Dissertation

Financial News and the Equity Market: A Machine Learning Approach to News Analysis

carried out for the purpose of obtaining the degree of Doctor technicae (Dr. techn.),

submitted at TU Wien

Faculty of Mechanical and Industrial Engineering

by

Dipl. Ing. Stefan SALBRECHTER

Mat.No.: 01327435

under the supervision of

Ao.Univ.Prof. Mag.rer.nat. Dr.rer.soc.oec. Dr.techn. Thomas Dangl

Institute of Management Science, E330

Reviewed by

Univ.-Prof. Mag. Günter Strobl, PhD

Institut für Finanzwirtschaft, Oskar-Morgenstern-Platz 1, 1090 Wien

and

Ao.Univ.Prof. Mag.rer.soc.oec. Dr.rer.soc.oec. Wolfgang Aussenegg

Institute of Management Science, E330

Place, Date

Signature

Affidavit

I declare in lieu of oath, that I wrote this thesis and carried out the associated research myself, using only the literature cited in this volume. If text passages from sources are used literally, they are marked as such.

I confirm that this work is original and has not been submitted for examination elsewhere, nor is it currently under consideration for a thesis elsewhere.

I acknowledge that the submitted work will be checked electronically-technically using suitable and state-of-the-art means (plagiarism detection software). On the one hand, this ensures that the submitted work was prepared according to the high-quality standards within the applicable rules to ensure good scientific practice "Code of Conduct" at the TU Wien. On the other hand, a comparison with other student theses avoids violations of my personal copyright.

Place, Date

Signature

Acknowledgements

At this point, I would like to sincerely thank everyone who has accompanied and supported me over the past few years. I'd like to especially thank my supervisor, Ao.Univ.Prof. Mag. DDr. Thomas Dangl. I value his time and the numerous conversations we've had. I also appreciate his sense of humor and his invaluable comments and advice, which have significantly influenced my academic work. I extend my deepest gratitude to my parents, Renate and Dietmar, who made my studies possible and have always supported me in important decisions. I would also like to thank my sisters, Ina and Bettina, who, along with my parents, have supported me in all situations, whether good or bad. And finally, from the bottom of my heart, I thank my girlfriend, Karina, who is always by my side and enriches my life day by day.

Danksagung

An dieser Stelle möchte ich mich herzlich bei allen bedanken, die mich über die letzten Jahre hinweg begleitet und unterstützt haben. Ein besonderer Dank gebührt meinem Betreuer Ao.Univ.Prof. Mag. DDr. Thomas Dangl. Ich schätze seine Zeit, die zahlreichen Gespräche, als auch seinen Sinn für Humor sowie seine wertvollen Kommentare und Ratschläge, die meine akademische Arbeit maßgeblich geprägt haben. Meinen tiefsten Dank richte ich zudem an meine Eltern, Renate und Dietmar, die mir das Studium ermöglicht und bei wichtigen Entscheidungen immer unterstützt haben. Ebenso danke ich meinen Schwestern, Ina und Bettina, die mir neben meinen Eltern in allen Lebenslagen, ob gut oder schlecht, beigestanden sind. Und schließlich danke ich von Herzen meiner Freundin, Karina, die stets an meiner Seite ist und mein Leben Tag für Tag bereichert.

Table of Contents

Abstract

This thesis explores the relationship between financial news and stock market movements. By applying advanced Natural Language Processing (NLP) techniques, we aim to gain deeper insights into the underlying dynamics present in the equity markets. The first part of this thesis addresses the efficiency of the stock market by examining how fast new information is incorporated into asset prices. Utilizing a self-trained BERT machine learning model, we estimate the sentiment of financial news. Our analysis identifies a positive correlation between news sentiment and next-day stock returns. Additionally, we observe that the market incorporates financial news into asset prices, usually within one day. The second part of the thesis focuses on the interplay between previous-day returns and overnight news. The results show over- and underreactions happening at market opening, which lead to a predictable and statistically significant pattern in asset returns – a reversal relative to the previous day's returns – on the following trading day. The third part of the thesis is devoted to the pricing of climate risks in the equity market. Again, NLP techniques are used to derive signals from public news that allow an estimation of firm specific climate risks. We are the first to document a positive and statistically significant risk premium on physical climate risk. We also document a regime-shift in the regulatory climate risk premium occurring around 2012 within a consistent framework. While the risk premium is positive prior to 2012, it turns negative thereafter. The analysis builds on a self-developed topic-modeling algorithm that utilizes Word2Vec embeddings and allows the generation of comprehensive topic clusters, as described in a technical note.

Kurzfassung

In dieser Dissertation wird der Zusammenhang zwischen Finanznachrichten und Aktienmarktbewegungen untersucht. Durch die Anwendung fortschrittlicher Techniken des Natural Language Processing (NLP) wollen wir bessere Einblicke in die zugrunde liegende Dynamik der Aktienmärkte gewinnen. Diese Arbeit basiert auf einem umfassenden Datensatz, welcher in Summe etwa 40 Millionen Nachrichten enthält, die im Zeitraum von Januar 1996 bis Juli 2021 von der Nachrichtenagentur Thomson Reuters veröffentlicht wurden. Im ersten Teil der Dissertation wird die Effizienz des Aktienmarkts untersucht, indem wir analysieren, wie schnell neue Informationen in Vermögenspreise einfließen. Dazu wird ein selbst trainiertes BERT-Modell genutzt, um das Sentiment von Finanznachrichten zu bestimmen. Unsere Ergebnisse zeigen eine positive Korrelation zwischen Nachrichtensentiment und Aktienrenditen des nächsten Tages. Zudem stellen wir fest, dass der Markt Finanznachrichten üblicherweise innerhalb eines Tages einpreist. Im zweiten Teil liegt der Fokus auf dem Zusammenspiel zwischen den Renditen des Vortages und Nachrichten, die über Nacht veröffentlicht werden. Hierbei beobachten wir Überreaktionen sowie Unterreaktionen zur Markteröffnung. Diese führen zu einem statistisch signifikanten, prädiktiven Muster in den Renditen des nachfolgenden Handelstages, einem "Reversal" relativ zu den Renditen des Vortages. Der dritte Teil befasst sich mit der Bepreisung von Klimarisiken am Aktienmarkt. Hierbei setzen wir erneut NLP-Techniken ein, um Signale aus öffentlichen Nachrichten zu extrahieren, die eine Schätzung der klimabezogenen Risiken einzelner Unternehmen ermöglichen. Erstmals in der Literatur dokumentieren wir eine positive und statistisch signifikante Risikoprämie für das physische Klimarisiko. Zudem konnten wir einen Regimewechsel in der regulatorischen Klimarisikoprämie feststellen. Dieser Regimewechsel ereignete sich etwa 2012: Während die Risikoprämie vor

Papers

- Salbrechter, S. (2021). Financial News Sentiment Learned by BERT: A Strict Out-of-Sample Study. *Available at SSRN: https://ssrn.com/abstract=3971880*
  Presented at the 36^{th} AWG Workshop

- Dangl, T. and Salbrechter, S. (2022). Overnight Reversal and the Asymmetric Reaction to News. *Available at SSRN: https://ssrn.com/abstract=4307675*
  Presented at the 2023 FMA European Conference in Aalborg and at the 2023 DGF Conference in Stuttgart

- Dangl, T. and Halling, M. and Salbrechter, S. (2023). Firm-specific Climate Risk Estimated from Public News. *Available at SSRN: https://ssrn.com/abstract=4575999*
  Presented at the 38^{th} AWG Workshop

- Dangl, T. and Salbrechter, S. (2023). Guided Topic Modeling with Word2Vec: A Technical Note. *Available at SSRN: https://ssrn.com/abstract=4575985*

1 Introduction

In recent years, several important breakthroughs were made in the fields of Natural Language Processing (NLP) and Artificial Intelligence (AI), including the development of models like ChatGPT, that have revolutionized the way we communicate, access information and interact with computers. At the heart of these advancements lies the development of transformer-based architectures, most notably BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer). These models, backed by massive amounts of data and cutting-edge infrastructure, have allowed machines to comprehend and generate human-like text with unprecedented accuracy. These advancements gave us, researchers in financial academia, new tools to study the interplay between the release of new information in the form of written texts and financial markets. While the literature on text analysis in the financial domain started to gain traction in the 2000s, early work is based on dictionary, count based approaches to transform news from written text into a machine readable numerical form (Tetlock, 2007; Loughran and McDonald, 2011; Bollen and Mao, 2011). These methods, however, have the disadvantage that the semantic similarity between words, as well as their context, is not taken into account. A first major breakthrough was achieved with Word2Vec (Mikolov et al., 2013). With this methodology words are not anymore considered as unrelated, orthogonal vectors, but are projected into a vector space that also captures the semantic similarity between words. The second major breakthrough was the development of the transformer architecture and the BERT model (Vaswani et al., 2017; Devlin et al., 2018) that takes the context specific meaning of words into account while enabling fast parallel processing.

In this thesis, I build on these advanced models to derive signals from financial news, in order to better understand the dynamics that are present in the U.S. equity market. To

be more precise, I start by investigating the short-term impact of news on stock returns. I thereby find a positive out-of-sample correlation between news sentiment and next-day stock returns. Also, the market reacts quickly to the release of financial news, which is largely incorporated into asset prices within one day. In a co-authored follow-up study, we focus on overnight news and the interplay with previous day returns. What we find is a predictable pattern in asset returns on the following trading day, that is caused by over- and underreactions of market participants to overnight news. With the third paper we contribute to the climate finance literature by using news to derive signals that allow an estimation of firm specific climate risks. We are the first to document a positive and statistically significant risk premium on physical climate risk. We also document a regime-shift in the regulatory climate risk premium occurring around 2012 within a consistent framework. While the risk premium is positive prior to 2012, it becomes negative afterwards. The analysis builds on a self-developed topic-modeling algorithm that utilizes Word2Vec embeddings and allows the generation of comprehensive topic clusters, as described in a technical note.

1.1 Methodology

1.1.1 Natural Language Processing (NLP)

Natural language processing (NLP) is a specialized discipline within computer science and artificial intelligence that gives computers the capability to understand and extract information from written text. In this thesis I employ two popular methods of the field, namely Word2Vec and BERT. Both are based on artificial neural networks that allow them to learn statistical patterns from data. What these models have in common, is the concept of mapping words into a vector space which allows them to capture the semantic similarity between words. In research paper 1 & 2, I use a self-trained version of BERT to estimate the sentiment of financial news articles. I explicitly pre-train the model on domain specific financial news data to account for the financial jargon and to ensure strict out-of-sample predictions. Although it was released in 2018, the BERT model remains one

of the top architectures for natural language understanding, including tasks like sentiment analysis. This is particularly noteworthy given its relatively small number of parameters compared to more recently released large language models. In the third research paper, we take a different approach and perform topic modeling to uncover climate related topics in almost 5 million news articles. This is achieved using a self-developed topic modeling algorithm, termed "Guided Topic Modeling with Word2Vec", which facilitates the generation of comprehensive topic word clusters within the word-embedding space from a self trained Word2Vec model. This methodology is described in more detail in a technical note, referred to as paper 4.

News Data and Data Preprocessing

This thesis is based on an extensive dataset of financial news published by Refinitiv, formerly known as Thomson Reuters. The dataset contains more than 40 million news items with the exact timestamp of publication and a complete tracking of update histories. The dataset covers the period from January 1996 to July 2021.

Word2Vec

Word2Vec, introduced by Mikolov et al. (2013), uses a shallow neural network model to convert words into numerical vectors. This transformative approach has changed how words are processed by machines. Instead of relying on the traditional sparse, high-dimensional representations like one-hot encoding, Word2Vec utilizes dense, continuous vector spaces. This allows words with similar meanings or contexts to be located in close proximity in the vector space. The primary motivation behind this approach is the distributional hypothesis, which posits that words that appear in similar contexts tend to have similar meanings. Word2Vec is available in two versions: Skip-gram, which aims to predict the context words from a given target word, and Continuous Bag of Words (CBOW), which seeks to predict a target word from its context. Hence, these are unsupervised learning algorithms as they require no labeled dataset.

BERT

BERT, which stands for Bidirectional Encoder Representations from Transformers was developed by researchers at Google AI in 2018. It has achieved state-of-the-art performance on various NLP tasks, including question answering, named entity recognition, and sentiment analysis. Unlike other models that process word sequences either from left-to-right or right-to-left, BERT is designed to consider context from both directions, leading to a more robust understanding of the semantic role of each word in a sentence. The BERT architecture is based on the transformer model and its key feature, the attention mechanism. The attention mechanism, as described in Equation (1.1), allows the model to determine which parts of the input sequence (represented by value vectors V) are most relevant for a given query vector Q (current word for which the algorithm determines the relevance of context words). It is computed with the inner product of each query vector with the key vectors K (key vectors correspond to all the words in the input sequence). When the inner product of a query and a key is high, it indicates that the corresponding key (and its value) is relevant to the query. The inner product is then scaled by the square root of the vector dimension d and transformed into a probability distribution by applying the softmax function. This results in the attention weights. These weights are then used to take a weighted sum of the values, producing the final output O of the attention mechanism.

$$O = \text{softmax}\left(\frac{QK'}{\sqrt{d}}\right)V \tag{1.1}$$

BERT uses multi-head attention, meaning it runs multiple attention operations in parallel, each looking at different parts or "subspaces" of the input. This allows the model to capture various aspects or types of relationships in the data. The attention mechanism is crucial in enabling BERT's bidirectional understanding of text. This ability to "attend" to distant words helps in capturing long-range dependencies and understanding complex sentence structures.

For a robust training of BERT, I follow Liu et al. (2019). They propose further optimizations regarding the hyperparameter choice, tokenizer and learning objective. The model is then trained in a two-step procedure: First, it undergoes unsupervised pre-training on large amounts of text data using the "masked language modeling" (MLM) task. This is then followed by task-specific supervised fine-tuning. In MLM, a certain fraction of the input tokens (about 15%) is randomly masked. The objective of the model is then to predict these masked tokens based on the context provided by the remaining unmasked tokens. Unlike Word2Vec, that assigns a fixed vector to each word, models trained with MLM produce contextualized word embeddings. This means the representation of a word varies based on its surrounding context. Once trained on the MLM task, the model can be fine-tuned for specific downstream NLP tasks, such as sentiment analysis. These tasks require much smaller datasets than the pre-training stage. This approach has proven to be highly effective, as the knowledge gained during MLM pre-training transfers well to more specific tasks.

1.1.2 Asset Pricing

Asset pricing is a central discipline in finance aiming to explore the factors and mechanisms that determine prices and returns of financial assets. At its core lies a fairly simple concept: the higher the risk associated with a security, the higher the expected return should be to compensate investors for bearing the risk. Several models have been developed to describe this relationship between risk and return. The most prominent is the Capital Asset Pricing Model (CAPM), a seminal framework which postulates that the only priced risk factor is the return on the market portfolio. Asset returns that have a high covariance with the market portfolio, characterized by a large beta coefficient, are riskier. Consequently, investors demand a higher risk premium for these assets. The CAPM became a central concept in asset pricing, earning William Sharpe the 1990 Nobel prize. The model is formalized in Equation (1.2). The expected return $E(R_i)$ of asset i is calculated by the sum of the risk-free rate R_f and the excess market return $E(R_m) - R_f$ multiplied by the beta coefficient β_i.

$$E(R_i) = R_f + \beta_i(E(R_m) - R_f) \tag{1.2}$$

Building upon the CAPM, Fama and French (1993, 2015) and Carhart (1997) introduce additional risk factors that improve the explainability of variations in stock returns compared to the CAPM. These additional risk factors are size, represented by SMB (small-minus-big), and value, denoted by HML (high-minus-low book-to-market) — these are risk factors of the three-factor model. Further, there's profitability, captured by RMW (robust-minus-weak), and investment, expressed as INV (conservative-minus-aggressive) — both are part of the five-factor model. Additionally, there is momentum, indicated by UMD (up-minus-down). Equation (1.3) generalizes the concept of a multi-factor asset pricing model: The excess return of stock i, denoted as $R_i - R_f$, is modeled by its factor loadings, $\boldsymbol{\beta_i}$, which is a $1 \times k$ vector, and the risk-factors, $\boldsymbol{F}$, which is a $k \times n$ matrix. Here, k represents the number of factors, and n is the number of observations. The term ϵ_i represents the disturbance or error term. For a model that perfectly explains variations in asset returns, the intercept term α should be zero. If, however, α is statistically different from zero, then the model does not fully explain the variation is asset returns.

$$R_i - R_f = \alpha + \boldsymbol{\beta_i}\boldsymbol{F} + \epsilon_i \tag{1.3}$$

While statistical factor models, as described above, explain the variation of asset returns relative to risk factors, factor pricing models go one step further and measure the risk premia attributable to these factors. In this thesis I apply the Fama-MacBeth (Fama and MacBeth, 1973) procedure to estimate the risk premia of risk factors. This can either be done by estimating the factor betas according to Equation (1.3) in a first step or by considering firm specific characteristics directly, such as the logarithm of the market capitalization or the book-to-market ratio. At each time step t, a cross-sectional regression is

estimated, as formulated in Equation (1.4).

$$R_{i,t} - R_{f,t} = \delta_{0,t} + \boldsymbol{\delta_t \beta_{i,t-1}} + \epsilon_{i,t} \tag{1.4}$$

The final step in the Fama-MacBeth procedure is to average these estimates over all time periods to get a single estimate for each factor risk premium.

1.2 Research Papers

1.2.1 Financial News Sentiment Learned by BERT: A Strict Out-of-Sample Study

This study contributes to the literature of sentiment analysis in the financial domain and also examines the efficiency of the stock market by measuring the speed of diffusion of new information into asset prices. Human language is very complex and highly dimensional (Gentzkow et al., 2019). This has posed significant challenges to researchers in computer science and natural language processing for decades. This changed with the introduction of the transformer architecture (Vaswani et al., 2017), that set this field on a trajectory which enabled the development of models that achieved massive advancements in the understanding and generation of human text. One of these models is the BERT model, a transformer based model that was pre-trained on large amounts of textual data. A 1:1 application of this model in the financial context, however, comes with two additional issues: First, the meaning of words is quite different in the financial domain compared to the general case. Second, the BERT model was trained in 2018, making an out-of-sample study prior to 2018 prone to look-ahead bias. To overcome these challenges, I pre-train and fine-tune a domain-specific BERT-like model strictly out-of-sample on the sentiment classification task. I therefore use historical financial news as a pre-training

dataset. Furthermore, I use the joint behavior of news articles and idiosyncratic stock returns to derive a dataset with sentiment annotations that is then used for the supervised training on the sentiment classification task. With the sentiment predictions at hand, I then investigate the short-term impact of financial news on individual stock returns over the period from 1996 to 2020. With daily return data of S&P 500 constituents, the analysis shows that financial news carry information that is not immediately reflected in equity prices. News is largely priced-in within one day, with diffusion varying across industries. A trading strategy that leverages the sentiment signal generates an average return per trade of 24.06 bps over an 18 year out-of- sample period.

1.2.2 Overnight Reversal and the Asymmetric Reaction to News

Financial markets react quickly to the release of new information in the form of financial news. This study examines the market's reaction to news that is released overnight - specifically, after the stock market closes on day $t-1$ and before it opens on day t. Furthermore, it explores the interplay between previous trading day returns (z-scores) and overnight news. This interplay reveals an interesting predictable pattern in the returns of the subsequent trading day – measured from market open to close on day t: We observe a statistically significant reversal relative to the previous-day return (as measured from market open to close on day $t-1$) if the previous-day return is large and news with strong sentiment, either positive or negative, is published overnight. This surprising finding is caused by over- and underreactions to the overnight news at market opening. On the one hand, we observe overreactions happening if the news sentiment "confirms" previous-day returns, i.e., positive news released after positive returns (or negative news released after negative returns). In these cases the market opens too high (or too low), which leads, on average, to a reversal during the subsequent trading day. On the other hand, we observe underreactions happening if the news sentiment does not "confirm" previous-day returns, i.e., positive news released after negative returns (or negative news released after positive returns). In these cases the market opens too low (or too high). Thus, the information is not immediately reflected in asset prices at market open, but diffuses into asset prices only during the subsequent trading day. We further differentiate between news on the

topics "analyst forecast" and "earnings report". Our findings suggest that investors tend to underreact to information provided by analysts as we observe no reversal on the subsequent trading day following such news. Instead, we find, on average, positive returns after positive analyst forecasts and negative returns after negative forecasts.

1.2.3 Firm-specific Climate Risk Estimated from Public News

With this study we contribute to the climate finance literature by documenting a positive and significant risk premium for physical climate risk. In addition, we find a regime shift for regulatory climate risk that occurred around 2012 and reconcile conflicting evidence in the literature. While the risk premium is positive in the earlier period, it becomes significantly negative after 2012. These findings were only made possible by using textual data, in the form of news articles, to derive firm-specific estimates of climate risks over a sufficiently long horizon of over 20 years. Related studies which use ESG ("Environmental, Social, and Governance") data to determine firm-specific climate risk exposures are limited by a relatively short period of ESG data availability, mostly available from 2010 onwards (Engle et al., 2020; Pástor et al., 2022). Thus, they lack observations prior to 2010 and consequently report results that are in contradiction with the literature that uses data other than ESG over earlier time periods. While Hsu et al. (2022) reports a positive premium for regulatory climate risk, estimated over the period 1996 to 2016 by using firm emissions data, Pástor et al. (2022) reports a negative risk premium over the period 2013 to 2020. We, in contrast, study the period 1996 to 2020 and are able to detect the shift from a positive to a negative regulatory climate risk premium within a consistent framework. This study is most closely related to that of Sautner et al. (2023a) and Berkman et al. (2021), who also use textual data, in the form of earnings-call transcripts and 10-K reports, to derive firm-specific climate risk estimates. We differ from Sautner et al. (2023a) in several ways: First, we use financial news in contrast to earnings-call transcripts. Second, we also differ in terms of the methodology as we propose a novel topic modeling algorithm termed "Guided Topic Modeling (GTM) with Word2Vec" to generate comprehensive topic word clusters while Sautner et al. (2023a) builds on the work of King et al. (2017). Third, we also differ in our findings as Sautner et al. (2023b) mostly finds insignificant climate risk

premia.

We then extend the set of firm-specific climate risk estimates from the one obtained by the direct exposures in the news to a broader set of climate risk estimates for about 9000 firms via the estimation of climate risk betas. We therefore construct zero investment, long/short portfolios: a low-minus-high regulatory climate risk portfolio (GMB, "green-minus-brown") and a high-minus-low physical climate risk portfolio. We then augment the market model by these climate risk portfolios and regress individual stock returns on these models to obtain climate risk beta estimates for each firm. Again, we form a green-minus-brown portfolio, this time, however, sorted by the climate risk beta. This portfolio constitutes a priced risk factor and shows a surprisingly strong correlation with an ESG-sorted benchmark portfolio.

1.2.4 Guided Topic Modeling with Word2Vec: A Technical Note

Although complex transformer-based models have achieved human-like performance in many NLP tasks, much simpler, dictionary-based approaches are still frequently used in financial academia due to their simplicity, interpretability and the advantage that no training dataset is required. However, a major drawback of these dictionary-based approaches is the creation of the dictionary itself. Hayes and Weinstein (1990) point out that recalling representative words from memory is a "near-impossible" task for humans. Furthermore, King et al. (2017) show that the selection of an incomplete keywords list can result in a severe selection bias. To overcome these challenges, we propose Guided Topic Modeling (GTM) with Word2Vec, an algorithm that enables the fast and flexible generation of comprehensive topic clusters (dictionaries) from (a pair of) seed words. It thereby leverages the Word2Vec algorithm and its ability to capture the semantic similarity between words, which translates into word vectors of related words appearing close to each other in the vector space. The iterative algorithm retrieves the most related words from the vector space to generate comprehensive topic clusters. This has the advantage that no further training dataset or expert knowledge is required as all the necessary information is already encoded in the word embeddings. Still, the algorithm is flexible and adjustable such that one can control the characteristics of the desired topic mappings. Internally, the algorithm

does not simply collect the words closest to the seed words but also adjusts its topic center accordingly, such that it converges towards an optimal center. In this way, we can extract additional information from the list of topic words – a similarity parameter (weight) that is higher for words closer to the topic center, i.e., important topic words, and lower for words that are more distant form the topic center, i.e., less important words.

Bibliography

Berkman, H., Jona, J., and Soderstrom, N. S. (2021). Firm-specific climate risk and market valuation. Available at SSRN 2775552.

Bollen, J. and Mao, H. (2011). Twitter mood as a stock market predictor. 44(10):91–94.

Carhart, M. M. (1997). On persistence in mutual fund performance. The Journal of finance, 52(1):57–82.

Dangl, T., Halling, M., and Salbrechter, S. (2023). Firm-specific climate risk estimated from public news. Available at SSRN 4575999.

Dangl, T. and Salbrechter, S. (2022). Overnight reversal and the asymmetric reaction to news. Available at SSRN 4307675.

Dangl, T. and Salbrechter, S. (2023). Guided topic modeling with word2vec: A technical note. Available at SSRN 4575985.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Engle, R. F., Giglio, S., Kelly, B., Lee, H., and Stroebel, J. (2020). Hedging climate change news. The Review of Financial Studies, 33(3):1184–1216.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. the Journal of Finance, 47(2):427–465.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of financial economics, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. Journal of financial economics, 116(1):1–22.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. Journal of political economy, 81(3):607–636.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3):535–74.

Hayes, P. J. and Weinstein, S. P. (1990). Construe/tis: A system for content-based indexing of a database of news stories. In IAAI, volume 90, pages 49–64.

Hsu, P.-H., Li, K., and TSOU, C.-Y. (2022). The pollution premium. The Journal of Finance.

King, G., Lam, P., and Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. American Journal of Political Science, 61(4):971–988.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. 66(1):35–65.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Pástor, L., Stambaugh, R. F., and Taylor, L. A. (2022). Dissecting green returns. Journal of Financial Economics, 146(2):403–424.

Salbrechter, S. (2021). Financial news sentiment learned by bert: A strict out-of-sample study. Available at SSRN 3971880.

Sautner, Z., van Lent, L., Vilkov, G., and Zhang, R. (2023a). Firm-level climate change exposure. Journal of Finance. forthcoming.

Sautner, Z., Van Lent, L., Vilkov, G., and Zhang, R. (2023b). Pricing climate change exposure. Management Science.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. 62(3):1139–1168.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

2 Financial News Sentiment Learned by BERT: A Strict Out-of-Sample Study

Financial News Sentiment Learned by BERT: A Strict Out-of-Sample Study

Stefan Salbrechter

September 13, 2023

I investigate the impact of financial news on equity returns and introduce a non-parametric model to generate a sentiment signal, which is then used as a predictor for short-term, single-stock equity return forecasts. I build on Google's BERT model and sequentially pre-train and fine-tune it using Thomson Reuters financial news data covering the period from 1996 to 2020.[1] With daily return data of S&P 500 constituents, the analysis shows that financial news carry information that is not immediately reflected in equity prices. News is largely priced-in within one day, with diffusion varying across industries. A trading strategy that leverages the sentiment signal generates an average return per trade of 24.06 bps over an 18 year out-of-sample period.

[1]I thank Thomson Reuters / Refinitiv for providing this comprehensive set of news data.

2.1 Introduction

The steady increase in the amount of available unstructured data in form of written text, accelerating growth in computing power together with advances in the design of algorithms have recently enabled important breakthroughs in the field of computational linguistics. Models, trained on large text repositories, have become powerful tools in extracting informative signals from text items. One of these breakthrough concepts is the natural language model BERT (Bidirectional Encoder Representations for Transformers), which has achieved excellent results in many NLP tasks (Devlin et al., 2018). This study borrows from these achievements by employing a BERT-based deep learning model for estimating sentiment in financial news. This sentiment signal is then used for short-term, single-stock equity return forecasts. I analyze the predictive quality of these forecasts and study return characteristics of portfolios that over- or underweight stocks according to their estimated sentiment.

The aim of this study is to analyze the following questions: (1) Is the proposed model able to extract a measure of sentiment from financial news that shows a positive out-of-sample correlation with future stock returns? In other words, does financial news contain predictive information about future stock returns? (2) How long does it take until new information is incorporated into stock prices? (3) Can prediction accuracy be improved by augmenting the text representation from BERT with a topic feature vector extracted from the article? (4) Does fresh news contain more predictive power than stale news? (5) Is it possible to implement a profitable trading strategy that derives trading signals solely from financial news sentiment?

While the pre-trained BERT model is available for direct and quick application, I argue that this model should not be used in an out-of-sample asset pricing study starting before 2018. This is because the model is trained on a huge corpus of text data that includes texts published up to 2018. Hence, there is potential look-ahead bias when extracting sentiment from historical text by mapping it on embeddings which incorporate information not yet available when the historical text has been released. Furthermore, Liu et al. (2020) point out that the broad, cross-domain data used to pre-train BERT negatively

affects performance on domain-specific tasks such as financial news. To overcome these issues, I pre-train a BERT-like model on financial news data only. Starting in 2002, the model is then re-trained sequentially every two years until 2017. The proposed version of BERT, which I refer to as FinNewsBERT, is a smaller version of BERT with only 18.95 million parameters, as opposed to the 110 million parameters of BERT base. I demonstrate that this smaller model saves computing resources and costs while accuracy remains competitive.[2]

In order to assess the model's ability to classify news into positive and negative sentiment and to evaluate whether these are associated with abnormal asset returns, I conduct an event study in Section 2.6.1. Furthermore, to evaluate whether the extracted sentiment measure is suitable for making investment decisions I perform an out-of-sample backtest ranging from 01-2002 to 01-2020. The trading strategies are simple: (i) Form a portfolio of stocks with positive sentiment signal (equally weighted, long-only strategy). (ii) Form a long-short strategy with positive weights in stocks with positive sentiment signal and negative weight in stocks with negative sentiment signal. I initiate trades at market open and generate sentiment signals using all news that is published between market close on day $t-1$ at 4:00pm and market open on day t at 9:30am (news window: 17.5h). I find that restricting the analysis to news categorized as "analyst forecast" considerably increases the return per trade generated by the strategies. Furthermore, stocks that have already realized a large negative return between the opening at day $t-1$ and date t and get a negative sentiment signal at the opening of day t tend to realize a negative excess return at day $t+1$, i.e., I identify the existence of a negative momentum, conditional on the negative sentiment signal. Focusing on texts with "analyst forecast" as assigned topic and regarding negative sentiment only if the z-value of the day-t return is below -1.96, the long-short strategy results in a monthly alpha of 6.46% and an R^2 of 5.56% according to the Fama French 5 factor model plus momentum. The corresponding Sharpe Ratio is 2.26 and the return per trade is 24.06 bps without considering transaction costs. In addition, I also investigate whether there is exploitable alpha left when trades are placed at market closing. I find that financial news is incorporated into asset prices very quickly. Using

[2]See, e.g., the study Turc et al. (2019) that confirms my findings.

Settings B from Table 2.4 but entering trades at market closing lowers the return per trade from 24.06 bps to 11.92 bps (see Table 2.4). In addition, I benchmark the model against FinBERT (Araci, 2019), a full-sized BERT model that is fine-tuned on Thomson Reuters news. Although the proposed model is less than one-fifth the size of FinBERT, it shows superior out-of-sample performance (see Table 2.9).

The remainder of the paper is composed as follows: Section 2.2 provides a review of the literature, Section 2.3 describes the data and its pre-processing. Section 2.4 describes the model architecture, and the training procedure. The risk adjusted portfolio benchmark is described in Section 2.5 and Section 2.6 contains the empirical analysis.

