort the stocks. While conceptually straightforward, the application is less so since there are many ways to measure the trend of a time series. Furthermore, the relatively sparse literature in finance does not provide any "gold standard" on how to measure it. One possibility is to simply take the last change, but when working with quarterly data, one already has to decide whether to take the quarterly or the year-over-year change. In this paper, I

[2] Note that I use quarterly data instead of annual data.

follow Akbas et al. (2017), who measure the trend in profitability with the following rolling regression for each firm i:[3]

$$GPQ_{i,q} = \alpha_i + \beta_i t + \lambda_1 D_1 + \lambda_2 D_2 + \lambda_3 D_3 + \epsilon_{i,q}, \tag{2}$$

where $GPQ_{i,q}$ is the quarterly gross profitability, $t = 1, 2, ..., 8$ represents a time index covering quarters q_{-7} to q, and D_1, D_2, and D_3 are quarterly dummy variables to account for potential seasonality in gross profits. Hence, β_i is the measure of the trend in gross profitability for firm i. It is important to note that this model is somewhat misspecified: Because it only covers the last eight quarters and includes a quarterly dummy variable, the solution to the regression could also be computed as the mean year-over-year change in gross profitability from the year before the previous year and to the previous year. This misspecification does not result in an invalid trend measure; however, it could be computed more efficiently, and one should be aware that it ignores any quarterly growth within the previous two years. Nevertheless, I use this trend measure to provide comparable results to Akbas et al. (2017). Furthermore, I also measure the trend in gross profitability as follows:[4]

$$GPQ_{i,q} = \alpha_i + \beta_i t + \epsilon_{i,q}. \tag{3}$$

Apart from the level and trend, I investigate the delta (Δ) of a firm's profitability, which is simply the difference between a firm's level of profitability and the corresponding industry's mean. Since results may heavily depend on the definition of industry, I use the Fama-French 12, 17, 30, and 49 industries as well as the first-level SIC industries. Furthermore, I compute the difference between a firm's level of profitability and the weighted mean profitability of a firm's competitors according to the scores of the Hoberg and Phillips (2010) TNIC database. The measures are compared using Fama-MacBeth regressions, where I control for value $\left(\text{Ln}\left(\frac{BE}{ME}\right)\right)$, size (Ln(ME)), short-term reversal ($\text{return}_{-1,0}$), and momentum ($\text{return}_{-12,-1}$). To be consistent with Novy-Marx (2013), the market value of equity (ME) is lagged by six months to avoid uptakes in momentum. All independent variables are winsorized at the 1%- and 99%-level, and the dependent variable (next month's return) is capped at 100% (0.34% of the observations) in order for results to not be mainly determined by a few outliers. Since Fama-MacBeth regressions weigh all stocks equally, comparing the results of value-weighted portfolio sorts is necessary to assess the economic viability of the effects. Note that I use the uncapped returns because they are multiplied with a small weight since they occur at micro-caps. For all portfolio sorts, I show the arithmetic mean as well as the geometric mean and the alpha of a Fama-French three-factor model regression:

$$r_t = \alpha + \beta \text{MKT}_t + \gamma \text{SMB}_t + \lambda \text{HML}_t + \epsilon_t, \tag{4}$$

[3] Note that I use gross profitability to describe the measurement of the trend, but it applies to all other profitability measures analogously.

[4] In the presence of seasonalities, this equation is only suitable to measure the trend if the same amount of observations for each calendar quarter (here two) is used. Furthermore, it only provides an adequate trend measure but would not yield any suitable predictions for the profitability of a specific quarter.

where r_t is the portfolios' monthly return which is regressed on the three factors (MKT: market, SMB: size, HML: value) that are taken from Kenneth French's website. The portfolio returns and the three-factor model alphas are tested with t-tests; in the case of geometric returns, the t-tests are conducted on the log of one plus the returns. All portfolios are market capitalization-weighted and rebalanced annually at the end of June.

III-B Data

All data is gathered from Center for Research in Security Prices (CRSP), *Compustat*, and CRSP-*Compustat* merged database. The dataset starts in January 1977 (quarterly data is not available before) and ends in December 2021. However, due to the lagging of some variables and to be consistent with Novy-Marx (2013), all portfolio sorts and asset pricing tests cover the period between June 1980 and December 2021 (except for the ones using the TNIC, which is only available from 1988 onwards). The data is merged on *LPERMNO* and a date key (year and month), which is formed from the data date. All balance sheet and income statement variables are lagged for three months to ensure availability. This merging procedure is slightly different from Novy-Marx (2013), who merges the data on a given fiscal year variable (*FYEAR*) from *Compustat* with the CRSP data starting at the end of June of the subsequent year. As a consequence, the lag of the variables can differ between firms in Novy-Marx (2013), whereas it is always three months here (plus the difference between the fiscal quarter end and the month end if it does not coincide). All financial firms (SIC code between 6000 and 6999) are excluded, and only primary listings are considered (*PRIUSA*=1). Furthermore, I use the Fama-French industry classifications (12, 17, 30, and 49 industries) from Kenneth French's website and the TNIC scores and the TNIC HHI from Hoberg and Phillips (2010, 2016), which are available on their website.

Table 1 shows the descriptive statistics for the monthly returns, raw in the first column and capped at 100% in the second. I use the capped returns for the Fama-MacBeth regressions so that the results are not purely driven by a few outliers at the upper end. The raw returns are used for the value-weighted portfolios where the outliers do not matter much since they occur at micro- caps and hence receive a tiny weight.

Table 2 shows the descriptive statistics of the level and trend variables as well as the difference between the level and the industry mean (i.e., the delta) according to the Fama-French 49 industries. The median profitability is positive across all measures except for E/BE (due to its nominator because negative book equity is excluded). It is worth noting that measures scaled by book equity are more dispersed, with standard deviations between 43.78% and 69.99%, than the ones scaled by total assets (17.00% to 30.90%). The trend measures clearly tend to be negative on average and have the same dispersion pattern as the level measures. The number of observations is slightly lower for the trend measures since they need eight-quarters of available data in order to be computed.

Table 3 shows the pairwise Spearman rank correlations, which already offer some key insights. First of all, GP/A is only moderately correlated with the other profitability measures, which are highly correlated with each other. Second, both trend measures (i.e., regressions with and without dummies) are highly correlated. Third, all trend measures are only very weakly positively correlated to the profitability level or the delta to the industry mean. And fourth, all level measures are moderately to highly related to the industry deltas.

Table 1: Descriptive Statistics - Returns

This table describes the monthly returns of the dataset covering the entire sample from June 1980 to December 2021. The first column covers the raw returns, and the second column the monthly returns capped at 100% (0.34% of the observations). I use the capped returns for the Fama-MacBeth regressions in order for results not to be purely driven by a few outliers at the upper end. The raw returns are used for the value-weighted portfolios where the outliers do not matter much since they occur at micro-caps and hence receive a tiny weight.

	Raw returns	Capped returns
Mean	13.38%	1.00%
Std.	110.02%	18.48%
Min.	−100.00%	−100.00%
25%-quantile	−7.80%	−7.80%
50%-quantile (median)	0.00%	0.00%
75%-quantile	7.90%	7.90%
Max.	149′999.00%	100.00%
N (in million)	2.04	2.04

Table 2: Descriptive Statistics - Level and Trend

This table describes the dataset covering the entire sample from June 1980 to December 2021. The level of profitability is the annual profitability (i.e., the last four quarters), and all variables are winsorized at the 1%- and the 99%-quantile. Note that the level and the difference to the industry mean are in percentage points, and the trend variables are in basis points.

		GP/A	GPSGA/A	GPSGARD/A	GPSGAI/BE	GPSGARDI/BE	E/BE
Level	Mean	32.42%	10.12%	6.53%	17.88%	10.46%	−13.87%
	Std.	30.90%	17.00%	19.95%	43.78%	48.84%	69.99%
	Min.	−74.30%	−68.20%	−94.94%	−204.35%	−269.71%	−465.14%
	25%-quantile	14.97%	5.35%	3.20%	7.38%	3.97%	−10.04%
	50%-quantile	30.38%	12.24%	10.60%	21.44%	18.06%	6.16%
	75%-quantile	48.63%	18.84%	16.53%	34.40%	29.83%	13.23%
	Max.	123.75%	46.30%	38.89%	172.99%	139.23%	58.06%
	N (in million)	1.68	1.40	1.40	1.13	1.00	1.54
Trend AJK	Mean	−1.28*bp*	−6.15*bp*	−1.91*bp*	14.36*bp*	−11.65*bp*	−25.09*bp*
	Std.	80.12*bp*	129.08*bp*	63.35*bp*	656.03*bp*	172.13*bp*	270.84*bp*
	Min.	−300.32*bp*	−478.42*bp*	−240.16*bp*	−2915.86*bp*	−861.28*bp*	−1502.02*bp*
	25%-quantile	−27.62*bp*	−55.41*bp*	−22.44*bp*	−121.27*bp*	−47.37*bp*	−50.80*bp*
	50%-quantile	−0.76*bp*	−2.02*bp*	−0.95*bp*	−1.46*bp*	−3.12*bp*	−4.10*bp*
	75%-quantile	23.29*bp*	46.14*bp*	17.87*bp*	118.63*bp*	33.57*bp*	27.77*bp*
	Max.	325.24*bp*	436.02*bp*	251.96*bp*	3319.96*bp*	697.46*bp*	1052.13*bp*
	N (in million)	1.46	1.24	1.20	0.78	0.76	1.27
Trend 2yr reg.	Mean	−0.99*bp*	−6.01*bp*	−1.72*bp*	15.97*bp*	−10.83*bp*	−24.46*bp*
	Std.	76.63*bp*	127.66*bp*	60.47*bp*	633.54*bp*	164.09*bp*	242.28*bp*
	Min.	−276.01*bp*	−458.50*bp*	−222.60*bp*	−2764.40*bp*	−810.87*bp*	−1′378.66*bp*
	25%-quantile	−28.24*bp*	−57.99*bp*	−22.90*bp*	−123.96*bp*	−47.71*bp*	−47.78*bp*
	50%-quantile	−0.88*bp*	−2.25*bp*	−1.14*bp*	−0.72*bp*	−3.09*bp*	−4.45*bp*
	75%-quantile	24.13*bp*	48.27*bp*	18.45*bp*	123.45*bp*	35.11*bp*	27.72*bp*
	Max.	307.90*bp*	420.13*bp*	240.58*bp*	3187.52*bp*	654.78*bp*	885.90*bp*
	N (in million)	1.46	1.24	1.20	0.78	0.76	1.27
Delta to mean of Fama-French 49 industries	Mean	0.38%	0.54%	0.79%	0.71%	4.44%	16.99%
	Std.	26.20%	16.38%	18.58%	62.84%	69.57%	91.46%
	Min.	−88.31%	−72.96%	−88.08%	−311.23%	−339.75%	−417.74%
	25%-quantile	−12.92%	−4.91%	−4.29%	−14.44%	−11.27%	−0.55%
	50%-quantile	−0.70%	1.82%	2.58%	2.61%	5.44%	14.23%
	75%-quantile	12.31%	8.76%	9.58%	20.03%	23.02%	34.92%
	Max.	85.05%	39.98%	46.41%	274.73%	320.62%	466.61%
	N (in million)	1.65	1.38	1.38	1.11	0.98	1.52

Table 3: Pairwise Spearman Correlations

This table shows the pairwise Spearman rank correlations of the different variables computed over the whole sample period (June 1980 to December 2021). All correlations are positive, and the strength of the correlation is graphically displayed with the intensity of the grey tones (the darker, the more positive, and the lighter, the more uncorrelated or even negatively correlated).

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
(1)	Level GP/A	1.00																							
(2)	Level GPSGA/A	0.53	1.00																						
(3)	Level GPSGARD/A	0.45	0.90	1.00																					
(4)	Level GPSGAI/ABE	0.47	0.89	0.81	1.00																				
(5)	Level GPSGARDI/BE	0.39	0.82	0.91	0.91	1.00																			
(6)	Level E/BE	0.40	0.73	0.82	0.70	0.79	1.00																		
(7)	Trend AJK GP/A	0.15	0.13	0.12	0.13	0.13	0.09	1.00																	
(8)	Trend AJK GPSGA/A	0.11	0.10	0.08	0.10	0.08	0.02	0.69	1.00																
(9)	Trend AJK GPSGRD/A	0.14	0.24	0.24	0.24	0.24	0.22	0.74	0.60	1.00															
(10)	Trend AJK GPSGAI/BE	0.02	−0.04	−0.06	0.01	−0.02	−0.14	0.41	0.65	0.35	1.00														
(11)	Trend AJK GPSGARDI/BE	0.12	0.25	0.26	0.27	0.28	0.23	0.67	0.52	0.88	0.45	1.00													
(12)	Trend AJK E/BE	0.11	0.24	0.25	0.23	0.25	0.37	0.48	0.39	0.68	0.21	0.71	1.00												
(13)	Trend 2yr GP/A	0.14	0.12	0.10	0.11	0.11	0.07	0.91	0.64	0.67	0.38	0.61	0.44	1.00											
(14)	Trend 2yr GPSGA/A	0.11	0.09	0.07	0.09	0.07	0.02	0.63	0.90	0.54	0.59	0.47	0.35	0.71	1.00										
(15)	Trend 2yr GPSGARD/A	0.13	0.22	0.22	0.22	0.22	0.20	0.67	0.56	0.89	0.33	0.78	0.62	0.75	0.62	1.00									
(16)	Trend 2yr GPSGAI/BE	0.02	−0.05	−0.06	0.00	−0.03	−0.15	0.38	0.60	0.32	0.91	0.41	0.20	0.45	0.67	0.39	1.00								
(17)	Trend 2yr GPSGARDI/BE	0.11	0.23	0.23	0.25	0.26	0.20	0.61	0.49	0.78	0.42	0.89	0.64	0.69	0.55	0.89	0.48	1.00							
(18)	Trend 2yr E/BE	0.11	0.23	0.24	0.22	0.24	0.35	0.46	0.38	0.65	0.21	0.67	0.88	0.52	0.43	0.72	0.26	0.75	1.00						
(19)	Δ Level FF49 GP/A	0.75	0.48	0.39	0.40	0.33	0.34	0.18	0.11	0.14	0.01	0.13	0.12	0.17	0.11	0.13	0.01	0.12	0.12	1.00					
(20)	Δ Level FF49 GPSGA/A	0.46	0.89	0.80	0.78	0.72	0.64	0.12	0.08	0.22	−0.05	0.23	0.22	0.10	0.07	0.20	−0.05	0.21	0.21	0.54	1.00				
(21)	Δ Level FF49 GPSGARD/A	0.39	0.81	0.82	0.71	0.72	0.66	0.10	0.06	0.22	−0.06	0.23	0.22	0.08	0.06	0.20	−0.06	0.21	0.22	0.48	0.92	1.00			
(22)	Δ Level FF49 GPSGAI/BE	0.36	0.64	0.59	0.70	0.64	0.50	0.10	0.07	0.18	−0.01	0.21	0.17	0.08	0.06	0.16	0.01	0.19	0.17	0.42	0.75	0.70	1.00		
(23)	Δ Level FF49 GPSGARDI/BE	0.32	0.60	0.60	0.63	0.63	0.52	0.09	0.05	0.18	−0.03	0.21	0.19	0.08	0.05	0.16	−0.03	0.19	0.18	0.38	0.71	0.76	0.90	1.00	
(24)	Δ Level FF49 E/BE	0.22	0.43	0.40	0.39	0.38	0.48	0.04	0.00	0.13	−0.11	0.14	0.24	0.03	−0.00	0.12	−0.11	0.13	0.23	0.27	0.55	0.62	0.60	0.71	1.00

IV Main Results

IV-A Level of Profitability

Table 4 shows that the Fama-MacBeth regressions are highly significant for all the measures of the level of profitability and persistently so across time. All regressions are controlled for value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. To improve the readability, full results, i.e., including the coefficients of the control variables, are only displayed in Table A.1 in the appendix. All coefficients are in percent; hence, 1.15 for GP/A means that a one percentage point higher GP/A is related to a 1.15 basis points higher return in the subsequent month over the entire sample. Or in other words: A one standard deviation increase in GP/A (30.90%) is related to an increase of 0.36% in the next month's return. Note that larger coefficients do not necessarily translate into more profitable strategies since the profitability measures have different distributions; hence, they are inadequate for comparing the different measures. The t-test values are better suited for this purpose and indicate that GPSGA/A and GPSGAI/BE are the strongest predictors for future returns (note that the polluting effect of R&D expenses is substantial in both cases). Comparing t-test values of regressions with different numbers of observations requires some caution since increasing the number of observations by n increases the t-values mechanically by $\sqrt{n}$ (hence, the t-test values are also naturally larger for the first subperiod covering the sample till the end of 2010 than for the subperiod after 2010). In this case, it even advocates the superiority of GPSGAI/BE, which already has the largest t-test value since its regressions additionally have the lowest number of observations. But be aware when comparing the different measures, that the varying number of observations does not stem from randomly missing data but rather from a particular subset, namely small, illiquid stocks (due to missing data) or firms with negative book equity (because of exclusion). Since the data on these stocks would likely have higher standard deviations, it would have an adverse effect on the t-values as the increase in the number of observations. Although the adjustment seems less problematic for different subperiods, one still has to consider that the data coverage has improved over time.

Since Fama-MacBeth regressions weigh all observations equally, comparing them with value-weighted portfolio sorts is necessary to assess the effects' economic relevance. Figure 1 shows the portfolio developments (highest quintile minus lowest quintile) graphically, and Table 5 displays the corresponding means, geometric means, and Fama-French three-factor model alphas and their respective t-test values. They confirm the previous finding that GPSGA/A and GPSGAI/BE are the strongest predictors for future returns not only for the whole sample period but both in the early and the late sample period.

Unlike in the remainder of this paper, I included E/A here since Ball et al. (2015) claim that earnings before extraordinary items have similar predictive power as GP/A when scaled with the same denominator. In their Table 3, they use Fama-MacBeth regressions to show that this is true when excluding micro-caps. Because the portfolios of Figure 1 are value-weighted, they are similar to Fama-MacBeth regressions that exclude micro-caps. And indeed, E/A experienced strong portfolio results despite significantly lower t-values in the early sample period compared to GP/A. Fama-MacBeth regressions excluding micro-caps (stocks that are below the 20th percentile of the New York Stock Exchange (NYSE) market capitalization distribution) for my sample period are displayed in Table A.2 in the appendix and suggest that the finding of Ball et al. (2015) is also primarily driven by the earlier sample period of 1963 to 1980, which is not included here. But the story really gets flipped on its head post-2010 when E/A had similar t-values in Fama-MacBeth regressions, including micro-caps but clearly underperforming portfolio results (Figure 1 and Table 5) and also weaker results in Fama-MacBeth regressions excluding micro-caps. I further elaborate on the role of the denominator in Subsection IV-D (decomposition 3) and conclude here that the choice of the nominator matters and more so in recent years since "clean" measures like GP/A, GPSGA/A, and GPSGAI/BE result in returns about twice as high as earnings based measures (see last part of Table 5).

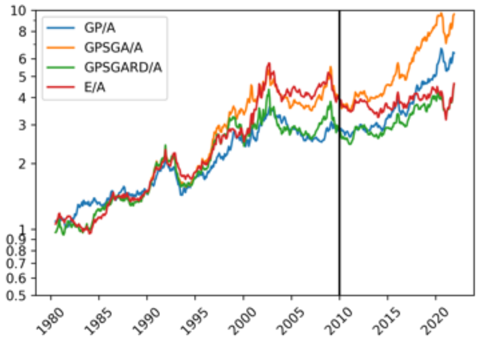
Table 4: Fama-MacBeth Regression Results - Level

This table shows the results of the Fama-Mac-Beth regressions for the level of profitability, where all measures are based on the previous four quarters. The first part covers the whole sample (June 1980 to December 2021), the second and third the time period before and after December 2010. All coefficients are in percent, and the values in brackets are the corresponding t-test values. For readability, the results for the control variables (value as Ln(BE/ME), size as Ln(ME), short-term reversal as past month return, and momentum as past twelve months return, skipping the most recent month) are not displayed here but can be found in Table A.1 in the appendix.

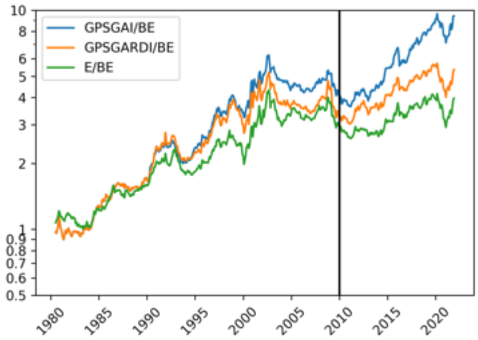
	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$	$\frac{E}{A}$
June 1980 to December 2021							
Profitability level	1.15	2.65	1.73	1.02	0.72	0.32	1.11
	(7.66)	(9.06)	(5.71)	(10.32)	(6.91)	(4.29)	(4.20)
N (in million)	1.50	1.25	1.25	1.00	1.00	1.54	1.54
June 1980 to December 2010							
Profitability level	1.20	2.73	1.79	1.12	0.80	0.28	1.01
	(7.32)	(8.40)	(4.81)	(9.21)	(6.05)	(2.97)	(3.11)
N (in million)	1.18	0.97	0.97	0.76	0.76	1.22	1.22
January 2011 to December 2021							
Profitability level	1.06	2.52	1.72	0.67	0.52	0.43	1.42
	(3.13)	(3.94)	(3.35)	(4.55)	(3.45)	(3.76)	(3.34)
N (in million)	0.32	0.27	0.27	0.23	0.23	0.32	0.32

Figure 1: Portfolios Sorted on Level of Profitability

These figures show the developments of the value-weighted portfolios sorted on the level of profitability (highest minus lowest quintile) between June 1980 and December 2021. Subfigure (a) displays the profitability measures which are scaled by the book value of assets, and Subfigure (b) the ones that are scaled by the book value of equity. Portfolios are rebalanced annually at the end of June. Table 5 summarizes the corresponding mean, geometric mean, and Fama-French three-factor model alphas.



(a) Profitability measures scaled by assets



(b) Profitability measures scaled by book equity

Table 5: Portfolio Results - Level

This table shows the results of the value-weighted portfolios sorted on the different measures of the level of profitability (highest minus lowest quintile). The first covers the whole sample, the second and third the time period before and after December 2010. Portfolios are rebalanced annually at the end of June. The percentage values are percent per month, and the values in brackets are the corresponding t-test values. FF3F α stands for Fama-French three-factor model alpha.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$	$\frac{E}{A}$
				June 1980 to December 2021			
Mean	0.41%	0.51%	0.37%	0.51%	0.40%	0.33%	0.36%
	(3.28)	(3.39)	(2.36)	(3.40)	(2.53)	(2.27)	(2.46)
Geomean	0.37%	0.46%	0.31%	0.45%	0.34%	0.28%	0.31%
	(2.98)	(3.00)	(1.96)	(3.02)	(2.14)	(1.91)	(2.10)
FF3F α	0.46%	0.63%	0.46%	0.57%	0.47%	0.36%	0.44%
	(3.58)	(4.14)	(2.95)	(3.78)	(2.92)	(2.48)	(2.96)
				June 1980 to December 2010			
Mean	0.31%	0.41%	0.32%	0.41%	0.37%	0.32%	0.39%
	(2.18)	(2.30)	(1.67)	(2.35)	(1.94)	(1.85)	(2.30)
Geomean	0.28%	0.35%	0.25%	0.36%	0.31%	0.26%	0.34%
	(1.92)	(1.97)	(1.31)	(2.04)	(1.58)	(1.53)	(1.98)
FF3F α	0.36%	0.53%	0.40%	0.50%	0.45%	0.34%	0.46%
	(2.43)	(2.98)	(2.08)	(2.83)	(2.31)	(1.94)	(2.66)
				January 2011 to December 2021			
Mean	0.68%	0.81%	0.49%	0.71%	0.42%	0.38%	0.32%
	(2.82)	(3.05)	(1.96)	(2.67)	(1.66)	(1.46)	(1.17)
Geomean	0.65%	0.76%	0.45%	0.66%	0.38%	0.33%	0.27%
	(2.67)	(2.87)	(1.81)	(2.49)	(1.50)	(1.30)	(1.00)
FF3F α	0.58%	0.75%	0.42%	0.58%	0.35%	0.38%	0.30%
	(2.30)	(2.67)	(1.61)	(2.06)	(1.31)	(1.38)	(1.03)

IV-B Trend of Profitability

Table 6 shows the results of the FamaMac-Beth regressions of the trend in profitability. The trend is measured as the coefficient of a regression over the past eight quarters, once with quarterly dummies (AJK trend) and once without (2yr regression). Over the whole sample period, the trend has statistically significant predictive power for future returns. However, the trend in profitability has lost its predictive power over the last decade and is even negative in most cases.

Figures 2 and 3 provide the most revealing evidence concerning the trend of profitability. Subfigures (a) show that value-weighted portfolios built based on the trend of all measures scaled by the book value of assets had consistently negative returns in the subperiod 2010-2020 and that this trend already started around 2006/2007. A closer inspection of the earlier sample period clearly shows that most of the positive predictive power of the profitability's trend measures scaled by the book value of assets mainly stems from the time period between 2000 and 2006. Subfigures (b) show that value-weighted portfolios built based on the trend of all measures scaled by the book value of equity developed flatly between 2010 and 2020 and did more or less so as well before, with the exception of GPSGAI/BE and E/BE when measured with the AJK trend (but the development of the portfolios based on the trend of GPSGAI/BE was very unstable). The corresponding means, geometric means, and Fama-French three-factor model alphas can be found in Tables 7 and 8. None of the mean and geometric mean returns is statistically significant over the whole sample or the early sample, not even speaking of the later sample period where most of them are negative.

As expected, due to the high pairwise correlations (0.88 to 0.91) between the coefficients from regressions with and without dummies, the portfolio results do not differ systematically between the two. While I clearly show that both regressions measuring the trend over the past eight quarters do not yield a reliable predictor, I cannot rule out that such a trend measure exists. But profitability measures are generally sticky, and deviations from one quarter to the next are probably rather noise than relevant information in most cases, making the current level of profitability the best estimator for future profitability. Nevertheless, if a reliable measure of the trend in profitability existed, one could compute a better estimate for expected future profitability when combining it with the current level of profitability (e.g., current level plus X times trend).[5]

[5] This was one of the initial ideas of this paper.

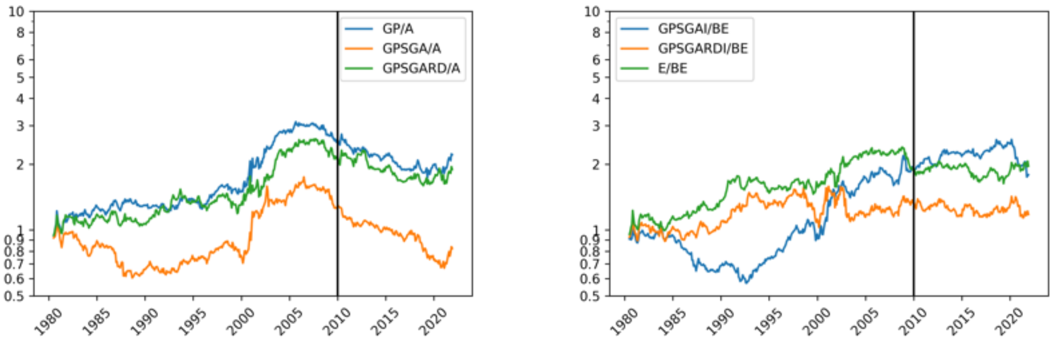
Table 6: Fama-MacBeth Regression Results - Trend

This table shows the results of the FamaMac-Beth regressions for the trend in profitability. The first part covers the whole sample (June 1980 to December 2021), the second and third the time period before and after December 2010. The trend is measured as the coefficient of a regression with quarterly dummies (AJK trend) and without dummies (2yr regression) over the previous eight quarters. All coefficients are in percent, and the values in brackets are the corresponding t-test values. For readability, the results for the control variables (value as Ln(BE/ME), size as Ln(ME), short-term reversal as past month return, and momentum as past twelve months return, skipping the most recent month) are not displayed here but can be found in Table A.3 and Table A.4 in the appendix.

	$\frac{\text{GP}}{\text{A}}$	$\frac{\text{GP-SGA}}{\text{A}}$	$\frac{\text{GP-SGA-RD}}{\text{A}}$	$\frac{\text{GP-SGA-I}}{\text{BE}}$	$\frac{\text{GP-SGA-RD-I}}{\text{BE}}$	$\frac{\text{E}}{\text{BE}}$
			June 1980 to December 2021			
AJK trend	11.38	10.17	10.17	1.11	9.02	2.10
	(3.61)	(5.45)	(2.13)	(2.92)	(4.81)	(1.73)
2yr regression trend	16.89	12.54	17.70	1.42	11.22	3.77
	(5.27)	(6.72)	(3.64)	(3.56)	(5.79)	(2.67)
N (in million)	1.30	1.10	1.07	0.78	0.76	1.26
			June 1980 to December 2010			
AJK trend	20.06	15.71	16.06	1.69	11.45	2.77
	(5.56)	(7.83)	(3.04)	(3.74)	(5.27)	(1.84)
2yr regression trend	26.71	18.45	25.71	2.08	15.01	5.15
	(7.32)	(9.22)	(4.79)	(4.37)	(6.76)	(2.94)
N (in million)	1.00	0.84	0.81	0.59	0.57	0.99
			January 2011 to December 2021			
AJK trend	−11.98	−6.32	−6.00	−0.61	2.58	0.25
	(−2.01)	(−1.54)	(−0.58)	(−0.90)	(0.72)	(0.13)
2yr regression trend	−9.84	−4.93	−4.55	−0.49	1.10	−0.01
	(−1.63)	(−1.21)	(−0.43)	(−0.69)	(0.29)	(−0.00)
N (in million)	0.30	0.25	0.25	0.19	0.19	0.27

Figure 2: Portfolios Sorted on AJK Trend of Profitability

These figures show the developments of the value-weighted portfolios sorted on the trend of profitability (highest minus lowest quintile) between June 1980 and December 2021. Subfigure (a) displays the profitability measures which are scaled by the book value of assets, and Subfigure (b) the ones that are scaled by the book value of equity. The AJK trend is measured as the coefficient of a regression with quarterly dummies over the previous eight quarters. Portfolios are rebalanced annually at the end of June. Table 7 summarizes the corresponding mean, geometric mean, and Fama-French three-factor model alphas.



(a) Profitability measures scaled by assets (b) Profitability measures scaled by book equity

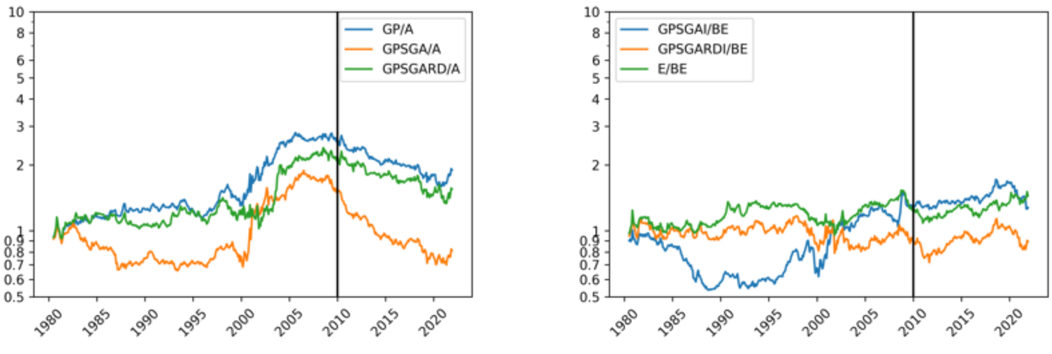
Table 7: Portfolio Results - AJK Trend

This table shows the results of the value-weighted portfolios sorted on the different AJK trend measures (highest minus lowest quintile). The AJK trend is measured as the coefficient of a regression with quarterly dummies over the previous eight quarters. The first part covers the whole sample (June 1980 to December 2021), the second and third the time period before and after December 2010. Portfolios are rebalanced annually at the end of June. The percentage values are percent per month, and the values in brackets are the corresponding t-test values. FF3F α stands for Fama-French three-factor model alpha.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$
June 1980 to December 2021						
Mean	0.20%	0.01%	0.18%	0.16%	0.09%	0.17%
	(1.55)	(0.05)	(1.27)	(1.24)	(0.59)	(1.42)
Geomean	0.16%	−0.04%	0.13%	0.12%	0.03%	0.14%
	(1.23)	(−0.29)	(0.93)	(0.93)	(0.22)	(1.13)
FF3F α	0.20%	0.01%	0.19%	0.16%	0.11%	0.19%
	(1.49)	(0.10)	(1.32)	(1.29)	(0.71)	(1.58)
June 1980 to December 2010						
Mean	0.29%	0.09%	0.27%	0.23%	0.12%	0.21%
	(1.86)	(0.51)	(1.59)	(1.52)	(0.66)	(1.41)
Geomean	0.25%	0.03%	0.22%	0.19%	0.06%	0.17%
	(1.57)	(0.20)	(1.29)	(1.25)	(0.32)	(1.15)
FF3F α	0.27%	0.06%	0.24%	0.22%	0.10%	0.21%
	(1.64)	(0.34)	(1.41)	(1.43)	(0.55)	(1.44)
January 2011 to December 2021						
Mean	−0.07%	−0.17%	−0.11%	−0.14%	−0.08%	−0.02%
	(−0.36)	(−0.84)	(−0.53)	(−0.75)	(−0.37)	(−0.10)
Geomean	−0.10%	−0.20%	−0.14%	−0.17%	−0.11%	−0.04%
	(−0.49)	(−0.97)	(−0.66)	(−0.86)	(−0.51)	(−0.22)
FF3F α	−0.07%	−0.16%	−0.09%	−0.15%	−0.10%	−0.05%
	(−0.32)	(−0.74)	(−0.41)	(−0.71)	(−0.43)	(−0.25)

Figure 3: Portfolios Sorted on 2yr Regression Trend of Profitability

These figures show the developments of the value-weighted portfolios sorted on the trend of profitability (highest minus lowest quintile) between June 1980 and December 2021. Subfigure (a) displays the profitability measures which are scaled by the book value of assets, and Subfigure (b) the ones that are scaled by the book value of equity. The 2yr regression trend is measured as the coefficient of a regression without quarterly dummies over the previous eight quarters. Portfolios are rebalanced annually at the end of June. Table 5 summarizes the corresponding mean, geometric mean, and Fama-French three-factor model alphas.