2.2 Literature Review

Over the past two decades, several studies have been published examining the influence of financial news on the stock market. Pioneering works thereby often rely on pre-specified word dictionaries (Tetlock, 2007; Loughran and McDonald, 2011; Bollen and Mao, 2011). Tetlock (2007) finds empirical evidence, that negative tone in a popular column of the Wall Street Journal predicts lower stock returns in the following trading days. These low returns are then followed by a reversal to fundamentals. In addition, Antweiler and Frank (2004) uses Naive Bayes and finds that messages posted on Yahoo Finance and Raging Bull help to predict subsequent trading volume. Other approaches include linear and non-linear text regression methods like support vector machines or Bayesian regression methods (Antweiler and Frank, 2004; Jegadeesh and Wu, 2013; Manela and Moreira, 2017). Most studies use the frequency-based bag-of-words approach to generate features from text (Rechenthin et al., 2013; Lee et al., 2014). However, this has the disadvantage of producing large, sparse vectors that are inefficient to compute. Also, it does not account for the semantic similarity between words. This changed with the introduction of Word2Vec by Mikolov et al. (2013), a method for generating dense word vectors from text that is able to capture the semantic relationships between words. Combining Word2Vec with deep learning models led to big improvements in the NLP space, including sentiment analysis. Severyn and Moschitti (2015) use word embeddings in combination with a deep

convolutional neural network (DCNN) for a sentiment analysis of tweets. With this approach a new state-of-the-art at the phrase level sub-task of the 2015 SemEval[3] challenge was achieved. In addition, Peng and Jiang (2015) applies Word2Vec to generate features from financial news, which serve as an input, along with historical price data features, for a deep neural network that is trained to predict future up- and downward movements in asset prices. This model achieves an accuracy of 52.44% in predicting the next day price direction. However, the test period is relatively short and covers only six month starting in June 2013. Furthermore, Cong et al. (2018) proposes a framework for analysing large amounts of textual data by generating a small number of textual factors. To do this, the authors transform words into dense vectors with Word2Vec and further reduce the dimensionality by using clustering techniques and topic modelling. Texts are then analyzed by calculating beta loadings on various textual factors. The authors identify several use cases for textual factors. One application is the prediction of macroeconomic variables using texts from the Wall Street Journal. As a result, the authors were able to reduce the out-of-sample RMSE by an average of 17% compared to models based on one-hot representations.

Another novel machine learning framework for the prediction of asset returns from financial news is proposed by Kelly et al. (2019). The authors use news data from the Dow Jones Newswires over a 38 year period ranging from 1989 to July 2017 along with stock data from the CRSP universe. The key feature of their SESTM approach, which stands for Sentiment Extraction via Screening and Topic Modeling, is the use of the common behaviour of financial news and stock returns to learn the sentiment of news articles. This idea is adopted in this paper for the annotation of news articles as positive, neutral or negative (see Section 2.3.1). The authors also examine the price reaction in relation to the novelty of news articles. They find that the impact of financial news sentiment on asset returns is 70% larger for fresh news than it is for stale news. While fresh news take four days to be fully reflected into stock prices, it takes just two days for stale news. They also find that the price responses are approximately four times larger for small and

[3]SemEval (the International Workshop on Semantic Evaluation) is an ongoing series that focuses on the evaluation of computer-based semantic systems. It is organized under the umbrella of SIGLEX, the Special Interest Group on the Lexicon of the Association for Computational Linguistics.

volatile stocks than for stocks with high market capitalisation. In the case of small stocks, it takes three days until the information is fully priced in. These results are consistent with the observations of Baker and Wurgler (2006), who observes that low-capitalisation, young, unprofitable, low-dividend-paying, high-volatility and high-growth companies are difficult to arbitrage or value according to traditional financial theory and are therefore very sensitive to investor sentiment. For large stocks in contrast, it takes only one day for the new information to be fully reflected in the prices. The authors argue, that stocks with high market capitalization tie up more investor capital and therefore receive more attention. In contrast, companies with lower market capitalization receive less attention, as fewer investors participate in these stocks. Kahneman (1973) further notes that people are limited in their ability to divide their attention between multiple tasks. As a result, individual investors devote only a limited portion of their attention to investing and an even smaller portion to reading the news of smaller companies. This affects their ability to react quickly to newly available information and to value stocks appropriately (Barber and Odean, 2013). In this paper, I solely consider stocks listed in the S&P 500. Since this index consists of the 500 largest companies in the US, I expect that the information contained in financial news is incorporated into stock prices within one day. The analysis in Section 2.6.1 confirms this assumption. Kelly et al. (2019) also implement a simple trading strategy that buys stocks if the news sentiment on the previous day is positive and sells stocks if the news sentiment is negative. They also form equally weighted and value weighted portfolios that are daily rebalanced. The realized Sharpe Ratios of the long/short portfolios are 4.29 for the equal weighted and 1.33 for the value weighted portfolio, without considering transaction costs. For comparison of these numbers with the results presented in this paper, it has to be considered that the CRSP universe consists of several thousand stocks, where the majority of them has a low market capitalisation. This means that the equally weighted portfolio largely benefits from the strong effect observed with small stocks. The value-weighted portfolio may therefore be better suited for comparison with the findings of this study, which are derived from stocks included in the S&P 500.

Moreover, in recent years, further important breakthroughs have been made in the field

of natural language processing (NLP). Based on the research by Vaswani et al. (2017), Devlin et al. (2018) developed BERT. What differentiates BERT from previous deep learning models is the use of transfer learning, the unsupervised pre-training on large datasets, followed by task-specific fine-tuning through supervised machine learning. Furthermore, unlike Word2Vec (Mikolov et al., 2013), BERT uses contextual word embeddings that allow identical words to take on different meanings and thus different vector representations depending on the context. As a result, BERT achieves excellent results in many NLP tasks and became the new stat-of-the art in the industry. Also, Araci (2019) proposes FinBERT, a BERT model that is further pre-trained on Thomson Reuters financial news and fine-tuned for the financial text classification task. The author however, does not further investigate the return predictability of financial news with this model. As part of this paper I complement his work by using FinBERT as a benchmark model (see Section 2.6.3).

2.3 Data and Data Preprocessing

For this work, I collect data from two sources. The first is a dataset of financial news published by Refinitiv (formerly Thomson Reuters) from January 1996 to February 2020.[4] From the second data source, Refinitiv Datastream, the daily price data of 1330 companies that were listed in the S&P 500 during this period are downloaded. For all subsequent tasks, I use adjusted open/close prices. Each article in the financial news dataset is assigned with several tags that include the timestamp, the language of the article, a list of topics, and in the case of business news, company ticker codes. These ticker codes are used, to extract all news articles that are related to those 1330 S&P 500 companies. For fine-tuning and inference, only news articles that contain either the company name or the ticker code in the headline are considered. The intention is that the content of these articles is more relevant to the associated company than other, more general news. These news articles are then converted to lower case and cleaned by removing all numbers, punctuation marks and brackets so that only letters remain. In addition, non-relevant

[4]This dataset contains more than 40 million news items with exact timestamp of publication and complete tracking of update histories.

data, which includes the author's contact information such as email addresses, phone numbers and hyperlinks, is also removed.

Thomson Reuters financial news stories are not published in one particular moment, rather their release evolves in several steps. First, a news alert is published which is followed by a news-break 5 to 20 minutes later. This is comprised of a headline and a short text. Another 20 to 30 minutes later, a news update is published with additional information. Further updates may be released successively as the story develops. In some cases, updates are released even days after the original news event. Consequently, using only the last updated status of a news article does not meet the need for fresh news. The objective therefore is to use those versions of news articles that appear as early as possible and contain as much information as possible. Considering trading hours from 9:30am to 4pm (ET) for the New York Stock Exchange (NYSE) and the Nasdaq Stock Market I am able to place trades either at market opening or closing. So, if a news article is published within stock market opening hours and is then followed by several updates until the evening, I only consider the last update that is published before the stock market closes. This allows opening a position in the corresponding asset at the market close of the same day. The same strategy is applied to news articles published before the market opens. In addition, all news articles that are published between 12am and 9:30am are denoted as pre-market news, all articles published between 9:30am and 4pm are denoted as market news and all articles published between 4pm and 12am are denoted as post-market news. Multiple news articles about one company that are published in either the pre-market, market or post-market hours are then combined into one document. Figure 2.1 (a) shows the distribution of news articles over the full period. It can be observed, that the majority of news articles is published within market times. Furthermore, a peak in 2008 to 2009 can be observed which is probably due to the financial crisis. Furthermore, I distinguish between fresh and stale news. Stale news are news articles that don't contain new information whereas fresh news contain new, previously unknown information. The algorithm used to detect fresh and stale news is explained in more detail in Appendix A.3. Figure 2.1 (b) shows the annual number of fresh and stale news used for fine-tuning and inference. The total number of combined news documents is 372,438 consisting of 299,043

fresh and 73,395 stale news documents.

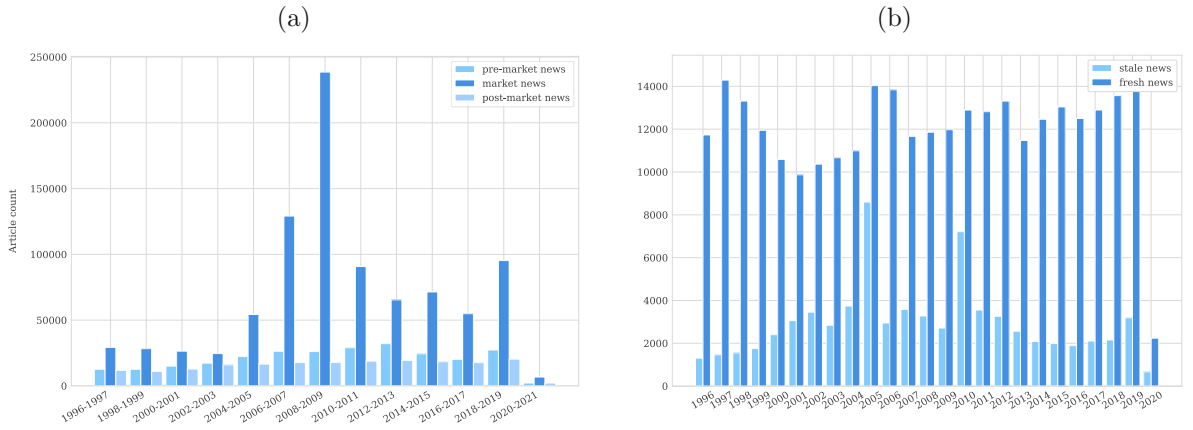


Figure 2.1: (a) Count of news articles within pre-market, market and post-market hours. (b) Annual number of fresh and stale news documents that is used for fine-tuning, validation and inference in the out-of-sample backtest.

2.3.1 Deriving Sentiment Annotations from Asset Returns

The news dataset from Thomson Reuters does not contain any labels with sentiment information. However, labelled data is necessary for supervised machine learning. Available datasets for learning the sentiment of financial texts include the Financial PhraseBank dataset from Malo et al. (2013) and the FiQA Sentiment dataset for financial opinion mining and question answering (WWW, 2018). Unfortunately, those datasets are rather small. Financial PhraseBank consists of 4845 hand labelled financial news articles and FiQA contains only 1174 financial news headlines and tweets. In this study, I adopt the approach presented in Kelly et al. (2019) and use the feedback of the stock market to learn the sentiment of financial news. Thus, I derive sentiment annotations from asset returns to obtain a large annotated dataset. This approach is based on the simple assumption that positive news is accompanied by positive returns and negative news is accompanied by negative returns. Since stock prices are influenced by many factors besides financial news, it is necessary to isolate the price effect of financial news. Therefore, I calculate idiosyncratic returns $IR_{i,t}$ using Formula 2.1. $R_{i,t}$ denotes the stock return of asset i at time t, $R_{f,t}$ is the risk-free rate at time t, $\beta_{i,t}$ is the market beta of asset i at time t and $R_{S\&P\,500,t}$ is the market return at time t.

$$IR_{i,t} = R_{i,t} - R_{f,t} - \beta_{i,t} * (R_{S\&P500,t} - R_{f,t}) \tag{2.1}$$

Leinweber and Sisk (2011) show that significant pre-news effects can occur due to well-informed market participants. Additionally, Kelly et al. (2019) find that it takes up to one day to fully incorporate financial news into asset prices in the case of large companies. Therefore, I calculate the mean idiosyncratic return $\bar{IR}_{i,t}$ of the day t idiosyncratic return (the day when a news article is published), the day $t-1$ idiosyncratic return and the day $t+1$ idiosyncratic return in order to capture the impact of news on stock returns properly. In addition, I also consider the volatility among the assets. Suppose I would label all news articles as positive if the idiosyncratic mean return is greater than 1%. As a result, I would get disproportionately more positive labels for high volatility stocks as opposed to low volatility stocks, regardless of news sentiment. To treat all firms equally, I calculate z-scores from the idiosyncratic means for each asset using Formula 2.2. The mean value μ and standard deviation σ are calculated over a rolling window of 505 days (approximately 2 years of trading days).

$$z_{i,t} = \frac{\bar{IR}_{i,t} - \mu_{i,t}}{\sigma_{i,t}} \tag{2.2}$$

Finally, I define barriers to obtain positive, neutral, and negative labels. If the z-value is greater than 1.4, I consider the corresponding news article as positive. If the z-value is below -1.4, I consider the news article as negative. If the z-value lies between -1 and 1, then the news article is labelled as neutral. This is illustrated in Figure 2.2. I introduce the gap of 0.4 standard deviations between those barriers for two reasons. First, it mitigates the problem of mislabeled data, as items associated with high (low) z-scores tend to be positive (negative) and items with z-scores closer to zero tend to be neutral. Second, it leads to a more balanced data set between the number of positive, neutral, and negative

observations.

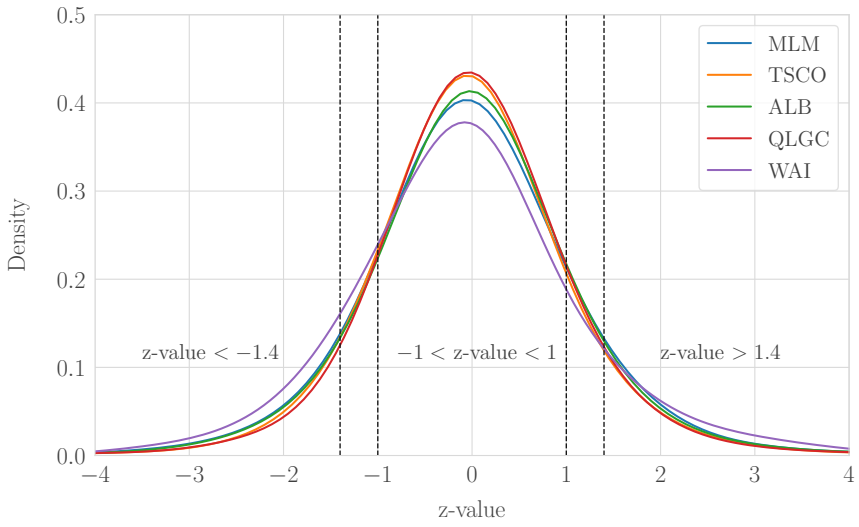


Figure 2.2: Density distribution of z-values and sentiment barriers for five constituents.

2.4 Text Classification Model

A major challenge in training the model is the high level of noise inherent in asset returns, which is reflected in a certain number of flawed training labels. To illustrate the problem, let's consider a positive-sounding news article about a company that reports high sales figures and solid profits. This, however, does not take the market's expectations into account. Thus, if analysts expect even higher earnings, the return on the following trading day may still be negative, leading to a negative annotation for this news article despite its positive-sounding content. In addition, macroeconomic influences affecting the overall market or individual sectors can also lead to mislabelled data. In order to still obtain a robust model, I use transfer learning (BERT) in combination with a robust loss function. Wang et al. (2019) shows that for training with noisy labels, performance can be drastically improved by using symmetric cross entropy loss as opposed to the commonly used cross entropy loss. The model, which is shown in Figure 2.3 consists of three parts. The first one is FinNewsBERT, that generates document embeddings (CLS token) from news articles. The second part, denoted as Text2Topic, categorizes news articles into pre-specified topics.

Both features are then combined into one feature vector. The third part of the model is a deep neural network (DNN) that receives the feature vector as input and classifies news articles into the three sentiment categories positive, neutral and negative.

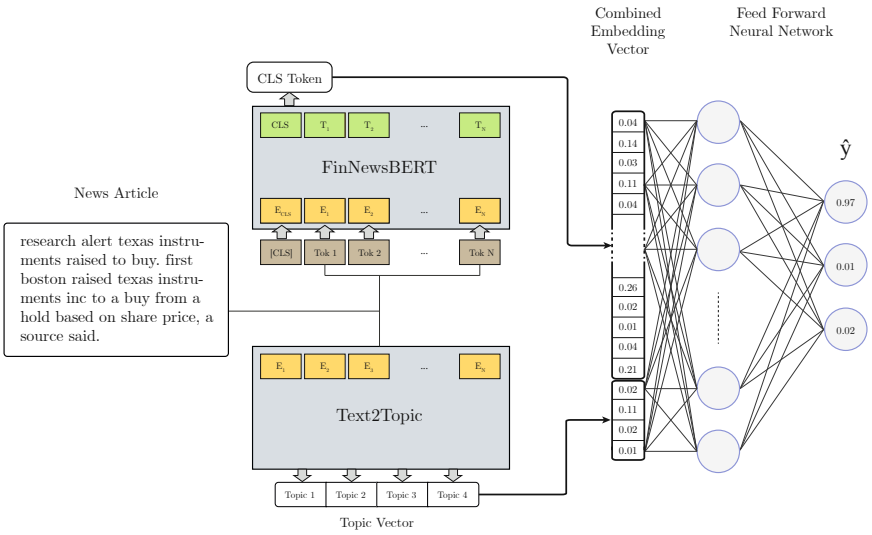


Figure 2.3: Architecture of the model consisting of the three main parts: a) FinNewsBERT, b) Text2Topic model and c) the feedforward neural network classifier.

2.4.1 Financial News BERT (FinNewsBERT)

For the implementation of BERT, I make use of the Hugging Face transformers library for Python (Wolf et al., 2020). This library contains a variety of pre-trained transformer models as well as the methods to configure models individually and pre-train them from scratch. For this study, I choose the configuration of the RoBERTa model, which I pre-train on the Thomson Reuters dataset. This model is based on BERT, with identical structure but further optimizations (Liu et al., 2019). The authors of RoBERTa adjust some hyperparameters, use a different tokenizer and remove the "next sentence prediction" task. Pre-training of the model is done with the "masked language modeling" (MLM) task which was introduced by Devlin et al. (2018). The masked language model randomly selects 15% of the input tokens. Out of this selection, 80% of the tokens are replaced by the MASK token, 10% are replaced by a random token, and 10% remain unchanged. The goal of the model is to predict the masked tokens solely based on their context.

The proposed version of BERT, which I call FinNewsBERT (Financial News BERT)

is much smaller than BERT base. BERT base consists of 12 hidden layers, 12 attention heads, an embedding size of 768 and a maximum sequence length of 512 tokens. This results in a total of 110 million parameters. This model, in contrast, has only 18.95 million parameters. It is composed of six hidden layers, four attention heads an embedding size of 256 and a maximum input sequence length of 256 tokens. BERT has a vocabulary size of 30,522 while FinNewsBERT has a vocabulary size of 30,257 words. Before the text data can be fed into the model, it must be converted into a machine-readable form. This task is called tokenization, which is the splitting of words into subwords that are then converted into token-ids via a lookup table. Subword tokenization is a modern tokenization approach that keeps frequently used words in it's original form, but splits less frequent words into meaningful subwords. The advantage of subword tokenization is that models work well with a moderate vocabulary size. In addition, this method allows the models to process unknown words by breaking them down into known subwords (Wolf et al., 2020). While the original BERT model uses a character level WordPiece tokenizer, I use the byte-level Byte-Pair Encoding (BPE) tokenizer, that was originally implemented in GPT-2 and also in RoBERTa.

2.4.2 Additional Topic Features

The financial news contained in the Thomson Reuters news dataset can be further categorized into different topics. Those include analyst forecasts, earnings announcements, news about mergers, product releases and others. Therefore, in addition to the features generated by BERT (CLS token), I also generate topic features. These feature vectors are then concatenated and used as a combined input for the feedforward classifier (see Figure 2.3). The initial hypothesis of further improving the classification accuracy with additional information about the news topic is empirically confirmed. The Sharpe ratio of the long/short portfolio improves by 47% (1.29 vs. 0.88) and the return per trade improves by 17% (11.74 bps vs. 10.05 bps) with the additional use of topic features (see Table 2.10 for details). Furthermore, the additional topic information allows further investigation into which topics contain the most predictive information. In the empirical analysis I find

that the return per trade of the long/short portfolio is 46% larger (17.15 bps vs. 11.74 bps) when trading signals are solely generated from news of the topic "analyst forecast" compared to signals generated from all news.

Text2Topic

Text2Topic is a universal algorithm that identifies whether the topic of a news article is similar to one of the predefined topics. It is based on Word2Vec (Mikolov et al., 2013), an algorithm that maps all words w contained in the vocabulary V to vectors $\mathbf{v}$ ($\forall \mathbf{w} \in \mathbf{V} : \mathbf{w} \mapsto \mathbf{v} \in \mathbb{R}^{\mathbf{n}}$). The vocabulary consists of 19,935 unique words in total and the vector size n (embedding size) is set to $n = 300$. With Word2Vec, words with similar meanings receive similar vector representations which is why the cosine similarity between vectors is a good measure for the semantic similarity between words. Table 2.1 shows the ten most similar words to the target word "raise" measured by cosine similarity.[5] With Text2Topic I am interested in finding news that belong to the topics "analyst forecast" and "earnings report".[6] Therefore, I pre-specify a short list of k topic words, each represented as a vector $\mathbf{t_i} \in \mathbb{R}^n$ with $i \in [1, k]$. Table 15 in Appendix A.5 shows the word support for each topic. In the next step, the cosine similarity $c_{i,j}$ is calculated between each topic word vector $\mathbf{t_i}$ and each word vector $\mathbf{d_j} \in \mathbb{R}^n$ contained in a news document with a length of l words and $j \in [1, l]$ (see Equation (2.3)). Doing this for each word pair combination results in the cosine similarity matrix $\mathbf{C} \in \mathbb{R}^{k,l}$.

$$\mathbf{c_{i,j}} = \frac{\mathbf{t_i} \cdot \mathbf{d_j}}{\|\mathbf{t_i}\|\|\mathbf{d_j}\|} \tag{2.3}$$

[5] I make use of the Python library Gensim (Řehůřek and Sojka, 2010) in order to train the Word2Vec model.

[6] I also define two other topics, which are "FED/monetary policy" and "corporate/strategy". These, however, receive little support in the data and are therefore not discussed further.

Next, I only consider the values of the matrix **C** that are greater than or equal to the threshold value of 0.45, since I am only interested in word pairs with a high cosine similarity. All values greater than 0.45 are then summed up and normalised by the square root of the news article's word count. The square root is used to account for the fact that longer news articles contain a greater proportion of non-relevant words and to avoid attenuating the score too much. This process is repeated for all topics and news documents. As a result I receive a data frame that contains four topic scores for each news article. The final topic features are calculated by normalising these topic scores for each topic in the interval [0,1] by dividing them by the maximum topic values.

raise	**Cosine Similarity**
raising	0.662
slash	0.632
cut	0.612
reduce	0.567
trim	0.563
revise	0.540
halve	0.531
add	0.517
boost	0.515
save	0.498

Table 2.1: This table shows the cosine similarities between the target word "raise" and it's ten most similar words.

2.4.3 Hyperparameters

For pre-training of the model I make use of the Hugging Face trainer function (Wolf et al., 2020). I train it with a batch-size of 128 and a maximum sequence length of 128 tokens for 100% of the steps. Training with longer sequences is expensive, since attention is quadratic to the sequence length (Devlin et al., 2018). The optimizer I use is Adam (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, epsilon = 1e-06 and a maximum learning rate of 1e-04 with 10,000 warm-up steps and a linear learning rate schedule. Furthermore, I use a dropout rate of 0.1 and the GELU activation function (Hendrycks and Gimpel, 2016) in all layers. In addition, I apply weight decay with a parameter value of 0.01.

2.4.4 Training

To obtain a strict out-of-sample backtest, I pre-train and fine-tune the model sequentially by re-training it every two years. While BERT is pre-trained on a total of 16GB of uncompressed text (Wikipedia + book corpus), this model is pre-trained on a maximum of 4GB, but domain-specific financial news articles (4GB of data is used for the model trained from 1996 to 2017). The first model is pre-trained over 400,000 steps with news data from January 1996 to December 2001. This model is then fine-tuned on the labeled data set for the sentiment classification task over the same time horizon and used for inference in the following two years. In the next step, data from January 1996 to December 2003 is used to pre-train and fine-tune the second model which is then used for inference from 2004 to 2005. This procedure is repeated until 2017 which results in a total of nine different models. Further details are shown in Appendix A.

Fine-Tuning

As described above, fine-tuning is done by sequentially re-training the model every two years. The validation dataset thereby consists of news articles that are published subsequent to the training period. In addition, the size of the validation set is set to 20% of the training set size, but limited to a maximum size of 20,000 news articles. Thus, if the model is trained with data from January 1996 to December 2015, the validation set starts in January 2016.

Although I constrain the number of neutral observations by considering only news articles associated with an $abs(z - value) < 1$ as neutral, instead of $abs(z - value) < 1.4$ (see Figure 2.2), the training and validation datasets are still imbalanced. The neutral class contains a larger number of observations than the others. To balance the data, I up-sample the minority classes.[7] The final number of observations in each class is determined by the minority class, which is up-sampled by a factor of two. The other classes are up-sampled to the same number of observations. Moreover, if the neutral class initially contains more than two times the observations of the minority class, the same amount of neutral news

[7]Up-sampling is a technique to randomly duplicate items until the desired number is reached.

is randomly selected from the pool of neutral news.

BERT can be fine-tuned end-to-end for a variety of downstream tasks with task specific input and output data (Devlin et al., 2018). For classification I use a deep neural network with three hidden layers. This network receives the concatenated feature vector as input (see Figure 2.3). The feature vector consists of the 256 dimensional CLS (or classification) token from FinNewsBERT and the 4 dimensional topic feature vector. The input layer and the three hidden layers of the neural network consist of $n = 260$ nodes each, which equals the length of the concatenated feature vector. Additionally, ReLu is used as an activation function and the dropout rate is set to 20% in all layers. I further use the optimizer AdamW with the parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $epsilon = 1e-08$ and $weight\ decay = 0.01$. The learning rate is warmed up over the first 15% of the training steps, with a maximum value of 5e-5 and linear learning rate decay of 0.01. The models are then trained over 3 epochs with a batch size of 32 and symmetric cross-entropy (SCE) loss (Wang et al., 2019).

As Wang et al. (2019) show, cross-entropy (CE) loss is not well suited for noisy labels. They find that deep neural networks trained with cross-entropy loss on noisy labels tend to overfit on easier to learn classes, while more difficult classes are underlearned. The authors therefore introduce symmetric cross-entropy learning (SL) for robust learning with noisy labels. Beside the traditional cross-entropy loss (Formula 2.4), the authors define the reverse cross-entropy (RCE) loss (Formula 2.5) where $q(k|\boldsymbol{x})$ denotes the ground truth class distribution of the sample $\boldsymbol{x}$ and $p(k|\boldsymbol{x})$ denotes the predicted distribution of class labels $k \in \{1, ..., K\}$. Combining both, cross-entropy loss and reverse cross-entropy loss, results in symmetric cross-entropy loss (SCE) (Formular 2.6) with the two hyperparameters α and β which are set to $\alpha = 0.5$ and $\beta = 3$ for fine-tuning.

$$l_{ce} = -\sum_{k=1}^{K} q(k|\boldsymbol{x})\ log\ p(k|\boldsymbol{x}) \tag{2.4}$$

$$l_{rce} = -\sum_{k=1}^{K} p(k|\boldsymbol{x})\ log\ q(k|\boldsymbol{x}) \tag{2.5}$$

$$l_{sce} = \alpha\, l_{ce} + \beta\, l_{rce} \tag{2.6}$$

Figure 2.4 shows the training and validation loss/accuracy for the last model that is trained on data from 1996 to 2017. The training and validation accuracies are 57.61% and 50.06%, respectively. Note, the accuracy of the random guessing baseline is 33%, since I predict the three classes positive, negative and neutral. However, these values still seem to be quite low compared to the results of BERT in NLP tasks under less noisy conditions. BERT base, for example, achieved 93.5% accuracy in the SST-2 (Stanford Sentiment Treebank) binary single-sentence classification task that is based on movie reviews (Devlin et al., 2018). In contrast to movie reviews however, the sentiment annotations of the financial news are derived from future asset returns that are noisy in their nature. Even a correct classification of a positive sounding news article as positive could result in a false prediction if the return on the next trading day is negative. So these numbers appear in a different light when they are seen less as the accuracy of predicting the right sentiment but more as the accuracy of predicting whether a stock will rise or fall the next day. If the model correctly predicts the sentiment of a news article as positive, it means that the asset return has also increased significantly, since the labels are derived from the asset returns. More specifically, accuracy indicates the model's ability to predict whether asset returns will rise above a certain threshold, which I previously defined as the barrier for labelling articles as positive, neutral or negative. Thus, if an article is predicted to be positive, but the associated label indicates it to be neutral, this does not necessarily mean that the prediction of the model is wrong in terms of the predictability of the price direction. This would be the case if the asset's return on the next day is slightly positive but below the positive barrier, resulting in a neutral label. The models ability in predicting future up- and down movements in asset prices is further discussed in Section 2.6.1.

Furthermore, Figure 2.4 also shows the training and validation loss. The loss seems to be quite high at first glance. However, the reason for those large values is the second term $\beta\, l_{rce}$ of the symmetric cross entropy loss (see Equation (2.6)). In addition, no

improvement in terms of validation accuracy and validation loss can be observed when trained over multiple epochs. Due to the pre-training of BERT and the up-sampling of minority classes, the accuracy after the first epoch is already relatively high. Moreover, I observe a slight decrease in accuracy and a slight increase in losses in the third epoch, indicating that the model is starting to overfit. Also, validation accuracy and loss is strongly influenced by the business cycle. Figure 2.5 shows the training and validation accuracy and loss for all models after training for three epochs. What is striking is the sharp drop in validation accuracy and the large spike in validation loss for the model trained with data from 1996 to 2007. This is because the subsequent period starting in 2008, which is used as the validation set, is marked by the financial crisis. I also observe that it takes about three years for the training accuracy to recover to pre-financial crisis levels. I assume that the data during the financial crisis contains a larger fraction of mislabelled observations, which weakens the accuracy in the following periods. However, I do not investigate further whether it is beneficial to exclude this period.

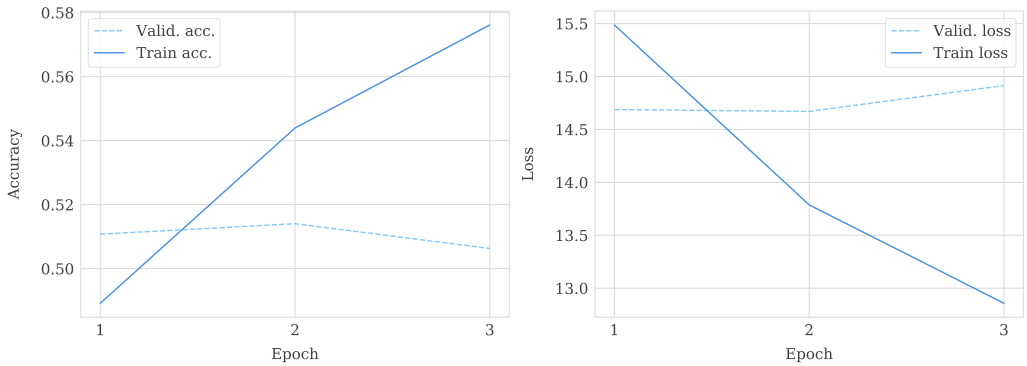


Figure 2.4: This figure shows the training and validation accuracy and loss for the ninth model, trained with data from 1996 to 2017.