(a) Profitability measures scaled by assets (b) Profitability measures scaled by book equity

Table 8: Portfolio Results - 2yr Regression Trend

This table shows the results of the value-weighted portfolios sorted on the different 2yr regression trend measures (highest minus lowest quintile). The 2yr regression trend is measured as the coefficient of a regression without quarterly dummies over the previous eight quarters. The first part covers the whole sample (June 1980 to December 2021), the second and third the time period before and after December 2010. Portfolios are rebalanced annually at the end of June. The percentage values are percent per month, and the values in brackets are the corresponding t-test values. FF3F α stands for Fama-French three-factor model alpha.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$
			June 1980 to December 2021			
Mean	0.16%	0.01%	0.14%	0.09%	0.03%	0.10%
	(1.36)	(0.04)	(0.97)	(0.70)	(0.20)	(0.94)
Geomean	0.13%	−0.04%	0.09%	0.05%	−0.02%	0.07%
	(1.05)	(−0.30)	(0.62)	(0.39)	(−0.16)	(0.66)
FF3F α	0.16%	−0.01%	0.14%	0.09%	0.03%	0.13%
	(1.33)	(−0.10)	(0.94)	(0.67)	(0.19)	(1.13)
			June 1980 to December 2010			
Mean	0.28%	0.12%	0.26%	0.11%	−0.00%	0.07%
	(1.87)	(0.68)	(1.48)	(0.72)	(−0.00)	(0.52)
Geomean	0.24%	0.06%	0.20%	0.07%	−0.06%	0.04%
	(1.59)	(0.37)	(1.17)	(0.43)	(−0.33)	(0.27)
FF3F α	0.25%	0.06%	0.21%	0.08%	−0.04%	0.07%
	(1.63)	(0.33)	(1.21)	(0.50)	(−0.24)	(0.54)
			January 2011 to December 2021			
Mean	−0.13%	−0.09%	−0.17%	−0.02%	0.07%	0.15%
	(−0.72)	(−0.47)	(−0.76)	(−0.14)	(0.32)	(0.87)
Geomean	−0.15%	−0.11%	−0.21%	−0.04%	0.04%	0.13%
	(−0.83)	(−0.59)	(−0.91)	(−0.25)	(0.19)	(0.76)
FF3F α	−0.05%	−0.15%	−0.10%	−0.11%	−0.01%	0.16%
	(−0.25)	(−0.74)	(−0.41)	(−0.59)	(−0.04)	(0.89)

IV-C Delta of Profitability Level to Industry Mean

Table 9 shows the results of the FamaMac-Beth regressions for the difference between a firm's level of profitability and the industry mean, where the industries are defined as the Fama-French 49 industries. All coefficients are estimated to be at least slightly lower in the later sample period, but the main driver of the lower t-test values is the lower number of observations compared to the full or early sample period. Nevertheless, the difference between a firm's level of profitability and the industry mean is a statistically significant predictor for future returns according to the FamaMac-Beth regressions across all different sample periods and profitability measures.

The value-weighted portfolios shown in Figure 4 and Table 10 show a more ambiguous picture. While holding stocks with profitability above their industry's mean was a profitable strategy till around 2000, it yielded losses between 2000 and 2010. Since 2010, there has been a tendency towards being profitable again, especially the portfolios based on the measures scaled by book equity.

These results are robust across different industry definitions. Table A.5 in the appendix summarizes the results of the FamaMac-Beth regressions for the Fama-French 12, 17, 30, and 49 industries as well as the first level SIC (Tables A.6, A.7, A.8, A.9, A.10, A.11 in the appendix show the complete results). Furthermore, I use the similarity scores of Hoberg and Phillips (2010)'s TNIC to compute the difference between a firm's profitability and the weighted mean profitability of its competitors.

Together with the results on the level of profitability, these results suggest that some industries are more profitable than others, and it is advantageous to have more exposure to them rather than having an industry-balanced portfolio that chooses the most profitable companies of each industry.

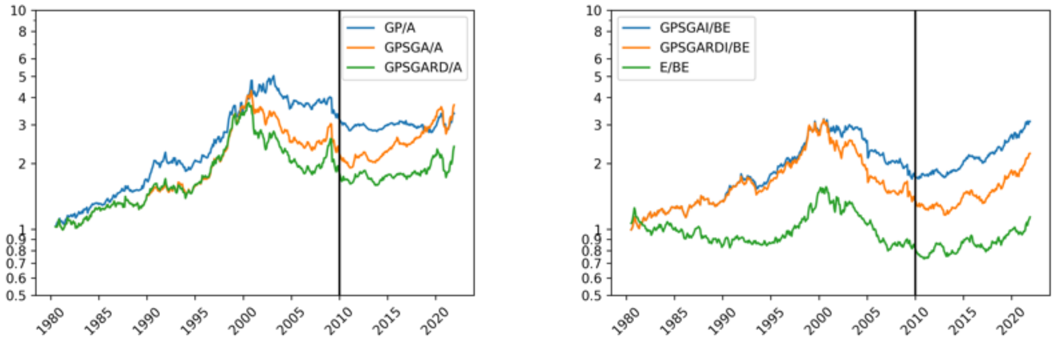
Table 9: Fama-MacBeth Regression Results - Delta to Fama-French 49 Industry Mean

This table shows the results of the Fama-MacBeth regressions for the delta between a firm's profitability and the industry mean, where the industries are defined as the Fama-French 49 industries. The first part covers the whole (June 1980 to December 2021), the second and third the time period before and after December 2010. All coefficients are in percent, and the values in brackets are the corresponding t-test values. For readability, the results for the control variables (value as Ln(BE/ME), size as Ln(ME), short-term reversal as past month return, and momentum as past twelve months return, skipping the most recent month) are not displayed here but can be found in Table A.9 in the appendix.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$
June 1980 to December 2021						
Delta to industry mean	1.26	2.51	1.94	0.55	0.44	0.23
	(10.34)	(9.44)	(7.77)	(9.26)	(8.00)	(4.97)
N (in million)	1.47	1.23	1.23	0.98	0.98	1.51
June 1980 to December 2010						
Delta to industry mean	1.38	2.59	2.03	0.68	0.54	0.22
	(9.82)	(8.44)	(6.56)	(8.54)	(7.18)	(3.63)
N (in million)	1.16	0.96	0.96	0.75	0.75	1.19
January 2011 to December 2021						
Delta to industry mean	0.95	2.25	1.69	0.28	0.25	0.25
	(3.92)	(4.23)	(4.45)	(4.28)	(4.40)	(4.45)
N (in million)	0.32	0.27	0.27	0.23	0.23	0.32

Figure 4: Portfolios Sorted on Delta to Mean of FF49 Industries

These figures show the developments of the value-weighted portfolios sorted on the delta between a firm's profitability and the industry mean (highest minus lowest quintile), where the industries are defined as the Fama-French 49 industries. The sample period covers June 1980 to December 2021. Subfigure (a) displays the profitability measures which are scaled by the book value of assets, and Subfigure (b) the ones that are scaled by the book value of equity. Portfolios are rebalanced annually at the end of June. Table 10 summarizes the corresponding mean, geometric mean, and Fama-French three-factor model alphas.



(a) Profitability measures scaled by assets

(b) Profitability measures scaled by book equity

Table 10: Portfolio Results - Delta to Mean of FF49 Industries

This table shows the results of the value-weighted portfolios sorted on the delta to the industry mean profitability (highest minus lowest quintile), where the industries are defined as the Frama-French 49 industries. The first part covers the whole (June 1980 to December 2021), the second and third the time period before and after December 2010. Portfolios are rebalanced annually at the end of June. The percentage values are percent per month, and the values in brackets are the corresponding t-test values. FF3F α stands for Fama-French three-factor model alpha.

	$\frac{\text{GP}}{\text{A}}$	$\frac{\text{GP-SGA}}{\text{A}}$	$\frac{\text{GP-SGA-RD}}{\text{A}}$	$\frac{\text{GP-SGA-I}}{\text{BE}}$	$\frac{\text{GP-SGA-RD-I}}{\text{BE}}$	$\frac{\text{E}}{\text{BE}}$
June 1980 to December 2021						
Mean	0.28%	0.30%	0.22%	0.26%	0.20%	0.06%
	(2.39)	(2.43)	(1.68)	(2.38)	(1.64)	(0.51)
Geomean	0.24%	0.26%	0.17%	0.23%	0.16%	0.03%
	(2.08)	(2.10)	(1.33)	(2.10)	(1.32)	(0.24)
FF3F α	0.36%	0.41%	0.32%	0.30%	0.25%	0.10%
	(3.10)	(3.26)	(2.49)	(2.78)	(2.08)	(0.89)
June 1980 to December 2010						
Mean	0.33%	0.23%	0.19%	0.19%	0.10%	−0.05%
	(2.24)	(1.50)	(1.19)	(1.40)	(0.68)	(−0.40)
Geomean	0.29%	0.18%	0.14%	0.16%	0.06%	−0.08%
	(1.95)	(1.21)	(0.87)	(1.14)	(0.40)	(−0.64)
FF3F α	0.43%	0.36%	0.34%	0.28%	0.22%	0.02%
	(2.91)	(2.37)	(2.17)	(2.07)	(1.48)	(0.13)
January 2011 to December 2021						
Mean	0.16%	0.58%	0.32%	0.47%	0.46%	0.34%
	(1.03)	(2.98)	(1.53)	(3.04)	(2.56)	(1.90)
Geomean	0.14%	0.56%	0.29%	0.46%	0.44%	0.32%
	(0.93)	(2.85)	(1.39)	(2.95)	(2.44)	(1.79)
FF3F α	0.14%	0.55%	0.21%	0.45%	0.44%	0.37%
	(0.91)	(2.70)	(0.98)	(2.73)	(2.31)	(1.95)

IV-D Profitability Decompositions

There are various possible decompositions of the profitability measures, and I first focus on one that sheds some light on the role of SG&A. While Novy-Marx (2013) claims that it has no incremental explanatory power beyond GP/A for future returns (Novy-Marx (2013), Table A2), Ball et al. (2015) show that *Compustat*'s SG&A variable cleaned by R&D expenses yields incremental explanatory power in Fama-MacBeth regressions (Ball et al. (2015), Table 6) and GPSGA/A has higher t-stats than GP/A when excluding micro-caps. Hence, the first decomposition is as follows:

$$\frac{\text{GPSGA}}{\text{A}} = \frac{\text{GP}}{\text{A}} - \frac{\text{SGA}}{\text{A}} = \frac{\text{REV}}{\text{A}} - \frac{\text{COGS}}{\text{A}} - \frac{\text{SGA}}{\text{A}}, \tag{5}$$

where REV stands for revenue and COGS for cost of goods sold. Table 11 shows the Fama-MacBeth regressions on the individual components separately and jointly in several combinations. While Table 11 includes the entire dataset, Table A.18 in the appendix excludes all stocks with a market capitalization below the 20th NYSE percentile, resulting in a drop of observations from 1.25 million to 0.44 million. The results of regressions (1) and (2) differ from the ones in Table 4 because only observations with no missing data on any variable are included (the number of observations is, therefore, the same for all 20 regressions). The t-stat of GP/A (10.83) is larger than the one of GPSGA/A (9.06) when including the micro-caps and smaller when excluding them (5.32 vs. 6.08), indicating that SG&A expenses are more important information regarding bigger companies. However, the difference is not very large and below statistical significance. Revenue-to-assets, also known as asset turnover, is also a clear predictor for future returns but significantly weaker than the other two profitability measures. Coefficients of regressions (4) and (5) are confusing at first glance since higher costs-to-assets are positive predictors, but this is simply because of an omitted variable bias, namely revenue-to-assets is missing.

It is crucial to be aware that the variables depend on each other, as the decomposition above shows. Note that also the last three variables on the right side of Equation 5 depend on each other because the cost of goods sold as well as SG&A scale with revenue. Therefore, the tables concerning the decompositions include the Variance Inflation Factors (VIFs) for each variable which are computed as 1/(1-R^2), where R^2 stems from a regression of the particular variable on all other dependent variables. The square root of the VIF measures how much the standard deviation of the coefficient estimate is larger due to the variable being correlated with the other right-hand side variables. Hence, the presence of moderate to strong multicollinearity may decrease t-stats significantly and result in not rejecting the null hypothesis despite it being true (also known as type 2 error). Furthermore, a high VIF indicates that the parameter estimates are very sensitive to small changes in the data.

Regression (6) shows that both GP/A and GPSGA/A contain significant information for future returns, which is not subsumed by the other measure. Regressions (7) and (8), on the other hand, show that the relevant information of REV/A and COGS/A is fully reflected in their difference (GP/A). There is no regression containing all three variables because there would be perfect collinearity among the independent variables (note that the regression would be solvable with the winsorized data but results in estimates entirely driven by the top and bottom 1% of the observations). Regression (9) is essential because it shows that SGA/A provides incremental information beyond GP/A, which remains so when also controlling for REV/A in regression (10). Unlike in regression (4), the sign indicates that higher costs indicate lower returns and vice versa, which seems more reasonable, and coefficients are also highly significant despite the high degree of collinearity (VIF between 5.97 and 9.51). This contradicts Novy-Marx (2013), where a t-stat of -2.05 is interpreted as insignificant. Applying high boundaries for t-stats naturally increases the risk for such type 2 errors of not falsifying wrong null hypotheses, which becomes even more likely in regressions with relatively high multicollinearity (which is the case according to the VIF scores).

Table 11: Fama-MacBeth Regressions - GPSGA/A Decomposition 1

This table shows the coefficients of the Fama-MacBeth regressions on GPSGA/A and its components according to the decomposition in Equation 5. All coefficients are in percent, the values in brackets the corresponding t-test values, and the values in square brackets the VIFs, which are a test for multicollinearity. For readability, the results for the control variables (value as Ln(BE/ME), size as Ln(ME), short-term reversal as past month return, and momentum as past twelve months return, skipping the most recent month) are not displayed here but can be found in Table A.17 in the appendix. Only observations not missing any variables are considered; hence n is 1.25 million for each regression. Table A.17 in the appendix displays the same table excluding micro-caps (and Table A.18 in the appendix also includes the results of the control variables).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
GPSGA/A	2.65					1.57				
	(9.06)					(5.01)				
	[1.58]					[1.88]				
GP/A		1.80				1.28	1.76	1.77	2.72	2.66
		(10.83)				(7.44)	(11.91)	(11.63)	(9.99)	(10.92)
		[1.77]				[2.78]	[3.29]	[2.10]	[8.13]	[9.51]
REV/A			0.33				0.01			0.04
			(6.16)				(0.25)			(0.70)
			[1.94]				[3.60]			[4.53]
COGS/A				0.20				0.02		
				(3.59)				(0.33)		
				[1.69]				[1.99]		
SGA/A					0.96				−1.40	−1.43
					(5.82)				(−5.06)	(−5.09)
					[1.72]				[5.97]	[6.20]
R^2	3.52%	3.51%	3.44%	3.34%	3.35%	3.81%	3.67%	3.67%	3.79%	3.96%
adj. R^2	3.32%	3.31%	3.23%	3.13%	3.15%	3.56%	3.43%	3.43%	3.55%	3.68%
	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
GPSGA/A	2.40	2.61	2.88	1.70	2.81	2.86				
	(8.80)	(9.21)	(9.76)	(5.45)	(10.53)	(10.13)				
	[1.63]	[1.59]	[1.61]	[1.92]	[1.86]	[1.65]				
GP/A										
REV/A	0.21			1.09	0.05		1.61	0.26	2.35	
	(4.38)			(7.24)	(0.92)		(11.53)	(4.38)	(10.24)	
	[2.14]			[32.12]	[4.47]		[16.05]	[3.87]	[67.06]	
COGS/A		0.15		−1.04		0.04	−1.57		−2.29	0.13
		(2.78)		(−6.54)		(0.86)	(−11.59)		(−10.70)	(2.31)
		[1.73]		[26.01]		[2.41]	[13.92]		[40.69]	[2.35]
SGA/A			1.20		1.10	1.15		0.52	−1.22	0.85
			(7.28)		(6.59)	(7.20)		(2.87)	(−4.55)	(5.19)
			[1.75]		[3.66]	[2.43]		[3.21]	[5.92]	[2.34]
R^2	3.78%	3.74%	3.80%	3.96%	3.97%	3.97%	3.66%	3.64%	3.93%	3.56%
adj. R^2	3.54%	3.50%	3.56%	3.68%	3.69%	3.69%	3.41%	3.40%	3.64%	3.32%

Regressions (11) to (16) can all be interpreted such that they confirm that GP/A provides incremental information beyond GPSGA/A. SGA/A is the direct link between GPSGA/A

and GP/A, hence the positive sign[6] and high t-stats for its coefficient in regressions (13) and (15). GP/A can also be constructed from REV/A and COGS/A in regression (14), which results in estimates for GPSGA/A which are close to regression (6). Because costs scale with revenue (COGS almost perfectly and SG&A also to a large degree), an increase by one percent of the top line (i.e., revenue) results in relative increases, which become smaller the further down one goes the income statement. Hence a change in REV/A or COGS/A has a more precise link to GP/A than to GPSGA/A and is therefore capturing the effect of GP/A in regressions (11) and (12). This is also the reason for the positive sign of COGS/A. Regressions (17) to (19) on the lowest level components (right side of Equation 5) can be analyzed analogously and confirm the previous statements. The R^2s and adjusted R^2s of the regressions confirm the conclusion from above: GP/A and GPSGA/A (together with the control variables) both explain a similar amount of the variance of future returns and explain more jointly.
The next decomposition investigates the role of margins and asset turnover (also known as the DuPont model or analysis):

$$\frac{\text{GPSGA}}{\text{A}} = \frac{\text{GPSGA}}{\text{REV}} \cdot \frac{\text{REV}}{\text{A}} = \left(\frac{\text{GP}}{\text{REV}} - \frac{\text{COGS}}{\text{REV}}\right) \cdot \frac{\text{REV}}{\text{A}}, \tag{6}$$

where REV stands for revenue, and when it is scaled by assets, it is known as asset turnover. The components scaled by revenue are margins (either profit margins or a cost margin in the case of COGS). Table 12 shows the Fama-MacBeth regressions of the components individually and in various combinations. Note that because the combinations containing only the first three components (GPSGA/A, GP/A, and REV/A) can be found in Table 11, they are not included in Table 12.

Regressions (1) to (3) show that margins have predictive power individually and that GPSGA/REV has much more than the gross margin (t-stat of 5.51 vs. 3.41). Regression (4) yields insignificant coefficients, but this specification suffers from an omitted variable bias, as the subsequent regressions indicate. Compared to regressions (1) to (4), regressions (5) to (8) control for asset turnover and show that both GPSGA/REV as well as the gross margin are relevant. Unlike results of regression (3), regression (7) and (8) indicate that SGA/REV itself is an insignificant factor (which is consistent with the fact that GPSGA/REV and GP/REV only have a minimally different t-stat in regressions (5) and (6)). According to regression (9), GPSGA/REV does not contain all the information in the gross margin, which is also indicated by the significant coefficient of SGA/REV in regression (10).[7]
Regressions (11) to (15) test the marginal information content of asset turnover and the margin components with respect to GPSGA/A, and regressions (16) to (20) do the same for GP/A. While asset turnover and gross margin do not provide additional information, which is not in GP/A (insignificant coefficients close to zero in regressions (11) to (15)), the

[6]Because one simply has to add SGA/A to GPSGA/A to get GP/A.
[7]GP/REV can be computed by subtracting SGA/REV from GPSGA/REV; therefore the negative coefficient of SGA/REV indicates that GP/REV has positive predictive power beyond that of GPSGA/REV.

Table 12: Fama-MacBeth Regressions - GPSGA/A Decomposition 2

This table shows the coefficients of the Fama-MacBeth regressions on GPSGA/A and its components according to the decomposition of Equation 6. All coefficients are in percent, the values in brackets the corresponding t-test values, and the values in square brackets the VIFs, which are a test for multicollinearity. For readability, the results for the control variables (value as Ln(BE/ME), size as Ln(ME), short-term reversal as past month return, and momentum as past twelve months return, skipping the most recent month) are not displayed here but can be found in Table A.19 in the appendix. Only observations not missing any variables are considered; hence n is 1.23 million for each regression. Table A.19 in the appendix displays the same table excluding micro-caps (and Table A.20 in the appendix also includes the results of the control variables).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
GPSGA/A										
GP/A										
REV/A					0.28	0.34	0.28	0.31	0.29	0.40
					(5.58)	(6.49)	(5.59)	(6.08)	(5.79)	(7.58)
					[2.10]	[2.00]	[2.12]	[2.13]	[2.12]	[2.13]
GPSGA/REV	0.61				0.62				0.51	1.49
	(5.51)				(5.66)				(3.48)	(9.94)
	[1.10]				[1.10]				[1.26]	[2.87]
GP/REV		0.24		0.15		0.39		0.33	0.22	
		(3.41)		(1.74)		(5.06)		(3.72)	(2.08)	
		[1.02]		[1.45]		[1.04]		[1.46]	[1.67]	
SGA/REV			−0.36	−0.31			−0.15	−0.12		1.33
			(−2.59)	(−1.88)			(−1.10)	(−0.78)		(6.56)
			[1.49]	[1.49]			[1.50]	[1.50]		[3.91]
R^2	3.43%	3.30%	3.45%	3.62%	3.72%	3.64%	3.73%	3.90%	3.88%	3.95%
adj. R^2	3.22%	3.09%	3.25%	3.37%	3.48%	3.39%	3.48%	3.61%	3.60%	3.66%
	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
GPSGA/A	2.49	2.37	2.39	2.85	2.80					
	(8.80)	(7.40)	(8.23)	(9.68)	(9.27)					
	[1.63]	[3.23]	[1.80]	[2.32]	[2.48]					
GP/A						1.76	1.50	1.82	1.77	1.91
						(11.72)	(9.36)	(10.95)	(11.94)	(11.84)
						[3.29]	[5.29]	[4.13]	[5.26]	[6.36]
REV/A	0.21	0.21	0.23	0.25	0.25	0.01	0.03	−0.00	−0.05	−0.08
	(4.47)	(4.45)	(4.68)	(5.08)	(5.17)	(0.30)	(0.66)	(−0.03)	(−1.08)	(−1.84)
	[2.14]	[2.22]	[2.15]	[2.24]	[2.24]	[3.60]	[4.55]	[3.81]	[4.42]	[4.77]
GPSGA/REV		−0.01					0.35			
		(−0.06)					(2.85)			
		[2.08]					[1.17]			
GP/REV			0.16		0.05			−0.09		−0.16
			(1.95)		(0.48)			(−0.83)		(−1.42)
			[1.53]		[1.56]			[1.29]		[1.76]
SGA/REV				0.48	0.50				−0.35	−0.38
				(3.39)	(2.78)				(−2.43)	(−2.49)
				[2.03]	[2.07]				[1.59]	[1.61]
R^2	3.82%	3.96%	3.96%	4.03%	4.18%	3.70%	3.95%	3.86%	3.96%	4.10%
adj. R^2	3.57%	3.67%	3.67%	3.74%	3.85%	3.46%	3.66%	3.57%	3.67%	3.77%

same cannot be said for GPSGA/A. While it does contain the GPSGA/REV margin, asset turnover, gross margin, and SG&A margin provide incremental information. Since GP/A contains at least most of the information in REV/A, the positive coefficients of asset turnover in regressions (11) to (15) also support the conclusion from the first decomposition that GPSGA/A and GP/A both provide distinct but highly relevant information for predicting future returns. Hence, the main finding of decomposition 2 is that margins are relevant, but the measures scaled by total assets contain at least most of their information.

The third decomposition is inspired by Ball et al. (2015) and investigates the role of the market value of equity (ME)-to-book assets (A) ratio:

$$\frac{\text{GPSGA}}{\text{A}} = \frac{\text{GPSGA}}{\text{ME}} \cdot \frac{\text{ME}}{\text{A}} = \left(\frac{\text{GP}}{\text{ME}} - \frac{\text{COGS}}{\text{ME}}\right) \cdot \frac{\text{ME}}{\text{A}}. \tag{7}$$

Ball et al. (2015) motivate this decomposition by the finding of Fama and French (1992) that ME/A is a priced factor. Fama and French (1992) interpret A/ME (note the switch of nominator and denominator) as market leverage (compared to A/BE, which is book leverage), but one should also keep in mind that it is how many assets (A) one gets control over for paying ME, i.e., it is a value measure. Results of the third decomposition can be found in Table 13, and the first regression shows that ME/A is an insignificant predictor, which is closely related to at least one control variable, as the VIF score indicates. A closer inspection of the results in Table A.21 in the appendix reveals that this is the case with Ln(BE/ME), i.e., the value measure. Fama and French (1992) point out that Ln(BE/ME) is equal to Ln(A/ME) minus Ln(A/BE) and that it is Ln(BE/ME), which is genuinely relevant for predicting future returns. But since this variable is already included in the control variables, also including ME/A is redundant, and results in Table 13 confirm this.
Despite this lack of insight from including ME/A, the decomposition is interesting due to its other components, especially GPSGA/ME and GP/ME. They offer an alternative measure of value (rather than profitability) that is unrelated (or at least not strongly) to the value control variable. Regressions (3) and (4) indicate that they predict future returns, even when controlling for ME/A (regressions (6) and (7)). According to regression (4), SGA/ME is a strong predictor, but this is mainly due to its strong relationship to GP/ME (see VIF at regressions (5), (16), and (21)), i.e., an omitted variable bias. Regression (10) indicates that also when using GP and GPSGA as value measures by scaling them with market equity, they both contain distinct relevant information.

Regressions (11) to (21) investigate whether these value measures provide incremental information beyond the profitability measures GP/A and GPSGA/A. One key observation is the fact that the predictive power of the profitability measures remains strong across all regressions. GP/ME adds significant information to both GPSGA/A and GP/A (regressions (14) and (19)); however, GPSGA/ME only to GP/A (regression (18) and not to GPSGA/A (regression (13)). Results excluding the micro-caps in Table A.22 in the appendix confirm most of the findings described above. However, as in the previous decompositions, excluding the micro-caps strengthens the predictive power of GPSGA compared to gross profits; hence, also of GPSGA/ME compared to GP/ME. E.g., GP/ME is no useful complement anymore to GP/A, while GPSGA/ME has a t-stat of 1.87 despite a VIF score of 3.07.
The decompositions do not consider R&D expenses as they are not the main focus of this paper. Nevertheless, I want to point out that they are an insignificant predictor when scaled by revenue, moderately significant when scaled by total assets (t-stat of 2.14 including micro-caps and 1.89 when excluding them), but highly significant when scaled by market equity (t-stat of 5.42 including micro-caps and 3.01 when excluding them). Note that this is even more impressive when considering that I only have 0.53 million observations on R&D expenses when including micro-caps and just 0.18 million when excluding them. Surprisingly, the R&D expenses scaled by total assets also have a significant coefficient when controlling for asset turnover. However, scaling by market equity leads to a t-stat twice the size in

Table 13: Fama-MacBeth Regressions - GPSGA/A Decomposition 3

This table shows the coefficients of the Fama-MacBeth regressions on GPSGA/A and its components according to the decomposition of Equation 7. All coefficients are in percent, the values in brackets the corresponding t-test values, and the values in square brackets the variance inflation factors, which are a test for multicollinearity. For readability, the results for the control variables (value as Ln(BE/ME), size as Ln(ME), short-term reversal as past month return, and momentum as past twelve months return, skipping the most recent month) are not displayed here but can be found in Table A.21 in the appendix. Only observations not missing any variables are considered; hence n is 1.25 million for each regression. Table A.21 in the appendix displays the same table excluding micro-caps (and Table A.22 in the appendix also includes the results of the control variables).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(8)	(10)	(11)
GPSGA/A											
GP/A											
ME/A	−0.02					−0.01	−0.00	−0.01	−0.00	−0.00	0.00
	(−0.89)					(−0.41)	(−0.08)	(−0.46)	(−0.08)	(−0.02)	(0.01)
	[3.08]					[3.11]	[3.10]	[3.13]	[3.14]	[3.14]	[3.14]
GPSGA/ME		0.89				0.87				0.44	0.84
		(5.39)				(5.35)				(2.37)	(5.12)
		[1.31]				[1.32]				[1.59]	[1.33]
GP/ME			0.45		0.61		0.44		0.60	0.38	
			(7.50)		(4.44)		(7.45)		(4.39)	(5.75)	
			[1.32]		[9.95]		[1.32]		[9.97]	[1.66]	
SGA/ME				0.41	−0.18			0.40	−0.18		0.40
				(6.17)	(−1.23)			(6.11)	(−1.19)		(6.14)
				[1.22]	[8.90]			[1.23]	[8.95]		[1.25]
R^2	3.28%	3.40%	3.39%	3.38%	3.67%	3.58%	3.57%	3.57%	3.85%	3.86%	3.87%
adj. R^2	3.08%	3.19%	3.18%	3.18%	3.42%	3.34%	3.33%	3.33%	3.56%	3.58%	3.59%
	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	
GPSGA/A	2.68	2.79	2.40	2.81	3.09						
	(9.26)	(8.94)	(8.03)	(9.77)	(9.94)						
	[1.60]	[2.23]	[1.61]	[1.61]	[2.10]						
GP/A						1.82	1.72	1.61	1.66	1.60	
						(10.93)	(10.38)	(8.59)	(9.03)	(8.62)	
						[1.77]	[2.58]	[2.65]	[3.53]	[3.87]	
ME/A	0.00	0.00	0.01	0.01	0.01	−0.03	−0.03	−0.03	−0.03	−0.03	
	(0.01)	(0.04)	(0.54)	(0.56)	(0.55)	(−1.24)	(−1.00)	(−0.98)	(−1.18)	(−1.02)	
	[3.14]	[3.14]	[3.16]	[3.16]	[3.16]	[3.09]	[3.18]	[3.11]	[3.16]	[3.17]	
GPSGA/ME		−0.10					0.45				
		(−0.60)					(2.83)				
		[1.83]					[1.43]				
GP/ME			0.33		−0.23			0.14		0.25	
			(5.54)		(−1.62)			(2.21)		(1.87)	
			[1.39]		[12.99]			[1.98]		[10.95]	
SGA/ME				0.44	0.65				0.12	−0.11	
				(6.73)	(4.34)				(1.59)	(−0.75)	
				[1.24]	[11.64]				[1.83]	[8.96]	
R^2	3.70%	3.92%	3.98%	3.99%	4.22%	3.70%	3.96%	3.95%	3.97%	4.21%	
adj. R^2	3.46%	3.64%	3.70%	3.71%	3.90%	3.46%	3.68%	3.67%	3.69%	3.89%	

all constellations, indicating that one should “buy R&D expenses as cheaply as possible” rather than measuring them in relation to some firm size variable like sales or total assets. The main conclusions of these three decompositions are the following: First, GP/A, and GPSGA/A both have distinct predictive power, which the other does not contain. Second, GP/A is a stronger predictor among small stocks, while GPSGA/A is stronger among larger stocks. Third, margins are predictors, but their information is more or less fully included in their product with asset turnover, i.e., in GP/A and GPSGA/A. Fourth, GP and GPSGA scaled by market equity is a value measure that yields incremental information beyond the profitability measures of GP/A and GPSGA/A, where again, GP/ME is stronger among smaller stocks while GPSGA/ME is stronger among larger stocks. Note that the choice of investigating GPSGA/ME was driven by the decomposition; however, one could argue for using GPSGAI/ME, i.e., further deducting interest payments, as a measure of value since it is a cleaner measure of the expected cash flow available to equity holders. In undisclosed results, I confirm that GPSGAI/ME has even slightly stronger predictive power[8] despite the fewer observations due to missing information on interest payments. On the other hand, taking interest payments into account may result in a less favorable measure because it could be biased towards older firms (which paid back most of their debt) and larger firms with more tangible assets (because of more favorable financing conditions).

V Moderating Variables

The two most widely studied moderating variables stem from the Fama-French tree-factor model: market capitalization (size factor) and the book-to-market ratio (value factor). Investigating anomalies among stocks of different sizes is crucial since many anomalies can only be found among small and, therefore, hard-to-arbitrage stocks.[9]

Stocks with a high book-to-market ratio are perceived to be value stocks, while the ones with low book-to-market ratios tend to be growth stocks. However, the book-to-market ratio has a few problems that have become more severe in recent years. On the one hand, low or even negative book equity can be caused by losses from operations or declines in asset values, i.e., troubled companies. On the other hand, paying out large amounts through dividends or share repurchases also lowers the book value of equity. But this payout policy is generally pursued by companies with large free cash flows and low investment needs, i.e., strong, healthy firms. Note that both cases would be perceived as growth stocks from their book-to-market ratio despite them maybe not having much growth.[10] Furthermore, companies with a lot of intangible assets created by themselves rather than bought (bought intangible assets may be held as assets on the balance sheet) have lower book equity compared to companies with the same amount of bought tangible assets.

Hence, the importance and validity of book-to-market ratio as a measure of value is falling victim to the progressing digitalization. Companies like *Apple*, *Microsoft*, *Alphabet*, and *Amazon* (the largest four of the S&P 500) can essentially never be value stocks due to their low book equity (caused by their share repurchases, dividends, and missing intangible assets on their balance sheet), almost independently of their price. E.g., even if *Apple* had been trading at a market cap of only USD 170 billion in September 2022, it would still not have belonged to the upper half of book-to-market stocks (i.e., the value stocks) with a book-to-market ratio of 0.3 (due to its book equity of just about USD 50 billion). Be aware that *Apple*'s net income in the 24 months before September 2022 was USD 195 billion, operating cash flow was USD 226 billion the company bought back shares worth USD 175 billion; i.e.,

[8] E.g., in regression (2) in Table 13, the t-stat of GPSGAI/ME would be 5.58 instead of 5.39.

[9] Small stocks are usually hard to arbitrage due to limited liquidity and big price impacts.

[10] Furthermore, negative book equity is often treated as missing (as it is in this paper) because the mechanics change: increases in market capitalization increase the book-to-market ratio of firms with positive book equity but decrease the book-to-market ratio of the ones with negative book equity (i.e., make it less negative).

it would have been the best value you could find in the market at USD 170 billion.

Now, one may argue that a low book-to-market ratio can still be a reasonable measure for growth stocks. Consider the example of *Philip Morris International* as anecdotal evidence against it: The company has carried negative book equity since 2013 and would therefore be perceived to be a growth company. However, the company (which produces tobacco products, a stagnating business for many years) had sales of USD 80 billion in 2013 and USD 81 billion in 2022 (unadjusted for inflation) with little variation in between. The market capitalization was roughly USD 150 billion in these ten years, annual operating cash flow ranged from USD 7.5 to twelve billion, and a total of USD 68 billion has been returned to shareholders through dividends and another USD eleven billion through share repurchases. Obviously, this fits the description of a value rather than a growth stock.

As the results of Subsection V-B show, strategies based on the book-to-market ratio have not yielded any positive returns on average. Together with the above-described issues, this asks for an alternative measure because, as Novy-Marx (2013) shows, value is an excellent complement to profitability strategies. Fortunately, two such measures were already discussed in Subsection IV-D: GP/ME and GPSGA/ME. Scaling profits with the market value of equity yields a value measure because it measures how much one has to pay per annual profit. Hence, it is an alternative to the book-to-market ratio, which measures how much one has to pay per book value of equity. The results of regression (18) in Table 13 and regression (18) in Table A.22 in the appendix show that GPSGA/ME contains information that complements GP/A. The results are not as clear for GPSGA/A (regression (13) in Table 13 and regression (13) in Table A.22 in the appendix), but this may be due to the relatively strong multicollinearity between ME/A, GPSGA/ME, and the control variable Ln(BE/ME). Additionally to the double sorts based on size (Subsection V-A) and value as BE/ME (Subsection V-B), I also double sort based on value as GPSGA/ME (Subsection V-C). The choice for GPSGA/ME compared to GP/ME was driven by the stronger results in regressions (12), (13), (18), and (19) in Table A.22 in the appendix. In Subsection V-D, I investigate the relationship between industry concentration and the profitability anomaly before I conclude in Section VI.