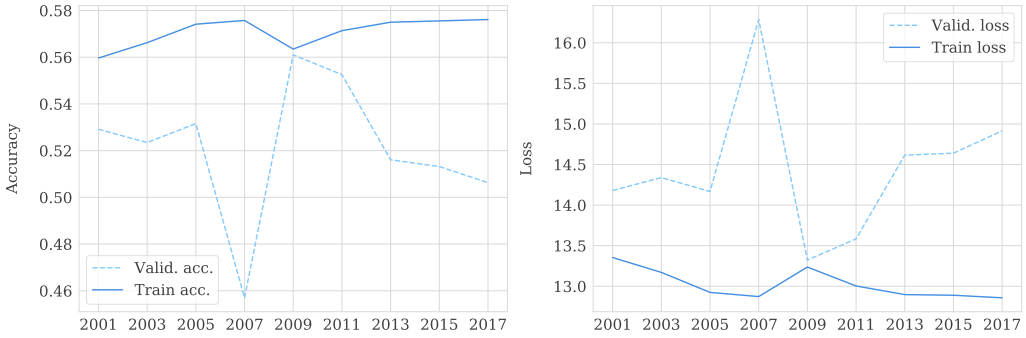


Figure 2.5: This figure shows the training and validation accuracy and loss for all models after training for three epochs.

2.5 Risk Adjusted Portfolio Benchmark

In this Section I briefly describe how I derive the benchmark, the CAPM-implied expected conditional return $x_{b(t)}$. According to the Capital Asset Pricing Model (CAPM), the expected return of an investment is determined by its beta with the efficient market portfolio. Within CAPM, it is assumed that all investors act rationally and have homogeneous expectations. As a consequence, all investors choose the same portfolio with the highest Sharpe ratio - the tangent portfolio which equals the market portfolio (Berk and DeMarzo, 2014). According to the CAPM, the expected return of an investment is calculated with Formula 2.7 and the financial market is in equilibrium when all assets lie on the security market line (SML). Deviations from the SML, denoted with α_i, occur when asset returns don't equal the CAPM expected returns. These inefficiencies can be caused either by investors not acting in a fully rational and unbiased manner. But also the emergence of new information in the form of financial news can change the expected return $E[R_i]$ of stocks, leading to a positive or negative alpha and thus a deviation from the SML. When the stock market is not efficient, investors can profit by buying stocks with positive alphas and selling stocks with negative alphas (Berk and DeMarzo, 2014). The hypothesis is that the proposed model is able to detect market inefficiencies via a sentiment analysis of financial news. To prove the models ability, a trading strategy is implemented with the objective to generate positive alpha by making trading decisions based

on the sentiment of financial news. Therefore it buys stocks associated with positive news, and sells stocks associated with negative news. A confirmation of this hypothesis would consequently provide evidence that the financial news from Thomson Reuters contains predictive information about future stock returns.

$$E^{\text{CAPM}}[R_i] = R_f + \beta_i * (E[R_m] - R_f) \tag{2.7}$$

$$\alpha_i = E[R_i] - E^{\text{CAPM}}[R_i] \qquad \text{and} \qquad \beta_i = \frac{Cov(R_i, R_m)}{Var(R_m)} \tag{2.8}$$

The expected excess portfolio return $E[R_P]$ - R_f equals the product of the portfolio beta β_P with the market risk premium (see Equation (2.9)). Since the portfolio allocation changes in every time step, the portfolio beta needs to be calculated on a daily basis with Formula 2.10. The covariance matrix $Cov(R_i, R_m)$, as well as the variance $Var(R_m)$ is calculated over a 2-year rolling window. The market return is the value weighted S&P 500 total return index. Also, the CAPM-implied expected conditional return $x_{b,t}$ and the excess portfolio return $x_{P,t}$ are calculated in each time step t (see Equation (2.11)).

$$E[R_P] - R_f = \beta_P * (E[R_m] - R_f) \tag{2.9}$$

$$\beta_{P,t} = \sum w_{i,t} * \beta_{i,t} \tag{2.10}$$

$$x_{P,t} = R_{P,t} - R_f \qquad\qquad x_{b,t} = \beta_{P,t-1} * (R_{m,t} - R_f) \tag{2.11}$$

The difference of the excess portfolio return and the risk adjusted benchmark return, the CAPM-implied expected conditional return, must be zero if the market portfolio is

in the CAPM equilibrium. However, if the market is not efficient, the risk adjusted outperformance $y_{(t)}$ is defined in Equation (2.12). This result is further tested for statistical significance with a t-test (Equation (2.13)) and the null hypothesis: The mean value of alpha is equal to the mean value of the error term ε_T which equals zero.

$$y_{(t)} = x_{P(t)} - x_{b(t)} = \alpha_{(t)} + \varepsilon_{(t)} \tag{2.12}$$

$$t = \frac{\bar{y_T} - \varepsilon_T}{\frac{\sigma_T}{\sqrt{n}}} \tag{2.13}$$

Symbol	Description
$E^{\text{CAPM}}[R_i]$	expected return of asset i according to CAPM
$E[R_i]$	expected return of asset i
$E[R_m]$	expected return of the market portfolio
$E[R_m] - R_f$	market risk premium
$\beta_i * (E[R_m] - R_f)$	risk premium for security i
$E[R_P]$	expected portfolio return
R_f	risk free rate
β_P	portfolio beta
$x_{b,t}$	CAPM-implied expected conditional excess return at time step t (risk adjusted benchmark)
$x_{P,t}$	excess portfolio return at time step t
$R_{P,t}$	portfolio return at time step t
$R_{m,t}$	return of the market portfolio at time step t
$\beta_{P,t}$	portfolio beta at time step t
$w_{i(t)}$	weight of the security i at time step t
$\varepsilon_{(t)}$	error term at time step t
$\bar{y_T}$	mean of the risk adjusted outperformance at time step T over the window of size n
σ_T	standard deviation evaluated at time step T over the window of size n
n	number of observations

Table 2.2: List of Symbols

2.6 Empirical Analysis

In this Section, I examine the model's ability to extract a meaningful sentiment signal from financial news and investigate whether this signal is associated with abnormal asset returns. Therefore I perform an event study in Section 2.6.1 by analyzing the impact of financial news for different sectors. In addition, I evaluate the model's ability to predict upward and downward movements of future asset prices in Section 2.6.1. I find a short-term negative momentum effect that lasts up to two days after the news release. This effect is described in more detail in Section 2.6.2. The subsequent sections deal with the ability of the model to predict future asset returns. In Section 2.6.3 I introduce the trading strategy and perform a factor analysis with Fama French factors in Section 2.6.3. In Section 2.6.3 I conduct backtests with different settings over the period from 01-2002 to 01-2020. I also compare the backtest performance between trading at market closing and trading at market opening in Section 2.6.3. This is followed by a comparison of FinNewsBERT with FinBERT in Section 2.6.3 and an analysis of the influence of the additional topic features in Section 2.6.4. All figures presented in this Section are strictly out-of-sample. The model is re-trained every two years and used for inference in the subsequent 2-year periods (see Section 2.4.4).

2.6.1 Event study

The event study in Figure 2.6 shows daily CAPM abnormal returns for 10 sectors around the release of financial news. Specifically, I consider fresh news published in the 17.5-hour time window between the market's close at 4 pm on day $t-1$ and its opening at 9:30 am on day t. It can be observed that news classified as positive (negative) is associated with large positive (negative) abnormal returns on day t. This indicates that FinNewsBERT is able to correctly predict the sentiment of financial news and its impact on asset returns. The largest abnormal returns can be observed in the same market opening to market opening interval where the news is published $r_t = \frac{p_t}{p_{t-1}} - 1$. Furthermore, significant abnormal returns can also be observed prior to the release of the news. Leinweber and Sisk (2011) finds similar pre-news effects and argues that the reasons are twofold. First, some investors

have better access to primary news sources, i.e. those reporters rely on when formulating news. Those sources include the Security and Exchange Commission (SEC), government and corporate sources as well as social media postings on platforms like twitter. Second, a subset of the news simply talks about previous price changes in assets without containing any predictive information.

Predictability, however, still exists on day $t+1$, but with much smaller magnitudes compared to day t. The sectors Industrials, Consumer Services, Basic Materials, Technology, Oil&Gas and Consumer Goods show the largest abnormal returns on day $t+1$. The sectors Financials and Utilities show the weakest return predictability on day $t+1$, which is why these sectors are excluded from the backtest.

Predicting Price Direction

I further investigate the ability of the model to predict upward and downward movements in asset prices. Therefore, I only consider news items that are predicted to be either positive or negative. If a news article published in the 17.5-hour window before market opening is predicted to be positive (negative) and the asset price rises (falls) from market opening on day t to market opening on day $t+1$, then the prediction is considered to be true positive (negative). Table 2.3 summarizes the results. The precision metric is more important than recall in terms of a trading strategy, as it indicates the proportion of predicted positive (negative) classes that are actually true positives (negatives). In other words, it is less problematic to miss potential trades by predicting them as neutral (low recall) than to incorrectly classify them as negative or positive (low precision). The precision of news filtered with Filter A in Table 2.3 is 51.40% for negative news and 52.03% for positive news, on the subsequent trading day ($t+1$) after the release of the news article. In addition, it can be observed that the mean returns on day $t+1$ is -13.77% p.a. for negative news and 33.03% p.a. for positive news. With an average return of 16.702% p.a. on day $t+1$ across all news, I can reject the null hypothesis that the mean return of positive and negative predictions is equal to the mean return across all news on a 1% significance level. If financial news is narrowed down by adding additional filters, as shown in Table 2.3, the classification metrics can be further improved. Filter B adds the

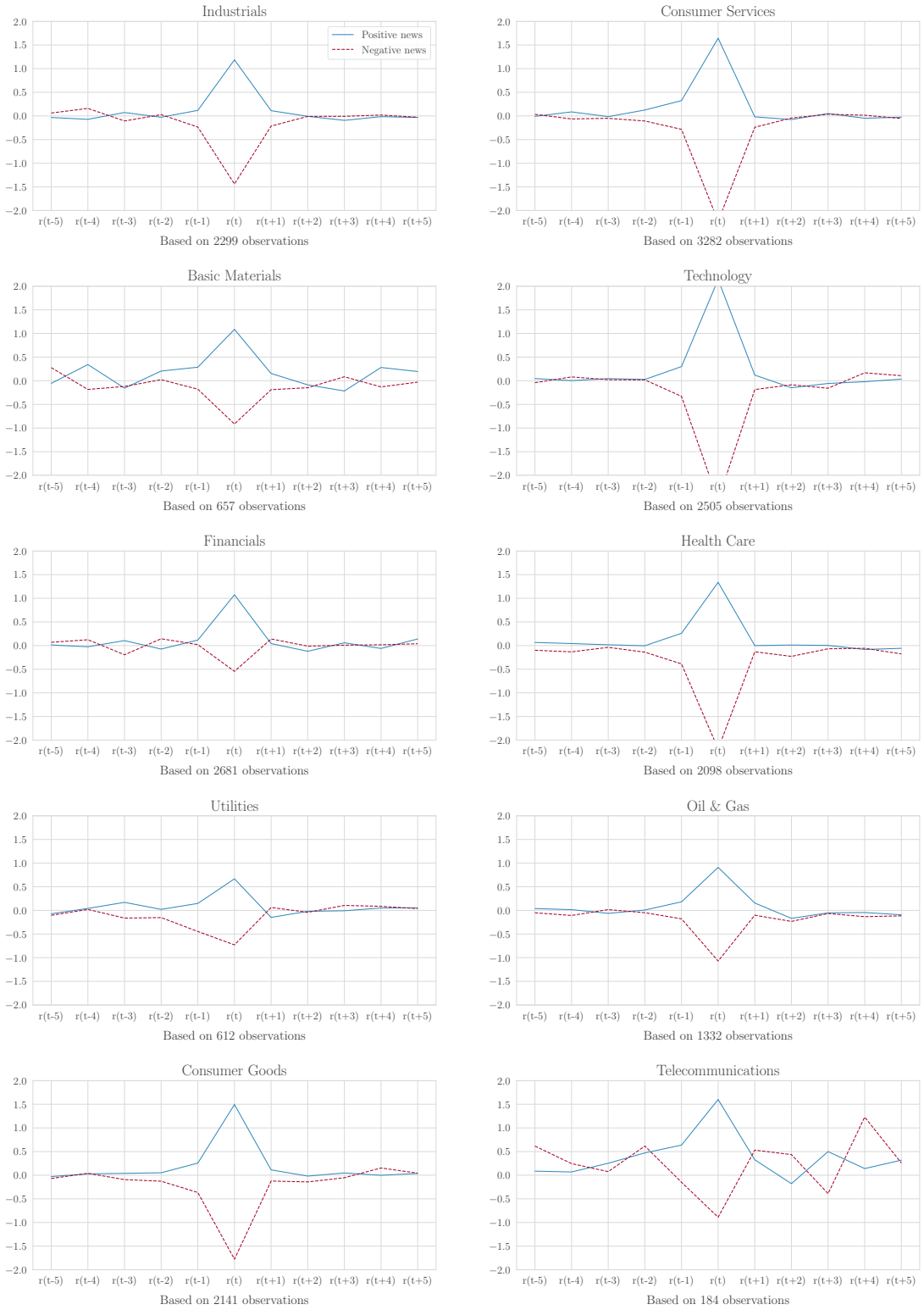


Figure 2.6: Event study of financial news showing CAPM abnormal returns measured from open-to-open (in percent) for 10 sectors. News articles published in the 17.5-hour time window between market close at 4 pm on day $t-1$ and market open at 9.30 am on day t over the period from 01-2002 to 01-2020 are classified into positive and negative news using FinNewsBERT with a confidence level of 95% and matched with market open returns from $t-5$ to $t+5$.

additional restriction to only consider news attributable to the topic "analyst forecast" and Filter C further adds the restriction to only consider negative news associated with a z-value < -1.96 (95% confidence interval) on day t. With Filter B, the precision is 52.36% for negative news and 52.78% for positive news. Filter C further improves the precision of negative news to 56.05% resulting in an average return of -96.19% p.a. on day t+1.

	Filter A		Filter B		Filter C	
	Neg. pred.	Pos. pred.	Neg. pred.	Pos. pred.	Neg. pred.	Pos. pred.
Precision	0.5140	0.5203	0.5236	0.5278	0.5605	0.5278
Recall	0.1646	0.1991	0.2838	0.3677	0.1786	0.7399
F1-score	0.2493	0.2880	0.3681	0.4335	0.2709	0.6161
Support	61916	65878	19746	20480	9680	10179
Avg. return p.a. (arithm.)	-13.77%	33.03%	-31.16%	47.15%	-96.19%	47.15%

Table 2.3: This table shows the classification metrics for three subsets of news data, determined by different filters. All numbers are related to the return on day t+1. Filter A includes all news (fresh & stale) that are classified with a confidence above 95%. Filter B, adds the additional restriction to only consider news with the topic "analyst forecast". Moreover, Filter C adds the restriction to only consider negative news associated with a z-value < -1.96 (95% confidence interval) on day t, positive news is not further restricted. In addition, the daily returns on t+1 are positive in 51.55% of the time for Filter A, 50.912% for Filter B and 51.256% for Filter C.

2.6.2 Short-term Momentum Effect

In this section I take a closer look at the short-term momentum effect induced by negative news. Figure 2.7 shows the average daily abnormal returns after news events. I find that stocks with significant negative returns before the news event (d to f) show large negative abnormal returns one to two days after the news event. In the case of fresh news, this effect lasts for two days and in the case of stale news it lasts for one day. To prove that this effect is caused by the news event and not just by the arbitrary price drop on day t, Figure 2.8 (a) shows the average abnormal daily returns subsequent to z-values that fall below the 95% confidence barrier (z-values < -1.96) on day t. Here, no momentum effect can be observed. On day t+1, the average return is positive in seven out of nine sub-periods, indicating a reversion to the mean. In contrast, Figure 2.8 (b) shows the combined effect of z-values < -1.96 on day t and the release of negative news in the 17.5h window before the market opens. Here, in eight out of nine sub-periods, returns are negative on both

day $t+1$ and day $t+2$. Only the period of the financial crisis from 2008 to 2009, shows deviations from this pattern.

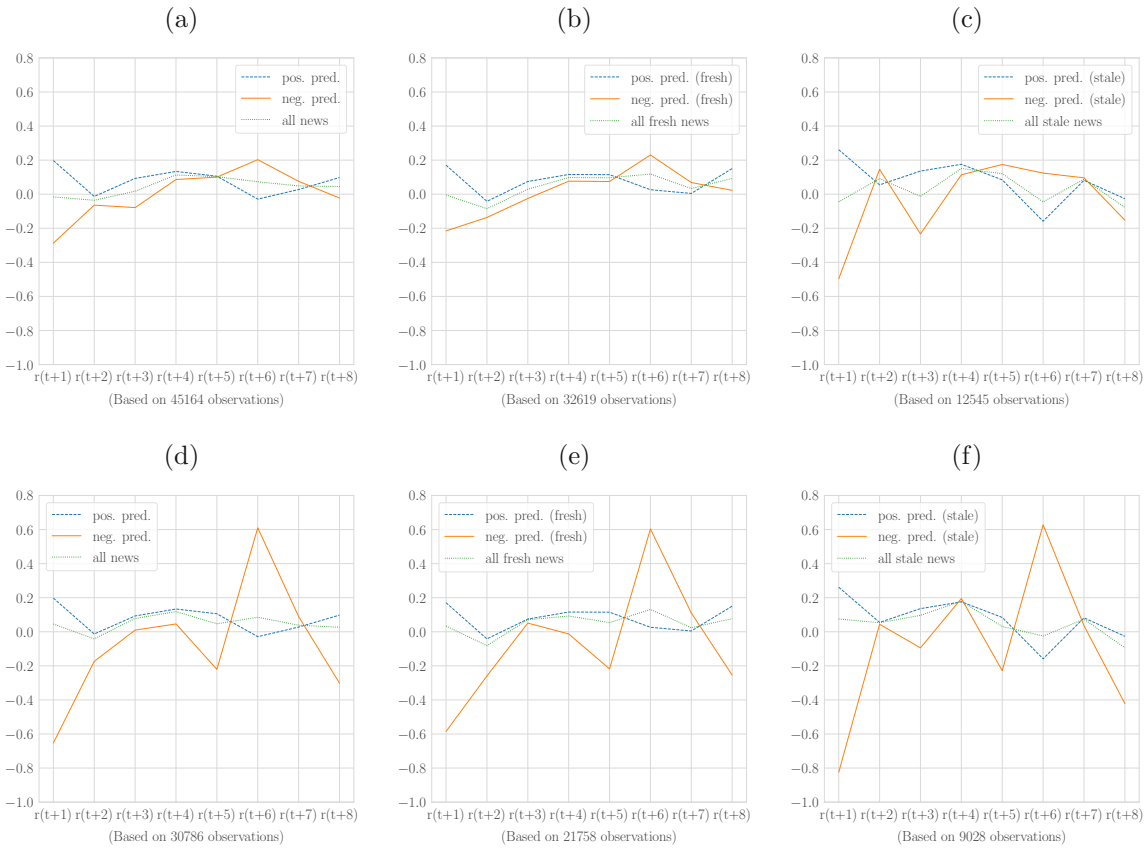


Figure 2.7: Event study showing the average daily abnormal returns (in percent) in the period 01-2002 to 01-2020 following a news event. The first row shows the abnormal returns for all-, fresh- and stale-news. The second row shows abnormal returns with data pre-filtered by only using negative news with a z-value < -1.96 on day t.

2.6.3 Return Predictions

To evaluate the economic value of the model, I further investigate the predictive power of the generated sentiment signal. Therefore I implement a trading strategy that goes long in assets with positive news sentiment and short in assets with negative news sentiment. Based on the selected assets, I then form equally weighted long and long/short portfolios that are rebalanced at a daily frequency. The trading algorithm is designed in a flexible way that allows us to examine and compare the results with different parameter settings. These settings include whether trades are initiated either at market open or close, the

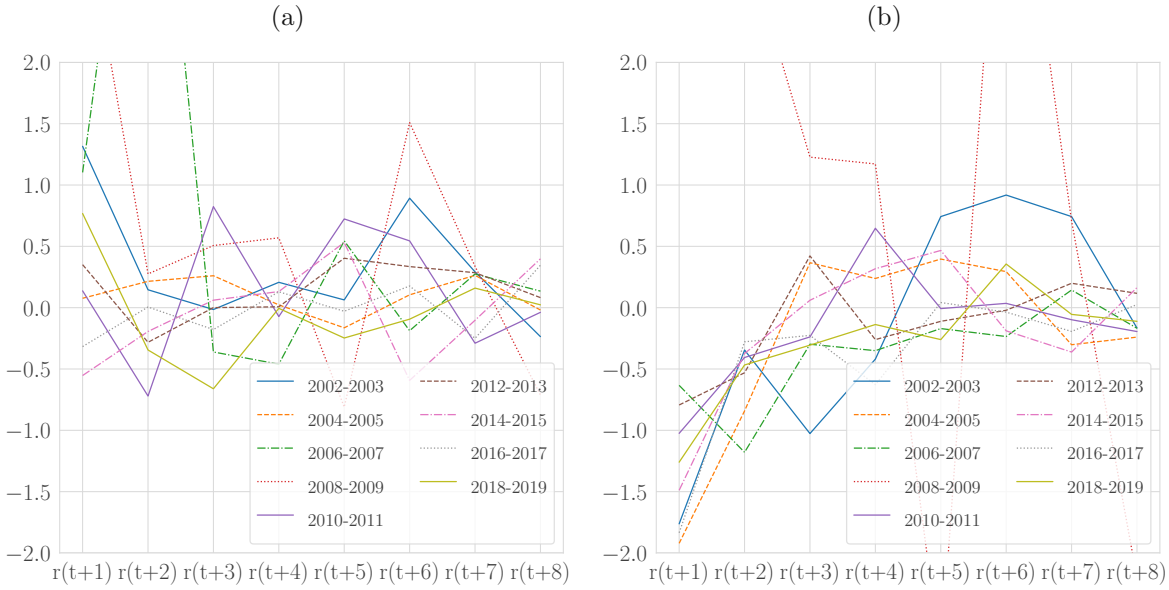


Figure 2.8: This graphs show the average abnormal returns (in percent) in nine 2-year sub-periods after (a) z-values < -1.96 on day t and (b) z-values < -1.96 on day t and release of negative news (fresh & stale) in the 17.5 hour window prior to market open on day t.

day on which trades are opened and closed, the time window of news considered for predictions, a restriction on maximum portfolio weighting and a restriction on portfolio size. In addition, it allows to filter for fresh or stale news as well as for news related to the topics "analyst forecast" or "earnings report". Besides, the sentiment signal is only used for asset allocation if the confidence, i.e. the probability that a news article is classified as positive or neutral, is above 95%. Table 2.4 summarises the parameter settings used in the following analyses and backtests.

	Settings A	Settings B	Settings C
News	Fresh	Fresh	All
Topics	All	Analyst forecast	All
News window	17.5h	17.5h	17.5h
Order time	MOO	MOO	MOO
Enter/Exit long (l)	t/t+1	t/t+1	t/t+1
Enter/Exit long (l/s)	t/t+1	t/t+1	t/t+1
Enter/Exit short (l/s)	t/t+1	t/t+1	t/t+1
Max. portfolio size	30	10	10
Max. portfolio weight	1	1	1
Z-value(t) neg./pos. news	$(-\infty, \infty)/(-\infty, \infty)$	$(-\infty, -1,96]/(-\infty, \infty)$	$(-\infty, \infty)/(-\infty, \infty)$

Table 2.4: This table shows different predefined settings for the trading strategy. The order time is either Market-On-Open (MOO) or Market-On-Close (MOC). The news window determines the time interval before the order time in which news is considered for predictions. If trades are initiated at market open (MOO) and the news window is 17.5h, then all news published between 4pm on day t–1 and 9.30am on day t are taken into account. The abbreviation (l) denotes the long portfolio and (l/s) denotes the long/short portfolio. In addition, all financial news associated with z-values within the intervals are considered by the trading strategy.

Factor analysis

A regression on the Fama French factor models shows significant alphas of 1.60% per month (19.20% p.a.) for the long portfolio and 5.01% per month (60.12% p.a.) for the long/short portfolio with respect to the Fama French five-factor model with momentum for Settings A. With Settings B, the monthly alpha of the long portfolio becomes 2.25% (27.00% p.a.) and 6.46% (77.56%) for the long/short portfolio (see Table 2.5).

	Settings A					
	FF3		FF5		FF5+MOM	
Portfolio	α	R^2	α	R^2	α	R^2
Long	1.78***	18.69%	1.61**	19.59%	1.60**	19.81%
Long/short	5.30***	9.07%	5.00***	11.01%	5.01***	11.06%
	Settings B					
Long	2.62***	16.64%	2.27***	17.40%	2.25***	18.10%
Long/short	6.42***	4.55%	6.48***	5.19%	6.46***	5.56%

Table 2.5: Monthly alphas (in %) and R^2s of the long and the long/short portfolio with respect to the Fama French three-factor model (FF3), the Fama French five-factor model and the Fama French five-factor model with momentum. **, *** corresponds to significance levels of 5% and 1%. The trading strategy is simulated with Settings A and B from Table 2.4.

Backtest

To further investigate the quality of the generated sentiment signal, I conduct an out-of-sample backtest over the 18-year period from 01-2002 to 02-2020. As a benchmark, I consider the CAPM-implied expected conditional return from Equation (2.11). I also set the initial capital to USD 100,000 and the transaction costs to zero. Figure 2.9 shows the backtest results for the long and the long/short strategy with Settings A from Table 2.4. It can be observed that both strategies outperform the benchmarks and the S&P 500 total return. The corresponding performance metrics are summarized in Table 2.6. The mean of the risk-adjusted outperformance α is 7.96 bps/day for the long portfolio and 18.41 bps/day for the long/short portfolio. Those values are significant on a 5% and 1% level with t-values of 2.51 for the long- and 4.59 for the long/short portfolio. Furthermore, the Sharpe Ratios are 0.88 for the long-, 1.29 for the long/short portfolio and 0.36 for the S&P 500 (calculated in excess to the risk-free rate). Additionally, Figure 2.10 (a) shows the ex-ante beta of the long and the long/short strategy, calculated over a rolling window of 30 days. Since the portfolio constituents change at a daily frequency, the ex-ante beta is also calculated at daily frequency. The average portfolio betas are 0.99 for the long- and 0.02 for the long/short portfolio. Figure 2.10 (b) shows the investment ratio of the long side and the short side of the long/short portfolio calculated over a 30-day rolling window.

If the investment ratio is equal to one, the entire capital is invested in equities, otherwise, if the investment ratio is less than one, the remaining capital is invested in a risk-free asset with a risk-free interest rate R_f. The investment ratio is derived by multiplying the equal asset weighting by the number of assets in the portfolio. In addition, Figure 2.11 shows the backtest with settings B from Table 2.4 with Sharpe Ratios of the long and long/short portfolios of 1.44 and 2.26 respectively (see Table 2.7). Furthermore, the risk-adjusted outperformance of the long/short portfolio increases to 29.94 bps/day with a return per trade of 24.06 bps.

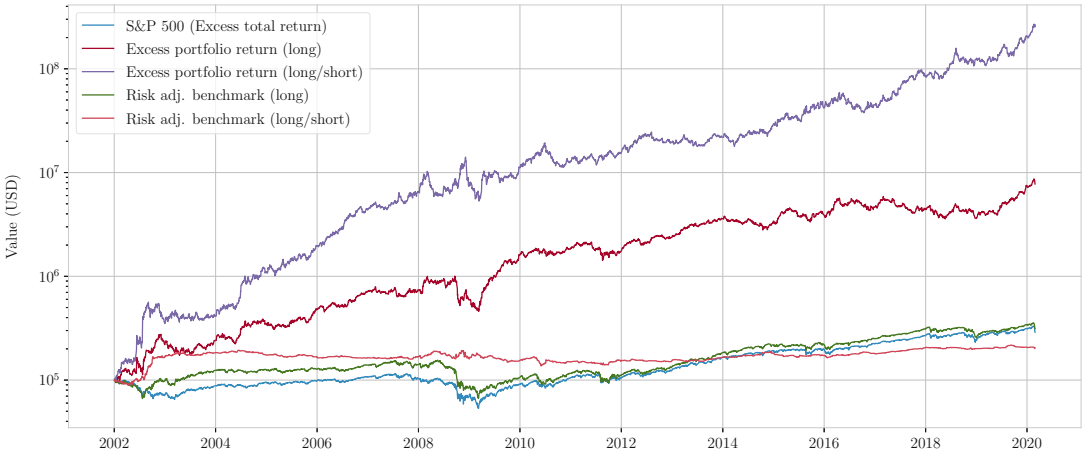


Figure 2.9: Backtest of the long- and the long/short portfolio with Settings A from Table 2.4. Benchmarks are the S&P 500 total return and the risk adjusted CAPM-implied expected conditional returns.

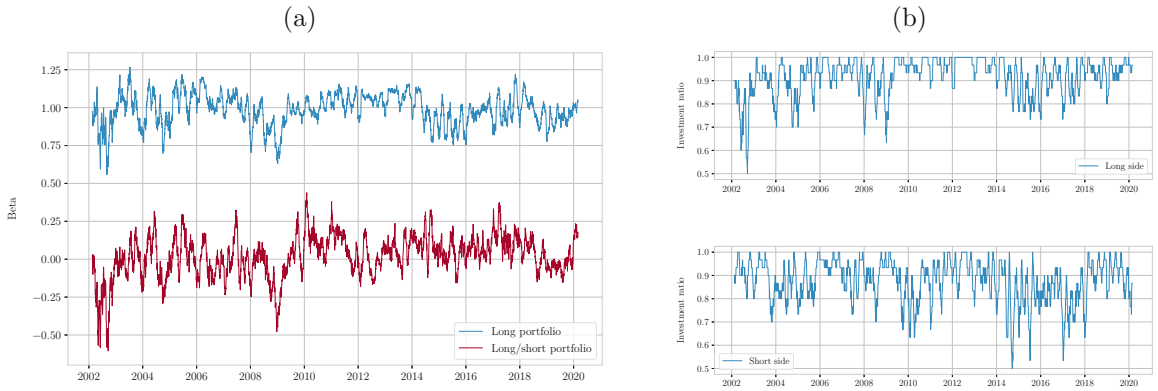


Figure 2.10: (a) Portfolio beta of the long- and the long/short portfolio with Settings A from Table 2.4 over a 30-day rolling window. (b) Investment ratio of the long side and the short side of the long/short portfolio calculated over a 30-day rolling window.

	Long	Long/Short	S&P 500
CAGR	27.04% p.a.	53.34% p.a.	6.57% p.a.
Std. dev.	30.78% p.a.	41.36% p.a.	18.33% p.a.
Sharpe Ratio	0.88	1.29	0.36
Beta	0.99	0.02	1.0
Alpha	7.96 bps/day**	18.41 bps/day***	
Trade count p.a.	834	1477	
Avg. portfolio size	3.50	Long: 3.50; Short: 2.70	
Daily turnover	95.13%	95.09%	
Return per trade	11.84 bps	11.74 bps	

Table 2.6: Performance metrics of the long- and the long/short portfolio with Settings A from Table 2.4 in the out-of-sample backtest period from 2002 to 2020. The buying and selling of an asset is considered as one trade and the metrics are calculated in excess to the risk-free rate. **, *** corresponds to significance levels of 5% and 1%.

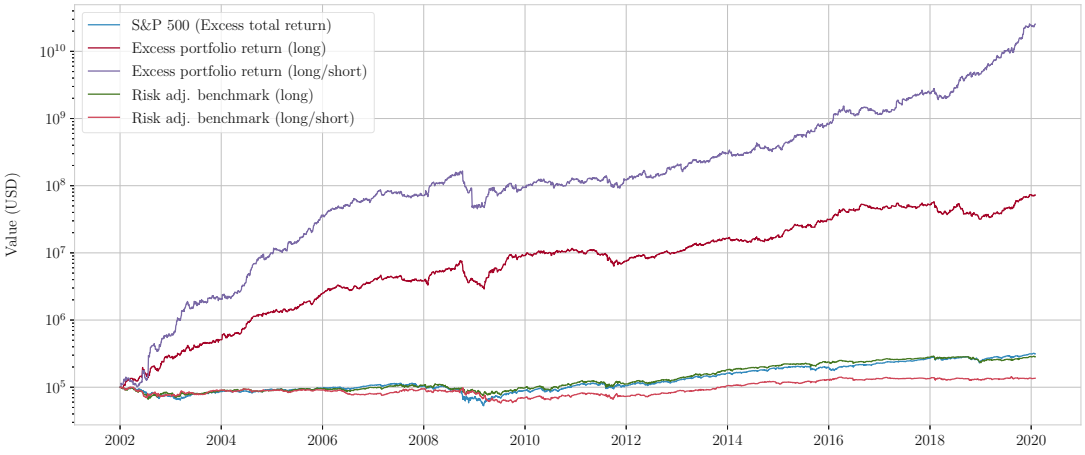


Figure 2.11: Backtest of the long- and the long/short portfolio with Settings B from Table 2.4. Benchmarks are the S&P 500 total return and the risk adjusted CAPM-implied expected conditional return.

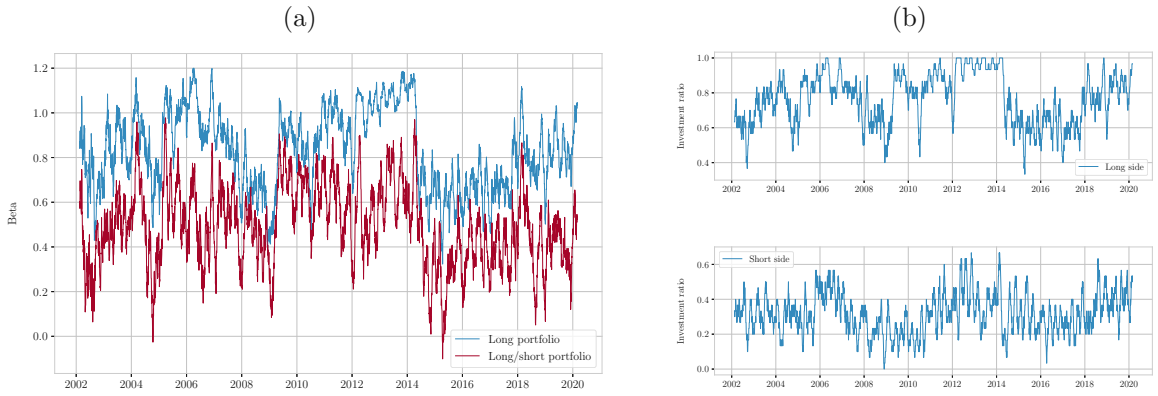


Figure 2.12: (a) Portfolio beta of the long- and the long/short portfolio with Settings B from Table 2.4 over a 30-day rolling window. (b) Investment ratio of the long side and the short side of the long/short portfolio calculated over a 30-day rolling window.

	Long	Long/Short	S&P 500
CAGR	44.04% p.a.	99.24% p.a.	6.56% p.a.
Std. dev.	30.58% p.a.	44.04% p.a.	18.33% p.a.
Sharpe Ratio	1.44	2.26	0.36
Beta	0.84	0.49	1.0
Alpha	13.43 bps/day***	29.94 bps/day***	
Trade count p.a.	454	553	
Avg. portfolio size	1.93	Long: 1.92; Short: 0.42	
Daily turnover	95.92%	96.34%	
Return per trade	18.31 bps	24.06 bps	

Table 2.7: Performance metrics of the long- and the long/short portfolio with Settings B from Table 2.4 in the out-of-sample backtest period from 2002 to 2020. The buying and selling of an asset is considered as one trade and the metrics are calculated in excess to the risk-free rate. *** corresponds to a significance level of 1%.

Trading at Market Closing

Trading at market closing significantly reduces performance compared to market opening. The Sharpe Ratio of the backtest performed with settings B from Table 2.4 drops from 2.26 to 0.90 and the return per trade drops from 24.06 bps to 11.92 bps when traded at market closing (see Table 2.6 and 2.8). The reason for this difference is the market's quick response to financial news. Information is incorporated into asset prices during market hours. As a consequence, few exploitable alpha remains at market close. Reducing the news window to 6.5 hours, i.e. taking into account only those news articles that are published within market opening hours, slightly increases the Sharpe Ratio to 1.00 and

the return per trade to 14.40 bps.

	News window 17.5h	News window 6.5h	News window 3h
CAGR	30.73% p.a.	33.16% p.a.	19.17% p.a.
Std. dev.	34.04% p.a.	33.25% p.a.	29.84% p.a.
Sharpe Ratio	0.90	1.00	0.64
Beta	0.53	0.45	0.31
Alpha	10.03 bps/day***	9.93 bps/day***	6.79 bps/day***
Trade count p.a.	705	320	175
Avg. portfolio size	Long: 2.46; Short: 0.53	Long: 1.07; Short: 0.26	Long: 0.58; Short: 0.14
Return per trade	11.92 bps	14.40 bps	13.73 bps
Daily turnover	95.91%	98.01%	98.5%

Table 2.8: Performance metrics for varying news window lengths of the long/short portfolio with trades initiated at market closing. Apart from the order time and news window, settings used for the backtest are identical to Settings B from Table 2.4. The news window determines the time interval before market close in which news is considered for predictions. The out-of-sample backtest period ranges from 2002 to 2020. The buying and selling of an asset is considered as one trade. *** corresponds to a significance level of 1%.

Comparison with FinBERT

In this section I compare FinBERT, proposed by Araci (2019), with FinNewsBERT. FinBERT is a BERT model in the base version with 110 million parameters, which is further fine-tuned on Thomson Reuters news data. FinBERT builds upon BERT that is pre-trained in 2018 on Wikipedia and a large corpus of books. Consequently, a backtest starting before 2018 would involve an in-sample bias due to the prior knowledge learned during pre-training. In order to compare both models out of sample, I start the backtest in 2018. I further use Settings C from Table 2.4 for the backtest and consider all sentiment classifications that exceed a confidence threshold of 95% for both FinNewsBERT and FinBERT. The backtest of the long/short portfolios is shown in Figure 2.13. Additionally, Table 2.9 highlights the performance metrics of the backtests. It can be observed that FinNewsBERT outperforms FinBERT in terms of absolute performance, risk adjusted performance and in terms of return per trade. Although our model is less than one-fifth the size of FinBERT, it shows superior out-of-sample performance.

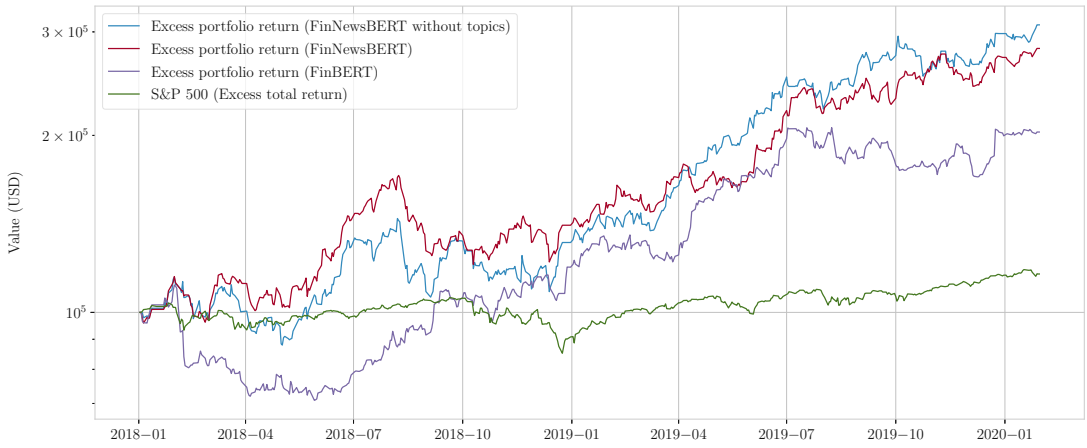


Figure 2.13: Comparison of the long/short portfolios generated with FinNewsBERT and FinBERT in the out-of-sample period from 01-2018 to 01-2020 with Settings C from Table 2.4.

	FinNewsBERT	FinBERT
CAGR	64.56% p.a.	40.54% p.a.
Std. dev.	32.03% p.a.	31.02% p.a.
Sharpe Ratio	2.02	1.31
Beta	0.05	0.17
Alpha	20.07 bps/day**	18.44 bps/day**
Trade count p.a.	1697	2034
Avg. portfolio size	Long: 3.93; Short: 2.79	Long: 6.11; Short: 1.91
Return per trade	8.99 bps	7.21 bps
Daily turnover	95.20%	95.18%

Table 2.9: Comparison of the performance metrics of the long/short portfolios generated with FinNewsBERT and FinBERT. The out-of-sample backtest period ranges from 01-2018 to 01-2020. The buying and selling of an asset is considered as one trade. ** corresponds to a significance level of 5%.

2.6.4 Topic Features

In this section, I investigate whether the additional topic features generated with Text2Topic increase the model's ability to produce valuable trading signals. Therefore, I sequentially fine-tune the model with identical hyperparameters, but without using the additional topic features. In order to compare both models I conduct backtests with Settings A from Table 2.4. As Table 2.10 shows, performance drastically improves with the additional use of topic features. The Sharpe Ratio increases by 47% from 0.88 to 1.29 and the return per trade increases by 17% from 10.05 bps to 11.74 bps in case of the long/short portfolio. Furthermore, the backtest in Figure 2.14 shows the equity curves in excess to the risk free rate and the risk adjusted benchmarks of the long/short portfolios for different subsets of news data. It can be observed that fresh news of the topic "analyst forecast" generates a better performance compared to fresh news of the topic "earnings report" and all fresh news. Since the observation count differs across the subsets of news data, the return per trade, shown if Table 2.11 is a better measure for comparing the different news subsets. The return per trade is highest for fresh news of the topic "analyst forecast".

	With Topic Features		Without Topic Features	
	Long Portfolio	Long/Short Portfolio	Long Portfolio	Long/Short Portfolio
CAGR	27.04% p.a.	53.34% p.a.	14.53% p.a.	38.21% p.a.
Std. dev.	30.78% p.a.	41.36% p.a.	30.66% p.a.	43.48% p.a.
Sharpe Ratio	0.88	1.29	0.47	0.88
Beta	0.99	0.02	0.98	0.03
Alpha	7.96 bps/day**	18.41 bps/day***	3.99 bps/day	14.57 bps/day***
Trade count p.a.	834	1477	795	1421
Avg. portfolio size	3.50	Long: 3.50; Short: 2.70	3.33	Long: 3.33; Short: 2.63
Daily turnover	95.84%	95.09%	95.14%	95.01%
Return per trade	11.84 bps	11.74 bps	9.82 bps	10.05 bps

Table 2.10: Comparison of the model with and without topic features. The out-of-sample backtest period ranges from 01-2002 to 01-2020 and is generated with Settings A from Table 2.4. Both models are fine-tuned with the same hyperparameters over three epochs. The buying and selling of an asset is considered as one trade. **, *** corresponds to significance levels of 5% and 1%

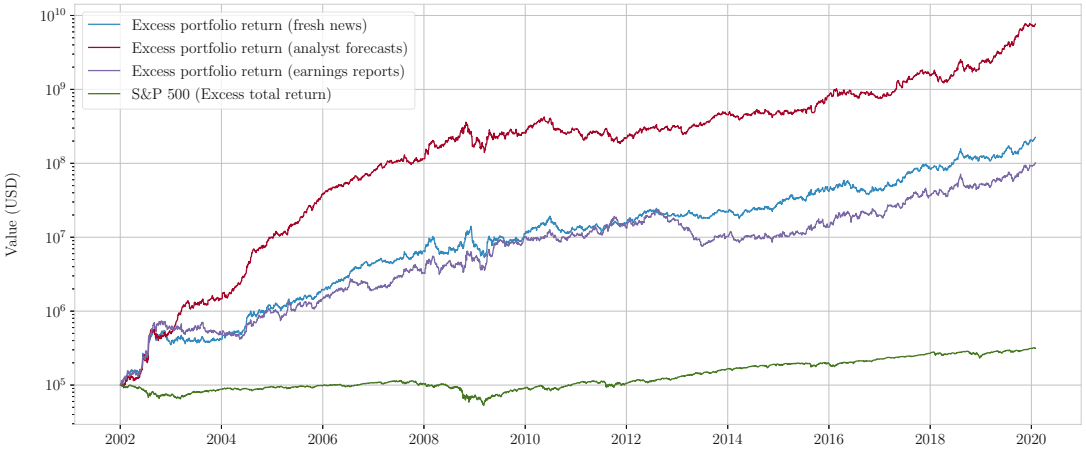


Figure 2.14: Backtest showing the excess portfolio returns and risk adjusted benchmarks of the long/short portfolios with Settings A from Table 2.4. The news are filtered for (a) fresh news, (b) fresh news of the topic "analyst forecast" and (c) fresh news of the topic "earnings report".

	Long Portfolio	Long/Short Portfolio
Fresh analyst forecasts	18.36	17.15
Fresh earnings reports	11.25	12.51
All fresh news	11.84	11.74

Table 2.11: Return per trade of the long and long/short portfolios in basis points corresponding to the backtest shown in Figure 2.14.

2.7 Conclusion

I contribute to the existing literature of financial sentiment analysis by investigating the ability of a BERT-based language model to generate a sentiment score from financial news articles for predicting short-term equity returns. To perform out-of-sample predictions prior to 2018, I train a BERT-based model from scratch on domain specific Thomson Reuters financial news data. I find that the model is able to extract a sentiment signal from financial news that is positively correlated with asset returns. While the information of financial news is incorporated into stock prices usually within one day, I find that it can take up to two days for fresh news. In contrast, I observe stronger abnormal returns for stale news on day t. The effects are amplified when negative news is accompanied with significantly negative returns on day t. Furthermore, I find that the prior categorisation of news articles into topics and providing this information in the form of additional features

further enhances the model's ability to predict future asset returns. Moreover, considering only a subset of data that comes with the topic "analyst forecast" results in the most accurate predictions of future asset returns. A backtest with the simple trading strategy, long (short) in assets with positive (negative) news and daily rebalancing results in a Sharpe Ratio of 1.44 for the long-only strategy and 2.26 for the long/short strategy over the out-of-sample period from 01-2002 to 01-2020. The realized alpha with respect to the FF5+MOM model is 27.02% p.a. for the long-only strategy and 77.56% p.a. for the long/short strategy. These values are significant at the 1% level. Furthermore, despite the fact that the proposed model is less than one fifth the size of BERT-base, it shows superior out-of-sample performance in comparison to FinBERT. In this paper, I focus on S&P 500 companies, all of which have large market capitalisations. I expect large improvements of the proposed results when smaller companies are also considered, since Kelly et al. (2019) shows that price reactions after news events are four times larger for small and volatile stocks than for large caps. I also assume that the placement of trades immediately after the release of financial news, within market opening hours, would further surpass the proposed results.

Bibliography

(2018). WWW '18: Companion Proceedings of the The Web Conference 2018, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. 59(3):1259–1294.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models.

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. 61(4):1645–1680.

Barber, B. M. and Odean, T. (2013). The behavior of individual investors. In Handbook of the Economics of Finance, volume 2, pages 1533–1570. Elsevier.

Berk, J. B. and DeMarzo, P. M. (2014). Corporate finance. The Pearson series in finance. Pearson, third edition edition.

Bollen, J. and Mao, H. (2011). Twitter mood as a stock market predictor. 44(10):91–94.

Cong, L., Liang, T., and Zhang, X. (2018). Textual Factors: A Scalable, Interpretable, and Data-driven Approach to Analyzing Unstructured Information.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Hendrycks, D. and Gimpel, K. (2016). Gaussian error linear units (gelus).

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. New phytologist, 11(2):37–50.

Jegadeesh, N. and Wu, D. (2013). Word power: A new approach for content analysis. 110(3):712–729. PII: S0304405X13002328.

Kahneman, D. (1973). Attention and effort, volume 1063. Citeseer.

Kelly, B. T., Ke, Z. T., and Xiu, D. (2019). Predicting returns with text data. (2019-10):54.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

Lee, H., Surdeanu, M., MacCartney, B., and Jurafsky, D. (2014). On the importance of text analysis for stock price prediction. In LREC, volume 2014, pages 1170–1175.

Leinweber, D. and Sisk, J. (2011). Event-driven trading and the "new news". 38(1):110–124.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., and Wang, P. (2020). K-bert: Enabling language representation with knowledge graph. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 2901–2908.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. 66(1):35–65.

Malo, P., Sinha, A., Takala, P., Korhonen, P., and Wallenius, J. (2013). Good debt or bad debt: Detecting semantic orientations in economic texts.

Manela, A. and Moreira, A. (2017). News implied volatility and disaster concerns. 123(1):137–162. PII: S0304405X16301751.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

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

Rechenthin, M., Street, W. N., and Srinivasan, P. (2013). Stock chatter: Using stock sentiment to predict price direction. 2(3-4):169–196.

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP

Frameworks, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Severyn, A. and Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In Baeza-Yates, R., Lalmas, M., Moffat, A., and Ribeiro-Neto, B., editors, Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15, pages 959–962. ACM Press.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. 62(3):1139–1168.

Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. (2019). Symmetric cross entropy for robust learning with noisy labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 322–330.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Appendices

A.1 Pre-training and fine-tuning of FinNewsBERT

As described in Section 2.4.4, the models are iteratively retrained every two years. Table 12 shows the number of training steps and the final training loss of the pre-training procedure for all nine models. Furthermore, Figure 15 shows the pre-training loss for the last model which is trained with data from 1996 to 2017.

Model	Steps	Loss
1996-2001	400,000	1.7752
1996-2003	405,000	1.7775
1996-2005	400,000	1.7853
1996-2007	210,000	1.8873
1996-2009	305,000	1.7977
1996-2011	410,000	1.7734
1996-2013	410,000	1.7938
1996-2015	450,000	1.8293
1996-2017	410,000	1.7844

Table 12: Number of pre-training steps and training loss of all nine FinNewsBERT models.

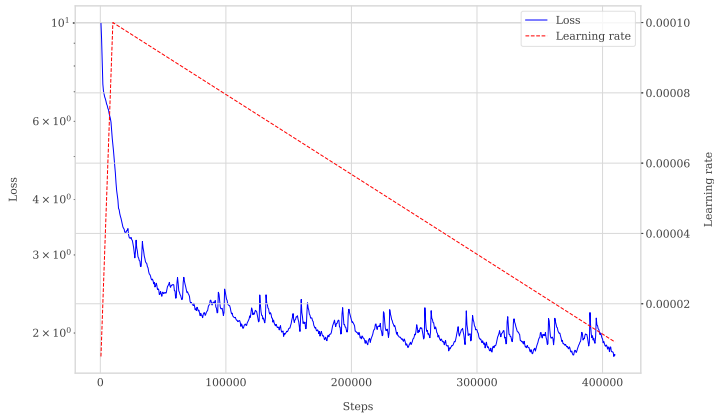


Figure 15: This Figure shows the pre-training loss of FinNewsBERT trained on data from 1996 to 2017 over 410,000 steps. After 410,000 steps the loss is 1.7844. In addition, the linear learning rate schedule is also displayed. The warm-up period is set to 10,000 steps with a maximum learning rate of 0.0001.

A.2 Test Data Sample

Timestamp	Ticker	News	Sentiment	Freshness	Topic 1	Topic 2	Topic 3	Topic 4
2018-01-04 15:45:02.187	CAT	buzz boeing caterpillar drive dow move to nyse...	0	fresh	0.110345	0.0	0.0	0.0
2018-01-04 15:45:02.921	CVS	buzz cvs health corp top gainer on on bln expe...	2	fresh	0.110345	0.163793	0.0	0.0
2018-01-04 15:45:04.421	HPQ	brief hp recalls batteries for notebook comput...	0	fresh	0.000000	0.0	0.0	0.0
2018-01-04 16:09:38.540	RRC	buzz range resources worst performer within bo...	1	fresh	0.334483	0.0	0.0	0.0
2018-01-04 16:09:53.180	HOLX	brief hologic says holders of pct convertible ...	0	fresh	0.000000	0.0	0.0	0.0
2018-01-04 16:29:37.894	DLTR	brief dollar tree says settled lawsuit relatin...	0	fresh	0.000000	0.172414	0.0	0.0

Table 13: This table shows an excerpt from the test data with annotated financial news and topic scores. The sentiment is denoted as follows: positive: 2, negative: 1, neutral: 0

As described in Section 2.4.4, news articles are merged on a daily basis. When multiple articles are merged, all headings are merged first at the beginning, then they are joined with the main texts.

"buzz boeing caterpillar drive dow move to nyse order imbalance shares on buy side dow jones industrial average breaks above level for first time ever on thurs boeing and caterpillar biggest contributors to the price weighted dow since the blue chip index closed above for first time on nov both industrial stocks have added more than points apiece over that time dow only required about points to eclipse because dji closed at on nov after huge move up on that day other main contributors to include goldman sachs united technologies home depot and chevron four of dow components have lost ground over that time travelers cos unitedhealth and intel nyse order imbalance shares on buy side"

"buzz cvs health corp top gainer on on bln expected cash flow boost nyse order imbalance shares on sell side shares of no drug store chain rose as much pct to touch month high at last up pct stock top gainer on and third biggest gainer on co expects tax overhaul to boost its cash flow by bln updates adj eps forecast to lower end of previous outlook and suspends share buyback to fund acquisition of aetna which is expected to close in end rivals walgreen down pct on weak retail sales and gross margin rite aid down pct after reporting smaller than expected revenue on wednesday in cvs down pct wba pct and rad pct nyse order imbalance shares on sell side"

"brief hp recalls batteries for notebook computers mobile workstations due to fire burn hazards nyse order imbalance shares on sell side jan consumer product safety commission hp recalls batteries for notebook computers and mobile workstations due to fire and burn hazards says recall involves about lithium ion batteries for hp notebook computers and mobile workstations says hp will provide free battery replacement services by an authorized technician says regarding recall hp received reports of battery packs overheating melting or charring including reports of property damage totaling source text for eikon nyse order imbalance shares on sell side"

"buzz range resources worst performer within bofa large cap coverage bofa merrill lynch says that within its large cap coverage range resources worst performer in due to factors such as lowered long term growth outlook weaker than expected natural gas prices cuts oil and gas producer rating to neutral from buy says company plans to generate free cash flow in sell some assets and do possible partial sale or jv of its southwest pennsylvania assets there may be some skepticism over

the extent it can accomplish this the company anadarko assets are northeast of the majority of activity in the stack scoop plays which may make it more difficult to find potential buyer bofa bofa says rrc lycoming marcellus assets positioned in less prospective of the play says co to receive value of roughly mln for the assets rrc shares closed down pct at of brokerages rate the stock buy or higher hold their median pt is drop of in months rrc lost pct of its value in while the energy index fell pct"

A.3 Determining the Freshness of News Articles

To determine the freshness of news, I compare the similarity of each news article with all articles published in the previous three days. As a measure of similarity I use a combination of the Jaccard similarity coefficient (Jaccard, 1912) and the cosine similarity between document embeddings (CLS tokens) generated by FinNewsBERT. Jaccard similarity measures the similarity of two sets of words (news articles A and B) by calculating the intersection between the two sets normalised by their union. This however has the disadvantage that news with similar meaning but different word usage can have a low Jaccard similarity coefficient. To counteract this, I also calculate cosine similarity between news articles, since document embeddings with similar meanings, have similar vector representations and thus a large cosine similarity. Only if the product of Jaccard similarity and cosine similarity between two articles is larger then a certain threshold, news articles are considered as similar and marked as stale.

A.4 Regression of the Realized Returns on the Predicted Sentiment

Figure 16 shows the realized returns of the out-of-sample test period from 2002 to 2020 relative to the predicted sentiment of both, fresh and stale news. The predicted sentiment value is defined as the difference between the probability of the positive class and the probability of the negative class. Thus, the predicted sentiment for an article classified as $P_{pos.} = 0.90$, $P_{neg.} = 0.01$ and $P_{neutral} = 0.09$ is $P_{pos.} - P_{neg.} = 0.90 - 0.01 = 0.89$. The

confidence level is set to 0.95 so that only news with abs(predicted sentiment) >= 0.95 is considered for the linear regression. Also, only news published in the 17.5-hour window prior to market opening on day t is taken into account. Plot (a) in Figure 16 shows the market open-to-open returns realised between $t-1$ and t. This includes a look ahead bias and is therefore not implementable in practice. However, it clearly shows the ability of the model to detect news articles that are associated with abnormal asset returns. Plot (b) displays the market open-to-open returns, realized from day t to $t+1$. The regression of returns on predicted sentiment results in an intercept of 3.747 bps/day and a slope of 9.529 basis points/day (see Table 14). Assuming an average predicted sentiment of 0.975 gives an theoretical annual return $r_{theor.}$ of 38.87% for the long-only portfolio with daily compounding (see Equation (14)).

$$r_{theor.} = (1 + (intercept + 0.975 * slope))^{252} - 1 \tag{14}$$

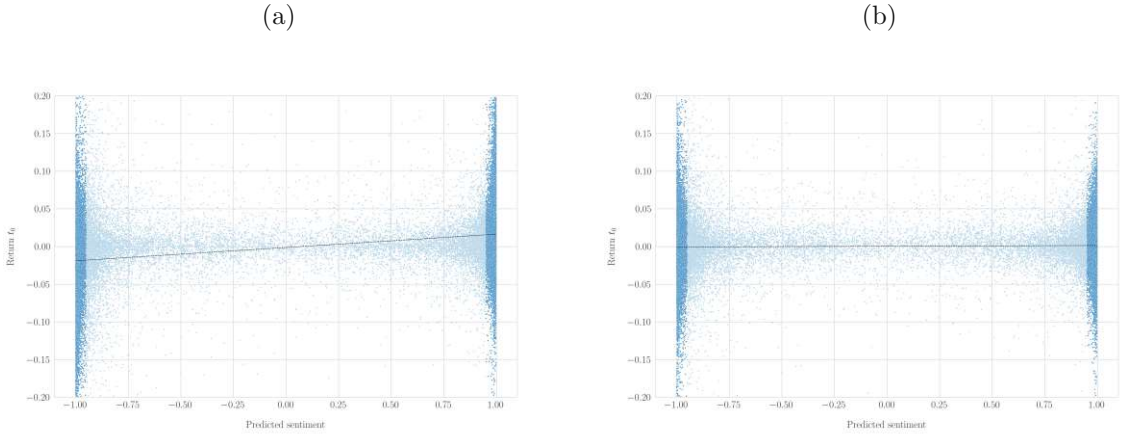


Figure 16: This figure shows the realized stock returns of the out-of-sample test period from 2002 to 2020 relative to the predicted sentiment. Plot (a) is subject to a look-ahead bias as it shows the realised returns for the same time period in which the news articles are published (day $t-1$ to t). Plot (b) shows the realized returns for the subsequent period from day t to $t+1$.

	Intercept (in bps)	Slope (in bps)	R^2
Return ($t-1$ to t)	-12.974***	176.706***	9.474e-02
Return (t to $t+1$)	3.747*	9.529***	5.784e-04

Table 14: Regression of the realized returns on the predicted sentiment. *, *** corresponds to a significance level of 5% and 0.1%.

A.5 Text2Topic

analyst forecast	**earnings report**	**Fed/Monetary Policy**	**Business/Strategic**
raise	eps	fed	business
cut	earnings	federal	strategy
buy	report	reserve	strategies
sell	reported	economy	management
hold	financial	unemployment	launch
upgrade	results	jobs	product
downgrade	quarter	inflation	operation
upgraded	annual	stimulus	service
downgraded	qtr	monetary	ceo
outperform	year	policy	announce
underperform	million	chairman	customer
analyst	ended	central	merging
analysts	operating	gdp	
estimate	net		
expect	income		

Table 15: This table shows the Text2Topic topics together with their word support.

Topic Scores

Figure 17 and 18 show the exposures of the topics "analyst forecast" and "earnings report" in the news over the period 1996 to 2020. Figure 18 shows regular spikes due to the quarterly release of earnings reports.

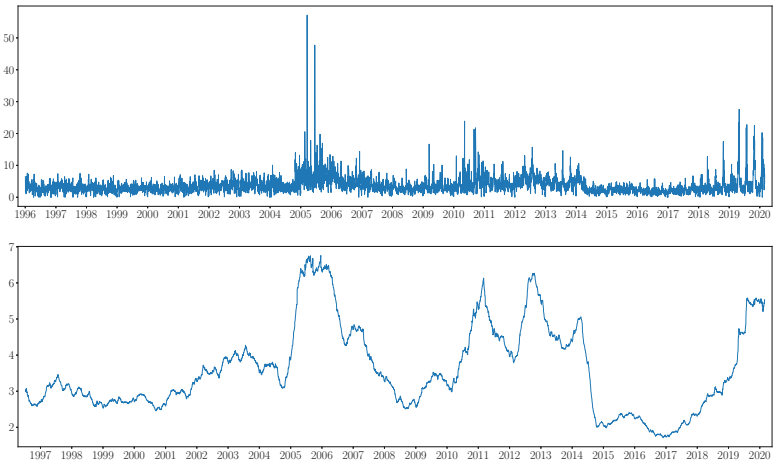


Figure 17: This figure shows the news exposure of the topic "analyst forecast" at the top and the six-month moving average at the bottom.

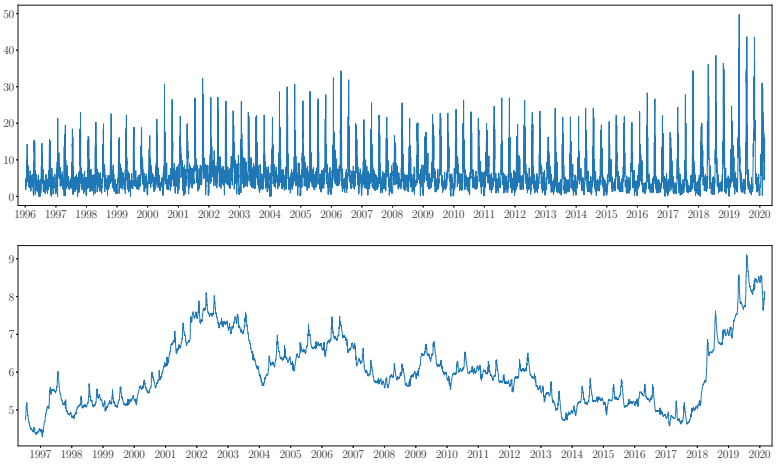


Figure 18: This figure shows the news exposure of the topic "earnings report" at the top and the six-month moving average at the bottom.