V-A Size

Tables 14, 15, and 16 show the double sorts for size and GP/A, GPSGA/A, and GPSGAI/BE (first sort on size, then profitability) over the whole sample period. The monthly returns of the 25 portfolios in the top left quadrant are in excess of the one-month risk-free rate. Profitability strategies yield consistently positive and statistically significant excess returns and three-factor model alphas across all sizes. In undisclosed results, I confirm that the Fama-French measure, which also deducts R&D expenses, yields slightly less profitable strategies than those shown in Table 16. Over the whole sample period, the return spread between more and less profitable stocks is larger among smaller stocks. Surprisingly, Tables A.23, A.24, and A.25 in the appendix show that this relationship has become U-shaped in

recent years, and the spread has become largest among big stocks.
The decreasing HML factor loadings of the GP/A profitability strategies with size are also documented in Novy-Marx (2013) and can also be found in the sample period after 2010. Strategies based on GPSGA/A and, to a smaller extent, also the ones based on GPSGAI/BE have a similar pattern. This indicates that profitability strategies choose companies with lower book-to-market ratios the bigger the companies are. Profitability strategies based on GP/A have no clear pattern concerning the SMB and the MKT factor, while the ones based on GPSGA/A and (to a slightly smaller extent) GPSGAI/BE load negatively, indicating that they tend to buy larger stocks and develop against the market direction.
The size strategies yield no statistically significant returns, but unlike in Novy-Marx (2013), they tend to be negative, which is most likely due to the later sample period and the fact that the size premium has become negative in recent years (e.g., the Fama-French SMB factor was -0.03% between the beginning of 2011 and the end of 2021). Naturally, the size strategies load significantly positively on the size factor. The increasing loadings on the value factor (HML) with profitability are consistent with the findings of Novy-Marx (2013), meaning that with increasing profitability, the book-to-market ratio becomes relatively higher in small companies compared to big companies. Combining profitability and size strategies does not yield any benefits: neither does it generate higher returns than the pure profitability strategies, nor does it decrease volatility (otherwise, the same returns would need to have higher t-values).
Note that returns are generally higher for the measures scaled by total assets since they also cover firms with negative book values; hence, a direct comparison between the returns of the 25 portfolios in Table 14 and Table 16 is difficult. But interestingly, there is an asymmetry between the two measures: For GP/A, the low portfolios are, on average, 0.14% p.m. below the mid (3) portfolio, and the high portfolios are 0.36% p.m. above it (GPSGA/A: -0.18% p.m. and +0.38% p.m.). For GPSGAI/BE, the low portfolios are on average 0.25% p.m. below the mind (3) portfolios, and the high portfolios are 0.20% p.m. above them. Hence, it seems that (high) GP/A or GPSGA/A particularly well detects undervalued firms while (low) GPSGAI/BE finds overvalued firms. This pattern is consistent across time and can also be found in the sample period after 2010 (see Tables A.23, A.24, and A.25).

Table 14: Double Sort on Size and GP/A

This table shows the value-weighted excess returns of the portfolios first sorted into quintiles based on size (market capitalization), then within the quintiles based on GP/A. Portfolios are rebalanced annually at the end of June and cover the whole sample from June 1980 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The size strategies are the small-minus-big portfolios, profitability strategies are the high-minus-low profitability (GP/A) portfolios, and the combined strategy is the small-high portfolio minus the big-low portfolio (column description of the profitability strategies applies). Table A.23 shows results for the subperiods before and after December 2010.

	GP/A					Profitability strategies					
Size	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Small	0.25%	0.29%	0.42%	0.65%	1.00%	0.75%	0.70%	0.76%	−0.04	−0.15	0.14
						(5.31)	(4.92)	(5.45)	(−1.13)	(−3.14)	(3.09)
2	0.24%	0.31%	0.40%	0.49%	0.72%	0.48%	0.45%	0.46%	0.02	0.02	−0.01
						(4.18)	(3.91)	(3.97)	(0.70)	(0.50)	(−0.15)
3	0.29%	0.39%	0.42%	0.43%	0.67%	0.38%	0.35%	0.32%	0.10	0.09	−0.08
						(3.43)	(3.16)	(2.88)	(3.96)	(2.36)	(−2.31)
4	0.38%	0.38%	0.50%	0.62%	0.85%	0.47%	0.43%	0.41%	0.12	0.13	−0.21
						(3.71)	(3.42)	(3.40)	(4.51)	(3.08)	(−5.19)
Big	0.46%	0.41%	0.58%	0.67%	0.87%	0.41%	0.36%	0.50%	−0.01	−0.15	−0.39
						(2.92)	(2.57)	(3.80)	(−0.18)	(−3.28)	(−8.91)
	Size strategies					Combined strategy					
Mean	−0.21%	−0.11%	−0.15%	−0.02%	0.13%	0.54%	0.46%	0.39%	0.06	0.99	0.03
	(−0.91)	(−0.75)	(−1.01)	(−0.12)	(0.68)	(2.94)	(2.50)	(3.05)	(2.15)	(22.52)	(0.64)
Geomean	−0.34%	−0.17%	−0.21%	−0.09%	0.04%						
	(−1.49)	(−1.13)	(−1.38)	(−0.53)	(0.21)						
α	−0.70%	−0.48%	−0.57%	−0.45%	−0.44%						
	(−4.19)	(−4.36)	(−5.62)	(−4.23)	(−3.63)						
β_{MKT}	0.10	−0.09	−0.06	−0.12	0.07						
	(2.70)	(−3.75)	(−2.60)	(−4.71)	(2.61)						
β_{SMB}	1.14	0.85	0.86	0.97	1.14						
	(20.04)	(22.64)	(24.55)	(26.73)	(27.54)						
β_{HML}	−0.12	0.14	0.32	0.50	0.41						
	(−2.22)	(3.82)	(9.26)	(14.12)	(10.29)						

Table 15: Double Sort on Size and GPSGA/A

This table shows the value-weighted excess returns of the portfolios first sorted into quintiles based on size (market capitalization), then within the quintiles based on GPSGA/A. Portfolios are rebalanced annually at the end of June and cover the whole sample from June 1980 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The size strategies are the small-minus-big portfolios, profitability strategies are the high-minus-low profitability (GPSGA/A) portfolios, and the combined strategy is the small-high portfolio minus the big-low portfolio (column description of the profitability strategies applies). Table A.24 shows results for the subperiods before and after December 2010.

	GPSGA/A					Profitability strategies					
Size	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Small	0.06%	0.34%	0.37%	0.52%	0.83%	0.77%	0.71%	0.91%	−0.16	−0.30	0.03
						(5.27)	(4.76)	(6.63)	(−4.94)	(−6.47)	(0.69)
2	0.07%	0.24%	0.20%	0.45%	0.64%	0.57%	0.55%	0.64%	−0.06	−0.06	−0.06
						(5.94)	(5.72)	(6.54)	(−2.73)	(−1.84)	(−1.87)
3	0.17%	0.38%	0.34%	0.31%	0.62%	0.45%	0.43%	0.50%	−0.03	−0.03	−0.15
						(4.37)	(4.13)	(4.86)	(−1.17)	(−0.89)	(−4.35)
4	0.21%	0.43%	0.33%	0.52%	0.70%	0.49%	0.46%	0.59%	−0.06	−0.11	−0.21
						(4.32)	(4.04)	(5.26)	(−2.39)	(−2.75)	(−5.72)
Big	0.39%	0.54%	0.44%	0.57%	0.79%	0.40%	0.34%	0.59%	−0.07	−0.29	−0.57
						(2.62)	(2.25)	(4.47)	(−2.34)	(−6.48)	(−13.08)
	Size strategies					Combined strategy					
Mean	−0.33%	−0.20%	−0.07%	−0.05%	0.04%	0.43%	0.39%	0.48%	−0.16	0.69	0.04
	(−1.67)	(−1.43)	(−0.53)	(−0.35)	(0.22)	(3.16)	(2.81)	(4.50)	(−6.61)	(18.88)	(1.06)
Geomean	−0.43%	−0.26%	−0.12%	−0.09%	−0.04%						
	(−2.20)	(−1.79)	(−0.88)	(−0.69)	(−0.21)						
α	−0.75%	−0.60%	−0.50%	−0.46%	−0.43%						
	(−4.89)	(−5.96)	(−4.85)	(−4.65)	(−4.09)						
β_{MKT}	−0.00	−0.06	−0.05	−0.05	−0.09						
	(−0.09)	(−2.71)	(−2.02)	(−2.20)	(−3.59)						
β_{SMB}	0.99	0.82	0.77	0.75	0.98						
	(18.89)	(23.53)	(21.80)	(21.97)	(27.06)						
β_{HML}	−0.00	0.22	0.31	0.28	0.60						
	(−0.00)	(6.42)	(9.02)	(8.43)	(17.10)						

Table 16: Double Sort on Size and GPSGAI/BE

This table shows the value-weighted excess returns of the portfolios first sorted into quintiles based on size (market capitalization), then within the quintiles based on GPSGAI/BE. Portfolios are rebalanced annually at the end of June and cover the whole sample from June 1980 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The size strategies are the small-minus-big portfolios, profitability strategies are the high-minus-low profitability (GPSGAI/BE) portfolios, and the combined strategy is the small-high portfolio minus the big-low portfolio (column description of the profitability strategies applies). Table A.25 shows results for the subperiods before and after December 2010.

	GPSGAI/BE					Profitability strategies					
Size	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Small	−0.05%	0.32%	0.40%	0.50%	0.63%	0.68%	0.64%	0.77%	−0.10	−0.24	0.07
						(5.07)	(4.65)	(5.93)	(−3.45)	(−5.37)	(1.71)
2	0.06%	0.20%	0.26%	0.39%	0.50%	0.43%	0.40%	0.43%	−0.00	−0.04	0.02
						(4.12)	(3.87)	(4.04)	(−0.04)	(−0.97)	(0.63)
3	0.07%	0.20%	0.31%	0.39%	0.43%	0.36%	0.33%	0.35%	0.03	−0.06	−0.05
						(3.49)	(3.26)	(3.36)	(1.27)	(−1.57)	(−1.32)
4	0.11%	0.29%	0.32%	0.39%	0.52%	0.41%	0.39%	0.42%	0.01	−0.05	−0.05
						(3.90)	(3.65)	(3.89)	(0.46)	(−1.26)	(−1.49)
Big	0.28%	0.29%	0.40%	0.60%	0.63%	0.35%	0.30%	0.51%	−0.11	−0.27	−0.27
						(2.49)	(2.14)	(3.80)	(−3.57)	(−5.87)	(−6.01)
	Size strategies					Combined strategy					
Mean	−0.34%	0.03%	−0.01%	−0.10%	−0.01%	0.34%	0.29%	0.23%	0.02	0.66	0.19
	(−1.78)	(0.22)	(−0.05)	(−0.73)	(−0.04)	(2.36)	(2.00)	(1.94)	(0.61)	(16.18)	(4.68)
Geomean	−0.43%	−0.02%	−0.05%	−0.15%	−0.08%						
	(−2.26)	(−0.11)	(−0.39)	(−1.07)	(−0.46)						
α	−0.86%	−0.38%	−0.40%	−0.60%	−0.60%						
	(−5.80)	(−3.53)	(−3.65)	(−6.02)	(−5.49)						
β_{MKT}	0.12	−0.02	−0.06	0.04	0.13						
	(3.59)	(−0.87)	(−2.47)	(1.91)	(5.15)						
β_{SMB}	0.90	0.67	0.63	0.72	0.93						
	(17.65)	(18.11)	(16.71)	(21.00)	(24.70)						
β_{HML}	0.11	0.19	0.29	0.38	0.45						
	(2.13)	(5.39)	(7.86)	(11.49)	(12.19)						

V-B Value as BE/ME

Tables 17, 18, and 19 show the double sorts for value as BE/ME and GP/A, GPSGA/A, and GPSGAI/BE (first sort on BE/ME, then profitability) over the whole sample period. The returns of the 25 portfolios in the top left quadrant are in excess of the one-month risk-free rate. Over the whole sample, profitability strategies tend to yield higher returns in lower BE/ME stocks (growth stocks), especially the measures scaled by total assets. However, in the later sample period, displayed in the second part of Tables A.26, A.27, and A.28, this tendency is not so clear anymore. The profitability strategies have no clear pattern concerning factor loadings, except for those based on GPSGA/A and GPSGAI/BE, which load negatively on the SMB factor, i.e., tend to buy big stocks (this is also consistently so across time).

Over the whole sample, the value premium is higher among stocks with low profitability. However, this pattern is not clear after 2010, and it is negative on average across this subperiod, which is consistent with the Fama-French HML factor having a -0.22% p.m. return over this time period. Naturally, the value strategies load strongly positively on the HML factor. But they also load positively on the SMB factor, indicating that the highest book-to-market stocks are smaller stocks and/or the lowest are big stocks. Furthermore, the value strategies load positively on the market factor, i.e., are against the common belief, not contrarian strategies. All these factor loading patterns regarding the value strategies can be found both before and after 2010.

Combining value and profitability strategies works well over the whole sample and yields higher profits than the average of them individually. However, the poor performance of value according to the book-to-market ratio in the sample period after 2010 yields combined strategies that perform worse than the average profitability strategy. Interestingly, the combination of BE/ME and profitability measures scaled by total assets yields a strategy neutral to the HML factor, while the combination with GPSGAI/BE yields one that heavily loads on the HML factor. Another difference is that the combination with measures scaled by total assets loads negatively on the SMB factor, but not the combination with GPSGAI/BE. Again, these factor loading patterns are consistent across time.

The asymmetry between the profitability measures computed with total assets and GPSGAI/BE described in the previous subsection is also present among the different BE/ME sorts: The low GP/A portfolios are on average 0.06% p.m. below the mind (3) portfolio and the high portfolios are 0.26% p.m. above it (GPSGA/A: -0.20% p.m. and +0.21% p.m.), while the low GPSGAI/BE portfolios are on average 0.31% p.m. below and high portfolios 0.19% p.m. above the mind (3) portfolios. Hence, high GP/A excels at detecting undervalued firms, while low GPSGAI/BE finds overvalued firms. This pattern is consistent across time and can also be found in the sample period after 2010 (see Tables A.26, A.27, and A.28).

Table 17: Double Sort on BE/ME and GP/A

This table shows the value-weighted excess returns for the portfolios first sorted into quintiles based on BE/ME, then within the quintiles based on GP/A. Portfolios are rebalanced annually at the end of June and cover the whole sample from June 1980 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The value strategies are the high-minus-low BE/ME portfolios, profitability strategies are the high-minus-low profitability (GP/A) portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies). Table A.26 shows results for the subperiods before and after December 2010.

	GP/A					Profitability strategies					
BE/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	0.14%	0.19%	0.25%	0.56%	0.65%	0.51%	0.46%	0.66%	−0.19	−0.17	0.03
						(3.35)	(2.98)	(4.49)	(−5.59)	(−3.45)	(0.69)
2	0.09%	0.24%	0.36%	0.52%	0.41%	0.32%	0.27%	0.38%	−0.09	0.07	0.01
						(2.32)	(1.98)	(2.73)	(−2.93)	(1.44)	(0.16)
3	0.01%	0.11%	0.28%	0.45%	0.48%	0.47%	0.42%	0.34%	0.09	0.29	0.18
						(3.28)	(2.91)	(2.43)	(2.95)	(6.01)	(3.92)
4	0.24%	0.26%	0.24%	0.32%	0.58%	0.34%	0.28%	0.17%	0.15	0.36	0.12
						(2.17)	(1.78)	(1.13)	(4.57)	(7.16)	(2.37)
High	0.62%	0.51%	0.29%	0.48%	0.71%	0.09%	−0.00%	0.20%	0.01	−0.10	−0.54
						(0.48)	(−0.01)	(1.07)	(0.31)	(−1.65)	(−8.79)
	Value strategies					Combined strategy					
Mean	0.48%	0.33%	0.04%	−0.07%	0.06%	0.57%	0.51%	0.66%	−0.12	−0.11	0.03
	(2.61)	(2.04)	(0.27)	(−0.50)	(0.45)	(3.55)	(3.17)	(4.09)	(−3.10)	(−1.94)	(0.64)
Geomean	0.40%	0.26%	−0.02%	−0.12%	0.02%						
	(2.14)	(1.65)	(−0.11)	(−0.85)	(0.12)						
α	0.14%	−0.11%	−0.44%	−0.50%	−0.33%						
	(0.85)	(−0.78)	(−3.31)	(−3.57)	(−2.41)						
β_{MKT}	−0.12	0.00	0.05	0.06	0.08						
	(−3.28)	(0.14)	(1.76)	(1.91)	(2.51)						
β_{SMB}	−0.00	0.05	0.11	0.28	0.07						
	(−0.01)	(0.93)	(2.40)	(5.80)	(1.50)						
β_{HML}	0.57	0.57	0.57	0.17	−0.01						
	(10.34)	(11.92)	(12.87)	(3.59)	(−0.13)						

Table 18: Double Sort on BE/ME and GPSGA/A

This table shows the value-weighted excess returns for the portfolios first sorted into quintiles based on BE/ME, then within the quintiles based on GPSGA/A. Portfolios are rebalanced annually at the end of June and cover the whole sample from June 1980 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The value strategies are the high-minus-low BE/ME portfolios, profitability strategies are the high-minus-low profitability (GPSGA/A) portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies). Table A.27 shows results for the subperiods before and after December 2010.

	GPSGA/A					Profitability strategies					
BE/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	0.11%	0.15%	0.49%	0.31%	0.58%	0.46%	0.40%	0.54%	−0.05	−0.37	−0.05
						(2.98)	(2.61)	(3.63)	(−1.41)	(−7.17)	(−0.95)
2	−0.15%	0.26%	0.19%	0.43%	0.53%	0.68%	0.63%	0.74%	−0.01	−0.15	−0.14
						(4.59)	(4.27)	(4.91)	(−0.43)	(−2.94)	(−2.85)
3	−0.03%	0.18%	0.22%	0.34%	0.41%	0.44%	0.39%	0.48%	−0.01	−0.08	−0.07
						(2.90)	(2.55)	(3.07)	(−0.42)	(−1.55)	(−1.44)
4	0.05%	0.34%	0.10%	0.21%	0.34%	0.28%	0.20%	0.21%	0.10	−0.19	0.07
						(1.60)	(1.14)	(1.19)	(2.56)	(−3.16)	(1.21)
High	0.45%	0.50%	0.45%	0.44%	0.64%	0.18%	0.07%	0.44%	−0.17	−0.43	−0.49
						(0.87)	(0.32)	(2.27)	(−3.75)	(−6.47)	(−7.48)
	Value strategies					Combined strategy					
Mean	0.34%	0.35%	−0.04%	0.13%	0.06%	0.52%	0.46%	0.54%	0.00	−0.28	0.06
	(1.76)	(1.91)	(−0.25)	(0.85)	(0.44)	(3.30)	(2.94)	(3.44)	(0.04)	(−5.28)	(1.15)
Geomean	0.25%	0.27%	−0.11%	0.07%	0.01%						
	(1.34)	(1.54)	(−0.66)	(0.47)	(0.09)						
α	−0.23%	−0.30%	−0.58%	−0.40%	−0.33%						
	(−1.27)	(−1.80)	(−3.79)	(−2.65)	(−2.38)						
β_{MKT}	0.17	0.26	0.13	0.19	0.05						
	(4.17)	(6.90)	(3.73)	(5.40)	(1.68)						
β_{SMB}	0.15	0.36	0.29	0.05	0.09						
	(2.47)	(6.28)	(5.53)	(0.92)	(1.86)						
β_{HML}	0.54	0.51	0.45	0.35	0.10						
	(8.95)	(9.18)	(8.90)	(6.98)	(2.18)						

Table 19: Double Sort on BE/ME and GPSGAI/BE

This table shows the value-weighted excess returns for the portfolios first sorted into quintiles based on BE/ME, then within the quintiles based on GPSGAI/BE. Portfolios are rebalanced annually at the end of June and cover the whole sample from June 1980 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The value strategies are the high-minus-low BE/ME portfolios, profitability strategies are the high-minus-low profitability (GPSGAI/BE) portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies). Table A.28 shows results for the subperiods before and after December 2010.

	GPSGAI/BE					Profitability strategies					
BE/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	−0.08%	0.18%	0.37%	0.47%	0.44%	0.52%	0.46%	0.52%	−0.03	−0.27	0.28
						(3.18)	(2.79)	(3.33)	(−0.89)	(−5.10)	(5.50)
2	−0.10%	0.18%	0.31%	0.22%	0.52%	0.63%	0.58%	0.66%	−0.01	−0.19	−0.02
						(4.46)	(4.12)	(4.67)	(−0.40)	(−3.95)	(−0.50)
3	0.12%	0.21%	0.16%	0.35%	0.50%	0.38%	0.33%	0.35%	0.04	−0.11	0.08
						(2.73)	(2.41)	(2.49)	(1.10)	(−2.33)	(1.69)
4	−0.08%	0.30%	0.18%	0.24%	0.38%	0.45%	0.38%	0.43%	0.06	−0.23	0.03
						(2.71)	(2.28)	(2.53)	(1.54)	(−3.94)	(0.55)
High	0.28%	0.30%	0.65%	0.33%	0.76%	0.48%	0.27%	0.49%	−0.01	−0.12	0.06
						(1.72)	(0.88)	(1.72)	(−0.22)	(−1.23)	(0.68)
	Value strategies					Combined strategy					
Mean	0.36%	0.11%	0.28%	−0.14%	0.32%	0.84%	0.69%	0.55%	0.15	0.16	0.82
	(1.41)	(0.48)	(1.65)	(−0.82)	(1.18)	(3.20)	(2.87)	(2.26)	(2.73)	(1.95)	(10.16)
Geomean	0.21%	−0.02%	0.21%	−0.22%	0.16%						
	(0.87)	(−0.09)	(1.23)	(−1.25)	(0.66)						
α	−0.26%	−0.47%	−0.24%	−0.65%	−0.29%						
	(−1.11)	(−2.09)	(−1.52)	(−4.01)	(−1.13)						
β_{MKT}	0.17	0.11	0.10	0.07	0.19						
	(3.11)	(2.21)	(2.74)	(1.80)	(3.17)						
β_{SMB}	0.29	0.38	0.34	0.36	0.44						
	(3.49)	(5.02)	(6.19)	(6.50)	(4.94)						
β_{HML}	0.75	0.71	0.50	0.51	0.53						
	(9.47)	(9.53)	(9.41)	(9.42)	(6.20)						

V-C Value as GPSGA/ME

The motivation for using GPSGA/ME as a measure for value is laid out at the beginning of Section V. When comparing the results of this subsection with the ones from the previous, one has to keep in mind that the sample of stocks is different due to the negative book equity stocks not being considered in the previous section. However, the value, profitability, and combined strategies are long and short and therefore rely on their ability to separate the stocks, i.e., are more or less independent of the average return of the stock sample. The value strategies based on GPSGA/ME perform significantly better on average than the ones based on BE/ME, both before and even more so after 2010. While the difference between cheaper and more expensive stocks is almost inexistent among the most profitable companies, it is the largest among the least profitable ones.

The combined strategies profit significantly from measuring value with GPSGA/ME instead of BE/ME, especially the combinations with GP/A (1.06% p.m. vs. 0.57% p.m.) and GPSGA/A (1.08% p.m. vs. 0.52% p.m.). The combination with GPSGAI/BE does not profit so much over the whole sample (0.89% p.m. vs. 0.84% p.m.), but it is significant in the sample after 2010 (0.74% p.m. vs. 0.32% p.m., see Table A.31 in the appendix). Furthermore, the relatively good performance of the high BE/ME and high GPSGAI/BE portfolio in Table 19 (0.76% p.m.) should be put into perspective by also considering the portfolios 4 in either direction, i.e., 0.33% p.m. to the left (lower GPSGAI/BE) and 0.38% p.m. above (lower BE/ME). Comparing this with deviations in the high GPSGA/ME and high GPSGAI/BE portfolio in Table 22 shows that the performance drop is much less substantial (from 0.68% p.m. to 0.55% p.m. or 0.50% p.m.).

The asymmetry described in the previous double sorts with size and value as BE/ME about GPSGAI/BE finding the poor future performers and the profitability measures scaled by total assets finding the strong future performers is less evident in double sorts with GPSGA/ME. Nevertheless, the low GPSGA/ME and low GPSGAI/BE portfolio has the worst return with -0.22% p.m. of all portfolios in this paper and could be combined with the second best portfolio, high GPSGA/ME and high GPSGA/A with a return of 0.95% over the whole sample, resulting in a long-short portfolio with an average return of 1.17% p.m. A combination with the best-performing portfolio of this paper, the small firm and high GP/A from Table 14 with 1.00% p.m. would also be possible and have the largest raw return at 1.22% p.m.; however, its three-factor alpha would undoubtedly be lower due to the exposure to the SMB factor.

Table 20: Double Sort on GPSGA/ME and GP/A

This table shows the value-weighted excess returns for the portfolios first sorted into quintiles based on GPSGA/ME, then within the quintiles based on GP/A. Portfolios are rebalanced annually at the end of June and cover the whole sample from June 1980 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The value strategies are the high-minus-low BE/ME portfolios, profitability strategies the high-minus-low profitability (GP/A) portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies). Table A.29 shows results for the subperiods before and after December 2010.

	GP/A					Profitability strategies					
GPSGA/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	−0.15%	0.06%	0.22%	0.45%	0.83%	0.99%	0.85%	1.16%	−0.08	−0.63	−0.28
						(4.26)	(3.69)	(5.30)	(−1.60)	(−8.35)	(−3.79)
2	0.08%	0.28%	0.50%	0.50%	0.63%	0.55%	0.50%	0.53%	0.10	−0.24	−0.12
						(3.74)	(3.37)	(3.61)	(2.88)	(−4.74)	(−2.55)
3	0.09%	0.28%	0.39%	0.51%	0.67%	0.58%	0.52%	0.65%	−0.06	−0.00	−0.11
						(3.90)	(3.51)	(4.27)	(−1.67)	(−0.07)	(−2.26)
4	0.28%	0.26%	0.55%	0.80%	0.81%	0.53%	0.44%	0.78%	−0.24	−0.10	−0.32
						(2.75)	(2.26)	(4.16)	(−5.62)	(−1.58)	(−5.10)
High	0.71%	0.59%	0.69%	0.63%	0.90%	0.20%	0.11%	0.14%	−0.01	0.26	0.22
						(1.08)	(0.63)	(0.75)	(−0.30)	(4.27)	(3.66)
	Value strategies					Combined strategy					
Mean	0.86%	0.53%	0.47%	0.18%	0.07%	1.06%	0.96%	1.15%	−0.15	−0.26	0.24
	(4.27)	(3.34)	(2.67)	(0.98)	(0.39)	(5.25)	(4.79)	(5.93)	(−3.38)	(−3.97)	(3.76)
Geomean	0.76%	0.47%	0.40%	0.10%	−0.01%						
	(3.76)	(2.94)	(2.24)	(0.52)	(−0.08)						
α	0.69%	0.21%	−0.00%	−0.32%	−0.34%						
	(3.66)	(1.46)	(−0.02)	(−1.87)	(−1.93)						
β_{MKT}	−0.13	−0.06	0.10	0.06	−0.07						
	(−3.11)	(−1.76)	(2.76)	(1.52)	(−1.66)						
β_{SMB}	−0.52	−0.26	−0.29	0.12	0.37						
	(−8.15)	(−5.22)	(−5.45)	(2.07)	(6.13)						
β_{HML}	0.02	0.32	0.58	0.62	0.51						
	(0.26)	(6.59)	(11.11)	(11.01)	(8.83)						

Table 21: Double Sort on GPSGA/ME and GPSGA/A

This table shows the value-weighted excess returns for the portfolios first sorted into quintiles based on GPSGA/ME, then within the quintiles based on GPSGA/A. Portfolios are rebalanced annually at the end of June and cover the whole sample from June 1980 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The value strategies are the high-minus-low BE/ME portfolios, profitability strategies are the high-minus-low profitability (GPSGA/A) portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies). Table A.30 shows results for the subperiods before and after December 2010.

	GPSGA/A					Profitability strategies					
GPSGA/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	−0.13%	0.16%	0.23%	0.25%	0.77%	0.89%	0.75%	1.21%	−0.16	−0.95	−0.52
						(3.75)	(3.06)	(6.06)	(−3.51)	(−13.88)	(−7.89)
2	0.15%	0.33%	0.18%	0.44%	0.48%	0.33%	0.27%	0.27%	0.17	−0.28	−0.21
						(2.11)	(1.71)	(1.81)	(4.89)	(−5.43)	(−4.08)
3	0.22%	0.26%	0.44%	0.30%	0.60%	0.39%	0.33%	0.46%	−0.01	−0.11	−0.26
						(2.61)	(2.26)	(3.09)	(−0.30)	(−2.11)	(−5.25)
4	0.47%	0.47%	0.30%	0.38%	0.64%	0.17%	0.08%	0.34%	−0.05	−0.37	−0.48
						(0.88)	(0.40)	(1.85)	(−1.22)	(−5.93)	(−8.02)
High	1.04%	0.45%	0.53%	0.52%	0.95%	−0.09%	−0.27%	0.16%	−0.17	−0.42	−0.45
						(−0.34)	(−0.97)	(0.65)	(−2.90)	(−4.93)	(−5.36)
	Value strategies					Combined strategy					
Mean	1.17%	0.29%	0.30%	0.27%	0.18%	1.08%	0.92%	1.28%	−0.22	−0.71	0.17
	(4.58)	(1.50)	(1.72)	(1.55)	(0.84)	(4.28)	(3.68)	(5.69)	(−4.31)	(−9.23)	(2.24)
Geomean	1.02%	0.20%	0.23%	0.19%	0.07%						
	(4.14)	(1.06)	(1.29)	(1.12)	(0.30)						
α	0.80%	−0.24%	−0.12%	−0.17%	−0.25%						
	(3.35)	(−1.31)	(−0.69)	(−1.07)	(−1.23)						
β_{MKT}	−0.05	0.14	0.02	−0.02	−0.06						
	(−0.98)	(3.27)	(0.43)	(−0.61)	(−1.27)						
β_{SMB}	−0.29	0.05	0.09	0.16	0.24						
	(−3.53)	(0.85)	(1.53)	(2.99)	(3.40)						
β_{HML}	0.61	0.54	0.39	0.59	0.68						
	(7.67)	(8.77)	(6.90)	(11.35)	(10.11)						

Table 22: Double Sort on GPSGA/ME and GPSGAI/BE

This table shows the value-weighted excess returns for the portfolios first sorted into quintiles based on GPSGA/ME, then within the quintiles based on GPSGAI/BE. Portfolios are rebalanced annually at the end of June and cover the whole sample from June 1980 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The value strategies are the high-minus-low BE/ME portfolios, profitability strategies are the high-minus-low profitability (GPSGAI/BE) portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies). Table A.31 shows results for the subperiods before and after December 2010.

	GPSGAI/BE					Profitability strategies					
GPSGA/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	−0.22%	0.06%	0.05%	0.01%	0.69%	0.90%	0.74%	1.27%	−0.27	−0.88	−0.42
						(3.58)	(2.88)	(5.84)	(−5.40)	(−11.77)	(−5.81)
2	0.00%	0.24%	0.12%	0.54%	0.42%	0.42%	0.37%	0.35%	0.19	−0.35	−0.19
						(2.99)	(2.65)	(2.68)	(6.33)	(−7.75)	(−4.27)
3	0.08%	0.01%	−0.00%	0.23%	0.51%	0.43%	0.36%	0.57%	−0.06	−0.30	−0.33
						(2.56)	(2.15)	(3.52)	(−1.68)	(−5.48)	(−6.19)
4	0.25%	0.18%	0.48%	0.23%	0.50%	0.24%	0.17%	0.37%	−0.04	−0.36	−0.30
						(1.40)	(0.97)	(2.16)	(−1.01)	(−6.20)	(−5.30)
High	0.46%	0.59%	0.45%	0.55%	0.68%	0.21%	0.11%	0.26%	0.05	−0.39	−0.22
						(1.07)	(0.57)	(1.32)	(1.05)	(−5.79)	(−3.37)
	Value strategies					Combined strategy					
Mean	0.68%	0.52%	0.41%	0.54%	−0.01%	0.89%	0.75%	1.04%	−0.16	−0.60	0.15
	(3.27)	(2.87)	(2.54)	(2.80)	(−0.04)	(3.69)	(3.13)	(4.60)	(−2.99)	(−7.71)	(1.93)
Geomean	0.57%	0.44%	0.34%	0.45%	−0.10%						
	(2.80)	(2.47)	(2.14)	(2.33)	(−0.53)						
α	0.46%	0.10%	−0.10%	0.11%	−0.55%						
	(2.33)	(0.56)	(−0.65)	(0.58)	(−2.96)						
β_{MKT}	−0.20	0.02	0.11	−0.00	0.12						
	(−4.44)	(0.41)	(3.16)	(−0.08)	(2.74)						
β_{SMB}	−0.20	−0.02	0.14	0.11	0.28						
	(−3.04)	(−0.39)	(2.72)	(1.82)	(4.39)						
β_{HML}	0.36	0.49	0.45	0.54	0.56						
	(5.54)	(8.39)	(8.83)	(8.86)	(9.00)						

V-D Industry Concentration

By far the most popular measure for industry concentration is the Herfindahl–Hirschman Index (HHI), named after Orris C. Herfindahl and Albert O. Hirschman.[11] It is simply the sum of the squares of each firm's market share (MS):

$$\mathrm{HHI} = \sum_{i=1}^{N} (\mathrm{MS})^2. \tag{8}$$

While the formula is simple, computing it in practice has its obstacles. Since data on privately held firms is generally not available, one often has the bias of only observing publicly traded companies. While the U.S. census now also publishes a HHI based on their economic survey, it has the drawback of only being conducted every five years and having a limited history. Apart from this issue, there is the problem of how to separate the companies into different industries, which was already discussed before.

Since the evidence on industry concentration is mixed, I only put part of the results in this paper. Table 23 shows the portfolios first sorted on the HHI based on the Fama-French 49 industries and then GPSGA/A (Tables A.32 and A.33 in the appendix show the same for GP/A and GPSGAI/BE). Theory would suggest that firms in industries with larger concentration could extract extra wealth from their market power, while those in industries

[11] Hirschman mentions the sum of the squares of each market participant's share as a measure of industry concentration in 1945, and Herfindahl in 1950 in his dissertation, apparently unaware of Hirschman's work.

Table 23: Double Sort on HHI (Fama-French 49 Industries) and GPSGA/A

This table shows the value-weighted excess returns for the portfolios first sorted into quintiles based on the HHI of the Fama-French 49 industries, then within the quintiles based on GPSGA/A. Portfolios are rebalanced annually at the end of June and cover the whole sample from June 1980 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The industry concentration strategies are the high concentration minus low concentration portfolios, profitability strategies the high-minus-low GPSGA/A portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies). Tables A.32 and A.33 in the appendix show the same for GP/A and GPSGAI/BE, and Table A.34 in the appendix for the sorts on the TNIC HHI from Hoberg and Phillips (2016).

	GPSGA/A					Profitability strategies					
HHI	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	0.18%	0.49%	0.58%	0.42%	0.91%	0.73%	0.63%	0.74%	0.06	−0.44	−0.10
						(3.64)	(3.15)	(3.79)	(1.43)	(−6.60)	(−1.52)
2	0.25%	0.51%	0.68%	0.48%	0.69%	0.45%	0.36%	0.61%	−0.07	−0.29	−0.46
						(2.47)	(2.00)	(3.58)	(−1.74)	(−4.95)	(−8.07)
3	0.32%	0.33%	0.47%	0.52%	0.54%	0.22%	0.15%	0.38%	−0.11	−0.43	−0.18
						(1.28)	(0.83)	(2.29)	(−2.89)	(−7.52)	(−3.24)
4	0.15%	0.15%	0.23%	0.48%	0.55%	0.40%	0.29%	0.60%	−0.12	−0.56	−0.31
						(1.98)	(1.45)	(3.23)	(−2.83)	(−8.81)	(−5.04)
High	0.16%	0.53%	0.50%	0.57%	0.90%	0.74%	0.62%	1.04%	−0.22	−0.41	−0.47
						(3.38)	(2.80)	(5.02)	(−4.67)	(−5.83)	(−6.79)
	Industry concentration strategies					Combined strategy					
Mean	−0.02%	0.04%	−0.09%	0.15%	−0.01%	0.72%	0.59%	0.88%	−0.09	−0.55	−0.18
	(−0.11)	(0.27)	(−0.52)	(0.98)	(−0.05)	(3.18)	(2.58)	(4.04)	(−1.90)	(−7.38)	(−2.48)
Geomean	−0.13%	−0.02%	−0.16%	0.09%	−0.10%						
	(−0.62)	(−0.13)	(−0.94)	(0.57)	(−0.51)						
α	−0.48%	−0.41%	−0.39%	−0.07%	−0.19%						
	(−2.39)	(−2.54)	(−2.25)	(−0.49)	(−1.02)						
β_{MKT}	0.13	0.18	−0.01	−0.08	−0.16						
	(2.80)	(4.80)	(−0.13)	(−2.42)	(−3.64)						
β_{SMB}	−0.13	−0.13	−0.27	−0.26	−0.10						
	(−1.94)	(−2.40)	(−4.68)	(−5.10)	(−1.64)						
β_{HML}	0.28	0.08	0.02	−0.04	−0.09						
	(4.17)	(1.45)	(0.40)	(−0.82)	(−1.40)						

with low concentration would suffer from the high competition. However, the empirical evidence is ambiguous, as the industry concentration strategies in Table 23 show. Although their returns have a stronger tendency towards being positive when using the TNIC HHI (Hoberg and Phillips, 2016) (results in Table A.34 in the appendix), they remain almost always below statistical significance. In undisclosed results, I confirm that this is also the case for the Fama-French 12, 17, and 30 industries, as well as the first-level SIC industries.[12] However, profitability separates poor from strong future performers within all industry concentrations, i.e., its value as a predictor is independent of industry concentration. One may find patterns like the fact that the high industry concentration and high profitability portfolio generally performs really well; however, this is almost entirely due to the profitability effect rather than the industry concentration. Hence, industry concentration may be a weak predictor for future stock performance, but I recommend combining profitability strategies with a value measured like GPSGA/ME or GPSGAI/ME instead of industry concentration.