3 Overnight Reversal and the Asymmetric Reaction to News

Overnight Reversal and the Asymmetric Reaction to News

Thomas Dangl Stefan Salbrechter

December 20, 2022

News released overnight has a significant directional impact on individual shares' opening prices, i.e., the market tends to open higher (lower) when news with positive (negative) sentiment is published. However, the market opening is not fully efficient due to over- or underreactions of market participants to the news, resulting in a predictable pattern of returns on the following trading day. In particular, we find that large daytime returns followed by overnight news with strong sentiment lead to a predictable return reversal during the subsequent trading day. This predictable reversal is present independent of the polarity of the news sentiment. Without overnight news, large previous-day returns only have marginal predictive power.

3.1 Introduction

We show that overnight news, i.e., news released during times when stock markets are closed, has a clear directional impact on the opening share price at the subsequent trading day. Overnight news with positive sentiment predicts a high opening price and overnight news with negative sentiment predicts a low opening price. This opening price is, however, not fully efficient in the sense that returns during the subsequent trading day are predictable. When there is no company relevant overnight news, predictability can hardly be detected.

Idiosyncratic returns of S&P 500 constituents reveal that investors do not simply over- or underreact to overnight news but that inefficiency is asymmetric. News releases after market close, which confirm the previous day's open-to-close return, i.e., good news after a positive open-to-close return or bad news after a negative open-to-close return, tend to come with an overshooting opening price on the next day. This overshooting reverses over the trading day in a predictable way. When news after market close opposes the previous day's open-to-close return, i.e., bad news after a positive open-to-close return or good news after a negative open-to-close return, the opening price does not fully reflect the new information. The return on the next trading day tends to make up this deficit and so it extends the direction of the overnight news release, which, again, results in a reversal relative to previous day's open-to-close return. In the absence of company-relevant overnight news, the opening price is –on average– efficient and we do not detect *exploitable* predictability of returns on the following trading day.

We determine news sentiment with a BERT-based language model, (see Salbrechter, 2021), which we train, strictly out-of-sample, on a dataset of 4 million financial news articles released between 1996 and 2020.[1][2]

Our contribution to the literature is threefold. First, as described above, we document the impact of overnight news sentiment on the market opening price and report a mispricing that leads to a predictable return reversal. The reaction to news sentiment is

[1] This corresponds to a total of 466 million words.

[2] We thank Refinitiv for providing us with the dataset.

apparently asymmetric, depending on the direction of the previous day's return in relation to news sentiment. Extending Boudoukh et al. (2019), who disregard news sentiment and find that the presence of overnight news increases the variance of overnight returns, we are able to measure the directional impact of news on the opening price and predict *occurrence*, *direction*, and *magnitude* of inefficient market opening that translates into a subsequent return reversal. In the absence of company specific overnight news, subsequent-day returns can hardly be predicted.

Second, we re-investigate the attention effect reported by Barber and Odean (2008) and Berkman et al. (2012), stating that increased investor attention (indicated by large absolute previous-day returns) leads to elevated prices at market opening followed by a reversal during the trading day. We detect this effect unconditional on news releases, however, in the absence of overnight news this effect is magnitudes smaller than the impact of news sentiment on asset prices. While the attention effect predicts that large previous-day returns lead to positive overnight returns, we report a dominant news-driven effect, suggesting a strong interaction between news sentiment and asset returns.

Third, we present a simple trading strategy that exploits overnight reversal by taking a long (short) position in the opening auction of stocks that experienced exceptionally negative (positive) idiosyncratic returns on the previous trading day only if we observe overnight news. Before transaction costs, the long/short strategy generates an average return per trade of 27.79 bps.

The remainder of this paper is composed as follows. In Section 3.2 we review the related literature, in Section 3.3 we describe the data sources and the performed data pre-processing steps and in Section 3.4 we briefly explain our BERT-based language model. In Section 3.5.1 we document the directional impact that overnight news has on (contemporaneous) overnight returns and in Section 3.5.2 we describe the return reversal observed in the subsequent daytime period. We run linear regressions in Section 3.5.3, in Section 3.5.4 we study the impact of news released during market opening hours and in Section 3.5.5 we investigate the influence of news topics. We present an out-of-sample backtest in Section 3.5.6 and in Section 3.6 we conclude.

3.2 Literature Review

Does financial news move stock prices? Boudoukh et al. (2019) study this research question and find that financial news has a significant and measurable impact on stock prices. They focus on detecting relevant news articles, i.e., news that move stock prices, from a stream of Dow Jones Newswire articles by utilizing state-of-the-art text analysis tools. The authors identify firm-relevant news articles and measure their impact on overnight and daytime returns by analyzing return volatilities. What they find is a strong (contemporaneous) link between the release of relevant news articles and elevated return volatility. The idiosyncratic variance thereby explained by public information accounts for approximately 49.6% (12.4%) of the total overnight (daytime) variance. They, however, do not determine news sentiment and therefore cannot make conclusions about the directional impact that these news has on overnight and daytime returns.

Beside Boudoukh et al. (2019), Greene and Watts (1996), Moshirian et al. (2012) and Jiang et al. (2012) also study the impact of overnight news on asset prices and volatility. The fact, that 95% of all earnings are reported outside of the regular U.S. trading hours makes the focus on overnight news even more relevant (Jiang et al., 2012; Michaely et al., 2014). Greene and Watts (1996) studies the impact of earnings announcements, released during trading and non-trading hours, on the NYSE and the NASDAQ exchange. In order to measure abnormal returns after earnings announcements, the authors implement a trading strategy where they go long (short) in assets that beat (miss) the analysts forecasts in earnings per share. They find that the opening price contains most of the price response. Since investors have more time to evaluate the news when the market is closed, the opening is usually more informative in comparison to the price response when earnings are announced during market opening hours. Moshirian et al. (2012) arrives at similar conclusions by studying the impact of overnight corporate announcements on the opening price for Australian Securities Exchange (ASX) listed stocks. The authors find that information asymmetry is reduced when overnight news is published, leading to a more efficient determination of the equilibrium price, while stock prices adjust quickly to released overnight announcements. The price response takes place to a great extent

during the pre-opening period and the first fifteen minutes after market opening. Jiang et al. (2012) examines price reactions triggered by earnings announcements that are published during non-trading hours. The authors document that during after-hours trading, which is primarily performed by institutional investors, price reaction shows a high degree of informational efficiency. In this study we confirm the large impact that information released overnight has on individual shares opening price. In addition, however, we also find a predictable pattern that realizes during the subsequent trading day. Thus, we argue that the opening price is not perfectly efficient but is exposed to investor over- and underreactions. Berkman et al. (2012) states that investors are more likely buyers of stocks that attract their attention. As a proxy for investor attention the authors consider the squared previous day returns. They argue that at days of high investor attention, retail investors tend to herd into stocks at market opening leading to elevated overnight returns, which in turn leads to a reversal during the subsequent trading day. Thus, the authors document an inefficient market opening, an overreaction, that reverses during the subsequent trading session. In this study, we confirm the existence of this attention effect. In addition, we show that the inefficiency caused by investor attention is much smaller than the inefficiency caused by the interplay of previous day returns and overnight news sentiment.

Other related research papers that examine the impact of financial news on the stock market using sentiment analysis include Tetlock (2007); Antweiler and Frank (2004); Loughran and McDonald (2011); Bollen and Mao (2011); Uhl et al. (2015); Kelly et al. (2019), among others.

3.3 Data and Data Preprocessing

We consider a total of 1122 constituents that are listed in the S&P 500 from 1996 to 2020. Also, we use daily open and close prices as well as payout- and split-adjusted close prices from Refinitiv Datastream. From adjusted close prices we calculate corresponding same-day adjusted open prices for each asset i as

$$p_{i,t}^{adj.open} = \frac{p_{i,t}^{open}}{p_{i,t}^{close}} \times p_{i,t}^{adj.close}, \tag{3.1}$$

where $p_{i,t}^{open}$ and $p_{i,t}^{close}$ are the open and the close price of asset i on day t, respectively, and $p_{i,t}^{adj.close}$ is the payout- and split-adjusted Datastream close price on that day. We calculate simple close-to-close returns $r_{i,t}$ from adjusted prices (total returns) and dissect them into overnight returns $r_{i,t}^{c}$, when markets are closed ($\text{close}_{(t-1)}$-to-$\text{open}_{(t)}$ total returns), and returns during market activity $r_{i,t}^{o}$ ($\text{open}_{(t)}$-to-$\text{close}_{(t)}$ total returns), as shown in Figure 3.1. Then we calculate the idiosyncratic return components relative to the market model, where βs are calculated from weekly returns over a rolling window of two years. The further analysis of this paper is fully based on idiosyncratic returns. When we use $t-1$ open-to-close returns as a predictive variable, we use z-scores $z_{i,(t-1)}$, calculated by dividing idiosyncratic open-to-close returns $r_{i,(t-1)}^{o}$ by the daily return volatility, estimated over a rolling window of 6 month.

Figure 3.1: Market opening and market closing hours.

The financial news data is provided by Refinitiv (formerly Thomson Reuters). This comprehensive dataset contains news published between January 1996 to February 2020.[3] Each news article is tagged with metadata containing ticker codes of the companies mentioned in the news. After matching news to the set of S&P 500 companies, we find 812 unique ticker codes in the news metadata. As we focus on company-specific news articles, we restrict our data to news articles that contain either a company name or a ticker code in the headline. The content of these articles is usually more relevant to the targeted

[3]The dataset contains more than 40 million news items with exact timestamp of publication and complete tracking of update histories. We thank Thomson Reuters for providing the dataset.

company than other, more general news. Then we feed each news article into a data cleaning pipeline, where the text is converted to lowercase and cleaned by removing all numbers, punctuation marks and brackets, so that only letters remain. In addition, irrelevant data such as the author's contact information, e.g. email addresses, phone numbers and hyperlinks, are also removed.

The release of Thomson Reuters financial news often occurs over several stages. First, a news alert is published which is followed by a news-break 5 to 20 minutes later. This is comprised of a headline and a short text. Another 20 to 30 minutes later, a news update is published with additional information. Further updates may be released successively as the story develops. In some cases, updates are released even days after the original news event. Consequently, using only the last updated status of a news article does not meet our need for a proper timing of the release of information. Our objective therefore is to use those versions of news articles that appear as early as possible and contain as much information as possible. As with the return data, we distinguish between news released during stock market closing hours and news released during stock market opening hours. If a news article is published during stock market closing (opening) hours and is then followed by several updates into the next trading session, we only consider the last update published before the market open (close). For the case when multiple news articles about the same company are published either during the stock market closing or opening hours, we combine them into a single news document. By doing so, we arrive at a total of 164,523 overnight and 109,952 daytime company-related news documents used in this study. An excerpt of news articles including their predicted sentiment is shown in Appendix B.6.

3.4 Determining News Sentiment

Language models made a big leap forward with the publication of the transformer model (Vaswani et al., 2017) and the idea of transfer learning popularized with the publication of the BERT model (Devlin et al., 2018). Since then, language models grew steadily in size and are trained with increasing amounts of textual data. However, a model trained with all today's available data would introduce a look-ahead bias if applied in an historical

context. A model would likely classify news articles dealing with a new virus variant discovered in China differently when also trained with text data containing all news stories and studies published during the recent pandemic. To rule out a look-ahead bias caused by the training data of the model, we train our own language model exclusively with historical news data. In order to be able to conduct an out-of-sample study over a period of 18 years, we retrain our model with new data every 2 years. The model we implement is a down-sized BERT-like model with a total of 18.95 million parameters.[4] This model is pre-trained exclusively on domain-specific Thomson Reuters financial news. Fine-tuning on the sentiment prediction task is performed on an annotated dataset with annotations automatically generated from the joint behavior of news stories and asset returns (z-scores) as described in Salbrechter (2021).

3.5 Results

3.5.1 Overnight News and Overnight Returns

In this section we study the impact of overnight news on overnight returns thereby conditioning on news sentiment. Boudoukh et al. (2019) document that the release of overnight news is associated with a significant increase in contemporaneous overnight return variance. They do not consider news sentiment, hence, they study the non-directional effect of the sheer presence of overnight news. We classify news sentiment as positive, neutral, or negative and, thus, are able to identify the directional effect of overnight news on the overnight return. The market tends to open significantly higher (lower) if positive (negative) news is released. Figure 3.2 plots the idiosyncratic mean returns of stocks with positive (red) and negative (blue) overnight news together with the mean overnight return of firms for which no news is released (solid line) during the hours where the stock market is closed. Mean overnight returns are calculated over all observations with z-values of previous-day returns exceeding (falling below) the threshold values indicated on the

[4]For a detailed description of the model architecture, hyperparameter choices and training settings see Salbrechter (2021).

abscissa of the plot. Specifically, negative news comes with an average overnight return of -101.42 bps (t-value = -41.39) while positive news come with an average return of 96.42 bps (t-value = 54.27) (see Table 3.1, Panel B and C).

Please note that the reported large predictive power of news sentiment on (contemporaneous) overnight returns can hardly be exploited by investors, since release occurs when the stock market is closed. The opening price (the overnight return) summarizes the aggregate reaction of investors to the news content and is, thus, a clear measure of how investors interpret the news. Sentiment classification is trained strictly out-of-sample. Consequently, we conclude that the sentiment assigned by the BERT-based model is a good representation of investors' news perception.

Results also indicate that previous-day returns are only weak predictors of overnight returns. Inspecting overnight returns of stocks without overnight news reveals the presence of the attention effect as reported by Berkman et al. (2012). I.e., unconditional on sentiment, large negative as well as large positive previous-day returns (allegedly creating attention) are followed by slightly positive overnight returns (investors tend to buy attention-grabbing stocks) in the range of 3.54 - 4.32 bps.[5] The directional effect of news sentiment is, however, many times larger and certainly dominates return figures. In particular, if attention is created by negative-sentiment overnight news, the opening price is not higher (as predicted by the attention effect) but significantly below the unconditional mean.

Figure 3.2 also reveals that the market reaction to positive as well as to negative sentiment news is influenced by previous-day returns. After days with extreme returns (z-values), the reaction to news sentiment seems to be mitigated. In Section 3.5.2 we identify the interplay of the z-value of previous-day returns and overnight news sentiment as a predictor of return reversal.

[5]See Table 3.1, Panel A (abs(z-value) $\geq$ 1.5).

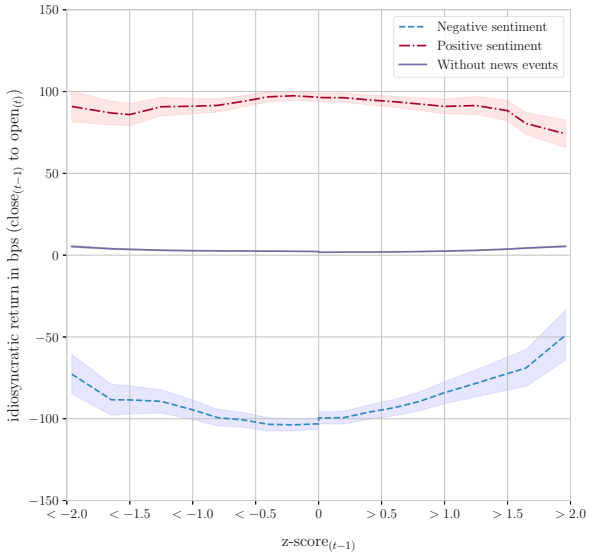


Figure 3.2: This plot shows the idiosyncratic mean returns in bps, measured from close$_{(t-1)}$ to open$_{(t)}$, conditional on both, the z-score at market close$_{(t-1)}$ and the predicted overnight news sentiment. The shaded area highlights the standard error of the returns. We consider all assets that are related to the published financial news articles and report the results over the time period from 2002 to 2020. For the correct interpretation of the results, it is important to note that we always consider all z-scores that are greater than > 0.0, > 0.5, > 1.0 etc. or smaller than < 0.0, < -0.5, < -1.0 etc.

z-score	$\leq$ -1.645	$\leq$ -1.5	$\leq$ -1.0	$\leq$ -0.2	$\geq$ 0.2	$\geq$ 1.0	$\geq$ 1.5	$\geq$ 1.645	(-∞, ∞)
Panel A: Without news events									
Mean	3.89***	3.54***	2.72***	2.36**	1.87*	2.49**	3.7***	4.32***	1.98**
SD	147.12	141.49	118.0	97.97	95.51	111.43	127.07	130.36	94.31
Std Err	0.49	0.42	0.23	0.11	0.11	0.22	0.37	0.43	0.07
t-value	7.87	8.43	11.75	21.64	17.33	11.34	9.95	10.09	29.39
Support	88,811	113,220	258,936	807,094	782,660	258,522	116,786	92,768	1,958,013
Panel B: Negative sentiment									
Mean	-88.42***	-88.47***	-94.57***	-103.75***	-99.35***	-84.07***	-72.36***	-69.05***	-101.42***
SD	476.29	465.51	431.71	410.12	409.48	438.1	479.03	493.18	407.63
Std Err	10.85	9.66	6.27	3.52	3.32	5.66	8.35	9.46	2.45
t-value	-9.32	-10.39	-15.95	-28.24	-25.3	-12.54	-7.19	-6.1	-41.39
Support	2,520	2,989	5,301	12,463	10,874	4,273	2,267	1,898	27,672
Panel C: Positive sentiment									
Mean	86.89***	85.94***	91.04***	97.45***	96.25***	90.92***	88.34***	80.56***	96.42***
SD	319.95	326.03	325.46	319.9	332.84	344.09	351.52	343.12	322.90
Std Err	7.26	6.73	4.68	2.78	2.78	4.5	6.26	6.67	1.78
t-value	11.97	12.77	19.45	35.01	34.65	20.2	14.1	12.08	54.27
Support	1,942	2,348	4,832	13,206	14,361	5,843	3,150	2,646	33,030

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3.1: Descriptive statistics of overnight returns measured from close$_{(t-1)}$ to open$_{(t)}$ (in bps) as displayed in Figure 3.2.

3.5.2 Overnight Reversal and the Asymmetric Reaction to News Sentiment

In this section, we discuss the impact of overnight news sentiment on subsequent daytime returns. Any predictability of subsequent daytime returns conditional on the overnight-news sentiment indicates that market open prices are not fully efficient after news is released. In contrast to the results documented in the previous section (contemporaneous overnight news and overnight returns), predictive information of overnight news can be exploited by entering a position in the opening auction.

We find a predictable reversal relative to the previous-day return which realizes during the trading day. This reversal is more pronounced the more extreme the previous-day return is. But this reversal is only present if overnight news is released. When extreme returns are followed by a night without company relevant news release, predictability is only marginal.

Figure 3.3 (a) and (b) illustrate the idiosyncratic returns conditional on $z_{i,(t-1)}$ and the overnight-news sentiment. Note, that in contrast to Figure 3.2 we aggregate the previous day's z-score $z_{i,(t-1)}$ into three subsamples according to it's value using the intervals (-∞, -0.5], (-0.5, 0.5) and [0.5, ∞). This allows for easier interpretation as well as the use of statistical tests of the difference in mean and median returns in these buckets. Figure 3.3 (a) shows as in the previous section that overnight news sentiment strongly influences contemporaneous overnight returns, while the predictive power of $z_{i,(t-1)}$ is only weak. Figure 3.3 (b) shows the joint impact of $z_{i,(t-1)}$ and overnight news sentiment on the return of the subsequent trading day, $r^o_{i,t}$, i.e., the predictable reversal.

When the previous-day return is positive with $z_{i,(t-1)}$ above 0.5 and overnight news with strong sentiment–positive or negative–is released, we find a predictable reversal during the subsequent trading session, $r^o_{i,t} < 0$. Mean returns are -10.13 basis points when news sentiment is positive and -19.86 basis points when news sentiment is negative (with t-values of -3.98 and -5.03 respectively, see Table 3.2, Panel B and C). For negative previous-day returns with $z_{i,(t-1)}$ below -0.5 we detect again a predictable return reversal in the subsequent trading session, $r^o_{i,t} > 0$. Mean returns are 13.05 bps after news with positive

sentiment and 10.87 bps after news with negative sentiment (with t-values of 4.68 and 2.56 respectively).

Without news release, if we solely condition on $z_{i,(t-1)}$, we observe an average open$_{(t)}$-to-close$_{(t)}$ return, $r^{o}_{i,(t)}$, of -4.03 bps if $z_{i,(t-1)}$ is above 0.5 and 0.67 bps if the z-score is below -0.5 (with t-values of -16.67 and 2.83 respectively, see Table 3.2, Panel A). Thus, in contrast to Berkman et al. (2012), who uses the squared return as a proxy for investor attention, we observe that the trading day reversal only exists after positive returns (z-scores). If the previous day's return is negative, the market tends to open slightly positive on average, but without a subsequent reversal.

Furthermore, if previous-day returns are comparably small (absolute z-values less than 0.5), we do not detect a significant reversal on the next trading day.

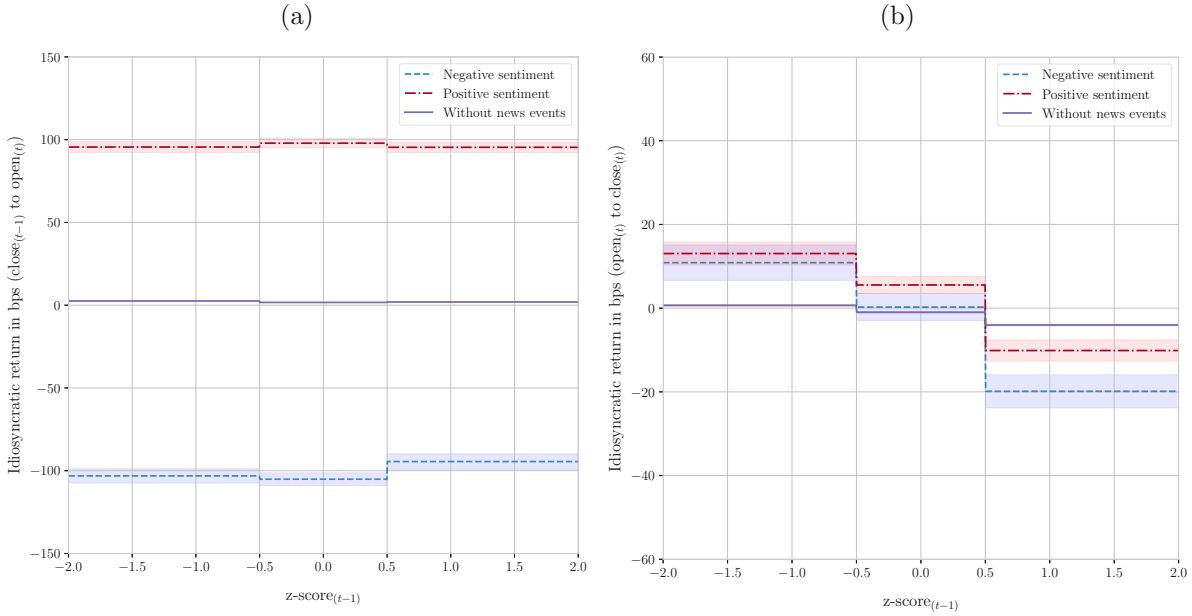


Figure 3.3: This plot shows the idiosyncratic mean returns in bps and the standard error, conditional on both, the z-scores measured at market close$_{(t-1)}$ and the predicted overnight news sentiment. We consider all assets that are related to the published financial news articles and report the results over the time period from 2002 to 2020. Figure (a) shows the idiosyncratic returns measured from close$_{(t-1)}$ to open$_{(t)}$, Figure (b) displays the idiosyncratic returns from open$_{(t)}$-to-close$_{(t)}$. The z-score subsamples we consider are the intervals: $(-\infty, -0.5], (-0.5, 0.5), [0.5, \infty)$.

	$\text{close}_{(t-1)}$ to $\text{open}_{(t)}$			$\text{open}_{(t)}$ to $\text{close}_{(t)}$		
z-score	(-∞, -0.5]	(-0.5, 0.5)	[0.5, ∞)	(-∞, -0.5]	(-0.5, 0.5)	[0.5, ∞)
Panel A: Without news events						
Mean	2.52***	1.66***	1.93***	0.67***	-0.99***	-4.03***
SD	103.55	83.92	99.97	176.05	153.09	176.65
Std Err	0.14	0.09	0.14	0.24	0.17	0.24
t-value	18.12	18.46	14.2	2.83	-5.96	-16.67
Support	553482	865875	538656	548654	857634	533999
Panel B: Negative sentiment						
Mean	-103.15***	-105.13***	-94.48***	10.87**	0.25	-19.86***
SD	411.92	397.24	416.09	406.84	334.46	350.54
Std Err	4.28	3.87	4.68	4.24	3.27	3.95
t-value	-24.08	-27.13	-20.2	2.56	0.08	-5.03
Support	9246	10510	7916	9196	10441	7870
Panel C: Positive sentiment						
Mean	95.54***	97.89***	95.39***	13.05***	5.54***	-10.13***
SD	318.06	324.97	324.62	269.14	232.26	260.42
Std Err	3.28	2.84	3.16	2.79	2.04	2.54
t-value	29.1	34.46	30.19	4.68	2.72	-3.98
Support	9387	13088	10555	9316	12988	10485

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3.2: Descriptive statistics of overnight and daytime returns (in bps) as displayed in Figure 3.3.

In each of the subsamples formed on $z_{i,(t-1)}$, $(-\infty, -0.5], (-0.5, 0.5), [0.5, \infty)$, we group time t daytime returns into three groups: Observations that come after positive overnight news, observations after negative returns, and observations without related company-specific news.

As a robustness check and as an alternative test to the simple t-tests provided in Table 3.2, we use a Kruskal-Wallis test for differences in median returns. For each of the $z_{i,(t-1)}$ intervals, we test for differences in medians of $r^o_{i,(t)}$ conditional on the news sentiment (positive, negative, no-news) and find a significant variation in bucket medians.[6] We complement the Kruskal-Wallis test with a post-hoc Dunn's test, which tests for pairwise differences in medians. The result of the Dunn's test confirms predictability as diagnosed by t-tests, the corresponding p-values are shown in Table 3.3. When comparing the returns following positive or negative overnight news to the no news case, we observe that returns significantly deviate from the no news case when $z_{i,(t-1)}$ is large ($\text{abs}(z_{i,(t-1)}) \geq 0.5$). Moreover, the test shows that median returns after positive and negative news differ only when

[6]The p-values of the Kruskal-Wallis test for the three buckets are: $p - value_{(-\infty,-0.5]} = 1.20e-15$, $p - value_{(-0.5,0.5)} = 0.019$, $p - value_{[0.5,\infty)} = 2.12e-15$.

$z_{i,(t-1)}$ is positive. In the other subsamples, the difference is small, indicating that the actual news sentiment has only minor influence on daytime returns following overnight news.[7] If the previous day's z-score is small ($\text{abs}(z_{i,(t-1)}) < 0.5$), returns differ from the no news case only if positive news is released. In the case of negative news, the returns are not different from the case without news, indicating an efficient market in this case.

Z-score	(-∞, -0.5]		(-0.5, 0.5)		[0.5, ∞)	
Sentiment	Negative	Positive	Negative	Positive	Negative	Positive
No News	0.041233**	0.000126***	0.885724	0.016725**	3.290547e-11***	0.001908***
Positive	0.209864		0.097544*		0.002795***	
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 3.3: P-values of the Dunn's test for the cases of negative overnight news sentiment, positive overnight news sentiment and no overnight news for the three subsamples. As an adjustment for the p-values we use "holm", a step-down method using Bonferroni adjustments. The Dunn's test is performed over the full period from January 2002 to February 2020. A subperiod Dunn's test is provided in Appendix B.1.

At this point we want to provide a possible explanation for the underlying effects leading to the observed return reversal. Inefficiencies at market opening are likely caused by one or a combination of several behavioral biases, including confirmation bias (Nickerson, 1998), attribution bias (Daniel et al., 1998), and availability bias (Kahneman et al., 1982). People influenced by the confirmation bias tend to search for evidence that supports their prior beliefs while neglecting contradicting information. The attribution bias (biased self-attribution) let investors become overconfident when public information confirms their prior private views. If, on the other hand, new information contradicts their private views, they tend to underweight this news. Due to the availability bias, people tend to overweight recent and salient information which leads to the attention grabbing effect documented by Barber and Odean (2008).

In order to interpret the observed effects, consider the four cases: (1) positive z-score & positive news, (2) positive z-score & negative news, (3) negative z-score & positive news and (4) negative z-score & negative news. For case (1) we observe an overreaction to positive news at market open which then reverses during the trading day. This overreaction

[7] In contrast, for analyst forecast news, we observe a stronger impact of overnight news sentiment on daytime returns as shown in Section 3.5.5.

may be caused by investors seeking for confirmation of their prior believes (confirmation bias), or by confirmation of their private signals (attribution bias) by financial news. As $z_{i,(t-1)}$ is positive, we assume that their aggregate believes and private signals are also positive. If their prior believes or their private signals are confirmed by positive overnight news, they are more likely buyers at market open. Also, a combination of large previous day returns and salient overnight news increases investor attention which in turn also attracts more buyers (availability bias). This combined elevated buying pressure tends to results in an overreaction at market open which then tends to reverse during the following trading day. In case (2), however, investors tend to neglect (confirmation bias) or underweight (attribution bias) this contradicting information. This likely causes an underreaction at market opening. In addition, the availability bias induces buying pressure due to the salient information. This slightly elevates the opening price which in turn adds up to a strong reversal during the trading day.

In cases (3) and (4) $z_{i,(t-1)}$ is negative. Thus, we assume that investors' aggregate believes and private signals are also negative. In case (3), investors again tend to neglect or underweight the contradicting information. This causes an underreaction (reduced buying pressure) at market opening and a subsequent reversal in the positive direction. For case (4) we observe an overreaction to negative news at market open which then reverses (in the positive direction) during the trading day. This overreaction may be again influenced by the confirmation and attribution bias. If their prior believes or their private signals are confirmed by negative overnight news, they are more likely sellers at market open. The availability bias on the other hand should again induce buying pressure due to the salient information. This slightly elevates the opening price which in turn should lead to a slight reduction of the reversal in cases (3) and (4).

Event Study

The event study shown in Figure 3.4 displays the average price responses for the cases (1) to (4). We set the initial price to 100 and observe a time window of seven days including open (o) and close (c) prices. An event is triggered by $\text{abs}(z_{i,(t-1)}) \geq 0.5$ (tIc) followed by either positive or negative news articles published during market closing hours from

close$_{(t-1)}$ to open$_{(t)}$ (to). From tIc to to a large price move in the direction of the news sentiment can be observed, as also shown in Figure 3.2. For the consecutive market open-to-close period (to to tc) declining prices can be observed in the case of positive z-scores, both for positive and negative news and increasing prices can be observed for negative z-scores, both for positive and negative news. Those results are in line with the observations in Figure 3.3.

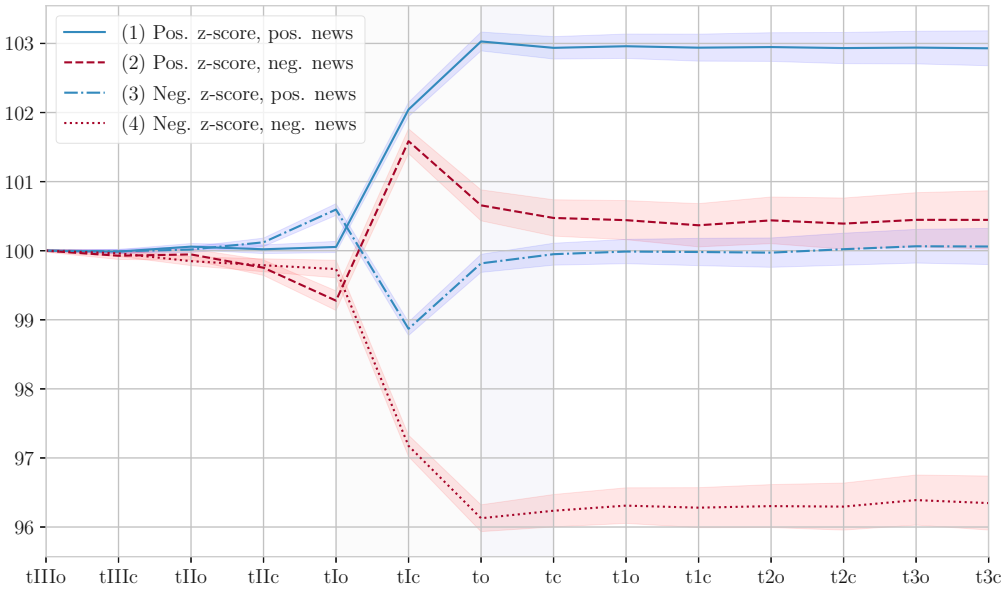


Figure 3.4: Event study showing the price response of the cases (1) to (4). Idiosyncratic daily returns are dissected into an overnight component (close-to-open) and into a daytime component (open-to-close). The returns are compounded starting with the close-to-open return three days prior to the news event $tIIIo$ until the open-to-close return $t3c$ three days after the news event. The z-scores are measured at tIc and the overnight news are released between tIc and to.

3.5.3 Regression Analysis

For the regression analysis we consider three predictive variables which are: (1) the z-score measured at market close $z_{(t-1)}$, (2) the overnight news sentiment[8] s_t and (3) the cross term $z_{(t-1)} \times abs(s_t)$. We regress (a) the idiosyncratic close$_{(t-1)}$-to-open$_{(t)}$ return

[8]Note, that the sentiment score is a numeric variable ranging from -1 for negative sentiment to 1 indicating positive sentiment.

and (b) the idiosyncratic open$_{(t)}$-to-close$_{(t)}$ return on these predictive variables. The regression results are presented in Table 3.4. Regression (a) shows that the overnight news sentiment has the greatest predictive power with a highly significant coefficient of 0.0108. Both the constant term and the cross term are significant as well, but with much smaller coefficients. The z-score of the previous close has no measurable impact on the opening price. The explained variance of this regression, R^2, is 2.19%. These results again show the large impact of overnight news on the opening price. For regression (b) we regress the idiosyncratic daytime return on the predictive variables and find that all coefficients are significant. The negative coefficient of $z_{(t-1)}$ indicates the existence of a slight reversal at daily frequencies. Furthermore, the coefficient of the overnight news sentiment is positive, while the coefficient of the cross term is negative. The significant cross term again underscores that overnight news sentiment, whether positive or negative, causes a reversal, irrelevant of the actual sentiment, due to investors over- and underreactions. This reversal is larger in magnitude for the combination of positive previous day's z-scores and negative overnight news. In regressions (c) to (e) we also control for (c) firm-fixed effects, (d) time-fixed effects and both, (e) firm and time fixed effects. We find that the results are robust after controlling for fixed effects.

	(a)	(b)	(c)	(d)	(e)
Const.	0.0002***	-0.0001***	-0.0001***	-0.0001***	-0.0001***
	(25.164)	(-10.638)	(-10.638)	(-10.645)	(-10.643)
$z_{(t-1)}$	-0.0000	-0.0002***	-0.0002***	-0.0002***	-0.0002***
	(-1.017)	(-17.736)	(-17.881)	(-17.774)	(-17.921)
s_t	0.0108***	0.0003***	0.0003***	0.0003***	0.0003***
	(227.678)	(4.452)	(3.880)	(4.851)	(4.337)
$z_{(t-1)}$ x abs(s_t)	-0.0001***	-0.0005***	-0.0005***	-0.0005***	-0.0005***
	(-3.230)	(-9.055)	(-8.993)	(-9.045)	(-8.972)
Fixed effects	None	None	Firm	Date	Firm & Date
Observations	2,323,806	2,323,806	2,323,806	2,323,806	2,323,806
R^2	0.0219	0.0002	0.0002	0.0002	0.0002
F Statistic	17339.7211***	168.58***	168.82***	169.99***	170.14***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 3.4: This Table shows the results of the regression $r_t = \beta_0 + \beta_1 \times z_{(t-1)} + \beta_2 \times s_t + \beta_3 \times (z_{(t-1)} \times abs(s_t)) + \epsilon$. The dependent variables are (a) the idiosyncratic close-to-open return and (b) to (e) the idiosyncratic open-to-close return. The regression is performed over the full period from January 2002 to February 2020. In addition, regression (c) controls for firm fixed-effects, regression (c) controls for time fixed-effects (19 yearly time periods) and regression (e) controls for both time and firm fixed-effects.

Regressions over Subperiods

We perform the same regression as above over the three subperiods: 01/2002 to 12/2007, 01/2008 to 12/2009 and 01/2010 to 02/2020 (see Table 3.5). For open-to-close returns (b) we note that the regression coefficients, as well as the R^2, are highest during the 2008 to 2009 financial crisis period. Furthermore, we observe that the coefficient of the cross-term (sentiment term) is larger (smaller) for the 2010 to 2020 period compared to the 2002 to 2007 period. Overall, the significance of the regression coefficients is strong.

	2002-2007		2008-2009		2010-2020	
	(a)	(b)	(a)	(b)	(a)	(b)
Const.	0.0003***	-0.0002***	0.0005***	0.0000	0.0000***	-0.0001***
	(26.4898)	(-10.2626)	(12.7150)	(0.6545)	(6.1106)	(-9.3683)
$z_{(t-1)}$	-0.0000***	-0.0002***	-0.0000	-0.0009***	0.0000**	-0.0001***
	(-3.7730)	(-8.9381)	(-0.3096)	(-16.1152)	(2.0379)	(-5.1588)
s_t	0.0109***	0.0007***	0.0133***	-0.0013***	0.0102***	0.0004***
	(127.7910)	(5.5920)	(57.6695)	(-3.5204)	(204.1787)	(6.1616)
$z_{(t-1)}$ x abs(s_t)	0.0003***	-0.0001	-0.0004**	-0.0019***	-0.0003***	-0.0003***
	(4.9931)	(-1.0504)	(-2.4884)	(-7.3597)	(-7.5418)	(-5.8074)
Observations	757,548	757,548	256,946	256,946	1,308,327	1,308,327
R^2	0.0215	0.0002	0.0128	0.0016	0.0309	0.0001
Adjusted R^2	0.0215	0.0002	0.0128	0.0016	0.0309	0.0001
Residual Std. Error	0.0108	0.0166	0.0191	0.0295	0.0090	0.0131
F Statistic	5542.7841***	39.2749***	1109.4759***	138.4413***	13907.4771***	36.7678***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3.5: This Table shows the results of the regression $r_t = \beta_0 + \beta_1 \times z_{(t-1)} + \beta_2 \times s_t + \beta_3 \times (z_{(t-1)} \times abs(s_t)) + \epsilon$ over different periods. The dependent variables are (a) the idiosyncratic close-to-open return and (b) the idiosyncratic open-to-close return.

Figure 3.5 shows the realization of the daytime returns in these subperiods. During the 2008 to 2009 financial crisis period, we observe the largest magnitudes in daytime returns (see Figure 3.5c). These findings suggest that the intensity of the observed effects is amplified during periods of high uncertainty. For the pre-financial crisis period, we observe a reversal after positive z-scores ($0 < z_{i,(t-1)} < 1.0$) and positive news (case 1)[9] (see Figure 3.5b). However, with large positive z-scores, idiosyncratic daily returns tend to be positive in this sample. The most persistent effect we observe in all three subperiods is the underreaction to news that contradicts the direction of the previous day's returns, i.e., negative news after positive z-scores and positive news after negative z-scores (cases 2 & 3). Again we perform a Dunn's test (see Appendix B.1) for the different subperiods (Panel B to C) including a test for the entire period where we exclude the financial crisis period from 01/01/2008 to 31/12/2009 (Panel A). The test shows that the reversal of case (4) is driven by the financial crisis, as the reversal after a negative z-score and negative

[9]We consider the four cases: (1) positive z-score & positive news, (2) positive z-score & negative news, (3) negative z-score & positive news and (4) negative z-score & negative news

news is only significant during this period. However, the effects of cases (1) to (3) remain persistent after excluding the financial crisis period (see Appendix B.1, Panel A).

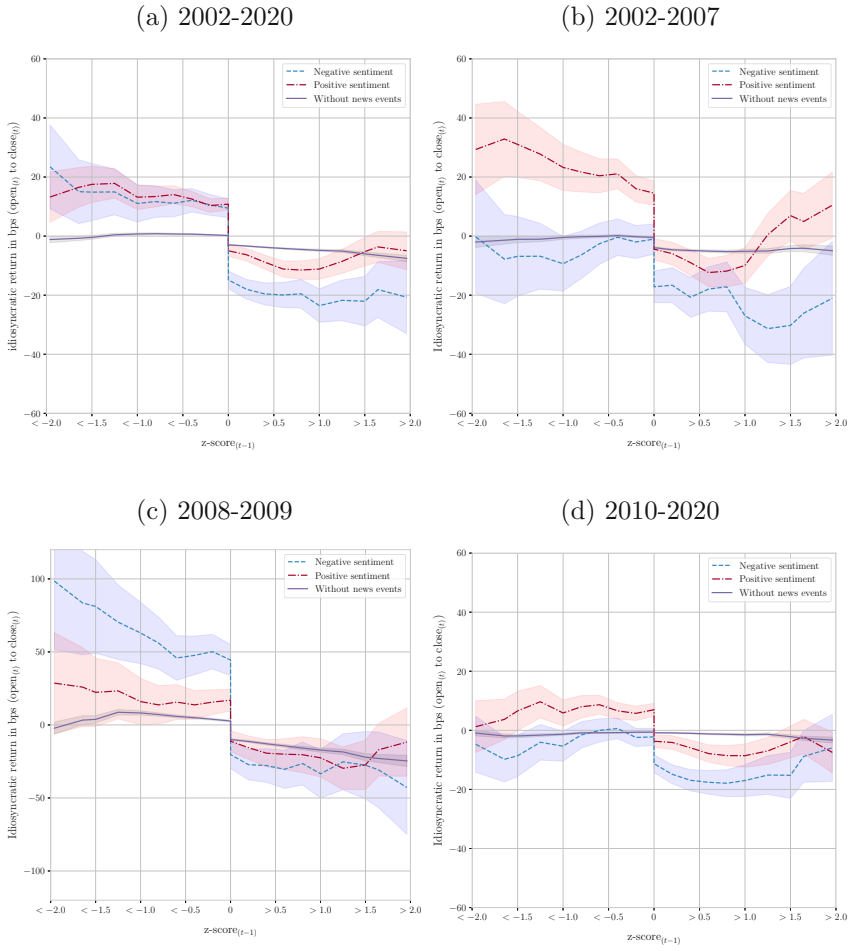


Figure 3.5: This plot shows the idiosyncratic mean returns in bps, measured from open$_{(t)}$ to close$_{(t)}$, conditional on both, the z-score at market close$_{(t-1)}$ and the predicted overnight news sentiment for different subperiods. The shaded area highlights the standard error of the returns. For the correct interpretation of the results, it is important to note that we always consider all z-scores that are greater than > 0.0, > 0.5, > 1.0 etc. or smaller than < 0.0, < -0.5, < -1.0 etc.

3.5.4 Daytime News and Overnight Returns

In this section, we examine the impact of news on asset returns for news articles published during the trading day. In particular, we investigate how well these news can explain both, idiosyncratic open$_{(t)}$-to-close$_{(t)}$ returns (daytime returns) as well as idiosyncratic close$_{(t)}$-to-open$_{(t+1)}$ returns (overnight returns). Therefore, we conduct a regression with the news sentiment as the only predictive variable and compare the results with the impact

of overnight news. The results are summarized in Table 3.6.

First, we analyze how well news sentiment can explain returns over the same period, i.e., the extent to which daytime news explains daytime returns and the extent to which nighttime news explains nighttime returns (see Table 3.6 a and b). The explained variance, R^2 is 1.709% for the daytime period and 5.925% for the overnight period. Hence, the variance explained by financial news is 3.5 times larger in the overnight period than in the daytime period. Also, the coefficient of the sentiment term is closely twice as large for the overnight period relative to the daytime period (0.01081 vs. 0.00505). We argue that the reasons for this result are twofold. On the one hand, French and Roll (1986) find that volatility is higher during the day when the market is open than during the night when the market is closed.[10] On the other hand, Jiang et al. (2012) reports that 95% of firm relevant announcements are made outside the regular trading hours. Consequently, the impact of news published outside the regular trading hours has a more predictable impact on overnight returns since the volatility is lower during the overnight period and the news released during this period more likely affect stock prices. The opposite is true for daytime news.

Second, we analyze the impact of news on the subsequent period, i.e., we measure the impact of daytime news on subsequent overnight returns and the impact of overnight news on subsequent daytime returns. We find that neither the market closing nor the market opening price is fully efficient as the news sentiment still remains a significant predictor variable (see Table 3.6 c and d). However, the coefficient of the sentiment term is 60% larger for overnight news compared to daytime news (0.00024 vs. 0.00015). Financial news released within a trading day are quickly incorporated into asset prices as Groß-Klußmann and Hautsch (2011) show. The authors examine the impact of company specific financial news on intraday trading activity using high frequency data and find a strong market response to relevant financial news within a window of 60 minutes prior and 120 minutes after the public arrival of news items. Hence, at market close news are already incorporated into asset prices to a great extent which is why the closing price has a higher degree of efficiency than the market opening price.

[10] Our data also show this pattern (see Figure 9 in Appendix B.4)

Furthermore, we analyze whether increased investor attention, due to the arrival of financial news, has an influence on returns on average. Therefore, we calculate the average return (constant term of the regression) for the same cases as above (Table 3.7 a to d), and also for the cases where no news is published, i.e., days where neither daytime nor overnight news is published (see Table 3.7 e and f). The results show that the average daytime return tends to be negative while the average overnight return tends to be positive (-1.5 bps vs. 1.6 bps).[11] In addition, we find that average daytime and overnight returns tend to be positive when news is published in the same period (Table 3.7 a and b) with mean returns of 1.20 bps and 5.80 bps respectively. Also, daytime and nighttime returns tend to be higher if news is released in the previous period (Table 3.7 c and d). The average overnight return is 113% (1.8 bps) larger if news is released in the preceding daytime period relative to the no news case (3.4 bps vs. 1.6 bps) and the average daytime return is 1.8 bps higher if news is released in the preceding overnight period (0.3 bps vs. -1.5 bps). Those results are in line with the findings of Berkman et al. (2012) who observes elevated overnight returns when investor attention in high. Our results suggest that the same is true for daytime returns when investor attention is increased due to salient overnight news.

	(a) daytime news & daytime returns	(b) overnight news & overnight returns	(c) daytime news & overnight returns	(d) overnight news & daytime returns
Const.	0.00039***	0.00022***	0.00035***	0.00003
	(5.59513)	(3.51179)	(8.71453)	(0.44073)
s_t	0.00505***	0.01081***	0.00015**	0.00024**
	(43.22846)	(100.86635)	(2.21665)	(2.34123)
Observations	107,388	161,525	105,297	160,198
R^2	0.01710	0.05926	0.00005	0.00003
Adjusted R^2	0.01709	0.05925	0.00004	0.00003
Residual Std. Error	0.02272	0.02521	0.01302	0.02366
F Statistic	1868.69975***	10174.02086***	4.91353**	5.48135**
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3.6: This Table shows the results of the regression $r_t = \beta_0 + \beta_1 \times s_t + \epsilon$. The dependent variables are the idiosyncratic open-to-close (daytime) return and the idiosyncratic close-to-open (overnight) return. The regression is performed over the full period from January 2002 to February 2020.

[11] This result is consistent with the findings of Cooper et al. (2008). The authors find that the U.S. equity premium is determined entirely by overnight returns, which tend to be positive, while daytime returns tend to be zero or slightly negative.

	(a) daytime news & daytime returns	(b) overnight news & overnight returns	(c) daytime news & overnight returns	(d) overnight news & daytime returns	(e) no news & daytime returns	(f) no news & overnight returns
Const.	0.00012*	0.00058***	0.00034***	0.00003	-0.00015***	0.00016***
	(1.74412)	(8.95070)	(8.55154)	(0.57300)	(-12.48961)	(23.65438)
Observations	107,388	161,525	105,297	160,198	1,980,897	1,849,881
R^2	0.00000	-0.00000	-0.00000	0.00000	0.00000	0.00000
Adjusted R^2	0.00000	-0.00000	-0.00000	0.00000	0.00000	0.00000
Residual Std. Error	0.02292	0.02599	0.01302	0.02366	0.01651	0.00892
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 3.7: This Table shows the results of the regression $r_t = \beta_0 + \epsilon$. The dependent variables are the idiosyncratic open-to-close (daytime) return and the idiosyncratic close-to-open (overnight) return. The regression is performed over the full period from January 2002 to February 2020.

3.5.5 Impact of the News Topic

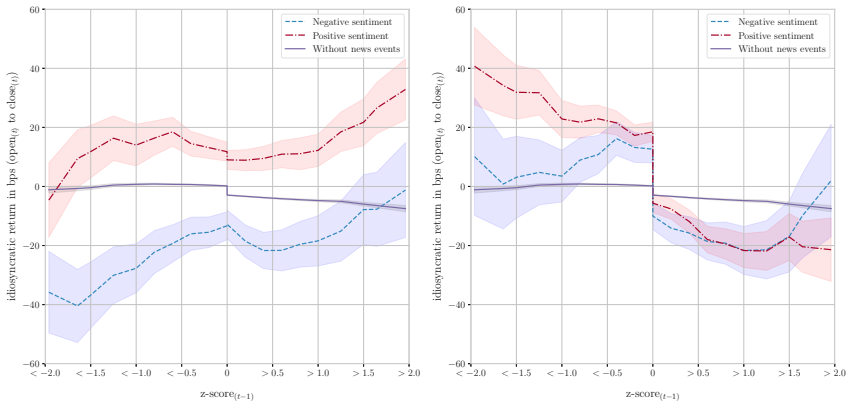
In this section, we examine how different news topics affect asset prices. Specifically, we consider the topics "analyst forecast" and "earnings report", i.e., we filter for news that are related to these topics.[12] Then we run regressions, as described in Section 3.5.3 for each subset of the news data. The results of the regressions are shown in Table 3.8. We find that the coefficient of the sentiment term is more than 4 times as large for analyst forecast news compared to general news (0.0013 vs. 0.0003). This causes quite a different pattern in open-to-close returns compared to the other subsets of news, as Figure 3.8 (a) shows. Analyst forecast news tends to drive return momentum. On average, negative news is associated with negative overnight and daytime returns, while positive news is associated with positive overnight and daytime returns. If we consider all news except analyst forecasts, see Figure 3.6 (c), we again observe a reversal which is stronger compared to the reversal observed in Section 3.5.2, since the contrarian effects attributable to analyst forecast news is eliminated. News marked as earnings reports show a strong reversal in the case of negative z-scores$_{(t-1)}$ and positive news, i.e., positive earnings surprises as Figure 3.6 (b) shows. This is also the case for positive z-scores and negative news, however, the effect disappears for z-scores greater than 1.5.

[12]In order to determine the news topic we use the Text2Topic approach as described in Salbrechter (2021). This is done by computing topic loadings for each news article by calculating the cosine distance between all words in a news article and the predefined topic words using word vectors generated with word2vec.

	(a)	(b)	(c)	(d)
Const.	-0.0001***	-0.0001***	-0.0001***	-0.0001***
	(-11.6554)	(-10.5624)	(-10.6178)	(-11.6332)
$z_{(t-1)}$	-0.0002***	-0.0002***	-0.0002***	-0.0002***
	(-16.6964)	(-16.5763)	(-17.6051)	(-17.8083)
s_t	0.0013***	0.0003***	0.0000	0.0003***
	(8.7235)	(3.2311)	(0.0645)	(2.9468)
$z_{(t-1)}$ x abs(s_t)	0.0004***	-0.0008***	-0.0008***	-0.0003***
	(3.9032)	(-9.1538)	(-13.1274)	(-4.3899)
Observations	2,099,597	2,122,527	2,226,202	2,203,272
R^2	0.0002	0.0002	0.0003	0.0002
Adjusted R^2	0.0002	0.0002	0.0003	0.0002
Residual Std. Error	0.0166	0.0169	0.0171	0.0168
F Statistic	120.9868***	139.5178***	199.2108***	127.4006***
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3.8: This Table shows the results of the regression $r_t = \beta_0 + \beta_1 \times z_{(t-1)} + \beta_2 \times s_t + \beta_3 \times (z_{(t-1)} \times abs(s_t)) + \epsilon$ for the following news subsets: (a) analyst forecasts, (b) earnings reports, (c) all news except analyst forecasts, (d) all news except earnings reports. The dependent variable is the idiosyncratic open-to-close return and the regression is performed over the full period from January 2002 to February 2020.

(a) Analyst forecasts (b) Earnings reports



(c) All news except analyst forecasts (d) All news except earnings reports

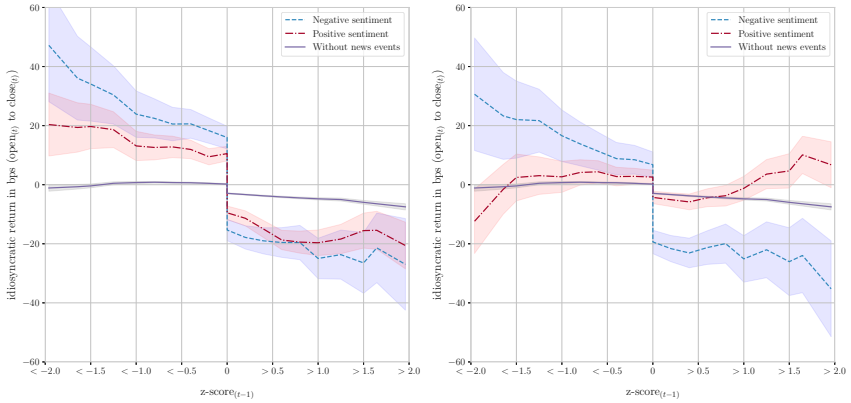


Figure 3.6: This plot shows the idiosyncratic mean returns measured from from open$_{(t)}$ to close$_{(t)}$ and the standard error, conditional on the z-scores measured at market close$_{(t-1)}$. We restrict the observations to news articles that are: (a) analyst forecasts, (b) earnings reports, (c) all news except analyst forecasts, (d) all news except earnings reports. The time period ranges from 2002 to 2020.

3.5.6 Backtest

To demonstrate that the observed effects are persistent and not solely driven by a small number of large impact events, we perform backtests utilizing the signals generated from financial news.

The trading logic we implement harnesses the inefficiencies caused by overnight news. We go long in stocks that either have a) positive overnight news of the topic analyst forecast and $z_{i,(t-1)} > 1.5$ or b) positive overnight news of the topic earnings report and $z_{i,(t-1)} <$ -1.0. In addition we short stocks with either c) negative overnight news of the

topic analyst forecast and $z_{i,(t-1)}$ < -1.5 or d) negative news in combination with $z_{i,(t-1)}$ > 1.0. This strategy is motivated by the findings of Sections 3.5.2 and 3.5.5. We enter trades at market open on day t and exit trades at market close on the same day. Moreover, the maximum weight is set to 100%, i.e., a maximum of 100% of the capital is invested in one asset within one trade,[13] portfolios are weighted equally.

The cumulative portfolio return of this strategy is shown in Figure 3.7, the corresponding statistics are displayed in Table 3.9. This strategy generates a return per trade of 26.17 bps in the long leg and 28.55 bps in the short leg with an average number of 1.87 trades/week in the long leg and 3.94 trades/week in the short leg. This results in a Sharpe ratio of 0.73 (0.85) in the long (short) portfolio and a Sharpe ratio of 1.31 in the long/short portfolio. The winning rate is 52.10% for trades in the long leg and 54.54% for trades in the short leg, i.e., the return on each trade is positive (negative) in 52.10% (54.54%) of the time.

	Performance (%)	CAGR (%)	SD (%)	Sharpe Ratio	Max. Drawdown (%)	Max. Drawdown (days)	Avg. nr. of trades per week	Avg. portfolio return (bps)	Avg. return per trade (bps)	Winning trades (%)
S&P500 Total Return	247.53	7.36	18.32	0.40	-58.67	1353.00				
S&P500 Price Index	142.86	5.19	18.32	0.28	-60.08	1516.00				
Long	1332.71	16.40	22.60	0.73	-57.88	1134.00	1.87	26.58	26.17	52.10
Short	5668.83	26.02	30.44	0.85	-54.38	832.00	3.94	-23.87	-28.55	54.54
Long/Short	91162.63	47.52	36.38	1.31	-59.75	279.00	5.82	32.01	27.79	

Table 3.9: Descriptive statistics of the backtest shown in Figure 3.7.

[13] A maximum weight of 100% is of course an extreme setting, by choosing a lower value, the portfolio volatility can be significantly reduced which in turn results in higher Sharpe ratios (see Appendix B.2). We have chosen the maximum weight of 100% to make the return per trade and the daily portfolio returns comparable.

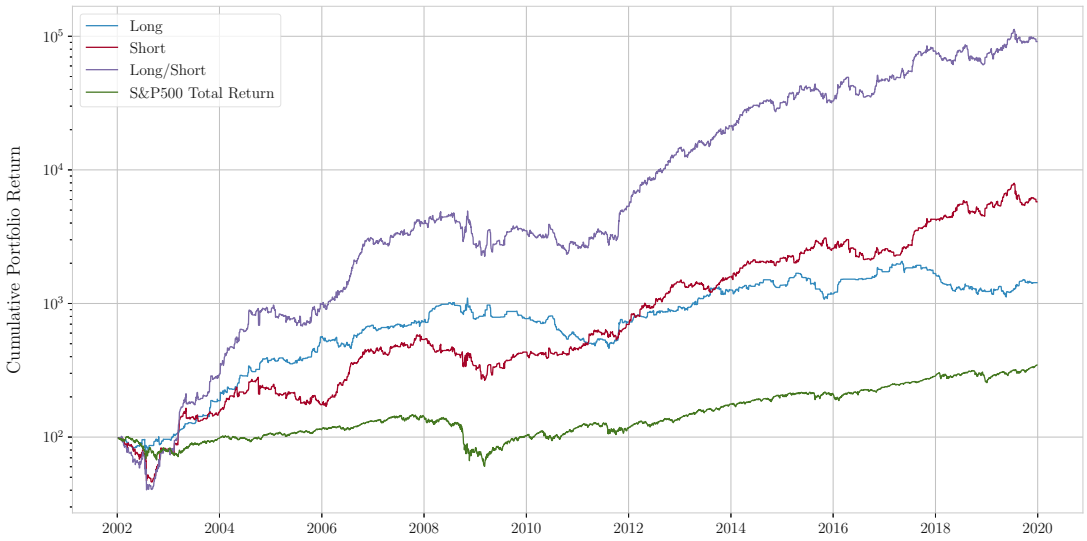


Figure 3.7: This figure shows the cumulative idiosyncratic returns of the long-, the short- and the long/short portfolio in comparison to the S&P 500 total return with transaction costs set to zero.

Table 3.9 reports the average daily portfolio return as well as the average return per trade. Most notably we observe that the average absolute return of the short portfolio is lower than the average return of all trades that enter the short portfolio. What this indicates is a tendency that the most profitable trades likely occur in clusters, i.e., on the same day. The equal weighting of returns then leads to an underweighting of these profitable trades. We examine this result in more detail by grouping returns according to the corresponding size of daily portfolios and compute descriptive statistics as shown in Table 3.10. Panel A shows the result for the entire period from 2002 to 2020, Panel B shows the results for the first subperiod from January 2002 to December 2009, and Panel C shows the results for the second subperiod from January 2010 to December 2019. We find that clustering occurs only for short trades and only in the first subperiod (Panel B). In this case, the return per trade and the winning rate increase quite substantially with increasing densities of trade signals and hence larger portfolio sizes. However, this pattern is not observed for returns of long trades in the first subperiod and also not for returns of long- and short trades in the second subperiod (Panel C).[14]

Assuming transaction costs of 10 basis points (20 bps per trade), which approximates

[14]Returns that enter the long and the short portfolio tend to be distributed evenly across the entire sample period (see Figure 8 in Appendix B.3).

the average costs of large asset managers,[15] this strategy generates substantial excess returns in the first subperiod, but only minor excess returns in the second subperiod, presumably due to the increased efficiency of capital markets over the past decade.

	Long				Short			
Portfolio size	Avg. return per trade	SD	Winning rate	#Trades	Avg. return per trade	SD	Winning rate	#Trades
Panel A: Full Period, Jan 2002 to Dec 2020								
1	28.93 bps	2.99%	52.35%	831	-14.52 bps	3.15%	53.20%	1156
2	14.57 bps	3.09%	52.56%	430	-33.99 bps	3.39%	55.83%	1030
3	39.15 bps	3.24%	49.49%	198	-38.29 bps	2.93%	55.61%	597
>3	26.64 bps	2.92%	52.51%	299	-33.66 bps	3.94%	54.09%	917
Panel B: Period Jan 2002 to Dec 2009								
1	45.69 bps	3.52%	56.37%	369	9.48 bps	3.63%	51.98%	531
2	51.81 bps	3.15%	56.33%	158	-67.17 bps	3.99%	57.68%	482
3	92.95 bps	3.55%	54.67%	75	-52.41 bps	2.95%	60.23%	264
>3	3.04 bps	3.40%	45.33%	75	-44.93 bps	5.34%	52.63%	342
Panel C: Period Jan 2010 to Dec 2019								
1	15.53 bps	2.48%	49.13%	462	-35.02 bps	2.67%	54.24%	625
2	-7.23 bps	3.04%	50.37%	272	-4.93 bps	2.72%	54.20%	548
3	6.52 bps	3.01%	46.34%	123	-27.07 bps	2.91%	51.95%	333
>3	34.57 bps	2.75%	54.91%	224	-26.97 bps	2.81%	54.96%	575

Table 3.10: We filter for trades that occur in portfolios of different sizes and report average returns per trade, standard deviations, winning rates and the number of trades.

3.6 Conclusion

In this study, we focus on information contained in financial news and its impact on the U.S. stock market. Previous research conducted in this area documents a quick response to news at the market opening (Greene and Watts (1996); Boudoukh et al. (2019); Jiang et al. (2012), among others). There is, however, a lack of literature analyzing the impact of overnight news by measuring news sentiment. We therefore train a modern BERT-based natural language model on the *Thomson Reuters* financial news database, which allows us to analyze the market reaction to news sentiment. We consider the U.S. market and a stock universe of S&P 500 companies. The interplay of previous-day returns with overnight news sentiment predicts *occurrence*, *direction*, and *magnitude* of an inefficient market opening, which in turn translates into a predictability of subsequent-day returns.

[15] Transaction costs of 10 bps are composed of bid-ask spreads, price impact and commissions as described in (Frazzini et al., 2018).

In particular, we document a predictable return reversal of previous day returns on days with company-relevant overnight news. When there is no overnight news, predictability is only marginal. This pattern comes from an asymmetric reaction of investors to news sentiment. Whenever overnight news sentiment confirms the direction of previous-day returns, the opening price tends to overreact to the news, a movement which is reverted on the subsequent day. When overnight news sentiment disagrees with the previous-day return, the market tends to under-react. A portion of the news content feeds into prices only during the subsequent day, resulting in a predictable return along the news sentiment. I.e., it also reverts the previous-day return. Hence, overnight news with strong sentiment predicts return reversal.

Bibliography

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. 59(3):1259–1294.

Barber, B. M. and Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. The review of financial studies, 21(2):785–818.

Berkman, H., Koch, P. D., Tuttle, L., and Zhang, Y. J. (2012). Paying attention: overnight returns and the hidden cost of buying at the open. Journal of Financial and Quantitative Analysis, 47(4):715–741.