[12] Note that the lower the number of industries gets, the more likely it becomes that sorting into quintiles is no longer possible, and one may have to switch to sorting into terciles.

VI Conclusion

This paper makes multiple contributions to the literature on profitability. First, I show that the predictive power of the trend in profitability described in Akbas et al. (2017) primarily stems from the time period 2000 to 2006, and it has been inexistent or in the opposite direction since at least 2010. The portfolio returns based on profitability's trend measures are below statistical significance over the investigated 40.5 years between June 1980 and December 2021. Second, I show that the level of most of the considered profitability measures remains a strong predictor for future returns. The empirical evidence suggests that gross profits (Novy-Marx, 2013) and gross profits minus SG&A (Ball et al., 2015), both scaled by total assets (GP/A and GPSGA/A), as well as gross profits minus SG&A and minus interest payments scaled by book equity (GPSGAI/BE, Fama and French (2015)) are the strongest predictors. Cleaning *Compustat*'s SG&A variable by re-adding R&D expenses not only improves the predictive power of GPSGA/A (as pointed out by (Ball et al., 2015)), it also improves the Fama-French measure GPSGAI/BE.

While Ball et al. (2015) try to establish that their profitability measure (GPSGA/A) is a better predictor than Novy-Marx (2013)'s GP/A, I show that both contain common information due to their close relationship, but that they also both provide distinct information related to future returns. Hence, neither one makes the other obsolete. The empirical evidence suggests that GP/A and GPSGA/A are the best measures for finding strong future performers (GP/A among small stocks and GPSGA/A among larger stocks), and GPSGAI/BE excels at finding poor future performers.

Another contribution of this paper is to show that difference between a firm's profitability and its industry's mean was a good predictor till around 2000, but it yields mixed results afterward and lags the absolute level of profitability. Industry concentration is at most marginally relevant; however, profitability measures separate weak and strong future performers independently of it.

Furthermore, I raise doubt about the usefulness of the book-to-market (BE/ME) ratio as a measure of value and show that an alternative, namely GPSGA-to-market equity, yields value-weighted portfolios with stronger performance before 2010, but especially afterward. I use this alternative value measure in double sorts with profitability and show that it also significantly improves the value-profitability combination compared to BE/ME. The resulting long-short strategies of double quintile sorts have returns of around 1% per month over the whole sample period, which is very impressive considering the annual portfolio rebalancing and the fact that Fama-French three-factor model alphas are even higher. Hence, this paper advises buying stocks with a solid wealth-generating technology for the firm (high GP/A or GPSGA/A) or the shareholders (high GPSGAI/BE) at a low price (low GP/ME, GPSGA/ME, or GPSGAI/ME).

Appendix Essay 3

Table A.1: Complete Fama-MacBeth Regression Results - Level

This table shows the complete Fama-MacBeth regression results for the level of profitability, where all measures are based on the previous four quarters. The first part covers the whole sample (June 1980 to December 2021), and the second and third the time period before and after December 2010. The control variables include value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. All coefficients are in percent, and the values in brackets are the corresponding t-test values.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$	$\frac{E}{A}$
June 1980 to December 2021							
Profitability level	1.15	2.65	1.73	1.02	0.72	0.32	1.11
	(7.66)	(9.06)	(5.71)	(10.32)	(6.91)	(4.29)	(4.20)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.44	0.37	0.34	0.40	0.35	0.33	0.34
	(6.62)	(6.22)	(6.08)	(6.76)	(6.36)	(5.41)	(6.03)
Ln(ME)	−0.04	−0.12	−0.11	−0.10	−0.09	−0.08	−0.09
	(−1.06)	(−3.55)	(−3.20)	(−2.92)	(−2.77)	(−2.16)	(−2.55)
$\text{return}_{-1,0}$	−4.43	−4.34	−4.35	−4.42	−4.42	−4.35	−4.39
	(−11.76)	(−11.48)	(−11.60)	(−11.54)	(−11.59)	(−11.57)	(−11.74)
$\text{return}_{-12,-1}$	0.47	0.47	0.49	0.51	0.52	0.47	0.46
	(3.44)	(3.46)	(3.61)	(3.59)	(3.69)	(3.50)	(3.44)
Intercept	1.03	1.51	1.56	1.51	1.53	1.53	1.60
	(2.73)	(4.47)	(4.77)	(4.47)	(4.63)	(4.64)	(4.98)
N (in million)	1.50	1.25	1.25	1.00	1.00	1.54	1.54
June 1980 to December 2010							
Profitability level	1.20	2.73	1.79	1.12	0.80	0.28	1.01
	(7.32)	(8.40)	(4.81)	(9.21)	(6.05)	(2.97)	(3.11)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.55	0.49	0.47	0.49	0.46	0.45	0.47
	(6.83)	(6.86)	(7.14)	(7.09)	(7.12)	(6.25)	(6.89)
Ln(ME)	−0.09	−0.16	−0.14	−0.15	−0.14	−0.11	−0.12
	(−1.79)	(−3.76)	(−3.40)	(−3.37)	(−3.19)	(−2.61)	(−2.88)
$\text{return}_{-1,0}$	−5.56	−5.47	−5.49	−5.62	−5.62	−5.45	−5.51
	(−13.23)	(−13.12)	(−13.40)	(−13.68)	(−13.81)	(−13.07)	(−13.36)
$\text{return}_{-12,-1}$	0.48	0.47	0.49	0.51	0.52	0.50	0.48
	(2.87)	(2.84)	(2.99)	(2.98)	(3.10)	(3.02)	(2.98)
Intercept	1.38	1.83	1.86	1.85	1.86	1.86	1.89
	(3.19)	(4.61)	(4.83)	(4.77)	(4.89)	(4.85)	(5.06)
N (in million)	1.18	0.97	0.97	0.76	0.76	1.22	1.22
January 2011 to December 2021							
Profitability level	1.06	2.52	1.72	0.67	0.52	0.43	1.42
	(3.13)	(3.94)	(3.35)	(4.55)	(3.45)	(3.76)	(3.34)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.13	0.04	−0.03	0.13	0.05	−0.01	−0.01
	(1.21)	(0.39)	(−0.31)	(1.22)	(0.52)	(−0.12)	(−0.10)
Ln(ME)	0.08	−0.02	−0.02	0.02	0.02	0.03	0.00
	(1.18)	(−0.33)	(−0.39)	(0.37)	(0.29)	(0.45)	(0.07)
$\text{return}_{-1,0}$	−1.22	−1.14	−1.10	−1.02	−1.03	−1.21	−1.21
	(−1.62)	(−1.47)	(−1.41)	(−1.23)	(−1.24)	(−1.59)	(−1.58)
$\text{return}_{-12,-1}$	0.44	0.47	0.48	0.51	0.51	0.40	0.39
	(1.95)	(2.02)	(2.03)	(2.03)	(2.00)	(1.79)	(1.72)
Intercept	0.09	0.65	0.77	0.55	0.61	0.62	0.81
	(0.11)	(1.01)	(1.24)	(0.781)	(0.92)	(0.95)	(1.30)
N (in million)	0.32	0.27	0.27	0.23	0.23	0.32	0.32

Table A.2: Fama-MacBeth Regression Results - Level (Excluding Micro-Caps)

This table shows the results of the Fama-MacBeth regressions for the level of profitability, where all measures are based on the previous four quarters. The first part covers the whole sample (June 1980 to December 2021), the second and third the time period before and after December 2010. The control variables include value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. All coefficients are in percent, and the values in brackets are the corresponding t-test values. The sample excludes micro-caps which are defined as stocks with a market capitalization below the 20th percentile of the NYSE market capitalization distribution.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$	$\frac{E}{A}$
			June 1980 to December 2021				
Profitability level	0.90	2.56	1.79	0.88	0.68	0.52	1.81
	(4.87)	(6.22)	(3.84)	(5.60)	(3.89)	(3.43)	(3.98)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.26	0.29	0.22	0.31	0.25	0.15	0.17
	(3.20)	(3.80)	(2.83)	(3.79)	(3.16)	(1.98)	(2.25)
Ln(ME)	−0.02	−0.03	−0.03	−0.02	−0.02	−0.03	−0.04
	(−0.45)	(−0.67)	(−0.77)	(−0.47)	(−0.45)	(−0.76)	(−1.03)
$\text{return}_{-1,0}$	−1.68	−1.65	−1.65	−1.83	−1.83	−1.62	−1.61
	(−3.42)	(−3.30)	(−3.35)	(−3.64)	(−3.69)	(−3.28)	(−3.29)
$\text{return}_{-12,-1}$	0.49	0.53	0.51	0.49	0.47	0.46	0.47
	(2.84)	(3.01)	(2.98)	(2.73)	(2.67)	(2.70)	(2.79)
Intercept	0.95	0.93	1.05	1.06	1.08	1.20	1.25
	(2.08)	(2.02)	(2.29)	(2.34)	(2.39)	(2.75)	(2.89)
N (in million)	0.53	0.44	0.44	0.37	0.37	0.55	0.55
			June 1980 to December 2010				
Profitability level	1.07	3.20	2.42	1.24	1.00	0.75	2.40
	(5.17)	(6.84)	(4.36)	(6.08)	(4.36)	(3.86)	(4.41)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.38	0.45	0.36	0.47	0.40	0.24	0.28
	(3.79)	(4.68)	(3.75)	(4.78)	(4.16)	(2.58)	(2.89)
Ln(ME)	−0.03	−0.04	−0.04	−0.03	−0.03	−0.05	−0.06
	(−0.60)	(−0.73)	(−0.88)	(−0.63)	(−0.62)	(−0.97)	(−1.14)
$\text{return}_{-1,0}$	−2.08	−2.24	−2.23	−2.44	−2.44	−2.01	−2.01
	(−3.67)	(−3.88)	(−3.94)	(−4.26)	(−4.31)	(−3.56)	(−3.58)
$\text{return}_{-12,-1}$	0.51	0.58	0.57	0.55	0.53	0.50	0.50
	(2.42)	(2.70)	(2.70)	(2.51)	(2.48)	(2.36)	(2.41)
Intercept	1.07	1.05	1.20	1.21	1.24	1.41	1.43
	(2.06)	(1.93)	(2.20)	(2.28)	(2.36)	(2.77)	(2.82)
N (in million)	0.39	0.31	0.31	0.25	0.25	0.40	0.40
			January 2011 to December 2021				
Profitability level	0.44	0.81	0.12	0.02	−0.08	−0.03	0.26
	(1.12)	(0.96)	(0.14)	(0.11)	(−0.43)	(−0.17)	(0.31)
$\text{Ln}\left(\frac{BE}{ME}\right)$	−0.07	−0.13	−0.17	−0.14	−0.17	−0.10	−0.11
	(−0.58)	(−1.19)	(−1.51)	(−1.01)	(−1.28)	(−0.95)	(−1.03)
Ln(ME)	0.01	0.00	0.01	0.01	0.01	0.02	0.00
	(0.17)	(0.02)	(0.09)	(0.18)	(0.20)	(0.23)	(0.01)
$\text{return}_{-1,0}$	−0.59	−0.01	−0.04	−0.11	−0.12	−0.53	−0.51
	(−0.59)	(−0.01)	(−0.04)	(−0.10)	(−0.12)	(−0.53)	(−0.51)
$\text{return}_{-12,-1}$	0.42	0.38	0.36	0.30	0.29	0.36	0.39
	(1.52)	(1.31)	(1.24)	(1.01)	(0.96)	(1.28)	(1.43)
Intercept	0.63	0.61	0.65	0.63	0.63	0.66	0.79
	(0.66)	(0.69)	(0.75)	(0.71)	(0.71)	(0.77)	(0.92)
N (in million)	0.15	0.13	0.13	0.12	0.12	0.15	0.15

Table A.3: Complete Fama-MacBeth Regression Results - AJK Trend

This table shows the complete Fama-MacBeth regression results for the trend in profitability, where the trend is measured as the coefficient of a two-year regression with quarterly dummies. The first part covers the whole sample (June 1980 to December 2021), and the second and third the time period before and after December 2010. The AJK trend is a two-year regression with quarterly dummy variables, and the 2yr regression is without dummies. The control variables include value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. All coefficients are in percent, and the values in brackets are the corresponding t-test values.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$
			June 1980 to December 2021			
AJK trend	11.38	10.17	10.17	1.11	9.02	2.10
	(3.61)	(5.45)	(2.13)	(2.92)	(4.81)	(1.73)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.36	0.37	0.36	0.38	0.37	0.37
	(5.49)	(6.08)	(5.94)	(6.05)	(5.87)	(5.47)
Ln(ME)	−0.04	−0.04	−0.04	−0.04	−0.05	−0.05
	(−0.97)	(−0.92)	(−1.08)	(−1.03)	(−1.21)	(−1.19)
$\text{return}_{-1,0}$	−4.33	−4.24	−4.26	−4.49	−4.50	−4.36
	(−11.27)	(−10.87)	(−10.95)	(−11.17)	(−11.24)	(−11.29)
$\text{return}_{-12,-1}$	0.42	0.46	0.47	0.53	0.47	0.44
	(3.03)	(3.22)	(3.31)	(3.53)	(3.17)	(3.16)
Intercept	1.41	1.43	1.46	1.42	1.48	1.45
	(3.95)	(4.12)	(4.23)	(4.20)	(4.39)	(4.13)
N (in million)	1.30	1.10	1.07	0.78	0.76	1.26
			June 1980 to December 2010			
AJK trend	20.06	15.71	16.06	1.69	11.45	2.77
	(5.56)	(7.83)	(3.04)	(3.74)	(5.27)	(1.84)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.45	0.49	0.48	0.52	0.50	0.46
	(5.59)	(6.66)	(6.53)	(6.95)	(6.71)	(5.72)
Ln(ME)	−0.08	−0.07	−0.07	−0.06	−0.07	−0.09
	(−1.67)	(−1.34)	(−1.48)	(−1.30)	(−1.48)	(−1.77)
$\text{return}_{-1,0}$	−5.50	−5.45	−5.48	−5.76	−5.78	−5.51
	(−12.87)	(−12.73)	(−12.85)	(−13.41)	(−13.52)	(−12.88)
$\text{return}_{-12,-1}$	0.42	0.45	0.46	0.56	0.48	0.44
	(2.45)	(2.59)	(2.69)	(3.14)	(2.70)	(2.68)
Intercept	1.80	1.75	1.78	1.68	1.75	1.79
	(4.40)	(4.34)	(4.43)	(4.30)	(4.49)	(4.43)
N (in million)	1.00	0.84	0.81	0.59	0.57	0.99
			January 2011 to December 2021			
AJK trend	−11.98	−6.32	−6.00	−0.61	2.58	0.25
	(−2.01)	(−1.54)	(−0.58)	(−0.90)	(0.72)	(0.13)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.11	0.04	0.03	0.01	0.01	0.09
	(1.04)	(0.38)	(0.31)	(0.05)	(0.09)	(0.84)
Ln(ME)	0.08	0.04	0.04	0.02	0.02	0.06
	(1.13)	(0.66)	(0.58)	(0.33)	(0.31)	(0.86)
$\text{return}_{-1,0}$	−1.01	−0.81	−0.81	−0.89	−0.89	−1.11
	(−1.31)	(−1.01)	(−1.01)	(−1.03)	(−1.02)	(−1.42)
$\text{return}_{-12,-1}$	0.44	0.50	0.49	0.45	0.45	0.41
	(1.87)	(2.01)	(2.01)	(1.59)	(1.62)	(1.68)
Intercept	0.35	0.56	0.60	0.70	0.71	0.52
	(0.49)	(0.82)	(0.89)	(1.05)	(1.08)	(0.74)
N (in million)	0.30	0.25	0.25	0.19	0.19	0.27

Table A.4: Complete Fama-MacBeth Regression Results - 2yr Regression Trend

This table shows the complete Fama-MacBeth regression results for the trend in profitability, where the trend is measured as the coefficient of a two-year regression without quarterly dummies. The first part covers the whole sample (June 1980 to December 2021), and the second and third the time period before and after December 2010. The AJK trend is a two-year regression with quarterly dummy variables, and the 2yr regression is without dummies. The control variables include value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. All coefficients are in percent, and the values in brackets are the corresponding t-test values.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$
June 1980 to December 2021						
2yr regression trend	16.89	12.54	17.70	1.42	11.22	3.77
	(5.27)	(6.72)	(3.64)	(3.56)	(5.79)	(2.67)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.36	0.37	0.36	0.39	0.36	0.36
	(5.51)	(6.12)	(5.98)	(6.10)	(5.82)	(5.41)
Ln(ME)	−0.04	−0.04	−0.04	−0.04	−0.05	−0.05
	(−0.95)	(−0.89)	(−1.07)	(−1.00)	(−1.25)	(−1.24)
$\text{return}_{-1,0}$	−4.33	−4.24	−4.26	−4.50	−4.50	−4.36
	(−11.29)	(−10.88)	(−10.96)	(−11.18)	(−11.24)	(−11.29)
$\text{return}_{-12,-1}$	0.41	0.45	0.45	0.53	0.46	0.43
	(2.94)	(3.17)	(3.17)	(3.53)	(3.09)	(3.10)
Intercept	1.41	1.43	1.47	1.42	1.48	1.46
	(3.95)	(4.12)	(4.24)	(4.19)	(4.41)	(4.17)
N (in million)	1.30	1.10	1.07	0.78	0.76	1.26
June 1980 to December 2010						
2yr regression trend	26.71	18.45	25.71	2.08	15.01	5.15
	(7.32)	(9.22)	(4.79)	(4.37)	(6.76)	(2.94)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.45	0.49	0.48	0.52	0.49	0.46
	(5.62)	(6.69)	(6.57)	(7.00)	(6.63)	(5.65)
Ln(ME)	−0.08	−0.06	−0.07	−0.06	−0.07	−0.09
	(−1.66)	(−1.32)	(−1.47)	(−1.27)	(−1.53)	(−1.83)
$\text{return}_{-1,0}$	−5.51	−5.46	−5.49	−5.77	−5.79	−5.51
	(−12.89)	(−12.74)	(−12.86)	(−13.42)	(−13.53)	(−12.88)
$\text{return}_{-12,-1}$	0.40	0.44	0.44	0.55	0.46	0.43
	(2.37)	(2.54)	(2.54)	(3.14)	(2.60)	(2.62)
Intercept	1.80	1.75	1.79	1.68	1.77	1.81
	(4.40)	(4.34)	(4.44)	(4.28)	(4.53)	(4.48)
N (in million)	1.00	0.84	0.81	0.59	0.57	0.99
January 2011 to December 2021						
2yr regression trend	−9.84	−4.93	−4.55	−0.49	1.10	−0.01
	(−1.63)	(−1.21)	(−0.43)	(−0.69)	(0.29)	(−0.00)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.11	0.04	0.03	0.01	0.01	0.09
	(1.04)	(0.40)	(0.32)	(0.06)	(0.13)	(0.85)
Ln(ME)	0.08	0.04	0.04	0.02	0.02	0.06
	(1.13)	(0.68)	(0.59)	(0.34)	(0.33)	(0.86)
$\text{return}_{-1,0}$	−1.00	−0.81	−0.80	−0.90	−0.88	−1.12
	(−1.30)	(−1.01)	(−1.00)	(−1.03)	(−1.01)	(−1.42)
$\text{return}_{-12,-1}$	0.44	0.50	0.48	0.45	0.46	0.41
	(1.85)	(2.00)	(1.97)	(1.59)	(1.66)	(1.69)
Intercept	0.35	0.56	0.59	0.69	0.70	0.52
	(0.49)	(0.82)	(0.88)	(1.04)	(1.06)	(0.74)
N (in million)	0.30	0.25	0.25	0.19	0.19	0.27

Table A.5: Fama-MacBeth Regression Results - Delta to Industry Mean

This table shows the results of the Fama-MacBeth regressions of the delta between a firm's profitability and the industry mean, where the industries are defined as the Fama-French 12, 17, 30, or 49 industries, the first level SIC codes or the weighted mean according to the Hoberg and Phillips (2010) similarity scores from Text-based Network Industrial Classification (TNIC). The first two parts cover the entire sample period (June 1980/1988 to December 2021), the next two parts cover the early (June 1980/1988 to December 2010), and the remaining parts cover the later sample period (January 2011 to December 2021). All coefficients are in percent, and the values in brackets are the corresponding t-test values. For readability, the results for the control variables (value as Ln(BE/ME), size as Ln(ME), short-term reversal as past month's return, and momentum as past twelve months' return, skipping the most recent month) are not displayed here but can be found in Tables A.6, A.7, A.8, A.9, and A.10 in the appendix.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$
			June 1980 to December 2021			
Delta to FF 12 industry mean	1.19	2.56	1.96	0.66	0.54	0.31
	(9.13)	(9.32)	(7.61)	(9.79)	(8.40)	(5.90)
Delta to FF 17 industry mean	1.14	2.59	1.83	0.70	0.52	0.26
	(8.38)	(9.36)	(6.66)	(9.19)	(7.00)	(4.82)
Delta to FF 30 industry mean	1.18	2.54	1.87	0.57	0.43	0.27
	(9.06)	(9.35)	(7.18)	(9.24)	(7.48)	(5.51)
Delta to FF 49 industry mean	1.26	2.51	1.94	0.55	0.44	0.23
	(10.34)	(9.44)	(7.77)	(9.26)	(8.00)	(4.97)
Delta to SIC industry mean	1.13	2.53	1.73	0.68	0.51	0.30
	(8.20)	(9.03)	(5.99)	(8.81)	(6.34)	(5.08)
N (in million)	1.44 - 1.50	1.20 - 1.25	1.20 - 1.25	1.08 - 1.12	0.96 - 1.00	1.48 - 1.54
			June 1988 to December 2021			
Delta to TNIC industry mean	1.06	1.89	1.37	0.55	0.38	0.11
	(8.93)	(6.98)	(5.19)	(7.57)	(5.07)	(2.40)
N (in million)	1.04	0.86	0.86	0.69	0.69	1.06
			June 1980 to December 2010			
Delta to FF 12 industry mean	1.30	2.66	2.06	0.75	0.62	0.31
	(8.82)	(8.54)	(6.52)	(8.67)	(7.21)	(4.50)
Delta to FF 17 industry mean	1.21	2.64	1.83	0.84	0.62	0.25
	(8.30)	(8.42)	(5.42)	(8.35)	(6.22)	(3.49)
Delta to FF 30 industry mean	1.30	2.64	1.95	0.68	0.52	0.27
	(8.79)	(8.44)	(6.02)	(8.45)	(6.61)	(4.22)
Delta to FF 49 industry mean	1.38	2.59	2.03	0.68	0.54	0.22
	(9.82)	(8.44)	(6.56)	(8.54)	(7.18)	(3.63)
Delta to SIC industry mean	1.23	2.61	1.74	0.81	0.58	0.27
	(8.28)	(8.29)	(4.90)	(8.14)	(5.57)	(3.60)
N (in million)	1.13 - 1.18	0.93 - 0.97	0.93 - 0.97	0.83 - 0.86	0.73 - 0.76	1.17 - 1.22
			June 1988 to December 2010			
Delta to TNIC industry mean	1.21	1.83	1.20	0.67	0.45	0.09
	(8.53)	(5.63)	(3.60)	(6.64)	(4.14)	(1.49)
N (in million)	0.78	0.64	0.64	0.50	0.50	0.80
			January 2011 to December 2021			
Delta to FF 12 industry mean	0.89	2.26	1.72	0.44	0.38	0.32
	(3.32)	(3.95)	(4.09)	(4.98)	(5.10)	(4.69)
Delta to FF 17 industry mean	0.95	2.46	1.90	0.38	0.33	0.24
	(3.07)	(4.28)	(4.18)	(4.17)	(3.66)	(4.65)
Delta to FF 30 industry mean	0.86	2.26	1.69	0.34	0.29	0.28
	(3.23)	(4.11)	(4.17)	(4.41)	(4.40)	(4.51)
Delta to FF 49 industry mean	0.95	2.25	1.69	0.28	0.25	0.25
	(3.92)	(4.23)	(4.45)	(4.28)	(4.40)	(4.45)
Delta to SIC industry mean	0.86	2.33	1.81	0.36	0.33	0.26
	(2.77)	(3.91)	(3.73)	(3.52)	(3.29)	(4.41)
N (in million)	0.31 - 0.32	0.27	0.27	0.25 - 0.26	0.23	0.31 - 0.32
Delta to TNIC industry mean	0.77	2.02	1.73	0.36	0.29	0.16
	(3.59)	(4.14)	(4.15)	(3.93)	(3.45)	(2.22)
N (in million)	0.26	0.22	0.22	0.19	0.19	0.26

Table A.6: Complete Fama-MacBeth Regression Results - Delta Fama-French 12 Industries

This table shows the complete Fama-MacBeth regression results for the delta between a firm's profitability measure and the mean profitability of the industry, where the industries are defined as the Fama-French 12 industries. The first part covers the whole sample (June 1980 to December 2021), and the second and third the time period before and after December 2010. The control variables include value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. All coefficients are in percent, and the values in brackets are the corresponding t-test values.

	$\frac{GP}{A}$	$\frac{GP-SGA}{A}$	$\frac{GP-SGA-RD}{A}$	$\frac{GP-SGA-I}{BE}$	$\frac{GP-SGA-RD-I}{BE}$	$\frac{E}{BE}$
June 1980 to December 2021						
Delta to industry mean	1.19	2.56	1.96	0.66	0.54	0.31
	(9.13)	(9.32)	(7.61)	(9.79)	(8.40)	(5.90)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.44	0.38	0.36	0.42	0.41	0.36
	(6.59)	(6.40)	(6.17)	(7.11)	(6.92)	(5.68)
Ln(ME)	−0.05	−0.12	−0.11	−0.07	−0.07	−0.07
	(−1.12)	(−3.31)	(−3.05)	(−2.01)	(−1.94)	(−1.72)
$\text{return}_{-1,0}$	−4.41	−4.33	−4.33	−4.40	−4.39	−4.34
	(−11.67)	(−11.46)	(−11.50)	(−11.40)	(−11.40)	(−11.48)
$\text{return}_{-12,-1}$	0.47	0.48	0.49	0.55	0.55	0.47
	(3.45)	(3.51)	(3.57)	(3.84)	(3.87)	(3.47)
Intercept	1.40	1.73	1.67	1.53	1.49	1.40
	(3.90)	(5.32)	(5.14)	(4.51)	(4.41)	(3.95)
N (in million)	1.50	1.25	1.25	1.00	1.00	1.54
June 1980 to December 2010						
Delta to industry mean	1.30	2.66	2.06	0.75	0.62	0.31
	(8.82)	(8.54)	(6.52)	(8.67)	(7.21)	(4.50)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.54	0.50	0.48	0.54	0.52	0.46
	(6.76)	(6.98)	(6.94)	(7.54)	(7.42)	(5.99)
Ln(ME)	−0.10	−0.16	−0.15	−0.12	−0.11	−0.12
	(−1.92)	(−3.65)	(−3.43)	(−2.60)	(−2.52)	(−2.46)
$\text{return}_{-1,0}$	−5.55	−5.46	−5.46	−5.60	−5.59	−5.44
	(−13.21)	(−13.10)	(−13.17)	(−13.56)	(−13.54)	(−12.94)
$\text{return}_{-12,-1}$	0.49	0.48	0.49	0.55	0.56	0.49
	(2.91)	(2.88)	(2.94)	(3.20)	(3.25)	(2.96)
Intercept	1.81	2.08	2.01	1.87	1.82	1.80
	(4.40)	(5.45)	(5.33)	(4.82)	(4.72)	(4.43)
N (in million)	1.18	0.97	0.97	0.76	0.76	1.22
January 2011 to December 2021						
Delta to industry mean	0.89	2.26	1.72	0.44	0.38	0.32
	(3.32)	(3.95)	(4.09)	(4.98)	(5.10)	(4.69)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.14	0.06	0.02	0.12	0.09	0.08
	(1.28)	(0.56)	(0.21)	(1.13)	(0.84)	(0.74)
Ln(ME)	0.09	−0.00	0.00	0.04	0.03	0.06
	(1.27)	(−0.01)	(0.02)	(0.65)	(0.57)	(0.98)
$\text{return}_{-1,0}$	−1.17	−1.15	−1.13	−1.00	−1.00	−1.21
	(−1.55)	(−1.48)	(−1.46)	(−1.19)	(−1.20)	(−1.59)
$\text{return}_{-12,-1}$	0.44	0.49	0.48	0.54	0.53	0.42
	(1.92)	(2.08)	(2.07)	(2.13)	(2.08)	(1.86)
Intercept	0.30	0.79	0.74	0.60	0.60	0.33
	(0.41)	(1.27)	(1.18)	(0.88)	(0.87)	(0.46)
N (in million)	0.32	0.27	0.27	0.23	0.23	0.32

Table A.7: Complete Fama-MacBeth Regression Results - Delta Fama-French 17 Industries

This table shows the complete Fama-MacBeth regression results for the delta between a firm's profitability measure and the mean profitability of the industry, where the industries are defined as the Fama-French 17 industries. The first part covers the whole sample (June 1980 to December 2021), and the second and third the time period before and after December 2010. The control variables include value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. All coefficients are in percent, and the values in brackets are the corresponding t-test values.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$
			June 1980 to December 2021			
Delta to industry mean	1.14	2.59	1.83	0.70	0.52	0.26
	(8.38)	(9.36)	(6.66)	(9.19)	(7.00)	(4.82)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.42	0.37	0.35	0.40	0.38	0.36
	(6.37)	(6.22)	(6.02)	(6.56)	(6.59)	(5.68)
Ln(ME)	−0.05	−0.12	−0.11	−0.08	−0.08	−0.07
	(−1.30)	(−3.47)	(−3.17)	(−2.24)	(−2.11)	(−1.79)
$\text{return}_{-1,0}$	−4.38	−4.30	−4.31	−4.37	−4.37	−4.31
	(−11.56)	(−11.32)	(−11.40)	(−11.33)	(−11.35)	(−11.36)
$\text{return}_{-12,-1}$	0.47	0.48	0.49	0.54	0.55	0.48
	(3.44)	(3.46)	(3.57)	(3.80)	(3.86)	(3.54)
Intercept	1.45	1.77	1.70	1.65	1.51	1.44
	(4.04)	(5.46)	(5.27)	(4.62)	(4.51)	(4.08)
N (in million)	1.46	1.22	1.22	0.97	0.97	1.50
			June 1980 to December 2010			
Delta to industry mean	1.21	2.64	1.83	0.84	0.62	0.25
	(8.30)	(8.42)	(5.42)	(8.35)	(6.22)	(3.49)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.53	0.50	0.48	0.52	0.50	0.47
	(6.62)	(6.91)	(7.02)	(7.32)	(7.29)	(6.10)
Ln(ME)	−0.10	−0.16	−0.14	−0.13	−0.12	−0.11
	(−2.03)	(−3.68)	(−3.35)	(−2.81)	(−2.66)	(−2.46)
$\text{return}_{-1,0}$	−5.53	−5.44	−5.45	−5.59	−5.58	−5.41
	(−13.12)	(−12.97)	(−13.12)	(−13.52)	(−13.54)	(−12.81)
$\text{return}_{-12,-1}$	0.49	0.48	0.49	0.54	0.55	0.50
	(2.91)	(2.85)	(2.97)	(3.15)	(3.23)	(3.01)
Intercept	1.85	2.10	2.01	1.91	1.85	1.83
	(4.50)	(5.52)	(5.36)	(4.96)	(4.84)	(4.53)
N (in million)	1.15	0.95	0.95	0.74	0.74	1.19
			January 2011 to December 2021			
Delta to industry mean	0.95	2.46	1.90	0.38	0.33	0.24
	(3.07)	(4.28)	(4.18)	(4.17)	(3.66)	(4.65)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.11	0.03	−0.02	0.08	0.05	0.08
	(1.06)	(0.29)	(−0.23)	(0.80)	(0.51)	(0.73)
Ln(ME)	0.07	−0.02	−0.02	0.03	0.03	0.06
	(1.13)	(−0.33)	(−0.42)	(0.56)	(0.52)	(0.96)
$\text{return}_{-1,0}$	−1.13	−1.08	−1.07	−0.93	−0.95	−1.16
	(−1.48)	(−1.39)	(−1.37)	(−1.11)	(−1.13)	(−1.52)
$\text{return}_{-12,-1}$	0.43	0.48	0.47	0.53	0.53	0.44
	(1.90)	(2.01)	(1.99)	(2.11)	(2.09)	(1.92)
Intercept	0.36	0.87	0.85	0.58	0.57	0.34
	(0.49)	(1.41)	(1.36)	(0.85)	(0.85)	(0.47)
N (in million)	0.31	0.27	0.27	0.23	0.23	0.31

Table A.8: Complete Fama-MacBeth Regression Results - Delta Fama-French 30 Industries