Bollen, J. and Mao, H. (2011). Twitter mood as a stock market predictor. 44(10):91–94.

Boudoukh, J., Feldman, R., Kogan, S., and Richardson, M. (2019). Information, trading, and volatility: Evidence from firm-specific news. The Review of Financial Studies, 32(3):992–1033.

Cooper, M. J., Cliff, M. T., and Gulen, H. (2008). Return differences between trading and non-trading hours: Like night and day. Available at SSRN 1004081.

Daniel, K., Hirshleifer, D., and Subrahmanyam, A. (1998). Investor psychology and security market under-and overreactions. the Journal of Finance, 53(6):1839–1885.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Frazzini, A., Israel, R., and Moskowitz, T. J. (2018). Trading costs. Available at SSRN 3229719.

French, K. R. and Roll, R. (1986). Stock return variances: The arrival of information and the reaction of traders. Journal of financial economics, 17(1):5–26.

Greene, J. T. and Watts, S. G. (1996). Price discovery on the nyse and the nasdaq: The case of overnight and daytime news releases. Financial Management, pages 19–42.

Groß-Klußmann, A. and Hautsch, N. (2011). When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. Journal of Empirical Finance, 18(2):321–340.

Jiang, C. X., Likitapiwat, T., and McInish, T. H. (2012). Information content of earnings announcements: Evidence from after-hours trading. Journal of Financial and Quantitative Analysis, 47(6):1303–1330.

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

Kelly, B. T., Ke, Z. T., and Xiu, D. (2019). Predicting returns with text data. (2019-10):54.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. 66(1):35–65.

Michaely, R., Rubin, A., and Vedrashko, A. (2014). Corporate governance and the timing of earnings announcements. Review of Finance, 18(6):2003–2044.

Moshirian, F., Nguyen, H. G. L., and Pham, P. K. (2012). Overnight public information, order placement, and price discovery during the pre-opening period. Journal of Banking & Finance, 36(10):2837–2851.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology, 2(2):175–220.

Salbrechter, S. (2021). Financial news sentiment learned by bert: A strict out-of-sample study. Available at SSRN.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. 62(3):1139–1168.

Uhl, M. W., Pedersen, M., and Malitius, O. (2015). What's in the news? using news sentiment momentum for tactical asset allocation. The Journal of Portfolio Management, 41(2):100–112.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Appendices

B.1 Subperiod Dunn's Test

Z-score	(-∞, -0.5]		(-0.5, 0.5)		[0.5, ∞)	
Sentiment	Negative	Positive	Negative	Positive	Negative	Positive
Panel A: Period 2002 to 2020, exlc. 2008 to 2009						
No News	0.459233	0.000431***	0.264434	0.010841**	2.219295e-10***	0.004286***
Positive	0.074001*		0.011481**		0.002547***	
Panel B: Period 2002 to 2007						
No News	0.932249	0.058648*	0.250728	0.161844	0.000152***	0.257995
Positive	0.213853		0.126007		0.035262**	
Panel C: Period 2007 to 2009						
No News	0.021630**	0.580282	0.203654	0.571911	0.024414**	0.377930
Positive	0.580282		0.643307		0.414816	
Panel D: Period 2010 to 2020						
No News	0.362556	0.005042***	0.632868	0.076837*	7.941048e-07***	0.009263***
Positive	0.293699		0.138845		0.026848**	
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 11: P-values of the Dunn's test for the cases of negative overnight news sentiment, positive overnight news sentiment and no overnight news for the three subsamples and multiple subperiods. As an adjustment for the p-values we use "holm", a step-down method using Bonferroni adjustments.

B.2 Backtest - Descriptive Statistics

Table 12 shows the descriptive statistics of a backtest with the same trading strategy as described in Section 3.5.6, but with a limited maximum asset weight of 50%. Thus, if the portfolio consists of only one asset, 50% of the capital is allocated to the strategy, the remaining capital is allocated to the risk-free asset. This strategy achieves Sharpe ratios of 0.83 (1.20) for the long (short) portfolio and 1.60 for the long/short portfolio.

	Performance (%)	CAGR (%)	SD (%)	Sharpe Ratio	Max. Drawdown (%)	Max. Drawdown (days)	Avg. nr. of trades per week	Avg. portfolio return (bps)	Avg. return per trade (bps)	Winning trades (%)
S&P500 Total Return	247.53	7.36	18.32	0.40	-58.67	1353.00				
S&P500 Price Index	142.86	5.19	18.32	0.28	-60.08	1516.00				
Long	588.55	11.63	14.06	0.83	-37.11	988.00	1.87	16.58	26.17	52.10
Short	5212.65	25.43	21.17	1.20	-31.20	186.00	3.94	-20.77	-28.55	54.54
Long/Short	30669.80	38.65	24.09	1.60	-37.17	186.00	5.82	24.61	27.79	

Table 12: Descriptive statistics of a backtest with the trading strategy described in Section 3.5.6 and a maximum asset weight of 50%.

B.3 Return Scatter-plots of the Backtest Strategy

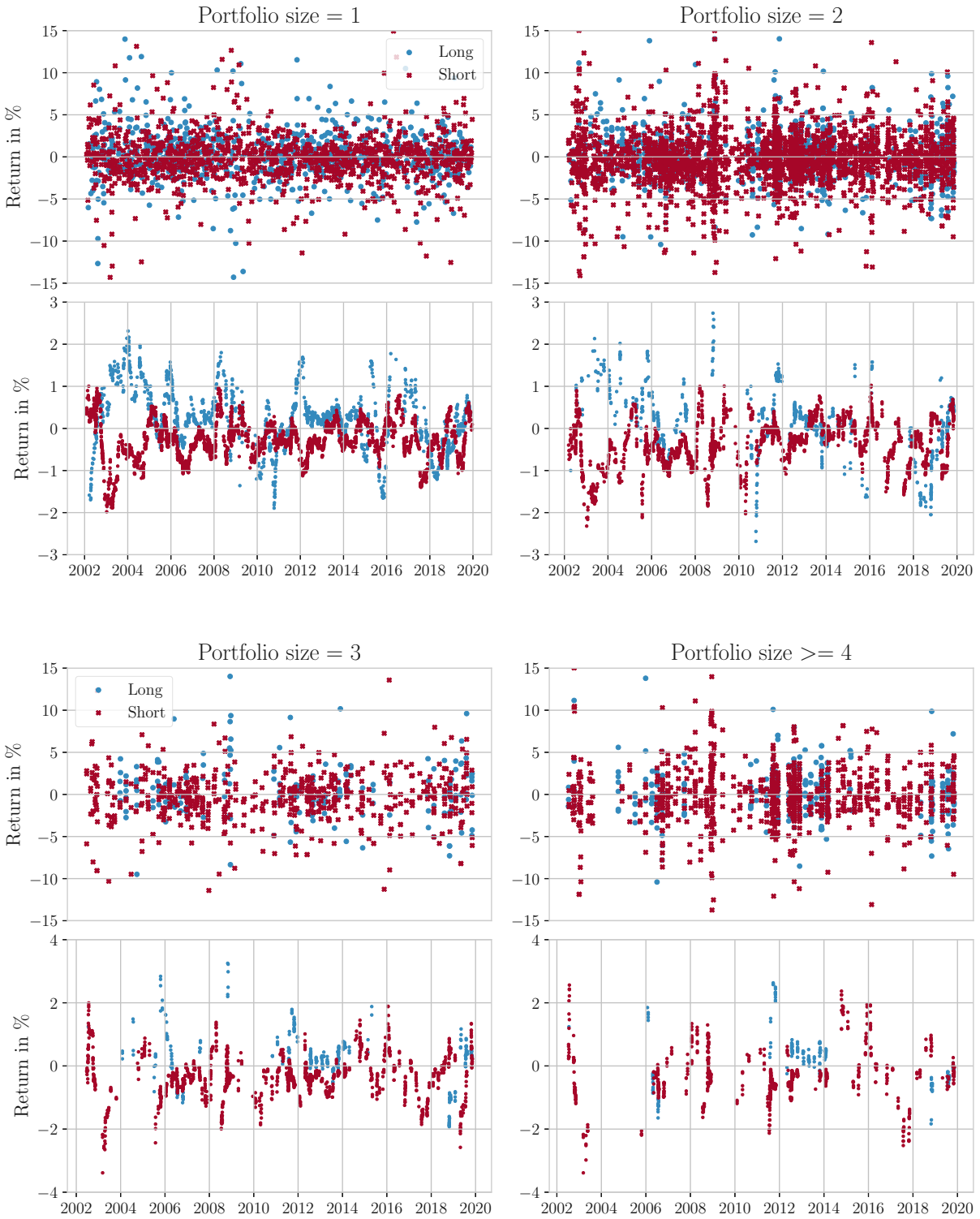


Figure 8: Realized idiosyncratic returns that enter the long- and the short portfolio, filtered for different portfolio sizes. The bottom graphs shows the return per trades averaged over a rolling window of a half year (127 trading days).

B.4 Return Variance: Overnight vs. Daytime

Figure 9 shows the distribution of (a) idiosyncratic overnight returns associated with positive and negative overnight news and (b) idiosyncratic daytime returns associated with positive and negative daytime news. It can be observed that the negative sentiment distribution is centered towards a negative mean, while the positive sentiment distribution is centered towards a positive mean. This pattern is stronger in the case of overnight returns, as overnight news tend to have a larger impact on overnight returns than daytime news has on daytime returns. Figure 9 (c) shows the distribution of the idiosyncratic overnight and daytime returns observed over the period 2002 to 2020. This figure shows that on average, overnight returns have a smaller variance than daytime returns. French and Roll (1986) also finds that daytime volatility is larger than overnight volatility. The authors argue that this is due to a) active trading during market opening times based on private signals and b) pricing errors that occur during trading hours. Since no trading happens during the overnight period, the average variance is smaller during this period. However, if we exclusively observe overnight and daytime returns that are directly affected by the release of news with strong sentiment, either positive or negative, the variance of overnight returns is almost identical with the variance of daytime returns (Figure 9 (d)). In Appendix B.5 we investigate in more detail the impact of news sentiment on return variances.

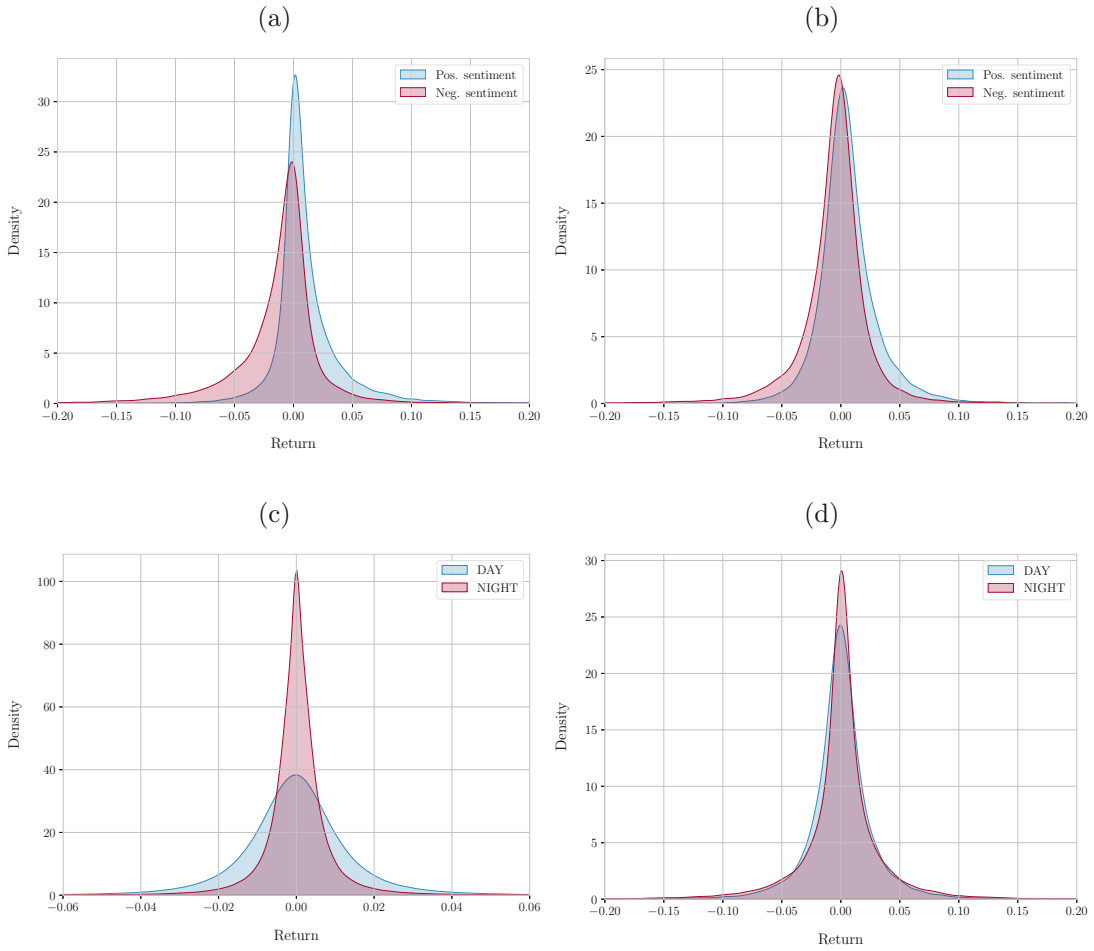


Figure 9: Density plots showing the distribution of (a) idiosyncratic overnight returns associated with positive and negative overnight news and (b) idiosyncratic daytime returns associated with positive and negative daytime news. Density plot (c) shows the distribution of all idiosyncratic overnight and daytime returns observed from 2002 to 2020 (Note the different scale on the abscissa!). Density plot (d) shows the distribution of overnight and daytime returns for observations where news with strong sentiment ($abs(sentiment) \geq 0.9$) is released in the overnight and in the daytime period, respectively.

B.5 Volatility Analysis

In this study we find a strong link between news arrival and asset returns. In addition, we now also quantify the impact of news on the return variance for both, the overnight and the daytime period. Thereby, we follow a similar approach to Boudoukh et al. (2019). As a measure of return volatility we consider the squared z-scores.[16] The z-scores are calculated

[16]While Boudoukh et al. (2019) sort daytime and overnight returns into percentiles separately for each stock and year to control for cross-sectional variation in total returns, we do not split for individual stocks but consider z-scores instead of stock returns to control for the cross-sectional variation in returns.

over a rolling window of 6 month for the daytime period using open-to-close returns $r^o_{i,t}$ and for the overnight period using close-to-open returns $r^c_{i,t}$. The squared z-scores are then assigned into percentiles which are the 20% most extreme values, the moderate 40% and the smallest 40%. Moreover, we define strong news, i.e., news with a clearly positive or negative sentiment (*abs(sentiment)* $\geq$ *0.9*). Columns one to three of Table 13 report the change of the quantiles relative to the unconditional expectation.[17] Daytime effects are displayed in Table 13, Panel A. The data shows that large price changes are 35.65% more likely if news is published and 84.42% more likely if strong news is published. Moreover, Panel B shows that overnight news releases make extreme price changes in overnight returns even more likely. If news (strong news) is published, it is 80.23% (182.77%) more likely to observe large price changes.[18] In order to determine whether the return variance significantly differs between news arrival and no news arrival we perform a variance ratios test.[19] The variance ratio is reported in Table 13. We note that the variance ratio is significant in all cases, which means that the variance upon news arrival is significantly different from the variance observed upon no news arrival. Moreover, the variance ratio is larger during the overnight period. Specifically, for strong news the variance ratio is five times larger (17.26 vs. 3.45) in the overnight period compared to daytime period. This is partly due to the fact that important announcements with large price impacts, such as earnings announcements, are mainly published during stock market closing hours.[20] Furthermore, the fact that the average variance is smaller for the overnight period (Figure 9c), but almost identical to the daytime period when news is published (Figure 9d) also explains the larger variance ratios for the overnight period.

[17] For example, if 40% of the observations, conditioned on strong news, are in the extreme 20% percentile, then the reported change would be 100% $((0.4/0.2) - 1)$.

[18] In this study we only consider firm-relevant news articles, which are news articles where either the company name or the ticker code is mentioned in the headline. The use of relevant news articles explains why news without strong sentiment is also associated with large return volatility.

[19] As described by Boudoukh et al. (2019).

[20] Our data includes 22.957 earnings related news published overnight in contrast to only 6.851 earnings news published during trading hours. The increasing release of earnings reports during market closing hours is in line with the findings in the literature (see Jiang et al. (2012) and Michaely et al. (2014)).

	Extreme 20% (%)	Moderate 40% (%)	Low 40% (%)	Var Ratio	Support
Panel A: Daytime					
Total	0.00	0.00	0.00	1.03***	3206603
No News	-1.20	0.20	0.40	1.00	3102406
News	35.65	-5.99	-11.84	1.97***	104197
Strong News	84.42	-14.55	-27.66	3.45***	33787
Strong News of Topic 1	94.97	-15.63	-31.86	3.11***	5452
Strong News of Topic 2	142.12	-26.91	-44.15	4.81***	6410
Panel B: Overnight					
Total	0.00	0.00	0.00	1.31***	3275806
No News	-4.07	0.76	1.27	1.00	3117794
News	80.23	-15.03	-25.08	7.37***	158012
Strong News	182.77	-39.22	-52.17	17.26***	50437
Strong News of Topic 1	159.35	-31.07	-48.60	7.31***	9811
Strong News of Topic 2	238.68	-53.64	-65.70	22.29***	22775

Note: *p<0.1; **p<0.05; ***p<0.01

Table 13: We transform daily returns into z-scores and sort these z-scores into percentiles, the extreme 20%, moderate 40% and low 40% for (A) the daytime period (open-to-close) and (B) the overnight period (close-to-open). In the first three columns we report the conditional change of percentile counts relative to the unconditional expectation. The fourth column reports the variance ratio - the variance of returns conditional on news arrival relative to the variance of no news observations. News of topic 1 are analyst forecasts and news of topic 2 are earnings reports.

Figure 10 shows the squared z-scores sorted into deciles ranging from q1, containing the 10% lowest values to q10, containing the 10% most extreme price changes. If we do not filter for news, the counts for each quantile are evenly distributed with a slight decline for extreme values (No News). Those extreme price changes are observed to a large proportion if we condition on the arrival of news. Extreme price changes are most likely if we condition on earnings reports with a large sentiment score (Strong News of Topic 2).

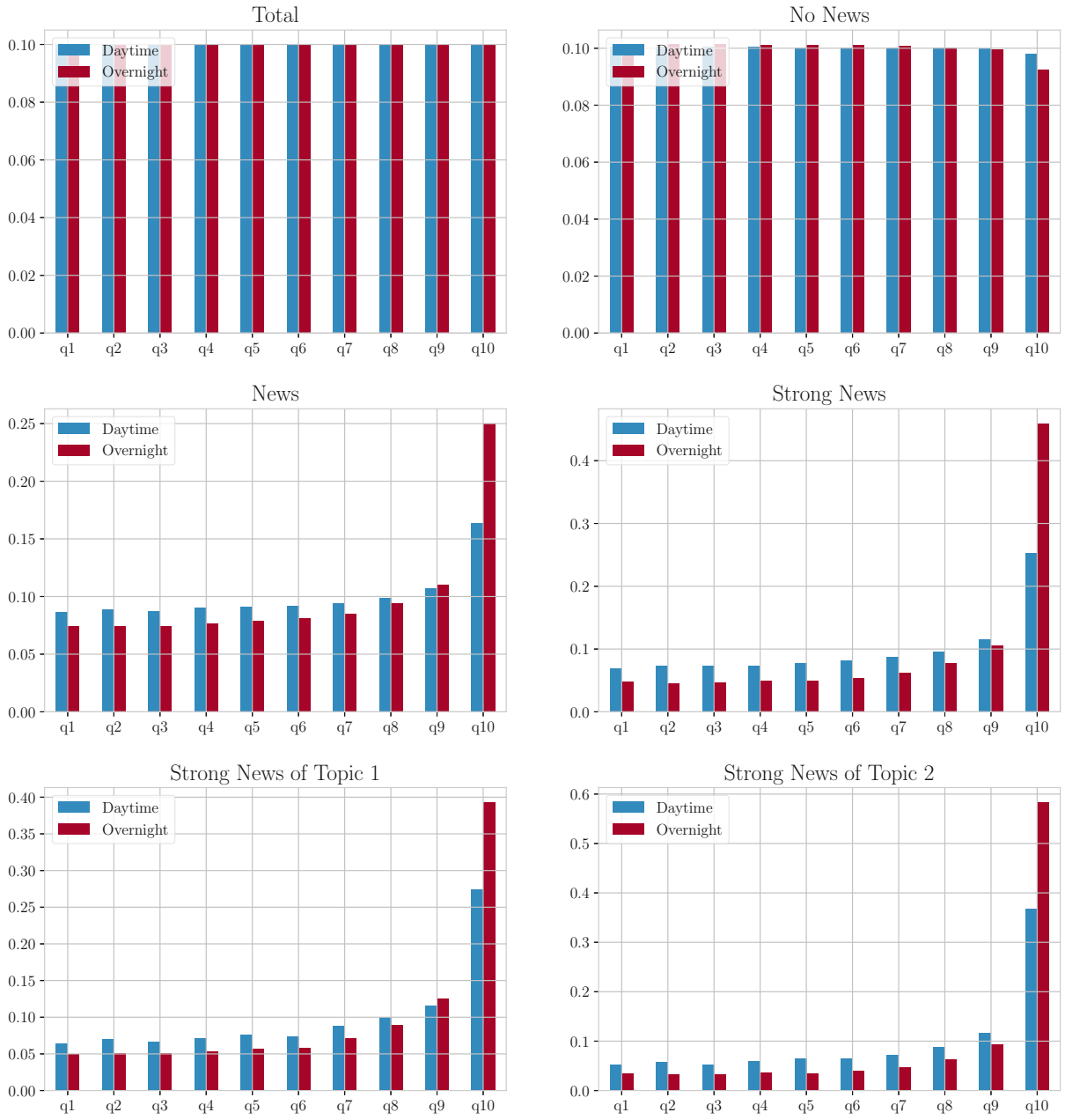


Figure 10: We sort the squared z-scores into deciles ranging from q1, containing the lowest values to q10, containing the most extreme price changes, separately for the daytime period and the overnight period. Strong news are news with a strong sentiment score ($abs(sentiment) \geq 0.9$).

B.6 News Data Excerpt

Timestamp	Ticker	Sentiment
21.06.2002 09:29	XEL	-0.99
... morgan stanley said it cut its rating of electricity and natural gas company xcel energy inc to equal weight from overweight on friday citing uncertainty over earnings and dividends and its subsidiary nrg energy inc morgan stanley said in research note there is chance that xcel which traditionally considers its dividend in june could decide next week to reduce its payments to shareholders the investment firm also said that xcel is facing pressure from ratings agencies over its affiliate nrg morgan stanley forecast that recent decision by connecticut regulators means the company forecast could come down further xcel said on tuesday that decision by connecticut regulators to deny rate increase could cut its income by million month the firm also said xcel earnings per share could be diluted as it expects the company to issue equity xcel closed thursday at per share ...		
12.05.2003 08:42	HIG	-0.96
... insurer hartford financial services group inc on monday said it would cut jobs or about percent of its staff and boost its asbestos reserves by billion the company based in hartford connecticut said it had loss of billion or share for the first quarter largely because of the reserve addition in the year earlier quarter it had profit of million or share new york may insurer hartford financial services group inc on monday said it would cut jobs or about percent of its staff and boost its reserves for asbestos claims by billion the company also said it would beef up its balance sheet by raising billion through stock and debt offerings and would pull out of the property casualty reinsurance business it said it was already in talks to sell its hartre unit as result of the asbestos reserve worth billion after tax hartford said it lost billion or share in the first quarter in the year earlier quarter ...		
20.10.2009 08:12	AAPL	0.99
... jp morgan expects apple annual revenue growth story to trend deep into double digit territory jp morgan expects increasing revenue growth out of apple in driven by sustained momentum in mac and iphone jp morgan says except for apple most other it hardware peers will have upside potential more on bottom line versus topline lance knobel is guest columnist the views expressed are his own he is an independent strategy advisor and writer based in the united states his professional site is www lknobel com by lance knobel berkeley calif oct here how bullish steve jobs and his colleagues at infinite loop in cupertino feel after blowout september quarter...		
19.01.2012 20:04	INTC	0.99
... intel corp shares were up percent after the bell as it reported results revenue also meets expectations gaap eps cents shares up after earnings report san francisco jan intel corp forecast quarterly revenue in line with wall street expectations as shortage of hard drives disrupts pc production in market already hobbled by shaky economy and growing preference for tablets intel said revenue in the current quarter would be billion plus or minus million analysts on average had expected current quarter revenue of billion according to thomson reuters the world leading chipmaker said revenue in the fourth quarter was billion up percent and slightly higher than the billion expected ...		

Table 14: This table shows excerpts of news from the Thomson Reuters dataset along with the predicted news sentiment (positive: $sentiment \geq 0.95$, or negative: $sentiment \leq -0.95$).

4 Firm-specific Climate Risk Estimated from Public News

Firm-specific Climate Risk Estimated from Public News

Thomas Dangl Michael Halling Stefan Salbrechter

September 19, 2023

We estimate firm-specific exposures to climate risk from public news covering a period of 20 years by applying a novel topic modeling algorithm. We differentiate between regulatory (or transition) and physical climate risks and document that financial markets price both risks. Our study is the first to find a positive and statistically significant risk premium for physical climate risk. For regulatory climate risk we find a regime shift occurring around the year 2012 reconciling the conflicting evidence in the literature. While the risk premium is positive in the earlier period, it becomes significantly negative in the later one. A long-short portfolio that is long "green" firms and short "brown" firms, as identified by their topic exposures in public news, constitutes a priced risk factor and shows a surprisingly strong correlation with an ESG-sorted benchmark portfolio.

4.1 Introduction

Do share prices reflect firms' exposures to regulatory and physical climate risk? The literature on ESG (Environmental, Social, and Governance) related asset pricing is growing very fast.[1] While most research tries to construct risk measures from a vast amount of data on the corporate climate footprint, we take a different route and deduct firms' risk exposures to climate risk from news.

We use news released via the Thomson Reuters newswire during the last 20 years to identify firms' climate risk exposure. In contrast to the common approach in academia as well as industry to use climate-related ESG-scores or emission data as proxies for these exposures, our machine-learning approach assigns firm-specific news texts to climate-related topics, and hence, is able to compute firm-specific measures of exposures to climate risk. Desirable features of using news are its high frequency, that it can be observed in real time, and that it covers a long history, as news archives start in the 1990s while other climate risk related databases have become available only during the 2010s. News also offers two important advantages content-wise. First, any aspect of a given topic that is potentially relevant for the readers of a news outlet will be covered by news. Thus, extracting information from news has the potential to capture the relevant aspects of a given topic in a very comprehensive manner. Second, news in many cases will also capture forward-looking aspects of the discussed problem and will not only rely on a backward-looking perspective.

However, those benefits of news come with a big disadvantage: text data is unstructured and high dimensional (Gentzkow et al., 2019), which is why it has to be transformed into a machine-readable form first. This involves additional challenges, such as high computational costs, difficulty in identifying and extracting useful information, or the lack of labels to train a machine-learning model. In this study, we apply a novel method which is fast, flexible and transparent. It also represents an unsupervised learning approach and, thus, does not require any labeled training dataset. Specifically, we propose Guided Topic Modeling (GTM), an algorithm to generate weighted lists of unigrams and bigrams, i.e.,

[1] We provide a literature review in Section 4.2.

individual words and 2-word phrases, that are most representative for a particular topic. The only information required are two (or more) seed words that describe a topic. These seed words are then mapped to word embeddings via a self-trained Word2Vec model.[2] Word embeddings represent high-dimensional vector representations that can be used to identify similar words in the vector space using vector algebra and different mathematical concepts of distance. The algorithm, however, does not simply collect words closest to the seed words but also learns about the topics' representation from the data and determines the optimal topic center in the vector space. In the end, the algorithm yields a similarity parameter (weight) that is higher for words closer to the topic center (i.e., words that are highly representative of a topic) and lower for words that are more distant form the topic center (i.e., less important words). While the proposed GTM algorithm can be used in any context, we specifically use it to identify words that cover different topics related to climate risks and opportunities, specifically, (i) regulatory climate risk (or transition risk), (ii) physical climate risk and (iii) sustainability (the idea of this third group of topics is to capture opportunities in the context of sustainability).

To measure firm-specific climate risk, we convert the unstructured news data into a numeric, structured format by calculating topic exposures: Once we have the topic word lists, we calculate the exposure of 4.95 million news articles to each climate-risk related topic. As news articles are tagged with metadata that include the associated companies, we are able to calculate company-specific topic exposures as the topic-weighted sum of words in all news articles related to a specific company. In principle, we observe this measure at the daily frequency. However, not every firm is covered in the news every day and, thus, we smooth these firm specific exposures over a rolling window of two years.

To evaluate the economic plausibility of these exposures, we first look carefully at industry distributions of firms that are exposed to regulatory and physical climate risks. We find that (i) *Electric, Gas, and Sanitary Services*, (ii) *Coal Mining* and (iii) *Petroleum Refining and Related Industries* have the highest exposures to regulatory climate risk during the sample period. For physical climate risk, we find that *Electric, Gas, and Sanitary Services* has the highest exposure followed by *Insurance Carriers*. *Oil and Gas*

[2]see, Dangl and Salbrechter (2023)

Extraction and *Food and Related Products* have the third and fourth highest exposure. In both cases, these industry exposures appear to be economically sensible suggesting that the news-based approach picks up relevant information about firm-specific climate risk exposures.

Equipped with these firm-level exposures, we then assess whether regulatory and physical climate risks are priced in equity markets. Using Fama-MacBeth regressions, in which we control for CAPM betas, market capitalization, book-to-market ratios, operating profitability, and investment, we find a statistically significant positive risk premium of 1.5% p.a. for physical climate risk. The risk premium is robust to the inclusion of sector or industry fixed effects and, thus, captures an effect that is tied to the individual firm. This result is particularly noteworthy, as our study is the first one to explicitly document that physical climate risk is priced in equity markets.

Looking at regulatory climate risk next, we find a more nuanced picture. Over the full sample period, the estimated risk premium is small and statistically insignificant. This result, however, is due to a regime shift in the risk premium occurring around 2012. If we split the sample roughly in half, we find a positive and statistically significant risk premium of 1.54% p.a. during the earlier years. Such a positive risk premium is consistent with the idea that stocks exposed to regulatory climate risk are riskier in financial terms (see, for example, Bolton and Kacperczyk, 2021; Hsu et al., 2022). During the latter years, however, the risk premium switches sign and becomes significantly negative with a point estimate of -2.56% p.a. While a negative risk premium seems counter-intuitive, it has been rationalized through large increases in the demand for green assets (see, for example, Pástor et al., 2022). In fact, this regime shift in the risk premium of regulatory climate risk, that we are able to document within a consistent framework due to the substantially longer time-series of data, provides an explanation for the ongoing controversy in the literature about the returns of green and brown investments and about the sign of the regulatory climate risk premium.