This table shows the complete Fama-MacBeth regression results for the delta between a firm's profitability measure and the mean profitability of the industry, where the industries are defined as the Fama-French 30 industries. The first part covers the whole sample (June 1980 to December 2021), and the second and third the time period before and after December 2010. The control variables include value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. All coefficients are in percent, and the values in brackets are the corresponding t-test values.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$
June 1980 to December 2021						
Delta to industry mean	1.18	2.54	1.87	0.57	0.43	0.27
	(9.06)	(9.35)	(7.18)	(9.24)	(7.48)	(5.51)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.43	0.38	0.36	0.41	0.40	0.36
	(6.42)	(6.28)	(6.09)	(6.83)	(6.68)	(5.62)
Ln(ME)	−0.05	−0.11	−0.10	−0.07	−0.07	−0.06
	(−1.15)	(−3.27)	(−2.95)	(−1.88)	(−1.77)	(−1.65)
$\text{return}_{-1,0}$	−4.38	−4.31	−4.32	−4.36	−4.36	−4.30
	(−11.54)	(−11.31)	(−11.37)	(−11.21)	(−11.22)	(−11.29)
$\text{return}_{-12,-1}$	0.47	0.48	0.49	0.56	0.57	0.48
	(3.44)	(3.47)	(3.55)	(3.93)	(3.98)	(3.51)
Intercept	1.41	1.73	1.65	1.50	1.46	1.41
	(3.92)	(5.31)	(5.09)	(4.41)	(4.32)	(3.96)
N (in million)	1.44	1.20	1.20	0.96	0.96	1.48
June 1980 to December 2010						
Delta to industry mean	1.30	2.64	1.95	0.68	0.52	0.27
	(8.79)	(8.44)	(6.02)	(8.45)	(6.61)	(4.22)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.54	0.50	0.48	0.53	0.51	0.46
	(6.58)	(6.85)	(6.86)	(7.32)	(7.25)	(5.90)
Ln(ME)	−0.10	−0.16	−0.14	−0.12	−0.11	−0.11
	(−1.97)	(−3.62)	(−3.34)	(−2.51)	(−2.38)	(−2.43)
$\text{return}_{-1,0}$	−5.53	−5.44	−5.45	−5.56	−5.55	−5.39
	(−13.08)	(−12.93)	(−13.05)	(−13.31)	(−13.31)	(−12.70)
$\text{return}_{-12,-1}$	0.49	0.48	0.49	0.57	0.58	0.50
	(2.90)	(2.85)	(2.93)	(3.31)	(3.37)	(2.98)
Intercept	1.83	2.08	2.00	1.86	1.81	1.82
	(4.44)	(5.46)	(5.31)	(4.80)	(4.69)	(4.49)
N (in million)	1.13	0.93	0.93	0.73	0.73	1.17
January 2011 to December 2021						
Delta to industry mean	0.86	2.26	1.69	0.34	0.29	0.28
	(3.23)	(4.11)	(4.17)	(4.41)	(4.40)	(4.51)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.13	0.05	0.02	0.09	0.07	0.08
	(1.24)	(0.53)	(0.18)	(0.90)	(0.67)	(0.81)
Ln(ME)	0.09	0.00	0.00	0.04	0.04	0.07
	(1.29)	(0.01)	(0.06)	(0.71)	(0.68)	(1.04)
$\text{return}_{-1,0}$	−1.13	−1.10	−1.09	−0.97	−0.98	−1.18
	(−1.48)	(−1.41)	(−1.40)	(−1.16)	(−1.17)	(−1.54)
$\text{return}_{-12,-1}$	0.44	0.48	0.48	0.53	0.53	0.43
	(1.92)	(2.05)	(2.05)	(2.09)	(2.08)	(1.90)
Intercept	0.28	0.77	0.72	0.52	0.51	0.29
	(0.38)	(1.24)	(1.13)	(0.75)	(0.75)	(0.40)
N (in million)	0.31	0.27	0.27	0.23	0.23	0.31

Table A.9: Complete Fama-MacBeth Regression Results - Delta Fama-French 49 Industries

This table shows the complete Fama-MacBeth regression results for the delta between a firm's profitability measure and the mean profitability of the industry, where the industries are defined as the Fama-French 49 industries. The first part covers the whole sample (June 1980 to December 2021), and the second and third the time period before and after December 2010. The control variables include value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. All coefficients are in percent, and the values in brackets are the corresponding t-test values.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$
June 1980 to December 2021						
Delta to industry mean	1.26	2.51	1.94	0.55	0.44	0.23
	(10.34)	(9.44)	(7.77)	(9.26)	(8.00)	(4.97)
$\mathrm{Ln}\left(\frac{BE}{ME}\right)$	0.44	0.39	0.37	0.42	0.41	0.38
	(6.62)	(6.43)	(6.21)	(7.02)	(6.86)	(5.85)
Ln(ME)	−0.05	−0.11	−0.10	−0.07	−0.07	−0.06
	(−1.17)	(−3.23)	(−2.99)	(−1.82)	(−1.75)	(−1.55)
$\text{return}_{-1,0}$	−4.37	−4.30	−4.30	−4.36	−4.35	−4.28
	(−11.56)	(−11.35)	(−11.38)	(−11.30)	(−11.28)	(−11.30)
$\text{return}_{-12,-1}$	0.47	0.48	0.49	0.56	0.56	0.49
	(3.44)	(3.50)	(3.57)	(3.90)	(3.95)	(3.60)
Intercept	1.43	1.74	1.68	1.50	1.47	1.42
	(3.97)	(5.32)	(5.16)	(4.42)	(4.34)	(3.98)
N (in million)	1.47	1.23	1.23	0.98	0.98	1.51
June 1980 to December 2010						
Delta to industry mean	1.38	2.59	2.03	0.68	0.54	0.22
	(9.82)	(8.44)	(6.56)	(8.54)	(7.18)	(3.63)
$\mathrm{Ln}\left(\frac{BE}{ME}\right)$	0.55	0.50	0.49	0.54	0.52	0.48
	(6.77)	(7.00)	(6.93)	(7.57)	(7.46)	(6.07)
Ln(ME)	−0.10	−0.16	−0.15	−0.11	−0.11	−0.11
	(−1.99)	(−3.58)	(−3.39)	(−2.46)	(−2.37)	(−2.34)
$\text{return}_{-1,0}$	−5.50	−5.42	−5.42	−5.55	−5.54	−5.37
	(−13.09)	(−12.98)	(−13.04)	(−13.44)	(−13.39)	(−12.70)
$\text{return}_{-12,-1}$	0.49	0.49	0.50	0.57	0.58	0.51
	(2.91)	(2.90)	(2.97)	(3.29)	(3.35)	(3.07)
Intercept	1.85	2.09	2.03	1.86	1.82	1.83
	(4.50)	(5.48)	(5.38)	(4.82)	(4.73)	(4.51)
N (in million)	1.16	0.96	0.96	0.75	0.75	1.19
January 2011 to December 2021						
Delta to industry mean	0.95	2.25	1.69	0.28	0.25	0.25
	(3.92)	(4.23)	(4.45)	(4.28)	(4.40)	(4.45)
$\mathrm{Ln}\left(\frac{BE}{ME}\right)$	0.14	0.06	0.03	0.09	0.07	0.10
	(1.33)	(0.59)	(0.31)	(0.88)	(0.69)	(0.92)
Ln(ME)	0.09	0.00	0.01	0.05	0.04	0.07
	(1.29)	(0.03)	(0.10)	(0.74)	(0.69)	(1.08)
$\text{return}_{-1,0}$	−1.15	−1.12	−1.11	−0.97	−1.00	−1.19
	(−1.52)	(−1.44)	(−1.43)	(−1.18)	(−1.19)	(−1.56)
$\text{return}_{-12,-1}$	0.43	0.48	0.48	0.53	0.53	0.44
	(1.88)	(2.04)	(2.04)	(2.08)	(2.07)	(1.92)
Intercept	0.30	0.78	0.73	0.52	0.52	0.31
	(0.40)	(1.25)	(1.14)	(0.84)	(0.75)	(0.42)
N (in million)	0.32	0.27	0.27	0.23	0.23	0.32

Table A.10: Complete Fama-MacBeth Regression Results - Delta SIC Industries

This table shows the complete Fama-MacBeth regression results for the delta between a firm's profitability measure and the mean profitability of the industry, where the industries are defined as the SIC industries (1: agriculture, forestry, and fishing, 2: mining, 3: construction, 4: manufacturing, 5: transportation, communications, electric, gas and sanitary service, 6: wholesale trade, 7: retail trade, 8: finance, insurance and real estate (excluded), 9: services, 10: public administration (excluded), 11: non-classifiable). The first part covers the whole sample (June 1980 to December 2021), and the second and third the time period before and after December 2010. The control variables include value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. All coefficients are in percent, and the values in brackets are the corresponding t-test values.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$
June 1980 to December 2021						
Delta to industry mean	1.13	2.53	1.73	0.68	0.51	0.30
	(8.20)	(9.03)	(5.99)	(8.81)	(6.34)	(5.08)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.42	0.36	0.34	0.40	0.38	0.35
	(6.43)	(6.16)	(6.07)	(6.74)	(6.65)	(5.63)
Ln(ME)	−0.05	−0.12	−0.10	−0.08	−0.08	−0.07
	(−1.19)	(−3.43)	(−3.13)	(−2.24)	(−2.14)	(−1.92)
$\text{return}_{-1,0}$	−4.40	−4.32	−4.33	−4.40	−4.40	−4.32
	(−11.65)	(−11.40)	(−11.54)	(−11.43)	(−11.48)	(−11.42)
$\text{return}_{-12,-1}$	0.47	0.48	0.49	0.54	0.54	0.48
	(3.44)	(3.47)	(3.59)	(3.76)	(3.82)	(3.53)
Intercept	1.41	1.74	1.66	1.54	1.49	1.45
	(3.93)	(5.36)	(5.19)	(4.57)	(4.50)	(4.13)
N (in million)	1.50	1.25	1.25	1.00	1.00	1.54
June 1980 to December 2010						
Delta to industry mean	1.23	2.61	1.74	0.81	0.58	0.27
	(8.28)	(8.29)	(4.90)	(8.14)	(5.57)	(3.60)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.53	0.49	0.47	0.51	0.49	0.46
	(6.69)	(6.87)	(7.13)	(7.32)	(7.42)	(6.22)
Ln(ME)	−0.10	−0.16	−0.14	−0.13	−0.12	−0.11
	(−1.93)	(−3.65)	(−3.29)	(−2.82)	(−2.66)	(−2.46)
$\text{return}_{-1,0}$	−5.54	−5.46	−5.48	−5.61	−5.61	−5.42
	(−13.19)	(−13.09)	(−13.33)	(−13.64)	(−13.72)	(−12.93)
$\text{return}_{-12,-1}$	0.49	0.48	0.50	0.54	0.56	0.51
	(2.92)	(2.89)	(3.03)	(3.17)	(3.26)	(3.05)
Intercept	1.81	2.07	1.97	1.89	1.82	1.81
	(4.39)	(5.43)	(5.28)	(4.91)	(4.79)	(4.48)
N (in million)	1.18	0.97	0.97	0.76	0.76	1.22
January 2011 to December 2021						
Delta to industry mean	0.86	2.33	1.81	0.36	0.33	0.26
	(2.77)	(3.91)	(3.73)	(3.52)	(3.29)	(4.41)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.11	0.02	−0.03	0.09	0.05	0.07
	(1.05)	(0.24)	(−0.33)	(0.85)	(0.52)	(0.66)
Ln(ME)	0.08	−0.01	−0.02	0.04	0.03	0.05
	(1.17)	(−0.27)	(−0.43)	(0.66)	(0.55)	(0.87)
$\text{return}_{-1,0}$	−1.17	−1.09	−1.08	−0.97	−0.98	−1.16
	(−1.54)	(−1.41)	(−1.39)	(−1.16)	(−1.18)	(−1.51)
$\text{return}_{-12,-1}$	0.43	0.47	0.46	0.52	0.50	0.43
	(1.88)	(1.96)	(1.93)	(2.03)	(1.98)	(1.86)
Intercept	0.33	0.84	0.84	0.57	0.59	0.39
	(0.46)	(1.36)	(1.35)	(0.82)	(0.87)	(0.54)
N (in million)	0.32	0.27	0.27	0.23	0.23	0.32

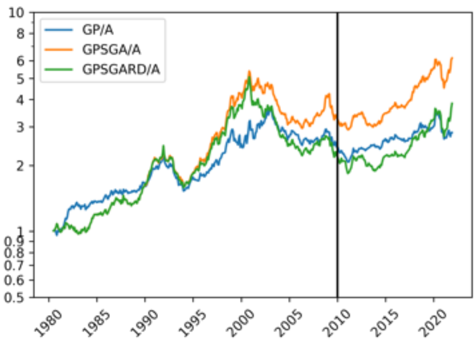
Table A.11: Complete Fama-MacBeth Regression Results - Delta TNIC Similarity

This table shows the complete Fama-MacBeth regression results for the delta between a firm's profitability and the weighted mean profitability of its competitors, where the weights are the Hoberg and Phillips (2010) similarity scores from their Text-based Network Industrial Classification (TNIC). The first part covers the whole sample (June 1988 to December 2021), and the second and third the time period before and after December 2010. Note that the start is at the end of June 1988 instead of June 1980 due to the availability of the TNIC data. The control variables include value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. All coefficients are in percent, and the values in brackets are the corresponding t-test values.

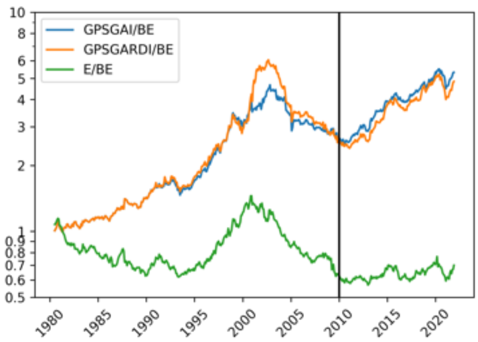
	$\frac{\text{GP}}{\text{A}}$	$\frac{\text{GP-SGA}}{\text{A}}$	$\frac{\text{GP-SGA-RD}}{\text{A}}$	$\frac{\text{GP-SGA-I}}{\text{BE}}$	$\frac{\text{GP-SGA-RD-I}}{\text{BE}}$	$\frac{\text{E}}{\text{BE}}$
			June1988 to December 2021			
Delta to industry mean	1.06	1.89	1.37	0.55	0.38	0.11
	(8.93)	(6.98)	(5.19)	(7.57)	(5.07)	(2.40)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.34	0.27	0.24	0.30	0.28	0.27
	(4.51)	(3.85)	(3.49)	(4.39)	(4.11)	(3.69)
Ln(ME)	−0.06	−0.12	−0.12	−0.09	−0.09	−0.07
	(−1.26)	(−2.85)	(−2.90)	(−2.03)	(−2.04)	(−1.57)
$\text{return}_{-1,0}$	−3.69	−3.54	−3.54	−3.63	−3.62	−3.71
	(−8.35)	(−8.02)	(−8.05)	(−7.99)	(−8.00)	(−8.39)
$\text{return}_{-12,-1}$	0.28	0.31	0.31	0.34	0.34	0.31
	(1.75)	(1.93)	(1.96)	(2.04)	(2.07)	(1.95)
Intercept	1.31	1.64	1.64	1.46	1.46	1.37
	(3.01)	(4.13)	(4.19)	(3.59)	(3.62)	(3.20)
N (in million)	1.04	0.86	0.86	0.69	0.69	1.06
			June1988 to December 2010			
Delta to industry mean	1.21	1.83	1.20	0.67	0.45	0.09
	(8.53)	(5.63)	(3.60)	(6.64)	(4.14)	(1.49)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.42	0.35	0.33	0.40	0.37	0.34
	(4.30)	(3.99)	(3.85)	(4.52)	(4.34)	(3.61)
Ln(ME)	−0.13	−0.18	−0.17	−0.15	−0.15	−0.15
	(−2.18)	(−3.27)	(−3.19)	(−2.70)	(−2.68)	(−2.44)
$\text{return}_{-1,0}$	−4.71	−4.55	−4.55	−4.70	−4.69	−4.69
	(−8.87)	(−8.69)	(−8.74)	(−8.94)	(−8.96)	(−8.84)
$\text{return}_{-12,-1}$	0.30	0.31	0.32	0.33	0.34	0.34
	(1.43)	(1.47)	(1.52)	(1.54)	(1.58)	(1.62)
Intercept	1.81	2.04	1.99	1.87	1.86	1.87
	(3.47)	(4.20)	(4.18)	(3.86)	(3.89)	(3.63)
N (in million)	0.78	0.64	0.64	0.50	0.50	0.80
			January 2011 to December 2021			
Delta to industry mean	0.77	2.02	1.73	0.36	0.29	0.16
	(3.59)	(4.14)	(4.15)	(3.93)	(3.45)	(2.22)
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.18	0.09	0.04	0.12	0.08	0.12
	(1.59)	(0.84)	(0.38)	(1.08)	(0.79)	(1.11)
Ln(ME)	0.09	0.00	−0.01	0.04	0.03	0.07
	(1.22)	(0.02)	(−0.24)	(0.59)	(0.52)	(1.00)
$\text{return}_{-1,0}$	−1.53	−1.42	−1.41	−1.37	−1.37	−1.62
	(−2.00)	(−1.81)	(−1.80)	(−1.63)	(−1.63)	(−2.11)
$\text{return}_{-12,-1}$	0.24	0.31	0.29	0.34	0.33	0.25
	(1.04)	(1.32)	(1.28)	(1.63)	(1.37)	(1.11)
Intercept	0.31	0.85	0.93	0.63	0.64	0.38
	(0.40)	(1.23)	(1.37)	(0.85)	(0.88)	(0.49)
N (in million)	0.26	0.22	0.22	0.19	0.19	0.26

Figure A.1: Portfolios Sorted on Delta to Mean of FF12 Industries

These figures show the developments of the portfolios sorted on the delta between a firm's profitability and the industry mean according to the Fama-French 12 industries between June 1980 and December 2021. Subfigure (a) displays the profitability measures which are scaled by total assets, and Subfigure (b) the ones that are scaled by the book value of equity. Table A.12 summarizes the corresponding mean, geomean and Fama-French three-factor alpha.



(a) Profitability measures scaled by assets



(b) Profitability measures scaled by book equity

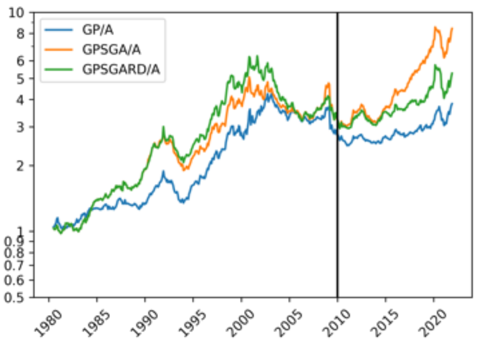
Table A.12: Portfolio Results - Delta to Mean of FF12 Industries

This table shows the results of the portfolios sorted on the delta to the industry mean profitability, where industries are defined as the Frama-French 12 industries. The first part covers the whole sample (June 1980 to December 2021), and the second and third the time period before and after December 2010. The percentage values are percent per month, and the values in brackets are the corresponding t-test values. FF3F α stands for Fama-French three-factor model alpha.

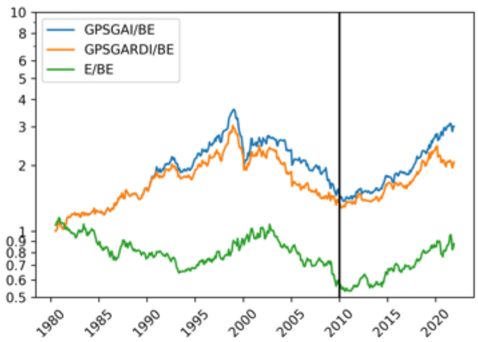
	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$
June 1980 to December 2021						
Mean	0.24%	0.41%	0.31%	0.37%	0.36%	−0.04%
	(2.07)	(3.13)	(2.37)	(3.21)	(2.83)	(−0.32)
Geomean	0.21%	0.37%	0.27%	0.34%	0.32%	−0.07%
	(1.77)	(2.79)	(2.04)	(2.90)	(2.52)	(−0.61)
FF3F α	0.31%	0.52%	0.39%	0.40%	0.38%	−0.01%
	(2.57)	(3.99)	(2.96)	(3.41)	(2.92)	(−0.05)
June 1980 to December 2010						
Mean	0.25%	0.35%	0.23%	0.29%	0.29%	−0.11%
	(1.72)	(2.21)	(1.47)	(2.08)	(1.85)	(−0.77)
Geomean	0.21%	0.30%	0.18%	0.26%	0.24%	−0.15%
	(1.45)	(1.92)	(1.18)	(1.81)	(1.58)	(−1.03)
FF3F α	0.29%	0.48%	0.33%	0.36%	0.33%	−0.03%
	(2.00)	(3.08)	(2.11)	(2.50)	(2.08)	(−0.23)
January 2011 to December 2021						
Mean	0.21%	0.61%	0.55%	0.54%	0.59%	0.19%
	(1.13)	(2.75)	(2.35)	(2.93)	(2.87)	(0.98)
Geomean	0.19%	0.58%	0.52%	0.51%	0.57%	0.17%
	(0.99)	(2.59)	(2.21)	(2.81)	(2.74)	(0.85)
FF3F α	0.20%	0.58%	0.46%	0.47%	0.52%	0.14%
	(1.02)	(2.47)	(1.85)	(2.48)	(2.41)	(0.68)

Figure A.2: Portfolios Sorted on Delta to Mean of FF17 Industries

These figures show the developments of the portfolios sorted on the delta between a firm's profitability and the industry mean according to the Fama-French 17 industries between June 1980 and December 2021. Subfigure (a) displays the profitability measures which are scaled by total assets, and Subfigure (b) the ones that are scaled by the book value of equity. Table A.13 summarizes the corresponding mean, geomean and Fama-French three-factor alpha.



(a) Profitability measures scaled by assets



(b) Profitability measures scaled by book equity

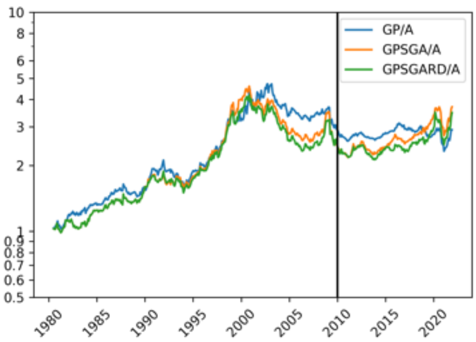
Table A.13: Portfolio Results - Delta to Mean of FF17 Industries

This table shows the results of the portfolios sorted on the delta to the industry mean profitability, where the industries are defined as the Frama-French 17 industries. The first part covers the whole sample (June 1980 to December 2021), and the second and third the time period before and after December 2010. The percentage values are percent per month, and the values in brackets are the corresponding t-test values. FF3F α stands for Fama-French three-factor model alpha.

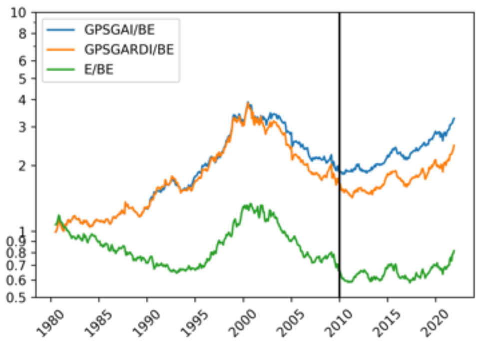
	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$
			June 1980 to December 2021			
Mean	0.31%	0.48%	0.38%	0.26%	0.18%	0.01%
	(2.55)	(3.45)	(2.69)	(2.16)	(1.55)	(0.06)
Geomean	0.27%	0.43%	0.33%	0.22%	0.15%	−0.03%
	(2.21)	(3.08)	(2.33)	(1.84)	(1.25)	(−0.23)
FF3F α	0.38%	0.60%	0.48%	0.25%	0.18%	0.04%
	(3.11)	(4.35)	(3.37)	(2.06)	(1.50)	(0.33)
			June 1980 to December 2010			
Mean	0.29%	0.35%	0.35%	0.13%	0.12%	−0.14%
	(1.95)	(2.15)	(2.07)	(0.92)	(0.86)	(−1.01)
Geomean	0.25%	0.30%	0.30%	0.09%	0.08%	−0.17%
	(1.63)	(1.82)	(1.76)	(0.63)	(0.60)	(−1.26)
FF3F α	0.37%	0.50%	0.45%	0.19%	0.19%	−0.07%
	(2.44)	(3.06)	(2.64)	(1.25)	(1.33)	(−0.53)
			January 2011 to December 2021			
Mean	0.34%	0.85%	0.50%	0.63%	0.37%	0.45%
	(2.10)	(3.46)	(1.90)	(3.40)	(1.74)	(2.14)
Geomean	0.32%	0.81%	0.45%	0.61%	0.34%	0.42%
	(2.00)	(3.33)	(1.74)	(3.30)	(1.59)	(1.99)
FF3F α	0.32%	0.86%	0.50%	0.52%	0.27%	0.45%
	(1.87)	(3.29)	(1.77)	(2.75)	(1.27)	(1.99)

Figure A.3: Portfolios Sorted on Delta to Mean of FF30 Industries

These figures show the developments of the portfolios sorted on the delta between a firm's profitability and the industry mean according to the Fama-French 30 industries between June 1980 and December 2021. Subfigure (a) displays the profitability measures which are scaled by total assets, and Subfigure (b) the ones that are scaled by the book value of equity. Table A.14 summarizes the corresponding mean, geomean and Fama-French three-factor alpha.



(a) Profitability measures scaled by assets



(b) Profitability measures scaled by book equity

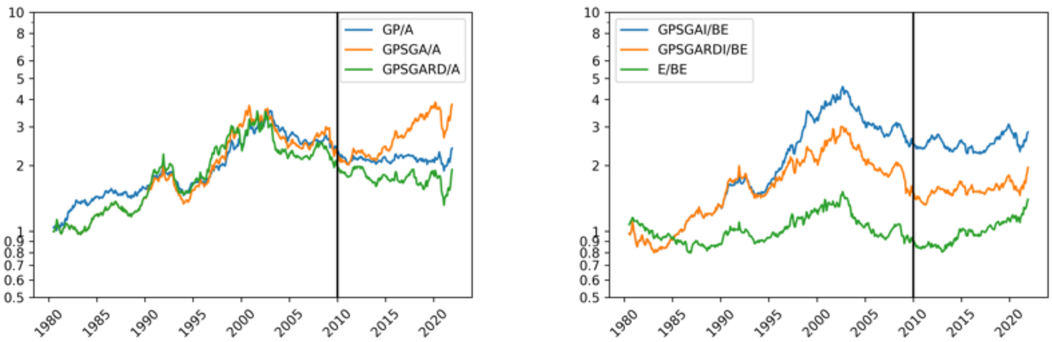
Table A.14: Portfolio Results - Delta to Mean of FF30 Industries

This table shows the results of the portfolios sorted on the delta to the industry mean profitability, where the industries are defined as the Frama-French 30 industries. The first part covers the whole sample (June 1980 to December 2021), and the second and third the time period before and after December 2010. The percentage values are percent per month, and the values in brackets are the corresponding t-test values. FF3F α stands for Fama-French three-factor model alpha.

	$\frac{GP}{A}$	$\frac{GP-SGA}{A}$	$\frac{GP-SGA-RD}{A}$	$\frac{GP-SGA-I}{BE}$	$\frac{GP-SGA-RD-I}{BE}$	$\frac{E}{BE}$
			June 1980 to December 2021			
Mean	0.25%	0.30%	0.29%	0.27%	0.22%	−0.01%
	(2.02)	(2.41)	(2.27)	(2.40)	(1.81)	(−0.10)
Geomean	0.21%	0.26%	0.25%	0.24%	0.18%	−0.04%
	(1.69)	(2.07)	(1.95)	(2.12)	(1.51)	(−0.38)
FF3F α	0.34%	0.42%	0.39%	0.32%	0.27%	0.01%
	(2.67)	(3.31)	(3.02)	(2.79)	(2.20)	(0.07)
			June 1980 to December 2010			
Mean	0.31%	0.26%	0.26%	0.29%	0.15%	−0.12%
	(1.95)	(1.71)	(1.74)	(2.26)	(1.01)	(−0.89)
Geomean	0.27%	0.22%	0.22%	0.26%	0.11%	−0.15%
	(1.64)	(1.42)	(1.46)	(2.03)	(0.74)	(−1.13)
FF3F α	0.41%	0.40%	0.38%	0.35%	0.25%	−0.05%
	(2.55)	(2.63)	(2.52)	(2.72)	(1.72)	(−0.40)
			January 2011 to December 2021			
Mean	0.11%	0.46%	0.41%	0.47%	0.46%	0.23%
	(0.67)	(2.24)	(1.68)	(2.73)	(2.38)	(1.18)
Geomean	0.09%	0.44%	0.37%	0.45%	0.44%	0.21%
	(0.56)	(2.10)	(1.55)	(2.64)	(2.26)	(1.06)
FF3F α	0.07%	0.44%	0.37%	0.48%	0.44%	0.24%
	(0.41)	(2.04)	(1.42)	(2.64)	(2.13)	(1.17)

Figure A.4: Portfolios Sorted on Delta to Mean of SIC Industries

These figures show the developments of the portfolios sorted on the delta between a firm's profitability and the industry mean according to the first-level SIC industries between June 1980 and December 2021. Subfigure (a) displays the profitability measures which are scaled by total assets, and Subfigure (b) the ones that are scaled by the book value of equity. Table A.15 summarizes the corresponding mean, geomean and Fama-French three-factor alpha.



(a) Profitability measures scaled by assets (b) Profitability measures scaled by book equity

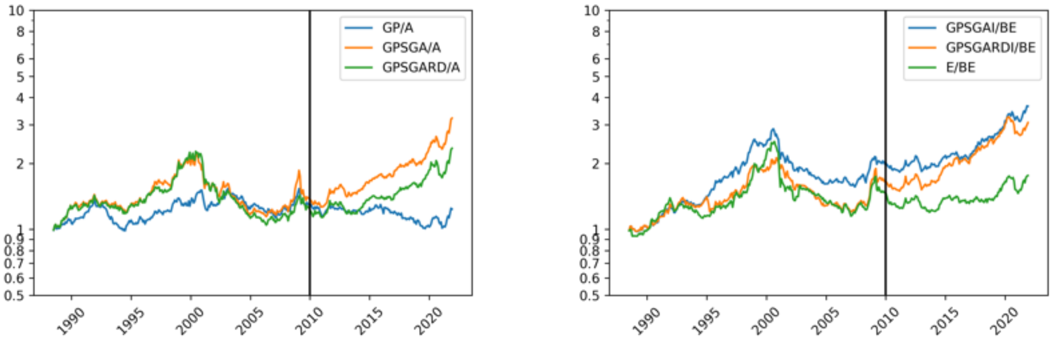
Table A.15: Portfolio Results - Delta to Mean of SIC Industries

This table shows the results of the portfolios sorted on the delta to the industry mean profitability, where industries are defined as the first-level SIC industries. The first part covers the whole sample (June 1980 to December 2021), and the second and third the time period before and after December 2010. The percentage values are percent per month, and the values in brackets are the corresponding t-test values. FF3F α stands for Fama-French three-factor model alpha.

	$\frac{GP}{A}$	$\frac{GP\text{-}SGA}{A}$	$\frac{GP\text{-}SGA\text{-}RD}{A}$	$\frac{GP\text{-}SGA\text{-}I}{BE}$	$\frac{GP\text{-}SGA\text{-}RD\text{-}I}{BE}$	$\frac{E}{BE}$
			June 1980 to December 2021			
Mean	0.20%	0.31%	0.18%	0.24%	0.17%	0.10%
	(1.93)	(2.36)	(1.28)	(2.11)	(1.39)	(0.89)
Geomean	0.18%	0.27%	0.13%	0.21%	0.13%	0.07%
	(1.67)	(2.02)	(0.93)	(1.82)	(1.08)	(0.62)
FF3F α	0.25%	0.41%	0.24%	0.29%	0.22%	0.10%
	(2.39)	(3.07)	(1.67)	(2.50)	(1.69)	(0.94)
			June 1980 to December 2010			
Mean	0.23%	0.24%	0.21%	0.27%	0.13%	−0.02%
	(1.85)	(1.56)	(1.28)	(1.98)	(0.82)	(−0.14)
Geomean	0.20%	0.20%	0.16%	0.24%	0.08%	−0.05%
	(1.61)	(1.27)	(0.98)	(1.73)	(0.54)	(−0.38)
FF3F α	0.30%	0.34%	0.24%	0.35%	0.17%	−0.00%
	(2.33)	(2.23)	(1.44)	(2.47)	(1.09)	(−0.01)
			January 2011 to December 2021			
Mean	0.14%	0.56%	0.13%	0.14%	0.23%	0.40%
	(0.80)	(2.24)	(0.47)	(0.68)	(1.10)	(2.17)
Geomean	0.12%	0.52%	0.08%	0.11%	0.20%	0.38%
	(0.68)	(2.06)	(0.30)	(0.54)	(0.97)	(2.06)
FF3F α	0.09%	0.52%	0.04%	0.05%	0.27%	0.31%
	(0.47)	(1.97)	(0.15)	(0.24)	(1.26)	(1.57)

Figure A.5: Portfolios Sorted on Delta to Mean of TNIC Industries

These figures show the developments of the portfolios sorted on the delta between a firm's profitability and the weighted mean according to the Hoberg and Phillips (2010) similarity scores from their TNIC. The sample period starts in June 1988 and ends in December 2021 (the start is not in June 1980 because the TNIC is not available before). Subfigure (a) displays the profitability measures which are scaled by total assets, and Subfigure (b) the ones that are scaled by the book value of equity. Table A.16 summarizes the corresponding mean, geomean and Fama-French three-factor alpha.



(a) Profitability measures scaled by assets

(b) Profitability measures scaled by book equity

Table A.16: Portfolio Results - Delta to Mean of TNIC Industries

This table shows the results of the portfolios sorted on the delta between a firm's profitability and the weighted mean according to the Hoberg and Phillips (2010) similarity scores from their TNIC. The first part covers the whole sample (January 1988 to December 2021), and the second and third the time period before and after December 2010. Note that the start is at the end of June 1988 instead of June 1980 due to the availability of the TNIC data. The percentage values are percent per month, and the values in brackets are the corresponding t-test values. FF3F α stands for Fama-French three-factor model alpha.

	$\frac{\text{GP}}{\text{A}}$	$\frac{\text{GP-SGA}}{\text{A}}$	$\frac{\text{GP-SGA-RD}}{\text{A}}$	$\frac{\text{GP-SGA-I}}{\text{BE}}$	$\frac{\text{GP-SGA-RD-I}}{\text{BE}}$	$\frac{\text{E}}{\text{BE}}$
January 1988 to December 2021						
Mean	0.08%	0.33%	0.25%	0.27%	0.31%	0.18%
	(0.69)	(2.28)	(1.78)	(2.37)	(2.40)	(1.28)
Geomean	0.05%	0.29%	0.21%	0.24%	0.28%	0.14%
	(0.46)	(1.97)	(1.49)	(2.16)	(2.13)	(0.97)
FF3F α	0.13%	0.43%	0.32%	0.29%	0.40%	0.26%
	(1.16)	(2.92)	(2.20)	(2.48)	(3.09)	(1.84)
January 1988 to December 2010						
Mean	0.09%	0.14%	0.09%	0.26%	0.19%	0.14%
	(0.62)	(0.71)	(0.52)	(1.72)	(1.14)	(0.72)
Geomean	0.06%	0.09%	0.05%	0.23%	0.15%	0.09%
	(0.43)	(0.44)	(0.27)	(1.51)	(0.92)	(0.44)
FF3F α	0.15%	0.27%	0.21%	0.35%	0.33%	0.29%
	(1.02)	(1.44)	(1.17)	(2.31)	(2.02)	(1.54)
January 2011 to December 2021						
Mean	0.07%	0.70%	0.58%	0.49%	0.52%	0.36%
	(0.41)	(3.42)	(2.86)	(2.56)	(2.57)	(2.08)
Geomean	0.05%	0.67%	0.55%	0.47%	0.49%	0.34%
	(0.30)	(3.32)	(2.75)	(2.44)	(2.43)	(1.97)
FF3F α	0.06%	0.64%	0.49%	0.36%	0.41%	0.34%
	(0.32)	(2.94)	(2.28)	(1.77)	(1.91)	(1.84)

Table A.17: Fama-MacBeth Regressions - GPSGA/A Decomposition 1 - Complete Results

This table shows the coefficients of the Fama-MacBeth regressions on GPSGA/A and its components according to the decomposition of Equation 5. All coefficients are in percent, the values in brackets the corresponding t-test values, and the values in square brackets the VIFs, which are a test for multicollinearity. The regressions control for value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. Only observations not missing any of the variables are considered; hence, n is 1.25 million for each regression.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
GPSGA/A	2.65					1.57				
	(9.06)					(5.01)				
	[1.58]					[1.88]				
GP/A		1.80				1.28	1.76	1.77	2.72	2.66
		(10.83)				(7.44)	(11.91)	(11.63)	(9.99)	(10.92)
		[1.77]				[2.78]	[3.29]	[2.10]	[8.13]	[9.51]
REV/A			0.33				0.01			0.04
			(6.16)				(0.25)			(0.70)
			[1.94]				[3.60]			[4.53]
COGS/A				0.20				0.02		
				(3.59)				(0.33)		
				[1.69]				[1.99]		
SGA/A					0.96				−1.40	−1.43
					(5.82)				(−5.06)	(−5.09)
					[1.72]				[5.97]	[6.20]
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.37	0.52	0.40	0.38	0.48	0.46	0.52	0.52	0.46	0.45
	(6.22)	(8.36)	(6.55)	(6.45)	(8.00)	(7.84)	(8.55)	(8.55)	(7.80)	(7.98)
	[1.72]	[1.63]	[1.62]	[1.62]	[1.80]	[1.80]	[1.63]	[1.63]	[1.81]	[1.89]
Ln(ME)	−0.12	−0.02	−0.03	−0.04	−0.00	−0.08	−0.02	−0.02	−0.08	−0.08
	(−3.55)	(−0.56)	(−0.71)	(−0.89)	(−0.05)	(−2.29)	(−0.56)	(−0.56)	(−2.18)	(−2.17)
	[2.36]	[2.33]	[2.58]	[2.36]	[2.14]	[2.91]	[2.64]	[2.64]	[2.89]	[3.18]
$\text{return}_{-1,0}$	−4.34	−4.41	−4.35	−4.30	−4.35	−4.47	−4.45	−4.45	−4.46	−4.50
	(−11.48)	(−11.67)	(−11.46)	(−11.31)	(−11.41)	(−11.90)	(−11.80)	(−11.81)	(−11.87)	(−12.02)
	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]
$\text{return}_{-12,-1}$	0.47	0.48	0.52	0.54	0.55	0.43	0.47	0.47	0.44	0.43
	(3.46)	(3.52)	(3.74)	(3.92)	(3.93)	(3.20)	(3.46)	(3.45)	(3.23)	(3.16)
	[1.09]	[1.08]	[1.07]	[1.07]	[1.08]	[1.09]	[1.08]	[1.08]	[1.09]	[1.09]
Intercept	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	(4.47)	(1.73)	(2.45)	(3.22)	(2.73)	(2.77)	(1.68)	(1.68)	(2.73)	(2.60)
R^2	3.52%	3.51%	3.44%	3.34%	3.35%	3.81%	3.67%	3.67%	3.79%	3.96%
adj. R^2	3.32%	3.31%	3.23%	3.13%	3.15%	3.56%	3.43%	3.43%	3.55%	3.68%

	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
GPSGA/A	2.40	2.61	2.88	1.70	2.81	2.86				
	(8.80)	(9.21)	(9.76)	(5.45)	(10.53)	(10.13)				
	[1.63]	[1.59]	[1.61]	[1.92]	[1.86]	[1.65]				
GP/A										
REV/A	0.21			1.09	0.05		1.61	0.26	2.35	
	(4.38)			(7.24)	(0.92)		(11.53)	(4.38)	(10.24)	
	[2.14]			[32.12]	[4.47]		[16.05]	[3.87]	[67.06]	
COGS/A		0.15		−1.04		0.04	−1.57		−2.29	0.13
		(2.78)		(−6.54)		(0.86)	(−11.59)		(−10.70)	(2.31)
		[1.73]		[26.01]		[2.41]	[13.92]		[40.69]	[2.35]
SGA/A			1.20		1.10	1.15		0.52	−1.22	0.85
			(7.28)		(6.59)	(7.20)		(2.87)	(−4.55)	(5.19)
			[1.75]		[3.66]	[2.43]		[3.21]	[5.92]	[2.34]
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.36	0.35	0.46	0.44	0.45	0.45	0.51	0.44	0.45	0.46
	(6.07)	(6.03)	(7.80)	(7.84)	(7.96)	(7.97)	(8.44)	(7.81)	(8.01)	(8.05)
	[1.72]	[1.73]	[1.81]	[1.88]	[1.89]	[1.89]	[1.63]	[1.89]	[1.89]	[1.89]
Ln(ME)	−0.11	−0.12	−0.08	−0.08	−0.08	−0.08	−0.02	−0.01	−0.07	−0.00
	(−3.09)	(−3.37)	(−2.37)	(−2.39)	(−2.36)	(−2.36)	(−0.55)	(−0.28)	(−1.93)	(−0.09)
	[3.20]	[3.11]	[2.91]	[3.20]	[3.21]	[3.22]	[2.62]	[2.82]	[3.09]	[2.60]
$\text{return}_{-1,0}$	−4.42	−4.40	−4.46	−4.50	−4.50	−4.50	−4.44	−4.42	−4.49	−4.40
	(−11.74)	(−11.69)	(−11.88)	(−12.02)	(−12.04)	(−12.04)	(−11.78)	(−11.74)	(−11.99)	(−11.63)
	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]
$\text{return}_{-12,-1}$	0.43	0.45	0.43	0.42	0.42	0.42	0.48	0.51	0.44	0.53
	(3.21)	(3.29)	(3.21)	(3.12)	(3.12)	(3.12)	(3.48)	(3.70)	(3.23)	(3.82)
	[1.09]	[1.09]	[1.09]	[1.09]	[1.09]	[1.09]	[1.08]	[1.08]	[1.09]	[1.08]
Intercept	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	(3.39)	(3.88)	(2.85)	(2.76)	(2.70)	(2.71)	(1.77)	(2.21)	(2.64)	(2.44)
R^2	3.78%	3.74%	3.80%	3.96%	3.97%	3.97%	3.66%	3.64%	3.93%	3.56%
adj. R^2	3.54%	3.50%	3.56%	3.68%	3.69%	3.69%	3.41%	3.40%	3.64%	3.32%

Table A.18: Fama-MacBeth Regressions - GPSGA/A Decomposition 1 - Complete Results (Excluding Micro-Caps)

This table shows the coefficients of the Fama-MacBeth regressions on GPSGA/A and its components according to the decomposition of Equation 5. All coefficients are in percent, the values in brackets the corresponding t-test values, and the values in square brackets the VIFs, which are a test for multicollinearity. The regressions control for value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. The sample excludes micro-caps, which are defined as stocks with a market capitalization below the 20th percentile of the NYSE market capitalization distribution. Only observations not missing any of the variables are considered; hence, n is 0.44 million for each regression.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
GPSGA/A	2.56					1.84				
	(6.22)					(3.93)				
	[3.74]					[6.20]				
GP/A		1.18				0.63	1.17	1.17	2.33	2.26
		(5.49)				(2.61)	(5.98)	(5.82)	(6.17)	(6.35)
		[2.77]				[6.01]	[4.52]	[3.16]	[18.15]	[19.06]
REV/A			0.19				0.00			0.05
			(3.16)				(0.08)			(0.83)
			[2.58]				[4.19]			[4.60]
COGS/A				0.11				0.00		
				(1.84)				(0.07)		
				[1.99]				[2.26]		
SGA/A					0.78				−1.65	−1.67
					(3.23)				(−3.90)	(−3.82)
					[2.23]				[11.16]	[11.41]
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.29	0.31	0.16	0.14	0.22	0.33	0.30	0.30	0.33	0.31
	(3.80)	(4.09)	(2.19)	(1.98)	(2.98)	(4.26)	(4.08)	(4.08)	(4.25)	(4.18)
	[2.91]	[2.45]	[2.25]	[2.24]	[2.79]	[2.99]	[2.49]	[2.49]	[2.99]	[3.10]
Ln(ME)	−0.03	0.01	0.00	−0.00	0.01	−0.02	0.01	0.01	−0.02	−0.01
	(−0.67)	(0.18)	(0.08)	(−0.02)	(0.19)	(−0.41)	(0.20)	(0.20)	(−0.39)	(−0.35)
	[4.10]	[3.15]	[3.52]	[3.21]	[3.16]	[4.31]	[3.62]	[3.63]	[4.24]	[4.89]
$\text{return}_{-1,0}$	−1.65	−1.76	−1.70	−1.66	−1.76	−1.84	−1.82	−1.83	−1.84	−1.91
	(−3.30)	(−3.58)	(−3.43)	(−3.33)	(−3.57)	(−3.75)	(−3.74)	(−3.74)	(−3.75)	(−3.92)
	[1.02]	[1.02]	[1.02]	[1.02]	[1.03]	[1.03]	[1.02]	[1.02]	[1.03]	[1.03]
$\text{return}_{-12,-1}$	0.53	0.50	0.49	0.50	0.50	0.49	0.48	0.49	0.49	0.47
	(3.01)	(2.93)	(2.83)	(2.90)	(2.92)	(2.87)	(2.84)	(2.84)	(2.89)	(2.78)
	[1.20]	[1.21]	[1.21]	[1.21]	[1.21]	[1.21]	[1.21]	[1.21]	[1.21]	[1.21]
Intercept	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	(2.02)	(1.35)	(1.71)	(2.07)	(1.83)	(1.59)	(1.30)	(1.30)	(1.59)	(1.47)
R^2	5.90%	6.08%	5.96%	5.79%	6.01%	6.61%	6.48%	6.48%	6.58%	6.98%
adj. R^2	5.35%	5.53%	5.40%	5.23%	5.45%	5.95%	5.81%	5.81%	5.91%	6.21%

	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
GPSGA/A	2.38	2.52	2.48	1.92	2.40	2.45				
	(6.20)	(6.25)	(6.13)	(4.02)	(6.28)	(6.12)				
	[4.12]	[3.81]	[3.95]	[6.22]	[4.13]	[3.96]				
GP/A										
REV/A	0.12			0.54	0.05		1.13	0.12	2.11	
	(2.06)			(2.38)	(0.84)		(5.81)	(2.01)	(6.10)	
	[3.05]			[55.96]	[4.58]		[26.16]	[4.39]	[145.14]	
COGS/A		0.09		−0.49		0.05	−1.12		−2.07	0.05
		(1.59)		(−2.06)		(0.84)	(−5.91)		(−6.26)	(0.89)
		[2.06]		[37.94]		[2.41]	[20.21]		[79.60]	[2.40]
SGA/A			0.61		0.52	0.57		0.52	−1.48	0.72
			(2.60)		(2.22)	(2.48)		(2.18)	(−3.53)	(3.13)
			[2.35]		[3.54]	[2.75]		[3.53]	[11.18]	[2.64]
$\text{Ln}\left(\frac{BE}{ME}\right)$	0.27	0.27	0.33	0.31	0.31	0.31	0.29	0.19	0.30	0.20
	(3.57)	(3.57)	(4.24)	(4.14)	(4.16)	(4.17)	(4.01)	(2.76)	(4.07)	(2.85)
	[2.92]	[2.95]	[2.99]	[3.11]	[3.10]	[3.10]	[2.48]	[2.84]	[3.10]	[2.88]
Ln(ME)	−0.02	−0.02	−0.02	−0.02	−0.01	−0.02	0.01	0.01	−0.01	0.01
	(−0.50)	(−0.57)	(−0.42)	(−0.38)	(−0.37)	(−0.38)	(0.21)	(0.19)	(−0.28)	(0.22)
	[4.95]	[4.97]	[4.31]	[4.97]	[4.96]	[4.98]	[3.60]	[4.19]	[4.76]	[3.91]
$\text{return}_{-1,0}$	−1.77	−1.74	−1.84	−1.91	−1.91	−1.91	−1.83	−1.84	−1.90	−1.83
	(−3.60)	(−3.52)	(−3.75)	(−3.92)	(−3.92)	(−3.92)	(−3.74)	(−3.76)	(−3.91)	(−3.75)
	[1.02]	[1.02]	[1.03]	[1.03]	[1.03]	[1.03]	[1.02]	[1.03]	[1.03]	[1.03]
$\text{return}_{-12,-1}$	0.48	0.49	0.49	0.47	0.47	0.47	0.48	0.47	0.47	0.48
	(2.81)	(2.86)	(2.87)	(2.77)	(2.77)	(2.77)	(2.83)	(2.79)	(2.77)	(2.82)
	[1.21]	[1.21]	[1.21]	[1.21]	[1.21]	[1.21]	[1.21]	[1.21]	[1.21]	[1.21]
Intercept	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	(1.60)	(1.74)	(1.59)	(1.47)	(1.47)	(1.47)	(1.31)	(1.59)	(1.49)	(1.71)
R^2	6.49%	6.40%	6.61%	7.01%	7.01%	7.02%	6.47%	6.49%	6.96%	6.43%
adj. R^2	5.82%	5.74%	5.94%	6.24%	6.24%	6.24%	5.80%	5.83%	6.18%	5.76%

Table A.19: Fama-MacBeth Regressions - GPSGA/A Decomposition 2 - Complete Results

This table shows the coefficients of the Fama-MacBeth regressions on GPSGA/A and its components according to the decomposition of Equation 6. All coefficients are in percent, the values in brackets the corresponding t-test values, and the values in square brackets the VIFs, which are a test for multicollinearity. The regressions control for value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. Only observations not missing any of the variables are considered; hence, n is 1.23 million for each regression.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
GPSGA/A										
GP/A										
REV/A					0.28	0.34	0.28	0.31	0.29	0.40
					(5.58)	(6.49)	(5.59)	(6.08)	(5.79)	(7.58)
					[2.10]	[2.00]	[2.12]	[2.13]	[2.12]	[2.13]
GPSGA/REV	0.61				0.62				0.51	1.49
	(5.51)				(5.66)				(3.48)	(9.94)
	[1.10]				[1.10]				[1.26]	[2.87]
GP/REV		0.24		0.15		0.39		0.33	0.22	
		(3.41)		(1.74)		(5.06)		(3.72)	(2.08)	
		[1.02]		[1.45]		[1.04]		[1.46]	[1.67]	
SGA/REV			−0.36	−0.31			−0.15	−0.12		1.33
			(−2.59)	(−1.88)			(−1.10)	(−0.78)		(6.56)
			[1.49]	[1.49]			[1.50]	[1.50]		[3.91]
$\mathrm{Ln}\left(\frac{BE}{ME}\right)$	0.34	0.41	0.34	0.35	0.34	0.41	0.36	0.38	0.36	0.44
	(5.71)	(6.66)	(6.08)	(6.28)	(5.70)	(6.74)	(6.37)	(6.79)	(6.03)	(7.98)
	[1.75]	[1.66]	[1.82]	[1.82]	[1.75]	[1.66]	[1.82]	[1.82]	[1.77]	[1.83]
Ln(ME)	−0.09	−0.05	−0.07	−0.07	−0.07	−0.03	−0.05	−0.05	−0.07	−0.05
	(−2.36)	(−1.20)	(−1.91)	(−1.91)	(−1.95)	(−0.82)	(−1.31)	(−1.27)	(−1.87)	(−1.40)
	[1.90]	[1.68]	[1.96]	[2.35]	[2.99]	[2.59]	[2.95]	[3.27]	[3.18]	[4.23]
$\mathrm{return}_{-1,0}$	−4.29	−4.25	−4.30	−4.32	−4.38	−4.37	−4.39	−4.41	−4.40	−4.47
	(−11.16)	(−11.00)	(−11.23)	(−11.32)	(−11.48)	(−11.40)	(−11.52)	(−11.64)	(−11.57)	(−11.81)
	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]
$\mathrm{return}_{-12,-1}$	0.54	0.58	0.55	0.55	0.49	0.52	0.50	0.50	0.49	0.48
	(3.84)	(4.09)	(3.93)	(3.92)	(3.52)	(3.71)	(3.62)	(3.61)	(3.51)	(3.48)
	[1.08]	[1.08]	[1.08]	[1.08]	[1.08]	[1.08]	[1.08]	[1.08]	[1.08]	[1.08]
Intercept	0.02	0.01	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.00
	(4.64)	(3.82)	(5.07)	(4.80)	(3.16)	(2.07)	(3.27)	(2.74)	(2.87)	(1.33)
R^2	3.43%	3.30%	3.45%	3.62%	3.72%	3.64%	3.73%	3.90%	3.88%	3.95%
adj. R^2	3.22%	3.09%	3.25%	3.37%	3.48%	3.39%	3.48%	3.61%	3.60%	3.66%

	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
GPSGA/A	2.49	2.37	2.39	2.85	2.80					
	(8.80)	(7.40)	(8.23)	(9.68)	(9.27)					
	[1.63]	[3.23]	[1.80]	[2.32]	[2.48]					
GP/A						1.76	1.50	1.82	1.77	1.91
						(11.72)	(9.36)	(10.95)	(11.94)	(11.84)
						[3.29]	[5.29]	[4.13]	[5.26]	[6.36]
REV/A	0.21	0.21	0.23	0.25	0.25	0.01	0.03	−0.00	−0.05	−0.08
	(4.47)	(4.45)	(4.68)	(5.08)	(5.17)	(0.30)	(0.66)	(−0.03)	(−1.08)	(−1.84)
	[2.14]	[2.22]	[2.15]	[2.24]	[2.24]	[3.60]	[4.55]	[3.81]	[4.42]	[4.77]
GPSGA/REV		−0.01					0.35			
		(−0.06)					(2.85)			
		[2.08]					[1.17]			
GP/REV			0.16		0.05			−0.09		−0.16
			(1.95)		(0.48)			(−0.83)		(−1.42)
			[1.53]		[1.56]			[1.29]		[1.76]
SGA/REV				0.48	0.50				−0.35	−0.38
				(3.39)	(2.78)				(−2.43)	(−2.49)
				[2.03]	[2.07]				[1.59]	[1.61]
$\mathrm{Ln}\left(\frac{BE}{ME}\right)$	0.37	0.36	0.38	0.42	0.42	0.52	0.47	0.52	0.47	0.47
	(6.20)	(6.03)	(6.36)	(7.36)	(7.46)	(8.64)	(8.08)	(8.58)	(8.39)	(8.32)
	[1.72]	[1.76]	[1.72]	[1.84]	[1.84]	[1.63]	[1.91]	[1.71]	[1.90]	[1.90]
Ln(ME)	−0.11	−0.11	−0.11	−0.10	−0.10	−0.02	−0.05	−0.02	−0.05	−0.05
	(−3.07)	(−3.13)	(−3.01)	(−2.86)	(−2.81)	(−0.51)	(−1.31)	(−0.51)	(−1.24)	(−1.25)
	[3.20]	[3.26]	[3.44]	[3.93]	[4.03]	[2.64]	[3.06]	[2.65]	[3.04]	[3.27]
$\mathrm{return}_{-1,0}$	−4.43	−4.43	−4.44	−4.48	−4.48	−4.45	−4.48	−4.46	−4.48	−4.49
	(−11.62)	(−11.65)	(−11.68)	(−11.81)	(−11.88)	(−11.65)	(−11.82)	(−11.69)	(−11.83)	(−11.88)
	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]	[1.01]
$\mathrm{return}_{-12,-1}$	0.44	0.43	0.43	0.42	0.42	0.48	0.46	0.47	0.46	0.45
	(3.16)	(3.12)	(3.17)	(3.09)	(3.09)	(3.44)	(3.30)	(3.42)	(3.31)	(3.29)
	[1.09]	[1.09]	[1.09]	[1.09]	[1.09]	[1.08]	[1.08]	[1.08]	[1.09]	[1.09]
Intercept	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	(3.33)	(3.38)	(3.10)	(2.67)	(2.55)	(1.66)	(2.21)	(1.70)	(2.58)	(2.71)
R^2	3.82%	3.96%	3.96%	4.03%	4.18%	3.70%	3.95%	3.86%	3.96%	4.10%
adj. R^2	3.57%	3.67%	3.67%	3.74%	3.85%	3.46%	3.66%	3.57%	3.67%	3.77%

Table A.20: Fama-MacBeth Regressions - GPSGA/A Decomposition 2 - Complete Results (Excluding Micro-Caps)

This table shows the coefficients of the Fama-MacBeth regressions on GPSGA/A and its components according to the decomposition of Equation 6. All coefficients are in percent, the values in brackets the corresponding t-test values, and the values in square brackets the VIFs, which are a test for multicollinearity. The regressions control for value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. The sample excludes micro-caps, which are defined as stocks with a market capitalization below the 20th percentile of the NYSE market capitalization distribution. Only observations not missing any of the variables are considered; hence, n is 0.43 million for each regression.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
GPSGA/A										
GP/A										
REV/A					0.21 (3.42) [2.93]	0.26 (3.93) [2.61]	0.17 (2.85) [2.81]	0.24 (3.75) [2.88]	0.26 (4.01) [2.94]	0.29 (4.37) [3.02]
GPSGA/REV	0.41 (2.12) [1.78]				0.67 (3.29) [1.87]				0.42 (1.50) [2.70]	1.18 (5.25) [2.51]
GP/REV		0.34 (2.34) [1.06]		0.26 (1.46) [2.67]		0.76 (4.47) [1.07]		0.72 (3.71) [2.75]	0.58 (2.51) [3.93]	
SGA/REV			−0.03 (−0.12) [2.34]	−0.07 (−0.24) [2.36]			0.10 (0.42) [2.35]	−0.11 (−0.38) [2.37]		0.93 (3.34) [3.15]
$\text{Ln}\left(\frac{\text{BE}}{\text{ME}}\right)$	0.15 (2.06) [2.63]	0.18 (2.58) [2.24]	0.15 (2.20) [2.79]	0.16 (2.37) [2.86]	0.16 (2.20) [2.64]	0.23 (3.23) [2.27]	0.15 (2.26) [2.80]	0.21 (3.12) [2.88]	0.22 (3.12) [2.74]	0.25 (3.66) [2.94]
Ln(ME)	−0.02 (−0.58) [3.33]	−0.01 (−0.25) [2.35]	−0.02 (−0.40) [3.23]	−0.02 (−0.47) [4.13]	−0.02 (−0.43) [5.45]	−0.00 (−0.02) [3.57]	−0.01 (−0.17) [4.89]	−0.01 (−0.22) [6.27]	−0.01 (−0.35) [6.03]	−0.01 (−0.23) [8.11]
$\text{return}_{-1,0}$	−1.70 (−3.40) [1.02]	−1.67 (−3.33) [1.02]	−1.71 (−3.44) [1.03]	−1.79 (−3.63) [1.03]	−1.82 (−3.68) [1.02]	−1.83 (−3.72) [1.03]	−1.86 (−3.80) [1.03]	−1.93 (−3.97) [1.03]	−1.90 (−3.89) [1.03]	−1.97 (−4.06) [1.03]
$\text{return}_{-12,-1}$	0.54 (3.06) [1.21]	0.56 (3.17) [1.21]	0.53 (3.03) [1.21]	0.54 (3.04) [1.21]	0.50 (2.86) [1.21]	0.51 (2.93) [1.21]	0.48 (2.79) [1.21]	0.49 (2.86) [1.21]	0.50 (2.88) [1.21]	0.49 (2.88) [1.21]
Intercept	0.01 (2.56)	0.01 (2.23)	0.01 (2.64)	0.01 (2.47)	0.01 (1.72)	0.01 (1.05)	0.01 (1.89)	0.01 (1.27)	0.01 (1.23)	0.00 (0.92)
R^2	5.90%	5.80%	6.03%	6.49%	6.57%	6.51%	6.70%	7.16%	7.10%	7.20%
adj. R^2	5.34%	5.24%	5.47%	5.83%	5.91%	5.84%	6.03%	6.38%	6.33%	6.42%

	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
GPSGA/A	2.36 (6.04) [4.12]	2.30 (5.22) [7.17]	2.08 (5.03) [5.00]	2.49 (6.47) [4.40]	2.29 (5.31) [5.29]					
GP/A						1.14 (5.76) [4.52]	0.93 (4.59) [6.52]	0.98 (4.01) [5.35]	1.26 (6.22) [7.26]	1.17 (4.81) [9.43]
REV/A	0.11 (1.97) [3.05]	0.11 (1.89) [3.84]	0.17 (2.71) [3.32]	0.12 (2.10) [3.05]	0.15 (2.47) [3.33]	0.00 (0.03) [4.19]	0.06 (0.97) [5.12]	0.06 (0.87) [4.35]	−0.05 (−1.03) [5.01]	−0.02 (−0.27) [6.28]
GPSGA/REV		−0.05 (−0.18) [3.13]					0.47 (2.17) [2.04]			
GP/REV			0.44 (2.38) [3.18]		0.23 (0.96) [3.30]			0.23 (0.98) [1.24]		0.18 (0.72) [3.57]
SGA/REV				0.45 (1.84) [2.41]	0.37 (1.20) [2.50]				−0.47 (−1.80) [2.86]	−0.52 (−1.83) [2.90]
$\text{Ln}\left(\frac{\text{BE}}{\text{ME}}\right)$	0.27 (3.55) [2.92]	0.26 (3.46) [3.06]	0.31 (4.04) [2.97]	0.31 (4.38) [3.23]	0.32 (4.45) [3.23]	0.29 (3.99) [2.49]	0.27 (3.71) [3.10]	0.29 (3.97) [2.61]	0.27 (3.76) [3.11]	0.27 (3.74) [3.11]
Ln(ME)	−0.02 (−0.51) [4.95]	−0.02 (−0.59) [5.50]	−0.02 (−0.43) [5.97]	−0.02 (−0.45) [6.01]	−0.02 (−0.45) [6.67]	0.01 (0.15) [3.62]	−0.01 (−0.24) [5.46]	0.00 (0.10) [3.64]	−0.01 (−0.21) [4.90]	−0.01 (−0.24) [6.55]
$\text{return}_{-1,0}$	−1.83 (−3.70) [1.02]	−1.87 (−3.81) [1.03]	−1.90 (−3.86) [1.03]	−1.96 (−4.02) [1.03]	−1.99 (−4.12) [1.03]	−1.89 (−3.85) [1.02]	−1.95 (−4.01) [1.03]	−1.93 (−3.97) [1.03]	−1.94 (−3.98) [1.03]	−1.99 (−4.12) [1.03]
$\text{return}_{-12,-1}$	0.50 (2.85) [1.21]	0.50 (2.86) [1.21]	0.50 (2.90) [1.21]	0.48 (2.79) [1.21]	0.49 (2.86) [1.21]	0.50 (2.89) [1.21]	0.49 (2.84) [1.21]	0.51 (2.93) [1.22]	0.48 (2.82) [1.21]	0.49 (2.88) [1.21]
Intercept	0.01 (1.64)	0.01 (1.68)	0.01 (1.19)	0.01 (1.37)	0.01 (1.17)	0.01 (1.36)	0.01 (1.42)	0.01 (1.15)	0.01 (1.83)	0.01 (1.62)
R^2	6.56%	6.99%	6.99%	7.18%	7.58%	6.56%	7.10%	6.97%	7.12%	7.55%
adj. R^2	5.89%	6.21%	6.21%	6.41%	6.70%	5.89%	6.32%	6.19%	6.35%	6.67%

Table A.21: Fama-MacBeth Regressions - GPSGA/A Decomposition 3 - Complete Results

This table shows the coefficients of the Fama-MacBeth regressions on GPSGA/A and its components according to the decomposition of Equation 7. All coefficients are in percent, the values in brackets the corresponding t-test values, and the values in square brackets the VIFs, which are a test for multicollinearity. The regressions control for value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. Only observations not missing any of the variables are considered; hence, n is 1.25 million for each regression.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
GPSGA/A											
GP/A											
ME/A	−0.02					−0.01	−0.00	−0.01	−0.00	−0.00	0.00
	(−0.89)					(−0.41)	(−0.08)	(−0.46)	(−0.08)	(−0.02)	(0.01)
	[3.08]					[3.11]	[3.10]	[3.13]	[3.14]	[3.14]	[3.14]
GPSGA/ME		0.89				0.87				0.44	0.84
		(5.39)				(5.35)				(2.37)	(5.12)
		[1.31]				[1.32]				[1.59]	[1.33]
GP/ME			0.45		0.61		0.44		0.60	0.38	
			(7.50)		(4.44)		(7.45)		(4.39)	(5.75)	
			[1.32]		[9.95]		[1.32]		[9.97]	[1.66]	
SGA/ME				0.41	−0.18			0.40	−0.18		0.40
				(6.17)	(−1.23)			(6.11)	(−1.19)		(6.14)
				[1.22]	[8.90]			[1.23]	[8.95]		[1.25]
$\text{Ln}\left(\frac{\text{BE}}{\text{ME}}\right)$	0.36	0.34	0.27	0.31	0.26	0.31	0.25	0.28	0.24	0.23	0.23
	(6.44)	(6.03)	(4.46)	(4.96)	(4.50)	(5.74)	(4.37)	(4.79)	(4.33)	(4.22)	(4.09)
	[2.76]	[1.84]	[1.74]	[1.79]	[1.89]	[2.84]	[2.94]	[2.88]	[2.95]	[2.96]	[2.96]
Ln(ME)	−0.04	−0.06	−0.00	0.01	−0.00	−0.06	−0.00	0.01	−0.01	−0.02	−0.01
	(−1.02)	(−1.65)	(−0.08)	(0.24)	(−0.13)	(−1.64)	(−0.09)	(0.22)	(−0.14)	(−0.44)	(−0.34)
	[1.77]	[2.27]	[2.18]	[2.11]	[2.49]	[2.46]	[2.26]	[2.23]	[2.63]	[2.71]	[2.71]
$\text{return}_{-1,0}$	−4.23	−4.22	−4.07	−4.12	−4.11	−4.24	−4.10	−4.13	−4.14	−4.15	−4.14
	(−11.19)	(−11.16)	(−10.79)	(−10.92)	(−11.02)	(−11.33)	(−10.99)	(−11.10)	(−11.23)	(−11.26)	(−11.24)
	[1.04]	[1.01]	[1.02]	[1.02]	[1.02]	[1.04]	[1.05]	[1.05]	[1.05]	[1.05]	[1.05]
$\text{return}_{-12,-1}$	0.58	0.56	0.65	0.65	0.64	0.55	0.64	0.65	0.63	0.62	0.62
	(4.30)	(4.11)	(4.96)	(4.91)	(4.94)	(4.22)	(5.07)	(5.08)	(5.06)	(4.94)	(5.00)
	[1.20]	[1.08]	[1.09]	[1.11]	[1.11]	[1.20]	[1.23]	[1.24]	[1.24]	[1.24]	[1.24]
Intercept	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	(3.97)	(3.80)	(2.51)	(2.71)	(2.56)	(3.81)	(2.52)	(2.75)	(2.58)	(2.70)	(2.60)
R^2	3.28%	3.40%	3.39%	3.38%	3.67%	3.58%	3.57%	3.57%	3.85%	3.86%	3.87%
adj. R^2	3.08%	3.19%	3.18%	3.18%	3.42%	3.34%	3.33%	3.33%	3.56%	3.58%	3.59%

	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
GPSGA/A	2.68	2.79	2.40	2.81	3.09					
	(9.26)	(8.94)	(8.03)	(9.77)	(9.94)					
	[1.60]	[2.23]	[1.61]	[1.61]	[2.10]					
GP/A						1.82	1.72	1.61	1.66	1.60
						(10.93)	(10.38)	(8.59)	(9.03)	(8.62)
						[1.77]	[2.58]	[2.65]	[3.53]	[3.87]
ME/A	0.00	0.00	0.01	0.01	0.01	−0.03	−0.03	−0.03	−0.03	−0.03
	(0.01)	(0.04)	(0.54)	(0.56)	(0.55)	(−1.24)	(−1.00)	(−0.98)	(−1.18)	(−1.02)
	[3.14]	[3.14]	[3.16]	[3.16]	[3.16]	[3.09]	[3.18]	[3.11]	[3.16]	[3.17]
GPSGA/ME		−0.10					0.45			
		(−0.60)					(2.83)			
		[1.83]					[1.43]			
GP/ME			0.33		−0.23			0.14		0.25
			(5.54)		(−1.62)			(2.21)		(1.87)
			[1.39]		[12.99]			[1.98]		[10.95]
SGA/ME				0.44	0.65				0.12	−0.11
				(6.73)	(4.34)				(1.59)	(−0.75)
				[1.24]	[11.64]				[1.83]	[8.96]
$\text{Ln}\left(\frac{\text{BE}}{\text{ME}}\right)$	0.36	0.37	0.26	0.26	0.28	0.47	0.44	0.42	0.43	0.41
	(6.39)	(6.72)	(4.59)	(4.60)	(5.08)	(8.29)	(8.14)	(6.91)	(6.91)	(6.99)
	[2.74]	[2.88]	[2.95]	[2.89]	[2.97]	[2.77]	[2.88]	[3.11]	[3.04]	[3.19]
Ln(ME)	−0.12	−0.13	−0.08	−0.08	−0.08	−0.02	−0.03	−0.01	−0.00	−0.01
	(−3.56)	(−3.58)	(−2.49)	(−2.23)	(−2.29)	(−0.56)	(−0.90)	(−0.22)	(−0.08)	(−0.20)
	[2.60]	[2.72]	[3.06]	[3.01]	[3.06]	[2.45]	[3.05]	[2.53]	[2.81]	[2.99]
$\text{return}_{-1,0}$	−4.37	−4.39	−4.27	−4.25	−4.28	−4.40	−4.42	−4.36	−4.39	−4.40
	(−11.70)	(−11.90)	(−11.65)	(−11.57)	(−11.78)	(−11.78)	(−11.95)	(−11.97)	(−12.03)	(−12.20)
	[1.04]	[1.04]	[1.05]	[1.05]	[1.05]	[1.04]	[1.04]	[1.06]	[1.06]	[1.06]
$\text{return}_{-12,-1}$	0.46	0.45	0.52	0.53	0.51	0.49	0.48	0.54	0.53	0.53
	(3.49)	(3.54)	(4.29)	(4.25)	(4.23)	(3.73)	(3.70)	(4.43)	(4.39)	(4.44)
	[1.22]	[1.23]	[1.26]	[1.26]	[1.26]	[1.20]	[1.20]	[1.26]	[1.26]	[1.26]
Intercept	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	(4.45)	(4.50)	(3.35)	(3.06)	(3.17)	(1.76)	(1.84)	(1.54)	(1.55)	(1.55)
R^2	3.70%	3.92%	3.98%	3.99%	4.22%	3.70%	3.96%	3.95%	3.97%	4.21%
adj. R^2	3.46%	3.64%	3.70%	3.71%	3.90%	3.46%	3.68%	3.67%	3.69%	3.89%

Table A.22: Fama-MacBeth Regressions - GPSGA/A Decomposition 3 - Complete Results (Excluding Micro-Caps)

This table shows the coefficients of the Fama-MacBeth regressions on GPSGA/A and its components according to the decomposition of Equation 7. All coefficients are in percent, the values in brackets the corresponding t-test values, and the values in square brackets the VIFs, which are a test for multicollinearity. The regressions control for value as Ln(BE/ME), size as Ln(ME), short-term reversal as the past month's return, and momentum as the past twelve months' return, skipping the most recent month. The sample excludes micro-caps, which are defined as stocks with a market capitalization below the 20th percentile of the NYSE market capitalization distribution. Only observations not missing any of the variables are considered; hence, n is 0.44 million for each regression.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
GPSGA/A											
GP/A											
ME/A	−0.02					0.00	0.01	−0.00	0.01	0.01	0.01
	(−0.51)					(0.11)	(0.23)	(−0.03)	(0.47)	(0.42)	(0.46)
	[3.98]					[4.07]	[4.01]	[4.06]	[4.07]	[4.08]	[4.08]
GPSGA/ME		1.72				1.65				1.56	1.68
		(4.91)				(4.95)				(3.80)	(4.92)
		[2.43]				[2.45]				[4.61]	[2.82]
GP/ME			0.49		1.65		0.45		1.59	0.09	
			(3.22)		(4.82)		(3.05)		(4.81)	(0.48)	
			[2.04]		[14.60]		[2.05]		[14.66]	[4.02]	
SGA/ME				0.36	−1.50			0.32	−1.46		0.09
				(1.95)	(−3.70)			(1.77)	(−3.66)		(0.48)
				[1.57]	[10.81]			[1.58]	[10.83]		[1.82]
$\text{Ln}\left(\frac{\text{BE}}{\text{ME}}\right)$	0.10	0.04	0.06	0.10	0.01	0.02	0.04	0.07	0.00	−0.00	−0.00
	(1.51)	(0.65)	(0.81)	(1.44)	(0.18)	(0.27)	(0.52)	(0.99)	(0.02)	(−0.02)	(−0.02)
	[4.11]	[3.04]	[2.48]	[2.81]	[3.06]	[4.90]	[4.27]	[4.76]	[4.92]	[4.93]	[4.93]
Ln(ME)	−0.00	−0.01	0.00	0.00	−0.01	−0.01	0.00	0.01	−0.00	−0.01	−0.01
	(−0.10)	(−0.32)	(0.07)	(0.11)	(−0.17)	(−0.27)	(0.10)	(0.13)	(−0.12)	(−0.14)	(−0.15)
	[2.24]	[5.47]	[4.11]	[3.78]	[5.55]	[5.51]	[4.14]	[3.79]	[5.59]	[5.73]	[5.75]
$\text{return}_{-1,0}$	−1.63	−1.47	−1.56	−1.64	−1.58	−1.56	−1.65	−1.73	−1.67	−1.67	−1.68
	(−3.27)	(−3.01)	(−3.17)	(−3.32)	(−3.26)	(−3.21)	(−3.39)	(−3.52)	(−3.48)	(−3.47)	(−3.48)
	[1.07]	[1.03]	[1.03]	[1.03]	[1.03]	[1.07]	[1.07]	[1.06]	[1.07]	[1.07]	[1.07]
$\text{return}_{-12,-1}$	0.53	0.59	0.58	0.55	0.59	0.57	0.56	0.53	0.56	0.56	0.56
	(3.10)	(3.52)	(3.45)	(3.22)	(3.59)	(3.44)	(3.35)	(3.16)	(3.45)	(3.43)	(3.43)
	[1.48]	[1.23]	[1.23]	[1.21]	[1.23]	[1.45]	[1.48]	[1.44]	[1.45]	[1.45]	[1.45]
Intercept	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	(2.35)	(1.73)	(1.70)	(2.00)	(1.53)	(1.67)	(1.66)	(1.95)	(1.47)	(1.49)	(1.48)
R^2	5.80%	5.85%	5.92%	5.90%	6.45%	6.32%	6.39%	6.39%	6.90%	6.89%	6.89%
adj. R^2	5.24%	5.29%	5.36%	5.34%	5.78%	5.65%	5.73%	5.73%	6.12%	6.11%	6.12%

	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
GPSGA/A	2.70	2.37	2.51	2.75	2.49					
	(6.25)	(4.94)	(5.83)	(6.36)	(5.15)					
	[3.81]	[4.79]	[4.04]	[3.84]	[4.72]					
GP/A						1.20	1.04	1.14	1.28	1.18
						(5.40)	(4.63)	(4.72)	(5.40)	(4.97)
						[2.77]	[4.14]	[3.96]	[5.50]	[5.62]
ME/A	−0.05	−0.03	−0.03	−0.03	−0.02	−0.03	−0.01	−0.02	−0.04	−0.02
	(−1.46)	(−1.08)	(−0.91)	(−1.00)	(−0.71)	(−0.98)	(−0.43)	(−0.80)	(−1.13)	(−0.65)
	[4.12]	[4.23]	[4.17]	[4.14]	[4.22]	[3.99]	[4.22]	[4.04]	[4.26]	[4.30]
GPSGA/ME		0.66					1.25			
		(1.87)					(3.80)			
		[3.07]					[2.73]			
GP/ME			0.29		0.53			0.05		1.32
			(2.04)		(1.50)			(0.37)		(4.09)
			[2.26]		[18.02]			[2.92]		[14.99]
SGA/ME				0.35	−0.25				−0.22	−1.66
				(1.94)	(−0.57)				(−1.22)	(−4.29)
				[1.58]	[12.64]				[2.33]	[11.01]
$\text{Ln}\left(\frac{\text{BE}}{\text{ME}}\right)$	0.19	0.14	0.13	0.16	0.13	0.23	0.15	0.21	0.24	0.18
	(2.76)	(2.02)	(1.81)	(2.26)	(1.84)	(3.39)	(2.22)	(2.86)	(3.29)	(2.41)
	[4.74]	[5.33]	[5.08]	[4.89]	[5.30]	[4.29]	[5.21]	[4.85]	[5.21]	[5.48]
Ln(ME)	−0.03	−0.03	−0.02	−0.02	−0.02	0.01	−0.00	0.01	0.00	−0.00
	(−0.71)	(−0.76)	(−0.53)	(−0.48)	(−0.56)	(0.17)	(−0.01)	(0.16)	(0.11)	(−0.11)
	[4.11]	[5.76]	[5.68]	[5.07]	[5.97]	[3.15]	[5.87]	[4.19]	[4.17]	[5.77]
$\text{return}_{-1,0}$	−1.66	−1.66	−1.71	−1.75	−1.76	−1.79	−1.77	−1.85	−1.88	−1.85
	(−3.36)	(−3.47)	(−3.54)	(−3.60)	(−3.72)	(−3.70)	(−3.73)	(−3.88)	(−3.91)	(−3.93)
	[1.06]	[1.07]	[1.07]	[1.06]	[1.07]	[1.07]	[1.07]	[1.07]	[1.06]	[1.07]
$\text{return}_{-12,-1}$	0.55	0.56	0.57	0.55	0.56	0.51	0.54	0.52	0.51	0.54
	(3.18)	(3.45)	(3.42)	(3.26)	(3.46)	(3.04)	(3.35)	(3.26)	(3.10)	(3.38)
	[1.44]	[1.45]	[1.45]	[1.44]	[1.45]	[1.47]	[1.45]	[1.48]	[1.44]	[1.45]
Intercept	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01
	(1.96)	(1.81)	(1.53)	(1.52)	(1.47)	(1.29)	(0.94)	(1.29)	(1.42)	(1.10)
R^2	6.47%	6.95%	7.03%	7.07%	7.55%	6.64%	7.14%	7.09%	7.07%	7.55%
adj. R^2	5.80%	6.18%	6.26%	6.30%	6.67%	5.98%	6.36%	6.32%	6.29%	6.67%

Table A.23: Double Sort on Size and GP/A - Subsamples Before and After 2010

This table shows the value-weighted excess returns of the portfolios first sorted into quintiles based on size (market capitalization), then within the quintiles based on GP/A. Portfolios are rebalanced annually at the end of June. The first part covers the sample period from June 1980 to December 2010, and the second part the period from January 2011 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The size strategies are the small-minus-big portfolios, profitability strategies are the high-minus-low GP/A portfolios, and the combined strategy is the small-high portfolio minus the big-low portfolio (column description of the profitability strategies applies).

June 1980 - December 2010											
	GP/A					Profitability strategies					
Size	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Small	0.17%	0.26%	0.41%	0.52%	0.97%	0.80%	0.76%	0.76%	0.00	−0.09	0.15
						(5.44)	(5.12)	(5.15)	(0.08)	(−1.91)	(2.91)
2	0.11%	0.21%	0.33%	0.35%	0.65%	0.53%	0.50%	0.50%	0.05	0.05	−0.01
						(4.27)	(4.05)	(3.91)	(1.85)	(1.22)	(−0.16)
3	0.16%	0.33%	0.28%	0.25%	0.58%	0.42%	0.39%	0.32%	0.14	0.14	−0.02
						(3.22)	(3.00)	(2.58)	(5.02)	(3.49)	(−0.39)
4	0.30%	0.30%	0.38%	0.47%	0.75%	0.45%	0.41%	0.39%	0.16	0.18	−0.15
						(2.99)	(2.73)	(2.79)	(5.00)	(3.87)	(−3.15)
Big	0.38%	0.33%	0.45%	0.47%	0.68%	0.29%	0.24%	0.43%	−0.02	−0.10	−0.32
						(1.81)	(1.51)	(2.74)	(−0.61)	(−2.00)	(−5.87)
	Size strategies					Combined strategy					
Mean	−0.21%	−0.07%	−0.04%	0.05%	0.30%	0.59%	0.50%	0.36%	0.08	1.03	0.08
	(−0.79)	(−0.38)	(−0.23)	(0.27)	(1.28)	(2.59)	(2.20)	(2.34)	(2.40)	(20.43)	(1.58)
Geomean	−0.34%	−0.13%	−0.10%	−0.02%	0.20%						
	(−1.29)	(−0.71)	(−0.56)	(−0.09)	(0.88)						
α	−0.83%	−0.62%	−0.65%	−0.63%	−0.50%						
	(−4.38)	(−4.82)	(−5.15)	(−4.82)	(−3.28)						
β_{MKT}	0.08	−0.08	−0.08	−0.14	0.11						
	(1.91)	(−2.81)	(−2.91)	(−4.66)	(3.09)						
β_{SMB}	1.13	0.85	0.86	0.96	1.14						
	(18.22)	(20.23)	(20.80)	(22.57)	(22.85)						
β_{HML}	−0.07	0.13	0.29	0.54	0.40						
	(−1.04)	(2.95)	(6.61)	(12.06)	(7.64)						
January 2011 - December 2021											
	GP/A					Profitability strategies					
Size	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Small	0.46%	0.38%	0.46%	1.00%	1.06%	0.60%	0.52%	0.85%	−0.16	−0.44	0.31
						(1.76)	(1.53)	(2.53)	(−1.89)	(−3.11)	(2.76)
2	0.60%	0.61%	0.60%	0.89%	0.93%	0.33%	0.29%	0.47%	−0.10	−0.14	0.12
						(1.28)	(1.11)	(1.75)	(−1.45)	(−1.23)	(1.32)
3	0.65%	0.55%	0.80%	0.93%	0.93%	0.28%	0.25%	0.21%	0.02	−0.08	−0.15
						(1.29)	(1.15)	(0.94)	(0.35)	(−0.84)	(−1.99)
4	0.61%	0.60%	0.85%	1.04%	1.13%	0.52%	0.49%	0.39%	0.06	−0.04	−0.24
						(2.26)	(2.12)	(1.63)	(0.98)	(−0.40)	(−3.08)
Big	0.66%	0.62%	0.92%	1.22%	1.40%	0.73%	0.68%	0.41%	0.13	−0.27	−0.60
						(2.62)	(2.43)	(1.70)	(2.16)	(−2.62)	(−7.50)
	Size strategies					Combined strategy					
Mean	−0.20%	−0.24%	−0.45%	−0.23%	−0.33%	0.40%	0.35%	0.36%	0.05	0.87	−0.06
	(−0.44)	(−0.88)	(−1.72)	(−0.74)	(−1.01)	(1.38)	(1.20)	(1.52)	(0.82)	(8.72)	(−0.75)
Geomean	−0.33%	−0.29%	−0.50%	−0.29%	−0.40%						
	(−0.74)	(−1.06)	(−1.89)	(−0.95)	(−1.22)						
α	−0.53%	0.02%	−0.35%	−0.11%	−0.09%						
	(−1.52)	(0.08)	(−2.05)	(−0.65)	(−0.54)						
β_{MKT}	0.21	−0.17	−0.01	−0.03	−0.08						
	(2.38)	(−3.17)	(−0.32)	(−0.63)	(−1.86)						
β_{SMB}	1.31	0.85	0.82	1.14	1.14						
	(8.87)	(9.37)	(11.50)	(16.23)	(15.71)						
β_{HML}	−0.36	0.23	0.38	0.32	0.55						
	(−3.11)	(3.20)	(6.73)	(5.69)	(9.55)						

Table A.24: Double Sort on Size and GPSGA/A - Subsamples Before and After 2010

This table shows the value-weighted excess returns for the portfolios first sorted into quintiles based on size (market capitalization), then within the quintiles based on GPSGA/A. Portfolios are rebalanced annually at the end of June. The first part covers the sample period from June 1980 to December 2010, and the second part the period from January 2011 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The size strategies are the small-minus-big portfolios, profitability strategies are the high-minus-low GPSGA/A portfolios, and the combined strategy is the small-high portfolio minus the big-low portfolio (values are according to the column description of the profitability strategies).

June 1980 - December 2010											
	GPSGA/A					Profitability strategies					
Size	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Small	−0.01%	0.24%	0.28%	0.44%	0.79%	0.80%	0.74%	0.91%	−0.14	−0.26	0.02
						(4.70)	(4.22)	(5.62)	(−3.93)	(−4.97)	(0.44)
2	−0.07%	0.11%	0.09%	0.35%	0.55%	0.62%	0.60%	0.70%	−0.06	−0.06	−0.09
						(5.74)	(5.57)	(6.39)	(−2.58)	(−1.65)	(−2.46)
3	0.03%	0.28%	0.24%	0.20%	0.49%	0.47%	0.44%	0.55%	−0.02	−0.02	−0.19
						(3.92)	(3.71)	(4.63)	(−0.77)	(−0.50)	(−4.71)
4	0.08%	0.35%	0.17%	0.43%	0.53%	0.45%	0.42%	0.56%	−0.04	−0.09	−0.22
						(3.46)	(3.22)	(4.33)	(−1.28)	(−2.01)	(−4.84)
Big	0.33%	0.47%	0.26%	0.46%	0.54%	0.20%	0.15%	0.53%	−0.11	−0.27	−0.61
						(1.15)	(0.82)	(3.37)	(−3.22)	(−5.31)	(−11.36)
	Size strategies					Combined strategy					
Mean	−0.34%	−0.23%	0.02%	−0.03%	0.25%	0.46%	0.41%	0.44%	−0.18	0.72	0.04
	(−1.40)	(−1.35)	(0.10)	(−0.17)	(1.22)	(2.69)	(2.37)	(3.45)	(−6.14)	(17.18)	(0.88)
Geomean	−0.44%	−0.29%	−0.03%	−0.08%	0.18%						
	(−1.88)	(−1.67)	(−0.19)	(−0.47)	(0.85)						
α	−0.90%	−0.81%	−0.56%	−0.65%	−0.51%						
	(−4.78)	(−6.63)	(−4.62)	(−5.24)	(−3.86)						
β_{MKT}	−0.03	−0.07	−0.09	−0.03	−0.06						
	(−0.75)	(−2.68)	(−3.20)	(−1.14)	(−2.08)						
β_{SMB}	0.99	0.82	0.73	0.72	1.00						
	(16.07)	(20.48)	(18.40)	(17.84)	(22.96)						
β_{HML}	0.01	0.20	0.27	0.30	0.65						
	(0.17)	(4.81)	(6.34)	(6.96)	(14.18)						

January 2011 - December 2021											
	GPSGA/A					Profitability strategies					
Size	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Small	0.25%	0.61%	0.60%	0.75%	0.93%	0.68%	0.63%	0.92%	−0.19	−0.52	0.14
						(2.41)	(2.21)	(3.52)	(−2.83)	(−4.74)	(1.63)
2	0.46%	0.60%	0.51%	0.74%	0.90%	0.44%	0.42%	0.53%	−0.07	−0.14	0.04
						(2.14)	(2.01)	(2.46)	(−1.28)	(−1.50)	(0.59)
3	0.55%	0.66%	0.61%	0.60%	0.97%	0.41%	0.39%	0.52%	−0.08	−0.17	0.03
						(1.97)	(1.85)	(2.38)	(−1.55)	(−1.88)	(0.46)
4	0.57%	0.67%	0.78%	0.78%	1.19%	0.62%	0.58%	0.77%	−0.16	−0.21	−0.11
						(2.59)	(2.45)	(3.25)	(−2.60)	(−2.11)	(−1.45)
Big	0.56%	0.74%	0.92%	0.85%	1.49%	0.93%	0.88%	0.72%	0.06	−0.47	−0.47
						(3.29)	(3.10)	(2.92)	(0.92)	(−4.57)	(−5.76)
	Size strategies					Combined strategy					
Mean	−0.31%	−0.13%	−0.33%	−0.11%	−0.56%	0.37%	0.34%	0.52%	−0.09	0.53	0.06
	(−0.92)	(−0.51)	(−1.15)	(−0.42)	(−1.88)	(1.71)	(1.58)	(2.59)	(−1.86)	(6.38)	(0.84)
Geomean	−0.39%	−0.17%	−0.38%	−0.15%	−0.62%						
	(−1.15)	(−0.67)	(−1.32)	(−0.59)	(−2.07)						
α	−0.45%	0.02%	−0.30%	0.13%	−0.25%						
	(−1.68)	(0.09)	(−1.57)	(0.87)	(−1.50)						
β_{MKT}	0.09	−0.07	0.05	−0.15	−0.15						
	(1.42)	(−1.48)	(0.93)	(−3.88)	(−3.55)						
β_{SMB}	1.05	0.81	0.87	0.95	1.01						
	(9.47)	(10.85)	(10.84)	(15.31)	(14.42)						
β_{HML}	−0.08	0.29	0.33	0.25	0.53						
	(−0.94)	(4.88)	(5.18)	(5.07)	(9.62)						

Table A.25: Double Sort on Size and GPSGAI/BE - Subsamples Before and After 2010

This table shows the value-weighted excess returns of the portfolios first sorted into quintiles based on size (market capitalization), then within the quintiles based on GPSGAI/BE. Portfolios are rebalanced annually at the end of June. The first part covers the sample period from June 1980 to December 2010, and the second part the period from January 2011 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The size strategies are the small-minus-big portfolios, profitability strategies are the high-minus-low GPSGAI/BE portfolios, and the combined strategy is the small-high portfolio minus the big-low portfolio (column description of the profitability strategies applies).

June 1980 - December 2010

	GPSGAI/BE					Profitability strategies					
Size	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Small	−0.18%	0.25%	0.27%	0.48%	0.61%	0.79%	0.74%	0.87%	−0.12	−0.21	0.04
						(5.20)	(4.83)	(5.98)	(−3.64)	(−4.44)	(0.78)
2	−0.12%	0.10%	0.08%	0.30%	0.38%	0.50%	0.47%	0.55%	−0.05	−0.05	−0.04
						(4.16)	(3.96)	(4.48)	(−1.83)	(−1.19)	(−0.86)
3	−0.04%	0.16%	0.17%	0.23%	0.34%	0.38%	0.36%	0.42%	0.01	−0.06	−0.11
						(3.16)	(2.96)	(3.44)	(0.33)	(−1.41)	(−2.49)
4	0.00%	0.17%	0.20%	0.30%	0.41%	0.41%	0.38%	0.41%	0.02	−0.01	−0.02
						(3.38)	(3.17)	(3.27)	(0.85)	(−0.29)	(−0.57)
Big	0.25%	0.20%	0.30%	0.38%	0.48%	0.23%	0.19%	0.36%	−0.06	−0.22	−0.18
						(1.46)	(1.17)	(2.28)	(−1.63)	(−4.24)	(−3.26)
	Size strategies					Combined strategy					
Mean	−0.43%	0.05%	−0.02%	0.09%	0.13%	0.36%	0.30%	0.12%	0.06	0.73	0.27
	(−1.87)	(0.34)	(−0.15)	(0.58)	(0.64)	(2.04)	(1.71)	(0.87)	(2.10)	(16.35)	(5.82)
Geomean	−0.52%	0.01%	−0.07%	0.05%	0.06%						
	(−2.30)	(0.05)	(−0.44)	(0.29)	(0.27)						
α	−1.18%	−0.49%	−0.61%	−0.58%	−0.67%						
	(−7.00)	(−3.75)	(−4.58)	(−4.68)	(−4.91)						
β_{MKT}	0.19	−0.06	−0.07	0.03	0.12						
	(4.87)	(−1.91)	(−2.38)	(1.14)	(4.03)						
β_{SMB}	0.95	0.64	0.60	0.68	0.95						
	(17.14)	(14.95)	(13.87)	(16.68)	(21.44)						
β_{HML}	0.23	0.16	0.31	0.37	0.45						
	(3.96)	(3.51)	(6.84)	(8.57)	(9.55)						

January 2011 - December 2021

	GPSGAI/BE					Profitability strategies					
Size	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Small	0.29%	0.51%	0.74%	0.55%	0.68%	0.39%	0.34%	0.46%	−0.04	−0.45	0.18
						(1.38)	(1.17)	(1.64)	(−0.54)	(−3.83)	(1.90)
2	0.57%	0.48%	0.78%	0.63%	0.81%	0.24%	0.21%	0.07%	0.15	−0.10	0.08
						(1.14)	(0.99)	(0.34)	(2.75)	(−1.11)	(1.13)
3	0.38%	0.33%	0.72%	0.83%	0.68%	0.30%	0.28%	0.24%	0.07	−0.17	0.11
						(1.51)	(1.38)	(1.14)	(1.32)	(−1.95)	(1.57)
4	0.40%	0.62%	0.67%	0.63%	0.82%	0.43%	0.40%	0.40%	0.00	−0.18	−0.07
						(1.95)	(1.81)	(1.74)	(0.02)	(−1.85)	(−0.96)
Big	0.39%	0.54%	0.70%	1.19%	1.07%	0.68%	0.62%	0.85%	−0.23	−0.35	−0.37
						(2.31)	(2.10)	(3.32)	(−3.57)	(−3.24)	(−4.36)
	Size strategies					Combined strategy					
Mean	−0.09%	−0.04%	0.04%	−0.64%	−0.38%	0.30%	0.26%	0.43%	−0.08	0.46	0.10
	(−0.28)	(−0.14)	(0.16)	(−2.34)	(−1.27)	(1.19)	(1.03)	(1.79)	(−1.26)	(4.58)	(1.31)
Geomean	−0.17%	−0.08%	−0.00%	−0.69%	−0.45%						
	(−0.50)	(−0.31)	(−0.00)	(−2.51)	(−1.45)						
α	−0.07%	−0.05%	0.13%	−0.64%	−0.46%						
	(−0.24)	(−0.29)	(0.71)	(−3.98)	(−2.48)						
β_{MKT}	−0.04	0.05	−0.05	0.06	0.15						
	(−0.48)	(1.11)	(−0.99)	(1.57)	(3.33)						
β_{SMB}	0.92	0.77	0.84	0.91	0.81						
	(7.18)	(9.82)	(10.73)	(13.60)	(10.50)						
β_{HML}	−0.07	0.22	0.17	0.34	0.48						
	(−0.68)	(3.55)	(2.82)	(6.42)	(7.81)						

Table A.26: Double Sort on BE/ME and GP/A - Subsamples Before and After 2010

This table shows the value-weighted excess returns of the portfolios first sorted into quintiles based on BE/ME, then within the quintiles based on GP/A. Portfolios are rebalanced annually at the end of June. The first part covers the sample period from June 1980 to December 2010, and the second part the period from January 2011 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The value strategies are the high-minus-low BE/ME portfolios, profitability strategies are the high-minus-low GP/A portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies).

June 1980 - December 2010											
	GP/A					Profitability strategies					
BE/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	−0.09%	−0.06%	0.06%	0.33%	0.53%	0.62%	0.56%	0.72%	−0.18	−0.14	0.08
						(3.32)	(3.00)	(3.91)	(−4.25)	(−2.25)	(1.23)
2	0.02%	0.05%	0.23%	0.37%	0.40%	0.37%	0.32%	0.35%	−0.05	0.15	0.10
						(2.18)	(1.87)	(1.99)	(−1.38)	(2.56)	(1.72)
3	−0.08%	−0.03%	0.19%	0.29%	0.29%	0.37%	0.33%	0.25%	0.06	0.31	0.13
						(2.41)	(2.13)	(1.66)	(1.80)	(6.23)	(2.43)
4	0.12%	0.27%	0.08%	0.34%	0.51%	0.39%	0.33%	0.21%	0.15	0.36	0.10
						(2.15)	(1.83)	(1.23)	(3.97)	(6.44)	(1.76)
High	0.52%	0.46%	0.23%	0.28%	0.59%	0.07%	−0.02%	0.20%	0.06	−0.05	−0.44
						(0.32)	(−0.09)	(0.95)	(1.15)	(−0.72)	(−6.04)
	Value strategies					Combined strategy					
Mean	0.61%	0.52%	0.16%	−0.05%	0.06%	0.68%	0.61%	0.67%	−0.06	−0.05	0.16
	(2.86)	(2.73)	(0.91)	(−0.27)	(0.36)	(3.45)	(3.10)	(3.35)	(−1.39)	(−0.70)	(2.26)
Geomean	0.53%	0.46%	0.10%	−0.11%	0.01%						
	(2.46)	(2.40)	(0.58)	(−0.59)	(0.05)						
α	0.04%	−0.06%	−0.53%	−0.67%	−0.47%						
	(0.22)	(−0.38)	(−3.45)	(−3.83)	(−2.78)						
β_{MKT}	−0.12	−0.06	0.04	0.10	0.11						
	(−2.74)	(−1.61)	(1.20)	(2.44)	(2.98)						
β_{SMB}	0.01	0.00	0.13	0.35	0.10						
	(0.17)	(0.04)	(2.60)	(6.11)	(1.73)						
β_{HML}	0.60	0.56	0.64	0.25	0.08						
	(9.16)	(9.54)	(12.02)	(4.17)	(1.28)						

January 2011 - December 2021											
	GP/A					Profitability strategies					
BE/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	0.77%	0.87%	0.76%	1.18%	0.98%	0.21%	0.17%	0.40%	−0.18	−0.31	−0.04
						(0.84)	(0.67)	(1.65)	(−2.90)	(−3.07)	(−0.54)
2	0.28%	0.76%	0.72%	0.92%	0.46%	0.17%	0.14%	0.30%	−0.13	−0.16	−0.12
						(0.81)	(0.67)	(1.40)	(−2.49)	(−1.83)	(−1.72)
3	0.26%	0.51%	0.55%	0.90%	0.99%	0.74%	0.66%	0.63%	0.16	0.09	0.36
						(2.24)	(2.01)	(1.93)	(1.99)	(0.63)	(3.35)
4	0.56%	0.23%	0.67%	0.26%	0.77%	0.21%	0.14%	0.07%	0.16	0.35	0.14
						(0.65)	(0.45)	(0.22)	(1.97)	(2.65)	(1.38)
High	0.88%	0.66%	0.46%	1.04%	1.04%	0.16%	0.05%	0.02%	−0.04	−0.22	−0.71
						(0.38)	(0.11)	(0.04)	(−0.39)	(−1.36)	(−5.55)
	Value strategies					Combined strategy					
Mean	0.12%	−0.21%	−0.30%	−0.14%	0.06%	0.27%	0.23%	0.43%	−0.18	−0.22	−0.19
	(0.32)	(−0.71)	(−1.04)	(−0.68)	(0.31)	(1.03)	(0.87)	(1.67)	(−2.75)	(−2.03)	(−2.28)
Geomean	0.03%	−0.26%	−0.35%	−0.16%	0.04%						
	(0.08)	(−0.90)	(−1.22)	(−0.81)	(0.18)						
α	0.37%	−0.43%	−0.44%	−0.16%	−0.01%						
	(1.03)	(−1.66)	(−1.58)	(−0.76)	(−0.07)						
β_{MKT}	−0.14	0.24	0.16	0.01	0.00						
	(−1.53)	(3.66)	(2.29)	(0.13)	(0.02)						
β_{SMB}	0.00	0.23	0.12	0.08	0.10						
	(0.02)	(2.11)	(1.03)	(0.87)	(1.09)						
β_{HML}	0.52	0.38	0.35	0.09	−0.15						
	(4.37)	(4.43)	(3.86)	(1.21)	(−2.12)						

Table A.27: Double Sort on BE/ME and GPSGA/A - Subsamples Before and After 2010

This table shows the value-weighted excess returns of the portfolios first sorted into quintiles based on BE/ME, then within the quintiles based on GPSGA/A. Portfolios are rebalanced annually at the end of June. The first part covers the sample period from June 1980 to December 2010, and the second part the period from January 2011 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The value strategies are the high-minus-low BE/ME portfolios, profitability strategies are the high-minus-low GPSGA/A portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies).

June 1980 - December 2010											
	GPSGA/A					Profitability strategies					
BE/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	−0.15%	0.01%	0.37%	0.09%	0.41%	0.56%	0.50%	0.63%	−0.03	−0.34	−0.01
						(3.01)	(2.69)	(3.47)	(−0.75)	(−5.67)	(−0.20)
2	−0.25%	0.15%	0.05%	0.39%	0.36%	0.61%	0.55%	0.67%	−0.02	−0.10	−0.11
						(3.35)	(3.07)	(3.62)	(−0.46)	(−1.56)	(−1.64)
3	−0.11%	0.04%	0.05%	0.19%	0.17%	0.28%	0.23%	0.43%	−0.10	−0.10	−0.21
						(1.61)	(1.30)	(2.41)	(−2.51)	(−1.64)	(−3.37)
4	−0.03%	0.24%	−0.03%	0.24%	0.24%	0.27%	0.20%	0.25%	0.08	−0.22	0.03
						(1.40)	(1.02)	(1.28)	(1.74)	(−3.40)	(0.41)
High	0.34%	0.39%	0.38%	0.35%	0.45%	0.12%	−0.00%	0.43%	−0.15	−0.39	−0.47
						(0.47)	(−0.00)	(1.80)	(−2.83)	(−4.99)	(−5.71)
	Value strategies					Combined strategy					
Mean	0.48%	0.38%	0.01%	0.26%	0.04%	0.60%	0.53%	0.58%	0.02	−0.24	0.12
	(2.08)	(1.68)	(0.05)	(1.39)	(0.24)	(3.11)	(2.79)	(3.05)	(0.38)	(−3.81)	(1.75)
Geomean	0.39%	0.30%	−0.06%	0.20%	−0.01%						
	(1.73)	(1.37)	(−0.30)	(1.06)	(−0.07)						
α	−0.27%	−0.52%	−0.73%	−0.36%	−0.47%						
	(−1.20)	(−2.58)	(−4.21)	(−1.93)	(−2.71)						
β_{MKT}	0.17	0.31	0.13	0.14	0.05						
	(3.39)	(6.73)	(3.17)	(3.39)	(1.23)						
β_{SMB}	0.16	0.40	0.33	0.00	0.10						
	(2.14)	(6.05)	(5.72)	(0.06)	(1.83)						
β_{HML}	0.58	0.69	0.56	0.33	0.13						
	(7.58)	(9.91)	(9.36)	(5.06)	(2.09)						
January 2011 - December 2021											
	GPSGA/A					Profitability strategies					
BE/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	0.84%	0.53%	0.81%	0.90%	1.04%	0.20%	0.15%	0.23%	−0.06	−0.48	−0.10
						(0.71)	(0.52)	(0.82)	(−0.85)	(−4.23)	(−1.14)
2	0.13%	0.55%	0.59%	0.53%	1.01%	0.88%	0.84%	0.76%	0.05	−0.38	−0.18
						(3.61)	(3.44)	(3.14)	(0.83)	(−3.76)	(−2.30)
3	0.19%	0.58%	0.67%	0.77%	1.08%	0.89%	0.83%	0.71%	0.18	−0.26	0.21
						(2.92)	(2.74)	(2.28)	(2.34)	(−1.98)	(2.03)
4	0.29%	0.63%	0.47%	0.13%	0.60%	0.31%	0.21%	0.14%	0.16	−0.13	0.12
						(0.80)	(0.53)	(0.33)	(1.58)	(−0.78)	(0.87)
High	0.78%	0.79%	0.63%	0.69%	1.15%	0.37%	0.26%	0.47%	−0.20	−0.64	−0.44
						(0.92)	(0.65)	(1.33)	(−2.24)	(−4.28)	(−3.75)
	Value strategies					Combined strategy					
Mean	−0.06%	0.27%	−0.18%	−0.22%	0.11%	0.31%	0.27%	0.26%	0.02	−0.42	−0.05
	(−0.17)	(0.90)	(−0.58)	(−0.78)	(0.53)	(1.17)	(1.00)	(0.95)	(0.30)	(−3.74)	(−0.57)
Geomean	−0.13%	0.21%	−0.25%	−0.27%	0.08%						
	(−0.39)	(0.71)	(−0.77)	(−0.95)	(0.40)						
α	−0.26%	−0.04%	−0.48%	−0.57%	−0.01%						
	(−0.81)	(−0.15)	(−1.55)	(−2.36)	(−0.05)						
β_{MKT}	0.22	0.25	0.25	0.32	0.08						
	(2.76)	(3.64)	(3.18)	(5.28)	(1.45)						
β_{SMB}	0.22	0.49	0.29	0.22	0.06						
	(1.61)	(4.33)	(2.27)	(2.17)	(0.69)						
β_{HML}	0.39	0.05	0.12	0.26	0.06						
	(3.72)	(0.61)	(1.17)	(3.28)	(0.76)						

Table A.28: Double Sort on BE/ME and GPSGAI/BE - Subsamples Before and After 2010

This table shows the value-weighted excess returns of the portfolios first sorted into quintiles based on BE/ME, then within the quintiles based on GPSGAI/BE. Portfolios are rebalanced annually at the end of June. The first part covers the sample period from June 1980 to December 2010, and the second part the period from January 2011 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The value strategies are the high-minus-low BE/ME portfolios, profitability strategies are the high-minus-low GPSGAI/BE portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies).

June 1980 - December 2010											
	GPSGAI/BE					Profitability strategies					
BE/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	−0.27%	−0.05%	0.18%	0.34%	0.30%	0.57%	0.50%	0.51%	−0.02	−0.26	0.31
						(2.78)	(2.42)	(2.62)	(−0.39)	(−4.08)	(4.50)
2	−0.19%	0.01%	0.16%	0.07%	0.39%	0.58%	0.53%	0.61%	−0.02	−0.15	−0.01
						(3.42)	(3.12)	(3.58)	(−0.44)	(−2.67)	(−0.22)
3	0.00%	0.04%	0.06%	0.26%	0.24%	0.24%	0.19%	0.20%	0.05	−0.10	0.06
						(1.50)	(1.23)	(1.25)	(1.39)	(−1.99)	(1.13)
4	−0.15%	0.23%	0.22%	0.16%	0.34%	0.50%	0.43%	0.51%	0.05	−0.24	−0.00
						(2.67)	(2.31)	(2.71)	(1.08)	(−3.93)	(−0.03)
High	0.17%	0.29%	0.58%	0.38%	0.75%	0.58%	0.39%	0.67%	−0.13	−0.12	0.01
						(2.01)	(1.09)	(2.28)	(−2.02)	(−1.28)	(0.14)
	Size strategies					Combined strategy					
Mean	0.44%	0.34%	0.39%	0.04%	0.45%	1.02%	0.92%	0.77%	−0.01	0.13	0.69
	(1.53)	(1.19)	(1.96)	(0.23)	(1.84)	(4.31)	(3.95)	(3.54)	(−0.24)	(1.81)	(9.15)
Geomean	0.31%	0.20%	0.32%	−0.02%	0.34%						
	(1.16)	(0.71)	(1.59)	(−0.12)	(1.42)						
α	−0.33%	−0.46%	−0.27%	−0.59%	−0.17%						
	(−1.17)	(−1.71)	(−1.41)	(−3.22)	(−0.70)						
β_{MKT}	0.12	0.11	0.08	0.01	0.01						
	(1.96)	(1.74)	(1.83)	(0.20)	(0.12)						
β_{SMB}	0.26	0.34	0.30	0.34	0.40						
	(2.83)	(3.85)	(4.67)	(5.57)	(5.03)						
β_{HML}	0.67	0.73	0.43	0.45	0.38						
	(6.93)	(7.85)	(6.36)	(7.02)	(4.54)						

January 2011 - December 2021											
	GPSGAI/BE					Profitability strategies					
BE/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	0.45%	0.83%	0.88%	0.84%	0.83%	0.38%	0.34%	0.51%	−0.07	−0.28	0.25
						(1.61)	(1.45)	(2.17)	(−1.11)	(−2.83)	(3.27)
2	0.14%	0.65%	0.74%	0.62%	0.90%	0.77%	0.73%	0.73%	0.02	−0.39	0.00
						(3.05)	(2.88)	(2.85)	(0.33)	(−3.64)	(0.06)
3	0.46%	0.69%	0.43%	0.61%	1.24%	0.78%	0.73%	0.90%	−0.06	−0.16	0.22
						(2.69)	(2.52)	(2.97)	(−0.77)	(−1.25)	(2.19)
4	0.13%	0.49%	0.08%	0.46%	0.47%	0.34%	0.25%	0.24%	0.09	−0.23	0.08
						(0.92)	(0.67)	(0.61)	(0.93)	(−1.39)	(0.62)
High	0.57%	0.33%	0.87%	0.18%	0.78%	0.20%	−0.06%	−0.35%	0.43	−0.25	−0.06
						(0.29)	(−0.09)	(−0.49)	(2.41)	(−0.83)	(−0.27)
	Size strategies					Combined strategy					
Mean	0.12%	−0.50%	−0.01%	−0.67%	−0.06%	0.32%	0.04%	−0.29%	0.69	0.03	0.85
	(0.23)	(−1.11)	(−0.03)	(−1.71)	(−0.08)	(0.44)	(0.06)	(−0.42)	(3.93)	(0.10)	(3.71)
Geomean	−0.06%	−0.62%	−0.09%	−0.77%	−0.34%						
	(−0.12)	(−1.43)	(−0.25)	(−1.95)	(−0.55)						
α	0.02%	−0.55%	0.03%	−0.84%	−0.85%						
	(0.03)	(−1.33)	(0.10)	(−2.38)	(−1.20)						
β_{MKT}	0.25	0.14	0.09	0.24	0.75						
	(2.10)	(1.39)	(1.22)	(2.66)	(4.26)						
β_{SMB}	0.28	0.65	0.44	0.37	0.31						
	(1.40)	(3.77)	(3.68)	(2.51)	(1.05)						
β_{HML}	0.92	0.54	0.67	0.55	0.60						
	(5.78)	(3.95)	(7.00)	(4.73)	(2.58)						

Table A.29: Double Sort on GPSGA/ME and GP/A - Subsamples Before and After 2010

This table shows the value-weighted excess returns of the portfolios first sorted into quintiles based on GPSGA/ME, then within the quintiles based on GP/A. Portfolios are rebalanced annually at the end of June. The first part covers the sample period from June 1980 to December 2010, and the second part the period from January 2011 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The value strategies are the high-minus-low GPSGA/ME portfolios, profitability strategies are the high-minus-low GP/A portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies).

June 1980 - December 2010

	GP/A					Profitability strategies					
GPSGA/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	−0.34%	−0.05%	−0.01%	0.20%	0.68%	1.02%	0.88%	1.26%	−0.12	−0.65	−0.21
						(3.74)	(3.25)	(4.90)	(−2.05)	(−7.69)	(−2.38)
2	−0.02%	0.17%	0.40%	0.44%	0.39%	0.41%	0.36%	0.40%	0.10	−0.20	−0.03
						(2.34)	(2.01)	(2.24)	(2.56)	(−3.41)	(−0.56)
3	−0.05%	0.17%	0.14%	0.34%	0.57%	0.62%	0.57%	0.61%	−0.01	0.06	0.01
						(3.79)	(3.50)	(3.64)	(−0.16)	(1.11)	(0.21)
4	0.06%	0.26%	0.42%	0.67%	0.80%	0.74%	0.66%	0.88%	−0.16	0.04	−0.16
						(3.51)	(3.21)	(4.15)	(−3.29)	(0.56)	(−2.23)
High	0.63%	0.48%	0.56%	0.61%	0.72%	0.09%	0.01%	0.05%	−0.07	0.25	0.13
						(0.42)	(0.03)	(0.22)	(−1.47)	(3.62)	(1.76)

	Size strategies					Combined strategy					
Mean	0.97%	0.53%	0.57%	0.41%	0.05%	1.06%	0.96%	1.10%	−0.15	−0.26	0.26
	(4.23)	(2.83)	(2.62)	(1.91)	(0.22)	(4.40)	(4.00)	(4.73)	(−2.86)	(−3.37)	(3.25)
Geomean	0.88%	0.47%	0.48%	0.33%	−0.04%						
	(3.82)	(2.49)	(2.23)	(1.52)	(−0.17)						
α	0.62%	−0.00%	−0.00%	−0.21%	−0.59%						
	(2.95)	(−0.01)	(−0.01)	(−1.03)	(−2.96)						
β_{MKT}	−0.08	−0.00	0.02	0.01	−0.03						
	(−1.63)	(−0.00)	(0.56)	(0.20)	(−0.67)						
β_{SMB}	−0.50	−0.23	−0.35	0.09	0.40						
	(−7.25)	(−4.02)	(−5.69)	(1.36)	(6.12)						
β_{HML}	0.13	0.40	0.52	0.52	0.47						
	(1.74)	(6.75)	(7.99)	(7.24)	(6.83)						

January 2011 - December 2021

	GP/A					Profitability strategies					
GPSGA/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	0.36%	0.37%	0.86%	1.15%	1.25%	0.89%	0.76%	0.54%	0.14	−0.43	−0.65
						(2.03)	(1.73)	(1.29)	(1.30)	(−2.45)	(−4.69)
2	0.37%	0.57%	0.79%	0.67%	1.30%	0.92%	0.88%	0.65%	0.15	−0.27	−0.35
						(3.60)	(3.45)	(2.60)	(2.35)	(−2.56)	(−4.20)
3	0.46%	0.59%	1.11%	0.99%	0.94%	0.48%	0.40%	0.53%	−0.12	−0.11	−0.35
						(1.43)	(1.19)	(1.58)	(−1.43)	(−0.81)	(−3.11)
4	0.87%	0.27%	0.91%	1.16%	0.84%	−0.04%	−0.17%	0.25%	−0.36	−0.58	−0.49
						(−0.08)	(−0.37)	(0.65)	(−3.71)	(−3.63)	(−3.93)
High	0.91%	0.90%	1.07%	0.69%	1.40%	0.49%	0.41%	0.46%	0.11	0.16	0.42
						(1.40)	(1.20)	(1.35)	(1.29)	(1.13)	(3.69)

	Size strategies					Combined strategy					
Mean	0.55%	0.53%	0.22%	−0.46%	0.15%	1.04%	0.96%	1.25%	−0.14	−0.27	0.20
	(1.33)	(1.77)	(0.73)	(−1.33)	(0.38)	(2.90)	(2.67)	(3.35)	(−1.49)	(−1.71)	(1.61)
Geomean	0.44%	0.47%	0.16%	−0.54%	0.05%						
	(1.04)	(1.57)	(0.54)	(−1.55)	(0.12)						
α	0.74%	0.83%	−0.08%	−0.49%	0.66%						
	(1.80)	(2.87)	(−0.33)	(−1.73)	(1.90)						
β_{MKT}	−0.25	−0.24	0.31	0.16	−0.28						
	(−2.42)	(−3.26)	(4.92)	(2.26)	(−3.13)						
β_{SMB}	−0.43	−0.28	−0.07	0.07	0.17						
	(−2.49)	(−2.33)	(−0.66)	(0.57)	(1.14)						
β_{HML}	−0.22	0.28	0.50	0.81	0.86						
	(−1.61)	(2.90)	(5.94)	(8.60)	(7.39)						

Table A.30: Double Sort on GPSGA/ME and GPSGA/A - Subsamples Before and After 2010

This table shows the value-weighted excess returns of the portfolios first sorted into quintiles based on GPSGA/ME, then within the quintiles based on GPSGA/A. Portfolios are rebalanced annually at the end of June. The first part covers the sample period from June 1980 to December 2010, and the second part the period from January 2011 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The value strategies are the high-minus-low GPSGA/ME portfolios, profitability strategies are the high-minus-low GPSGA/A portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies).

June 1980 - December 2010

	GPSGA/A					Profitability strategies					
GPSGA/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	−0.36%	0.06%	0.10%	0.08%	0.53%	0.88%	0.75%	1.27%	−0.12	−0.91	−0.51
						(3.28)	(2.72)	(5.52)	(−2.36)	(−12.14)	(−6.36)
2	−0.00%	0.29%	−0.01%	0.33%	0.28%	0.28%	0.21%	0.32%	0.16	−0.30	−0.23
						(1.44)	(1.08)	(1.67)	(3.63)	(−4.80)	(−3.44)
3	0.13%	0.22%	0.33%	0.14%	0.42%	0.29%	0.24%	0.46%	−0.07	−0.11	−0.32
						(1.70)	(1.40)	(2.74)	(−1.95)	(−1.98)	(−5.42)
4	0.31%	0.40%	0.19%	0.23%	0.43%	0.13%	0.03%	0.42%	−0.08	−0.35	−0.55
						(0.54)	(0.12)	(1.89)	(−1.62)	(−4.82)	(−7.19)
High	1.13%	0.32%	0.56%	0.51%	0.87%	−0.25%	−0.47%	0.14%	−0.24	−0.47	−0.54
						(−0.77)	(−1.32)	(0.45)	(−3.31)	(−4.49)	(−4.87)
	Size strategies					Combined strategy					
Mean	1.48%	0.26%	0.45%	0.43%	0.35%	1.23%	1.06%	1.42%	−0.21	−0.72	0.09
	(4.74)	(1.20)	(2.23)	(2.10)	(1.32)	(3.99)	(3.47)	(5.09)	(−3.24)	(−7.88)	(0.95)
Geomean	1.32%	0.17%	0.38%	0.35%	0.22%						
	(4.42)	(0.82)	(1.88)	(1.73)	(0.86)						
α	0.85%	−0.39%	−0.12%	−0.20%	−0.27%						
	(2.87)	(−1.90)	(−0.61)	(−1.05)	(−1.07)						
β_{MKT}	0.04	0.08	0.03	−0.04	−0.08						
	(0.52)	(1.70)	(0.73)	(−0.85)	(−1.43)						
β_{SMB}	−0.24	−0.01	0.10	0.15	0.20						
	(−2.51)	(−0.22)	(1.49)	(2.43)	(2.40)						
β_{HML}	0.63	0.50	0.34	0.56	0.59						
	(6.09)	(7.04)	(4.78)	(8.66)	(6.83)						

January 2011 - December 2021

	GPSGA/A					Profitability strategies					
GPSGA/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	0.50%	0.44%	0.58%	0.74%	1.43%	0.93%	0.75%	1.09%	−0.27	−1.10	−0.46
						(1.84)	(1.43)	(2.60)	(−2.53)	(−6.23)	(−3.33)
2	0.58%	0.42%	0.72%	0.73%	1.05%	0.47%	0.43%	0.16%	0.20	−0.23	−0.20
						(2.00)	(1.85)	(0.70)	(3.46)	(−2.37)	(−2.57)
3	0.46%	0.39%	0.75%	0.74%	1.12%	0.66%	0.60%	0.38%	0.18	−0.21	−0.19
						(2.20)	(2.02)	(1.22)	(2.39)	(−1.62)	(−1.90)
4	0.93%	0.66%	0.60%	0.80%	1.22%	0.29%	0.22%	0.16%	0.03	−0.59	−0.30
						(0.87)	(0.66)	(0.50)	(0.33)	(−4.55)	(−2.86)
High	0.80%	0.84%	0.46%	0.57%	1.17%	0.37%	0.29%	0.26%	0.01	−0.34	−0.35
						(1.08)	(0.85)	(0.76)	(0.09)	(−2.40)	(−3.06)
	Size strategies					Combined strategy					
Mean	0.29%	0.40%	−0.12%	−0.17%	−0.27%	0.66%	0.55%	1.16%	−0.35	−0.81	0.44
	(0.72)	(0.90)	(−0.34)	(−0.52)	(−0.67)	(1.57)	(1.29)	(3.16)	(−3.76)	(−5.28)	(3.66)
Geomean	0.18%	0.28%	−0.20%	−0.24%	−0.37%						
	(0.45)	(0.67)	(−0.57)	(−0.73)	(−0.93)						
α	0.86%	0.15%	0.12%	−0.04%	0.03%						
	(2.44)	(0.36)	(0.38)	(−0.12)	(0.08)						
β_{MKT}	−0.35	0.28	−0.10	−0.00	−0.08						
	(−3.99)	(2.69)	(−1.20)	(−0.04)	(−0.93)						
β_{SMB}	−0.46	0.34	−0.05	0.17	0.29						
	(−3.15)	(1.96)	(−0.37)	(1.38)	(2.04)						
β_{HML}	0.79	0.49	0.66	0.65	0.91						
	(6.81)	(3.54)	(6.09)	(6.65)	(8.17)						

Table A.31: Double Sort on GPSGA/ME and GPSGAI/BE - Subsamples Before and After 2010

This table shows the value-weighted excess returns of the portfolios first sorted into quintiles based on GPSGA/ME, then within the quintiles based on GPSGAI/BE. Portfolios are rebalanced annually at the end of June. The first part covers the sample period from June 1980 to December 2010, and the second part the period from January 2011 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The value strategies are the high-minus-low GPSGA/ME portfolios, profitability strategies are the high-minus-low GPSGAI/BE portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies).

June 1980 - December 2010

	GPSGAI/BE					Profitability strategies					
GPSGA/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	−0.42%	−0.07%	−0.03%	−0.14%	0.52%	0.95%	0.79%	1.34%	−0.24	−0.84	−0.36
						(3.25)	(2.63)	(5.19)	(−4.18)	(−9.98)	(−4.07)
2	−0.10%	0.18%	−0.06%	0.42%	0.25%	0.35%	0.30%	0.40%	0.16	−0.39	−0.24
						(2.04)	(1.73)	(2.52)	(4.39)	(−7.49)	(−4.29)
3	−0.11%	0.02%	−0.07%	0.09%	0.43%	0.54%	0.47%	0.71%	−0.06	−0.26	−0.29
						(2.77)	(2.50)	(3.72)	(−1.41)	(−4.16)	(−4.35)
4	0.10%	0.09%	0.34%	0.04%	0.36%	0.27%	0.19%	0.37%	0.02	−0.30	−0.19
						(1.28)	(0.91)	(1.79)	(0.35)	(−4.44)	(−2.68)
High	0.33%	0.52%	0.49%	0.47%	0.53%	0.20%	0.09%	0.30%	0.05	−0.37	−0.21
						(0.81)	(0.37)	(1.22)	(0.96)	(−4.60)	(−2.48)
	Size strategies					Combined strategy					
Mean	0.75%	0.59%	0.52%	0.62%	0.00%	0.95%	0.80%	1.08%	−0.14	−0.60	0.11
	(3.09)	(2.91)	(2.88)	(2.73)	(0.02)	(3.27)	(2.79)	(3.95)	(−2.25)	(−6.67)	(1.22)
Geomean	0.64%	0.52%	0.46%	0.52%	−0.09%						
	(2.69)	(2.62)	(2.56)	(2.33)	(−0.40)						
α	0.35%	−0.03%	−0.08%	0.07%	−0.69%						
	(1.55)	(−0.15)	(−0.45)	(0.32)	(−3.03)						
β_{MKT}	−0.19	0.03	0.06	−0.10	0.11						
	(−3.69)	(0.61)	(1.44)	(−1.97)	(2.06)						
β_{SMB}	−0.22	0.02	0.14	0.07	0.25						
	(−2.95)	(0.39)	(2.40)	(0.99)	(3.42)						
β_{HML}	0.32	0.50	0.36	0.47	0.47						
	(4.08)	(7.40)	(5.70)	(6.38)	(6.06)						

January 2011 - December 2021

	GPSGAI/BE					Profitability strategies					
GPSGA/ME	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	0.36%	0.44%	0.27%	0.42%	1.14%	0.78%	0.61%	0.96%	−0.29	−0.96	−0.54
						(1.55)	(1.21)	(2.25)	(−2.72)	(−5.38)	(−3.82)
2	0.29%	0.43%	0.62%	0.88%	0.90%	0.61%	0.58%	0.26%	0.25	−0.21	−0.15
						(2.68)	(2.53)	(1.16)	(4.47)	(−2.20)	(−2.04)
3	0.60%	−0.03%	0.18%	0.60%	0.74%	0.14%	0.06%	−0.02%	0.02	−0.48	−0.46
						(0.40)	(0.16)	(−0.05)	(0.19)	(−3.61)	(−4.37)
4	0.69%	0.42%	0.88%	0.76%	0.87%	0.18%	0.11%	0.20%	−0.13	−0.47	−0.47
						(0.56)	(0.36)	(0.72)	(−1.86)	(−4.00)	(−5.03)
High	0.84%	0.79%	0.35%	0.77%	1.10%	0.26%	0.18%	0.12%	0.05	−0.50	−0.22
						(0.77)	(0.54)	(0.35)	(0.62)	(−3.58)	(−2.03)
	Size strategies					Combined strategy					
Mean	0.49%	0.34%	0.08%	0.35%	−0.04%	0.74%	0.62%	1.08%	−0.24	−0.66	0.29
	(1.20)	(0.86)	(0.24)	(0.90)	(−0.10)	(1.70)	(1.41)	(2.56)	(−2.27)	(−3.73)	(2.07)
Geomean	0.38%	0.24%	0.01%	0.25%	−0.14%						
	(0.95)	(0.59)	(0.02)	(0.66)	(−0.35)						
α	0.92%	0.45%	−0.08%	0.12%	0.08%						
	(2.31)	(1.14)	(−0.28)	(0.34)	(0.24)						
β_{MKT}	−0.29	−0.02	0.23	0.27	0.05						
	(−2.91)	(−0.16)	(3.23)	(2.98)	(0.61)						
β_{SMB}	−0.16	−0.24	−0.03	0.22	0.30						
	(−0.97)	(−1.46)	(−0.26)	(1.46)	(2.16)						
β_{HML}	0.52	0.58	0.66	0.52	0.83						
	(3.91)	(4.42)	(7.03)	(4.45)	(7.62)						

Table A.32: Double Sort on HHI (Fama-French 49 Industries) and GP/A

This table shows the value-weighted excess returns for the portfolios first sorted into quintiles based on the HHI of the Fama-French 49 industries, then within the quintiles based on GP/A. Portfolios are rebalanced annually at the end of June and cover the whole sample from June 1980 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The industry concentration strategies are the high concentration minus low concentration portfolios, profitability strategies are the high-minus-low GP/A portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies).

	GP/A					Profitability strategies					
HHI	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	0.72%	0.51%	0.57%	0.50%	0.91%	0.20%	0.07%	−0.04%	0.34	0.24	−0.19
						(0.91)	(0.32)	(−0.21)	(7.38)	(3.56)	(−2.89)
2	0.54%	0.25%	0.40%	0.59%	0.83%	0.30%	0.22%	0.44%	−0.13	0.01	−0.24
						(1.67)	(1.24)	(2.48)	(−3.17)	(0.14)	(−4.18)
3	0.15%	0.25%	0.53%	0.49%	0.66%	0.50%	0.42%	0.65%	−0.13	−0.23	−0.17
						(2.79)	(2.36)	(3.66)	(−3.13)	(−3.81)	(−2.82)
4	0.12%	0.19%	0.33%	0.46%	0.57%	0.45%	0.32%	0.67%	−0.12	−0.54	−0.42
						(1.99)	(1.42)	(3.14)	(−2.37)	(−7.32)	(−5.89)
High	0.45%	0.63%	0.69%	0.57%	0.92%	0.46%	0.38%	0.73%	−0.26	−0.10	−0.32
						(2.50)	(2.02)	(4.06)	(−6.30)	(−1.56)	(−5.40)
	Industry concentration strategies					Combined strategy					
Mean	−0.27%	0.12%	0.12%	0.07%	0.00%	0.20%	0.09%	0.16%	0.13	−0.03	−0.28
	(−1.22)	(0.65)	(0.70)	(0.38)	(0.01)	(0.97)	(0.40)	(0.78)	(2.82)	(−0.44)	(−4.11)
Geomean	−0.39%	0.03%	0.05%	−0.01%	−0.07%						
	(−1.71)	(0.18)	(0.27)	(−0.04)	(−0.41)						
α	−0.90%	−0.55%	−0.29%	−0.25%	−0.13%						
	(−4.31)	(−3.39)	(−1.78)	(−1.47)	(−0.79)						
β_{MKT}	0.39	0.44	0.14	0.07	−0.20						
	(8.24)	(11.85)	(3.64)	(1.81)	(−5.50)						
β_{SMB}	0.07	0.22	0.14	−0.29	−0.27						
	(0.95)	(3.91)	(2.42)	(−4.89)	(−4.92)						
β_{HML}	0.04	−0.00	−0.15	−0.16	−0.09						
	(0.58)	(−0.05)	(−2.70)	(−2.89)	(−1.72)						

Table A.33: Double Sort on HHI (Fama-French 49 Industries) and GPSGAI/BE

This table shows the value-weighted excess returns for the portfolios first sorted into quintiles based on the HHI of the Fama-French 49 industries, then within the quintiles based on GPSGAI/BE. Portfolios are rebalanced annually at the end of June and cover the whole sample from June 1980 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The industry concentration strategies are the high concentration minus low concentration portfolios, profitability strategies the high-minus-low GPSGAI/BE portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies).

	GPSGAI/BE					Profitability strategies					
HHI	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	−0.09%	0.17%	0.30%	0.58%	0.78%	0.87%	0.78%	0.92%	−0.03	−0.39	0.02
						(4.51)	(4.12)	(4.89)	(−0.65)	(−5.98)	(0.24)
2	0.00%	0.28%	0.39%	0.49%	0.64%	0.64%	0.56%	0.83%	−0.11	−0.33	−0.40
						(3.62)	(3.21)	(4.95)	(−2.84)	(−5.78)	(−7.12)
3	0.03%	0.27%	0.40%	0.23%	0.51%	0.49%	0.43%	0.63%	−0.10	−0.33	−0.17
						(3.13)	(2.74)	(4.15)	(−2.82)	(−6.42)	(−3.36)
4	0.03%	0.04%	0.08%	0.25%	0.48%	0.46%	0.35%	0.73%	−0.25	−0.50	−0.22
						(2.24)	(1.70)	(3.88)	(−5.73)	(−7.68)	(−3.50)
High	0.17%	0.12%	0.43%	0.63%	0.81%	0.64%	0.52%	0.74%	−0.05	−0.36	−0.16
						(2.94)	(2.37)	(3.45)	(−1.08)	(−4.95)	(−2.20)
	Industry concentration strategies					Combined strategy					
Mean	0.26%	−0.04%	0.13%	0.05%	0.03%	0.89%	0.79%	1.07%	−0.14	−0.57	−0.09
	(1.17)	(−0.27)	(0.81)	(0.27)	(0.16)	(4.30)	(3.82)	(5.50)	(−3.07)	(−8.59)	(−1.37)
Geomean	0.14%	−0.11%	0.07%	−0.03%	−0.04%						
	(0.63)	(−0.67)	(0.42)	(−0.18)	(−0.23)						
α	0.00%	−0.49%	−0.21%	−0.29%	−0.18%						
	(0.00)	(−3.05)	(−1.31)	(−1.59)	(−1.14)						
β_{MKT}	−0.08	0.14	−0.02	0.03	−0.10						
	(−1.57)	(3.73)	(−0.49)	(0.70)	(−2.88)						
β_{SMB}	−0.20	−0.04	−0.09	−0.09	−0.18						
	(−2.69)	(−0.74)	(−1.64)	(−1.40)	(−3.34)						
β_{HML}	0.06	0.14	0.20	0.02	−0.11						
	(0.84)	(2.53)	(3.73)	(0.32)	(−2.08)						

Table A.34: Double Sort on TNIC HHI and GP/A, GPSGA/A and GPSGAI/BE

This table shows the value-weighted excess returns for the portfolios first sorted into quintiles based on the TNIC HHI (Hoberg and Phillips, 2016), then within the quintiles based on GP/A (first part), GPSGA/A (second part) and GPSGAI/BE (third part). Portfolios are rebalanced annually at the end of June and cover the whole sample from June 1988 to December 2021. The percentage values are percent per month, the values in brackets are the corresponding t-test values, and α stands for the Fama-French three-factor model alpha. The industry concentration strategies are the high concentration minus low concentration portfolios, profitability strategies are the high-minus-low GP/A portfolios, and the combined strategy is the high-high portfolio minus the low-low portfolio (column description of the profitability strategies applies).

	GP/A					Profitability strategies					
HHI	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	0.50%	0.46%	0.22%	0.30%	0.70%	0.20%	0.06%	0.02%	0.28	0.10	−0.40
						(0.77)	(0.23)	(0.09)	(4.77)	(1.17)	(−4.94)
2	0.45%	0.38%	0.67%	0.69%	0.86%	0.41%	0.32%	0.54%	−0.09	−0.36	−0.43
						(1.91)	(1.47)	(2.71)	(−1.80)	(−5.43)	(−6.58)
3	0.60%	0.51%	0.53%	0.55%	0.80%	0.20%	0.11%	0.44%	−0.24	−0.42	−0.28
						(0.93)	(0.49)	(2.31)	(−5.41)	(−6.73)	(−4.53)
4	0.55%	0.58%	0.60%	0.73%	0.80%	0.25%	0.17%	0.48%	−0.25	−0.19	−0.24
						(1.24)	(0.82)	(2.45)	(−5.42)	(−2.84)	(−3.64)
High	0.56%	0.48%	0.65%	0.70%	1.04%	0.48%	0.38%	0.63%	−0.14	−0.29	−0.28
						(2.18)	(1.74)	(2.98)	(−2.79)	(−4.19)	(−4.07)
	Industry concentration strategies					Combined strategy					
Mean	0.06%	0.02%	0.43%	0.40%	0.34%	0.54%	0.44%	0.48%	0.12	−0.27	−0.09
	(0.29)	(0.09)	(2.23)	(1.78)	(1.34)	(2.51)	(2.09)	(2.23)	(2.37)	(−3.83)	(−1.31)
Geomean	−0.03%	−0.06%	0.36%	0.29%	0.21%						
	(−0.13)	(−0.31)	(1.83)	(1.28)	(0.81)						
α	−0.39%	−0.40%	0.14%	0.35%	0.22%						
	(−1.92)	(−2.05)	(0.73)	(1.57)	(0.93)						
β_{MKT}	0.26	0.25	0.13	−0.21	−0.16						
	(5.44)	(5.39)	(2.89)	(−4.01)	(−2.79)						
β_{SMB}	0.03	0.13	0.00	−0.14	−0.36						
	(0.46)	(2.00)	(0.06)	(−1.90)	(−4.55)						
β_{HML}	0.19	−0.12	−0.34	−0.12	0.31						
	(2.88)	(−1.89)	(−5.52)	(−1.63)	(3.93)						
	GPSGA/A					Profitability strategies					
HHI	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	0.16%	0.29%	0.27%	0.36%	0.70%	0.54%	0.39%	0.62%	−0.02	−0.21	−0.52
						(1.98)	(1.42)	(2.34)	(−0.32)	(−2.41)	(−5.93)
2	0.32%	0.30%	0.42%	0.45%	0.92%	0.60%	0.51%	0.69%	−0.02	−0.48	−0.41
						(2.86)	(2.44)	(3.54)	(−0.33)	(−7.41)	(−6.49)
3	0.52%	0.46%	0.54%	0.40%	0.73%	0.20%	0.08%	0.42%	−0.17	−0.66	−0.43
						(0.87)	(0.30)	(2.10)	(−3.52)	(−9.96)	(−6.53)
4	0.31%	0.40%	0.67%	0.63%	0.81%	0.50%	0.42%	0.60%	−0.04	−0.52	−0.22
						(2.50)	(2.07)	(3.20)	(−1.01)	(−8.33)	(−3.63)
High	0.38%	0.55%	0.56%	0.57%	0.95%	0.57%	0.47%	0.77%	−0.13	−0.62	−0.46
						(2.53)	(2.08)	(3.86)	(−2.85)	(−9.51)	(−7.13)
	Industry concentration strategies					Combined strategy					
Mean	0.22%	0.26%	0.29%	0.21%	0.25%	0.79%	0.65%	1.05%	−0.22	−0.65	−0.34
	(0.99)	(1.23)	(1.47)	(1.29)	(0.96)	(2.99)	(2.44)	(4.33)	(−3.92)	(−8.12)	(−4.26)
Geomean	0.12%	0.17%	0.21%	0.15%	0.11%						
	(0.54)	(0.80)	(1.07)	(0.96)	(0.41)						
α	0.05%	−0.05%	0.11%	0.01%	0.19%						
	(0.20)	(−0.21)	(0.58)	(0.06)	(0.76)						
β_{MKT}	−0.09	0.08	−0.02	−0.04	−0.20						
	(−1.66)	(1.67)	(−0.53)	(−1.12)	(−3.41)						
β_{SMB}	−0.02	0.11	−0.20	−0.12	−0.43						
	(−0.22)	(1.50)	(−3.10)	(−2.19)	(−5.18)						
β_{HML}	0.13	0.07	−0.18	0.07	0.18						
	(1.72)	(0.98)	(−2.85)	(1.33)	(2.21)						
	GPSGAI/BE					Profitability strategies					
HHI	Low	2	3	4	High	Mean	Geomean	α	β_{MKT}	β_{SMB}	β_{HML}
Low	0.20%	0.05%	0.33%	0.46%	0.36%	0.16%	0.03%	0.32%	−0.14	−0.15	−0.48
						(0.65)	(0.14)	(1.35)	(−2.44)	(−1.88)	(−6.11)
2	0.25%	0.18%	0.37%	0.34%	0.79%	0.54%	0.45%	0.66%	−0.08	−0.49	−0.24
						(2.58)	(2.17)	(3.37)	(−1.69)	(−7.67)	(−3.76)
3	0.29%	0.26%	0.35%	0.41%	0.62%	0.33%	0.24%	0.41%	−0.03	−0.53	−0.16
						(1.59)	(1.14)	(2.13)	(−0.76)	(−8.29)	(−2.54)
4	0.20%	0.22%	0.58%	0.47%	0.62%	0.42%	0.35%	0.45%	0.03	−0.46	−0.18
						(2.26)	(1.86)	(2.57)	(0.67)	(−7.90)	(−3.08)
High	0.17%	0.46%	0.26%	0.56%	0.64%	0.47%	0.35%	0.60%	−0.11	−0.53	−0.05
						(1.96)	(1.47)	(2.65)	(−2.10)	(−7.06)	(−0.64)
	Industry concentration strategies					Combined strategy					
Mean	−0.03%	0.42%	−0.07%	0.10%	0.28%	0.44%	0.29%	0.74%	−0.31	−0.47	−0.31
	(−0.12)	(1.92)	(−0.31)	(0.51)	(1.38)	(1.66)	(1.08)	(2.99)	(−5.26)	(−5.77)	(−3.77)
Geomean	−0.15%	0.32%	−0.17%	0.02%	0.20%						
	(−0.61)	(1.49)	(−0.75)	(0.11)	(0.97)						
α	−0.09%	0.24%	−0.18%	−0.05%	0.19%						
	(−0.38)	(1.11)	(−0.81)	(−0.27)	(0.97)						
β_{MKT}	−0.19	−0.07	−0.13	−0.06	−0.17						
	(−3.33)	(−1.44)	(−2.40)	(−1.31)	(−3.72)						
β_{SMB}	0.06	0.07	−0.06	−0.13	−0.32						
	(0.78)	(0.99)	(−0.82)	(−2.13)	(−5.02)						
β_{HML}	−0.26	−0.07	−0.18	−0.25	0.17						
	(−3.21)	(−0.97)	(−2.47)	(−4.04)	(2.72)						

References

Akbas, F., Jiang, C., Koch, P. D., 2017. The trend in firm profitability and the cross-section of stock returns. The Accounting Review 92, 1–32.

Altman, E. I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance 23, 589–609.

Balakrishnan, K., Bartov, E., Faurel, L., 2010. Post loss/profit announcement drift. Journal of Accounting and Economics 50, 20–41.

Ball, R., Gerakos, J., Linnainmaa, J. T., Nikolaev, V. V., 2015. Deflating profitability. Journal of Financial Economics 117, 225–248.

Barth, M. E., Elliott, J. A., Finn, M. W., 1999. Market rewards associated with patterns of increasing earnings. Journal of Accounting Research 37, 387–413.

Campbell, J. Y., Hilscher, J., Szilagyi, J., 2008. In search of distress risk. Journal of Finance 63, 2899–2939.

Chan, L. K., Karceski, J., Lakonishok, J., 2007. Analysts' conflicts of interest and biases in earnings forecasts. Journal of Financial and Quantitative Analysis 42, 893–913.

Chan, L. K. C., Lakonishok, J., Sougiannis, T., 2001. The stock market valuation of research and development expenditures. Journal of Finance 56, 2431–2456.

Chiang, C.-h., Dai, W., Fan, J., Hong, H., Tu, J., 2019. Robust measures of earnings surprises. The Journal of Finance 74, 943–983.

Doyle, J. T., Lundholm, R. J., Soliman, M. T., 2006. The extreme future stock returns following i/b/e/s earnings surprises. Journal of Accounting Research 44, 849–887.

Fama, E. F., French, K. R., 1992. The cross-section of expected stock returns. Journal of Finance 47, 427–465.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.

Fama, E. F., French, K. R., 1996. Multifactor explanations of asset pricing anomalies. Journal of Finance 51, 55–84.

Fama, E. F., French, K. R., 2006. Profitability, investment and average returns. Journal of Financial Economics 82, 491–518.

Fama, E. F., French, K. R., 2008. Dissecting anomalies. The Journal of Finance 63, 1653–1678.

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Fama, E. F., MacBeth, J. D., 1973. Risk, return, and equilibrium: empirical tests. Journal of Political Economy 81, 607–636.

Foster, G., Olsen, C., Shevlin, T., 1984. Earnings releases, anomalies, and the behavior of security returns. Accounting Review pp. 574–603.

Haugen, R. A., Baker, N. L., 1996. Commonality in the determinants of expected stock returns. Journal of Financial Economics 41, 401–439.

Hoberg, G., Phillips, G., 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. The Review of Financial Studies 23, 3773–3811.

Hoberg, G., Phillips, G., 2016. Text-based network industries and endogenous product differentiation. Journal of Political Economy 124, 1423–1465.

Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: an investment approach. The Review of Financial Studies 28, 650–705.

Hou, K., Xue, C., Zhang, L., 2020. Replicating anomalies. The Review of Financial Studies 33, 2019–2133.

Linnainmaa, J. T., Roberts, M. R., 2018. The history of the cross-section of stock returns. The Review of Financial Studies 31, 2606–2649.

Malmendier, U., Shanthikumar, D., 2007. Are small investors naive about incentives? Journal of Financial Economics 85, 457–489.

Mikhail, M. B., Walther, B. R., Willis, R. H., 2007. When security analysts talk, who listens? The Accounting Review 82, 1227–1253.

Novy-Marx, R., 2013. The other side of value: The gross profitability premium. Journal of Financial Economics 108, 1–28.

Ohlson, J. A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research pp. 109–131.

Piotroski, J. D., 2000. Value investing: The use of historical financial statement information to separate winners from losers. Journal of Accounting Research pp. 1–41.

Richardson, S., Teoh, S. H., Wysocki, P. D., 2004. The walk-down to beatable analyst forecasts: The role of equity issuance and insider trading incentives. Contemporary Accounting Research 21, 885–924.

Soliman, M. T., 2008. The use of dupont analysis by market participants. The Accounting Review 83, 823–853.

Statement of Authorship / Selbständigkeitserklärung

I hereby declare that I have written this thesis independently and have not used any sources other than those indicated. I have marked all co-authorships as well as all passages taken verbatim or in spirit from sources as such. I am aware that otherwise the Senate is entitled to withdraw the title awarded on the basis of this thesis in accordance with Article 36 paragraph 1 letter o of the Law of September 5, 1996 on the University.

Ich erkläre hiermit, dass ich diese Dissertation selbständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Koautorenschaften sowie alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe o Gesetzes vom 5. September 1996 über die Universität zum Entzug des aufgrund dieser Dissertation verliehenen Titels berechtigt ist.

Jan Pichler — Bern, 31st July, 2023