Next, we evaluate whether the identified climate exposures are priced in the cross-section of all firms and not only among those covered in the media. We add the monthly return series of a *green-minus-brown* portfolio (GMB portfolio), that is long firms with

high exposures to sustainability (i.e., a portfolio of firms offering opportunities in terms of sustainability) and short firms that show high exposures to regulatory climate risk, to the market model and calculate climate betas. The obtained climate betas can be interpreted as estimates of companies' sustainability levels — high betas indicate green firms, while low betas indicate brown firms. Importantly, the climate beta as a measure of regulatory climate risk is not limited to companies explicitly mentioned in the news. Instead, by using data of all firms in the CRSP-Compustat universe, climate betas are available for a total of 9000 firms over the period Jan. 31, 2002 to Dec. 31, 2020. Another important advantage of this approach, compared to the literature, is that green firms are explicitly identified as firms with sustainability-related opportunities instead of being just classified as firms with no or little exposure to climate risk.[3]

In the case of physical climate risk betas, we need to follow a slightly different approach, as in this case we do not have a way to directly identify those firms with low exposures to physical risks. As a consequence, we construct a long-short portfolio by going long in firms with high exposures to physical risks and shorting all other stocks. Then we, again, calculate physical climate risk betas for the universe of stocks, as described above.[4]

When using these betas and thus extending our analysis to the universe of U.S. equities, we basically confirm the effects described above: We observe downward sloping cumulative returns of the GMB portfolio from 2002 to 2012 followed by strong positive returns until 2020. This outperformance from 2012 onwards cannot be explained by classic risk factors as we obtain an alpha of 9.60% p.a. when regressing the GMB portfolio on the Fama-French 5-factor model plus momentum. We further show that extending factor models by the climate risk factors, i.e., the GMB- and the high-minus-low physical climate-risk portfolio, significantly improves the explainability of asset return variations, as reflected in reductions of the GRS (Gibbons et al., 1989) test statistic. A final, noteworthy result is that the monthly returns of our climate-beta based GMB portfolio and the returns of the GMB portfolio of Pástor et al. (2022), which is constructed in a completely different way

[3]This aspect is related to the broader issue of whether no information is good or bad. See, for example, Engle et al. (2020) who classify days with no news about climate risk are low risk days.

[4]To validate the regulatory- and physical climate risk betas, we build portfolios and construct long-short strategies. Comparing the returns of these beta-based portfolio sorts to the returns of news-exposure-based sorts yields correlations of 0.6 and higher for monthly and quarterly returns.

using ESG data from MSCI, show a surprisingly large correlation of 0.64. This provides external validation of our approach, as it shows that we are able to extract climate-related information that is comparable to the one captured by E-related MSCI scores, but exclusively from news.

The remainder of this paper is structured as follows: In Section 4.2 we provide a review of the related literature, in Section 4.3 we describe the data and in Section 4.4 we give a detailed description of our methodology. This includes the topic modeling algorithm (Section 4.4.1), a visualization of the obtained topics (Section 4.4.2) as well as a description of all necessary steps to obtain company specific news indices (Section 4.4.3). Thereafter we present the results (Section 4.5), starting by highlighting the firm- and industry-specific topic exposures, which we use in Section 4.5.2 to form climate risk portfolios. In Section 4.5.3 we present the results of the Fama-MacBeth cross-sectional regressions and in Section 4.5.4 we calculate climate risk beta coefficients which we use to form beta-sorted climate risk portfolios. In Section 4.5.5, we validate our results by first showing the correlation with an ESG-sorted benchmark portfolio (Section 4.5.5) and second by calculating the exposure to well-known risk factors (Section 4.5.5). Thereafter, we present factor model extensions by including our climate risk factors (Section 4.5.5) and lastly show in Section 4.5.5 that climate betas predict future news flows.

4.2 Related Literature

This paper relates to a quickly growing literature that evaluates different ways to learn about the climate risk exposures of firms and examines whether climate risks as prized in equity markets. Related studies frequently rely on traditional data sources to measure individual firm's climate risk such as ESG data (Engle et al., 2020; Pástor et al., 2022; Seltzer et al., 2022) or emissions data (Bolton and Kacperczyk, 2021; Ardia et al., 2020; Hsu et al., 2022). Engle et al. (2020) use ESG data from Sustainalytics and MSCI to measure individual firms' exposures to climate risks. By implementing a mimicking portfolio strategy, the authors attempt to dynamically hedge climate change risk as measured by innovations of a climate news series that is extracted from the Wall Street Journal news

feed. Bolton and Kacperczyk (2021) report a positive, statistically significant, transition-risk premium for high-emission firms using firm-level emissions data from Trucost covering 77 countries and 14,400 firms from 2005 to 2018. The risk premium is more pronounced when controlling for industry fixed effects, suggesting that industries with high emission levels have earned low returns. Similarly, we also observe an increase in transition-risk from 1.54% p.a. (t-value: 1.61) to 1.75% p.a. (t-value: 2.34) after controlling for industry fixed effects over the period 2002 to 2012 (see the Fama-MacBeth regressions in Table 4.7). For the period 2012 to 2020 in contrast, we observe a risk premium that is negative, but insignificant. Also, Bolton and Kacperczyk (2021) observe no statistically significant transition-risk premium associated with emission levels for North America over the shorter 2014 to 2017 period - the two years before and after the 2015 Paris climate agreement - while Asia experienced a sharp increase in the risk premium for the two years following the conference.

Furthermore, Seltzer et al. (2022) documents that US firms with poor environmental profiles, as measured by their ESG rating, as well as high carbon emission firms, as measured by data from the Carbon Disclosure Project (CDP), are associated with lower credit ratings and higher yield spreads. The authors further find that this effect is amplified for firms located in states with stricter enforcement of regulations, i.e., firms that are more likely affected by regulations. Choi et al. (2020) documents a climate related attention effect - people update their beliefs about climate change when they are personally exposed to warmer than usual temperatures. As a consequence carbon-intensive firms underperform relative to low carbon emission firms on these days.

Traditional data sources usually have the disadvantages of being backward looking, time-lagging as they are only available at low frequencies and limited in historical coverage since the reporting of emission data became only mandatory with the signing of the Mandatory Reporting of Greenhouse Gases Rule of the Environmental Protection Agency in 2010. One alternative way to gather valuable climate information is text data in the form of earnings call transcripts or news data. Sautner et al. (2023a) applies a keyword discovery algorithm to generate lists of climate related keywords from a short list of initial keywords that describe different climate topics, namely *climate change opportunity*,

physical- and *regulatory climate risk*. They use earnings call transcript data to measure the individual firm's climate risk exposure by counting the overlapping words between transcripts and climate topics. By conducting a variance analysis, the authors find large firm-level variations in exposure measures, i.e., variations also exists among firms within the same industry. Furthermore, the authors provide evidence that climate exposures are priced in the options and equity markets. A conditional factor constructed from innovations in climate change exposures is positively correlated with higher uncertainty and thus, higher returns.[5] In a follow-up study Sautner et al. (2023b) test whether climate risks are priced in the cross-section of S&P500 firms. By performing a Fama-MacBeth regression over the period Jan. 2008 to Dec. 2020 the authors find an insignificant risk premia for *climate change opportunity*, *regulatory-* and *physical climate risks*. Only when using proxies for expected returns do the risk premia for *climate change opportunity* and *regulatory climate risk* become significant; however, at low margins with risk premia being below 0.23% p.a. Furthermore, the authors find weak support of topic exposures in the earnings call transcript data before 2008 which is why they exclude earlier years. In particular, the exposure to *physical climate risk* is close to zero in the vast part of their sample, leading to insufficient variation in the variable which in turn results in an ill-defined estimation problem. Therefore, the authors do not find a risk premium for physical climate risk. In contrast, our financial news dataset is not affected by such limitations. With almost 5 million news articles, we find sufficient support for all climate topics in the data from January 2002 onwards. In contrast to Sautner et al. (2023b), we also study a much larger sample of firms and develop our own guided topic modeling approach to extract relevant, firm-specific information from the news. Given this setup, we find starkly different results, as we document a significantly positive risk premium for physical climate risk, and time-varying risk premia for regulatory climate risk.

Another strand of the literature extracts climate risk related information from 10-K reports. Kölbel et al. (2020) classify climate related texts in these reports into the categories transition- and physical climate risk by using a fine-tuned BERT model named

[5] In addition, the authors show that climate change exposure predicts job creation in green technologies as well as green patents.

ClimateBERT and analyze how mandatory regulatory disclosures affect the CDS term structure. The authors document opposing effects of disclosing transition- and physical risks: when transition risks are disclosed, the CDS spreads tend to increase, due to increased uncertainty, while when physical climate risks are disclosed, they tend to decrease due to reduced uncertainty. Similarly, Berkman et al. (2021) utilize a firm-specific measure for climate risk based on 10-K reports and find that increased climate risk reduces firms' valuations. The authors report a negative impact on the prices of firms with high climate risk when investor attention to climate risk, triggered by catastrophic events such as floods or hurricanes, is high.

4.3 Data and Data Preprocessing

This research paper is based on a comprehensive dataset of news articles published via the news agency Thomson Reuters. It contains over 40 million news items, each linked to metadata with exact timestamps of publication, topic- and geography codes as well as ticker codes for firm-related news. The dataset covers the period from January 1996 to July 2021. We restrict our analysis on news written in English language, which sum up to a total of 12.42 million articles as well as to U.S. news, i.e., news tagged with the U.S. geography code, which finally results in a dataset of 4.95 million news articles. In a next step we clean the raw news data by excluding author information such as phone numbers, email addresses and URLs. We transform the text to lowercase and remove numbers, parentheses, and additional information added by Thomson Reuters such as notes and keywords. Also, we form multi-word expressions (bigrams) by training and applying the Phrases model available in the Gensim Python package, (see Řehůřek and Sojka, 2010). Furthermore, we collect security data from CRSP and select all stocks with share codes 10 and 11. Since small firms, as measured by their market capitalization, have only a minor news exposure, we exclude firms with a market capitalization below the median market capitalization across all CRSP stocks in each month. Later, in Section 4.5.4 we relax this restriction and include all stocks with a market cap above 5 million in each month.

4.4 Empirical Methodology

In contrast to papers that use environmental scores from ESG data providers as a measure of firms' sustainability (see, e.g., Pástor et al., 2022, who rank firms along the environmental (E) score obtained from MSCI ESG data), or studies that utilize emissions data (see, e.g., Bolton and Kacperczyk, 2021; Hsu et al., 2022), we approach this problem from an entirely different angle. We deduct a firm's exposure to various types of climate risk from its presence in related news articles. Firms are considered the more sustainable (i.e., "green") the higher the frequency at which they are mentioned in news articles that discuss topics positively associated with environmental friendliness. Conversely, they are considered less sustainable (i.e., "brown") when they are often mentioned in articles that are related to various types of climate risk (where we distinguish between topics associated to regulatory / transitional climate risk and to physical climate risk).

Keyword matching is a simple, unsupervised classification technique that does not require the training of a model. By counting the occurrences of selected words describing a topic in a text document, an exposure metric is obtained that explains how strongly the selected topic is represented in the text. Studies that apply such an approach include Baker et al. (2016); Engle et al. (2020); Ardia et al. (2020); Sautner et al. (2023a) among others. The main challenge thereby is to determine lists of words (consisting of uni- or bigrams) that are most representative of describing certain topics. Authors often rely on pre-specified dictionaries (Baker et al., 2016) as the manual creation of comprehensive lists of representative words is considered a "near-impossible" task for humans (Hayes and Weinstein, 1990). King et al. (2017) argue that the human brain has limited abilities to recall keywords, which prevents us from manually creating comprehensive, unbiased word lists. To circumvent a manual generation of word lists, Sautner et al. (2023a) rely on a keyword discovery algorithm developed by King et al. (2017).

We, in contrast, apply *Guided Topic Modeling with Word2Vec (GTM)* (see Dangl and Salbrechter, 2023), our novel approach that enables the fast generation of comprehensive topic clusters. One unique aspect of GTM is that each obtained topic word is associated with a weight parameter. Words that are closer to the topic center, i.e., are more

representative of a certain topic, receive a higher weight than more distant words that are less representative. With the obtained topic word lists we perform weighted keyword matching over all 4.95 million news articles. We count the number of words in each article that overlap with the words contained in a topic and then multiply the count by the associated weights. The sum of these weighted word counts gives us a score (loading) that indicates how strongly a topic is represented in a given news article. To make the topic loadings comparable over news articles and topics, we adjust the loadings for differences in news article lengths and word frequencies. Finally, we generate company specific topic indices. Therefore, we select all news articles related to a firm of interest and aggregate the loadings on individual news to daily scores. We use the information reflected in the company-specific topic indices to identify green and brown firms as well as firms exposed to physical climate risks, as described in the following sections.

4.4.1 Guided Topic Modeling

Guided Topic Modeling (GTM) is based on vector representations of words (word embeddings) which we obtain from a Word2Vec (Mikolov et al., 2013) model. The model is pre-trained on a total of 10 million news articles[6] (2.5 billion words) of the Thomson Reuters news dataset, covering the period from January 1996 to December 2017.

The Word2Vec model translates words into dense vector representations that capture semantic similarities between words. Words that appear in similar contexts tend to have similar meanings and thus, receive embeddings that point in similar directions in the embedding space. To avoid data sparsity we train Word2Vec with a rather low embedding dimension of n=64. The quality of topics is determined by our clustering algorithm and the Wor2Vec hyperparameters, especially the embedding size, the Word2Vec algorithm (CBOW is preferred over skip-gram) and the window size. Choosing a too high embedding size or a too narrow window will result in topic word clusters that are too specific and do not generalize well. Also, we augment the obtained word embeddings with information about word polarity and transform all embedding vectors to unit length.

With the word embeddings at hand we perform topic clustering in the embedding

[6]We do not limit the training data to news with the U.S. geographic code, but use all English news.

space. However, in contrast to Angelov (2020); Sia et al. (2020); Grootendorst (2022) we do not run a clustering algorithm that detects a certain number of latent topic clusters (dense areas) hidden in the embedding space. Instead, we generate topics based on seed words. With this approach, we are able to generate an unlimited number of topic clusters without being limited to the output of a clustering algorithm that has no guarantee that a topic of interest will actually be identified as a latent topic. The algorithm takes as input a list of two or more seed words, each associated with a weight parameter. The vectors associated with the seed words span an initial plane in the embedding space. All word vectors contained in the embedding space are projected on the plane and the word with the smallest projection angle, i.e. the word that is closest to the plane, is added to the topic cluster. Thereafter, the location of the plane is adjusted to minimize the residual sum of squares to all topic vectors. Thus, the topic center is not defined by the seed words but the algorithm iteratively finds a optimal topic center (i.e. the final location of the plane). Next, all remaining word vectors are projected onto the adjusted plane and the procedure continues until a specified cluster size is reached. After the topic generation is finished, we project all topic vectors onto the final plane to calculate the weights of the topic words. The weight is calculated by the Frobenius norm of the two projection coefficients. Thus, the weight is 1 if a vector lies in the plane and 0 if a vector would be orthogonal to the plane. Consequently, words located closer to the topic center (final location of the plane) receive larger weights than more distant words.

In addition, the parameters and hyperparamters of the clustering algorithm allow to control the characteristics of the generated topic clusters. The weight associated with each seed word controls whether the final topics center lies closer to seed word A or seed word B. In addition, negative seed words can be defined to avoid unwanted terms in the topic cluster. A gravity parameter can be used to drag the topic center closer to the location defined by the initial seed words. For a detailed description of the clustering algorithm we refer to Dangl and Salbrechter (2023).

4.4.2 Topics

We use GTM to generate subtopics that capture aspects of the three themes: regulatory climate risk, physical climate risk, and sustainability. Regulatory climate risk, often denoted as transition risk, emerges from new laws and regulations that could harm companies' profits due to mandatory investments in greener production facilities or due to penalty payments like carbon taxes. Physical climate risk is the risk of destruction of firms' assets (production facilities, real estate, farmland, assets) due to extreme weather events, floods, droughts or hurricanes. Sustainability captures green technologies and key concepts that define environmentally friendly businesses. Each of the three main topics is comprised of several subtopics – four subtopics for regulatory climate risk and sustainability and eight subtopics for physical climate risk – which are shown in Figure 4.1. The weight of each word in a topic is visualized by the font size, i.e., words closer to the topic center appear larger, more distant words appear smaller. The seed words used in GTM to generate these topics are shown in Table 4.1.

	Positive Seed Words (Weight)			Negative Seed Words (Weight)	
(1) Sustainability					
Subtopic 1	renewable_energy (1.0)	clean_energy (1.0)		fossil_fuel (-0.2)	
Subtopic 2	environmentally_friendly (1.0)	sustainable (1.0)	modernising (1.0)	car (-0.5)	carbon (-0.5)
Subtopic 3	environmentally_friendly (1.0)	eco-friendly (1.0)		burning (-0.5)	
Subtopic 4	solar_power (1.0)	wind_power (1.0)		fossil_fuel (-0.4)	
(2) Regulatory Climate Risk (Transition Risk)					
Subtopic 1	eco-tax (1.0)	carbon_tax (1.0)			
Subtopic 2	regulation (1.0)	carbon_tax (1.0)			
Subtopic 3	carbon_pollution (1.0)	carbon_tax (1.0)			
Subtopic 4	polluter (1.0)	carbon (1.0)	emissions (1.0)	emissions_reduction (-0.2)	
(3) Physical Climate Risk					
Subtopic 1	storm (1.0)	hurricane (1.0)			
Subtopic 2	heat_wave (1.0)	drought (1.0)		cold_weather (-0.5)	
Subtopic 3	wildfires (1.0)	bushfire (1.0)		cold_weather (-0.5)	fires (-0.5)
Subtopic 4	water_scarcity (1.0)	drought (1.0)		heavy_rains (-0.5)	epidemic (-0.5)
Subtopic 5	flood (1.0)	heavy_rain (1.0)			
Subtopic 6	sea-level (1.0)	flood (1.0)			
Subtopic 7	blizzard (1.0)	ice_storm (1.0)		hurricane (-0.5)	hot_weather (-0.5)
Subtopic 8	melting_ice (1.0)	drought (1.0)			

Table 4.1: Seed words used in GTM to generate subtopics. The GTM algorithm takes as input a list of two or more seed word with positive weight. To further guide the topic in a desired direction, we also define negative seed words if needed.

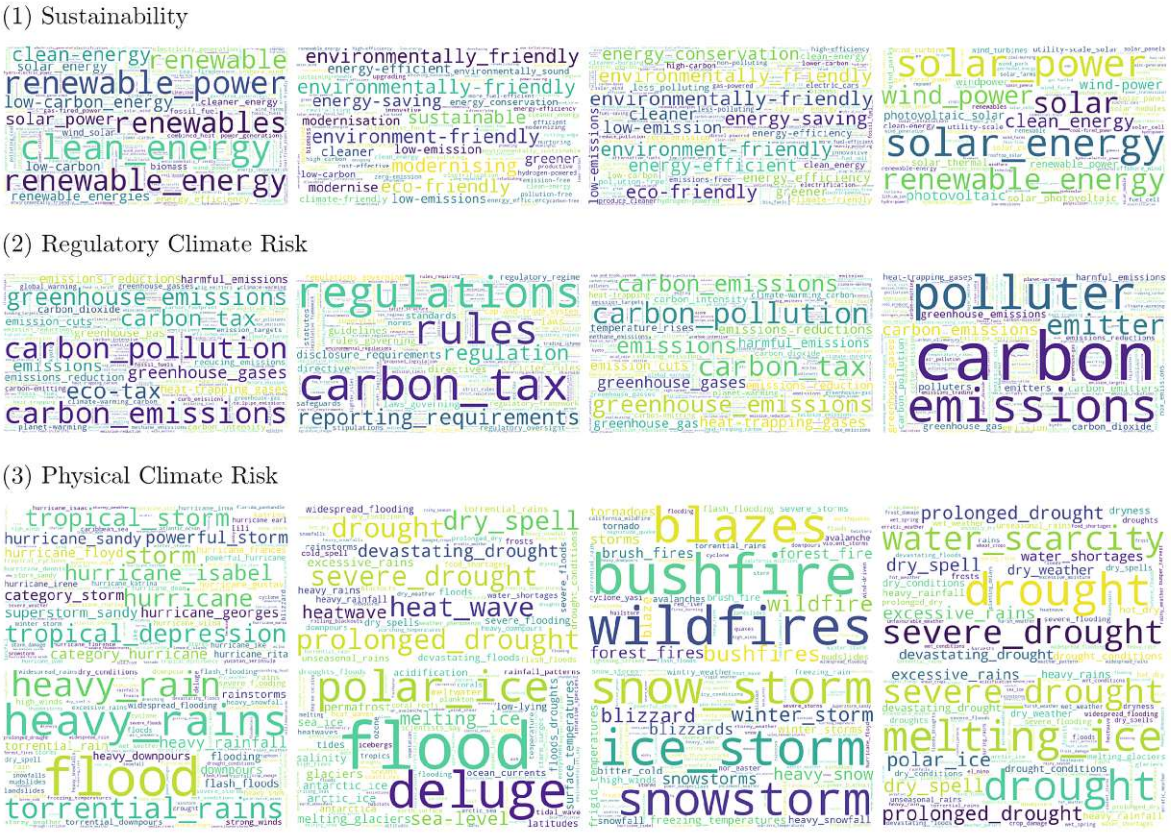


Figure 4.1: Subtopics generated with GTM and the seed words shown in Table 4.1. The font size of the words indicate the distance of each word to the topic center. Words close to the center appear large, more distant words are smaller.

4.4.3 News Indices

We calculate the coverage of these subtopics in news articles by counting the number of words in each article that overlap with the words contained in a topic. These counts are then multiplied by the weights, i.e., the importance of the words. We limit the count, i.e., the contribution of a single word in a subtopic, to a maximum value of 4 to avoid a high score driven exclusively by one or a few words. The sum of these weighted word counts finally gives us a score (loading) that indicates how strongly a topic is represented in a given news article. As each topic is comprised of multiple subtopics we calculate the final topic loading of a given news article as the mean of the individual subtopic loadings.

Scaling of the News Indices

The different news indices have different magnitudes as the words assigned to each topic occur with different frequencies in the text corpus. Topics composed of words that occur more frequently tend to have higher news exposures and thus higher values on average than topics containing less frequent words. In addition, news articles with a high word count, as opposed to short news articles, would receive a disproportionate topic loading, overstating the relevance of long news articles. To make the topic indices comparable, we adjust the topic loading of each news article with two adjustment parameters, g_{freq} and g_{len}. To calculate g_{freq}, we first count how often each word w_i appears in the corpus (c_i) and then calculate the average count $\bar{c}$ over all words contained in the vocabulary $V = \{w_1, w_2, w_3, ..., w_N\}$ (Equation (4.1)). Then, we calculate the average word count $\bar{c}_k$ over all words of topic C_k of size $|C_k|$ $\forall k \in \{1, 2, ..., K\}$ (Equation (4.2)). The adjustment parameter g_{freq} is then calculated according to Equation (4.3).

$$\bar{c} = \frac{\sum_{i=1}^{N} c_i}{N} \tag{4.1}$$

$$\bar{c}_k = \frac{\sum_{c_i \in C_k} c_i}{|C_k|} \tag{4.2}$$

$$g_{freq,k} = \frac{\bar{c}}{\bar{c}_k} \tag{4.3}$$

Next, we calculate the word count l_j of each news article $\mathcal{D}_j$ contained in the news corpus $\mathcal{D}$ as well as the average article length $\bar{l}$ over all news articles. The adjustment parameter g_{len} is calculated according to Equation (4.4). We use the logarithm to avoid over penalizing long news articles. Finally, the topic loading $L_{k,j}$ of topic k on news article j is scaled by Equation (4.5).

$$g_{len,j} = \frac{log(\bar{l})}{log(l_j)} \tag{4.4}$$

$$\tilde{L}_{k,j} = L_{k,j} \times g_{freq,k} \times g_{len,j} \tag{4.5}$$

Visualization of the News Indices

With the scaled topic loadings, calculated over the 4.95 million news articles, we can now visualize the intensities of the individual topics over time. We therefore aggregate the topic loadings at the monthly frequency by summing them up across all news articles published in a given month. In Figure 4.2 we plot the news index of the physical climate risk topic and in Figure 4.3 we plot the news indices of the topics regulatory climate risk and sustainability.

The physical climate risk news index (Figure 4.2) shows a clear seasonal pattern as it peaks mostly during the months August and September at hurricane season in the U.S. The extreme media coverage of severe hurricanes creates these large peaks that dominate the plot. Other natural disasters such as wildfires, snowstorms, or droughts do not receive as much media attention as hurricanes, making these events difficult to locate in this plot. In Figure 4.3 we plot the news indices of regulatory climate risk and sustainability. Regulatory climate risk usually peaks during month where a climate conference takes place. The largest peaks are caused by the Kyoto climate conference (COP 3) in 1997, COP 6-2 in Bonn, COP 13 in Bali, COP 15 in Copenhagen and COP 21 in Paris. Also, we observe that the initial relative low coverage of regulatory climate risk increases from 2005 to 2007. From then on, it fluctuates more or less around the same level. At the start of the pandemic in early 2020, both news indices show a sharp decline, as the media attention was centered towards the pandemic. However, later in 2020 we observe a sharp recovery in both indices that slightly exceeds the prior levels.

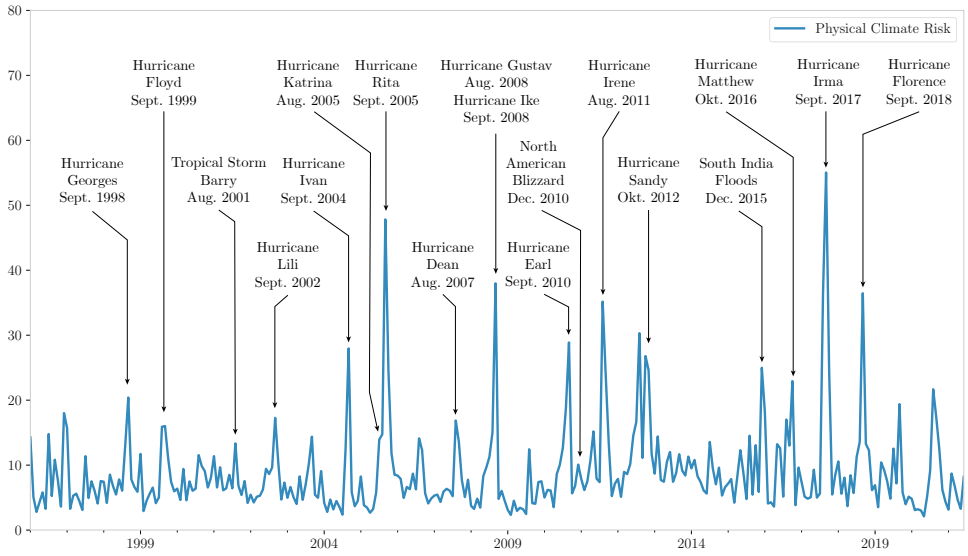


Figure 4.2: News index showing the monthly aggregate exposure of the topic Physical Climate Risk in Thomson Reuters news over the period Jan. 1996 to July 2021.

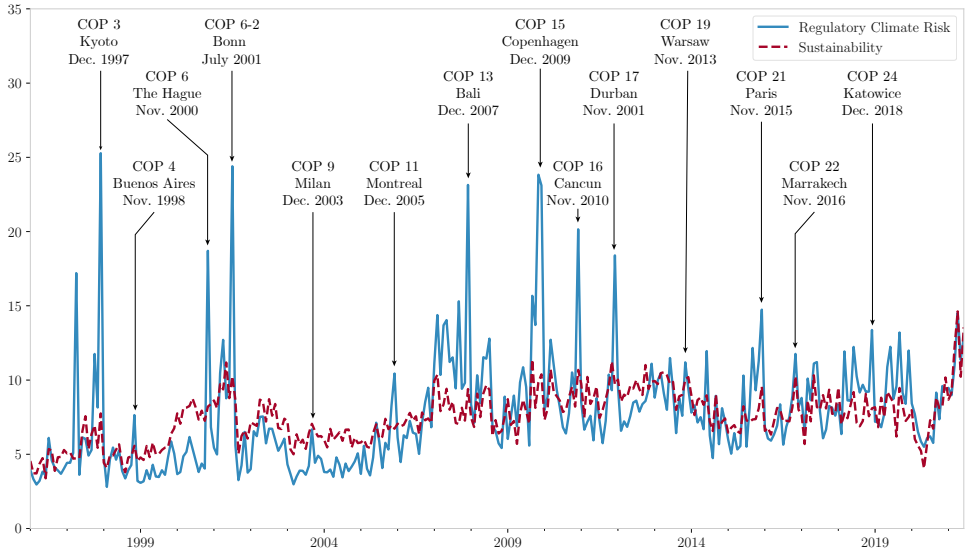


Figure 4.3: News index showing the monthly aggregate exposure of the topics Regulatory Climate Risk (Transition Risk) and Sustainability in Thomson Reuters news over the period Jan. 1996 to July 2021.

Company-specific News Indices

To obtain company specific news indices $I_{t,k,p}$, we aggregate topic loadings $\tilde{L}_{k,j}$ to daily scores by summing the loadings of all news documents $\mathcal{D}_{t,p}$ published on day t and attributable to firm p (Equation (4.6)). From the 4.95 million news articles associated with the U.S. geography code, we find 2.14 million news articles tagged with at least one ticker

code.

$$I_{t,k,p} = \sum_{j \in \mathcal{D}_{t,p}} \tilde{L}_{k,j} \tag{4.6}$$

Companies have different levels of coverage, as news about companies with high market capitalization and high media attention is published more frequently than news about companies with low market capitalization and low attention. As a consequence, these large, high attention firms receive much higher loadings on the topic indices. By selecting the firms with the highest topic loading we would not necessarily select the firms with the highest climate risk or the most sustainable firms, but also the firms with the highest media attention. Therefore, we adjust the daily topic index at the individual firm level by the number of news articles published for each company on a given day $|\mathcal{D}_{t,p}|$ according to Equation (4.7).

$$\bar{I}_{t,k,p} = \frac{I_{t,k,p}}{|\mathcal{D}_{t,p}|} \tag{4.7}$$

Sustainability and transition risk encompass concepts that are often discussed in the news together. Transition risks arise as companies transition towards lower environmental footprints. Thus, discussions about this transition can give both green and brown companies a loading on these opposing topics. Experiments show that this can lead to an inaccurate classification of green and brown companies. To improve the accuracy of our methodology, we first classify news articles into the categories *regulatory climate risk news* and *sustainability news*. As we know the loadings of each news article on all topics, we use this information for the classification. We define a threshold parameter γ that we set to 1.5 and classify a news article as *regulatory climate risk news* if the loading on the regulatory climate risk topic (k=2) is greater than the loading on the sustainability topic

(k=1) multiplied by γ.[7] Analogously we classify a news article as *sustainability news* if the loading on the sustainability topic (k=1) is greater than the loading on the regulatory climate risk topic (k=2) multiplied by γ (Equation (4.8)). With this classification we obtain 254.317 *regulatory climate risk news* and 414.391 *sustainability news*. Also, we find 89.750 firm-specific news that load on physical climate risk. News articles about physical climate risks are usually explicit, which is why there is no need to classify them into physical climate risk news in advance. Note that all these news articles can be tagged with more than one firm, since news often affects multiple firms simultaneously.

$$Category = \left\{ \begin{array}{ll} \textit{regulatory climate risk news}, & \text{for } \tilde{L}_{k=2,j} > \gamma \times \tilde{L}_{k=1,j} \\ \textit{sustainability news}, & \text{for } \tilde{L}_{k=1,j} > \gamma \times \tilde{L}_{k=2,j} \end{array} \right\} \tag{4.8}$$

4.5 Results

In this Section, we report our main empirical results. Given the multi-step approach that we require to extract signals from news and relate them to equity returns, we proceed as follows. In Section 4.5.1, we first document the firms and industries most exposed to regulatory climate risk, physical climate risk and sustainability. We do this according to our news-based methodology in order to establish the validity of our identification strategy. In Sections 4.5.2 and 4.5.3, we use firm-specific, news-based topic exposures to assess whether climate-related risks are priced in equity markets. In Section 4.5.4, we extend the sample of firms beyond those explicitly covered in the news by calculating climate-risk related betas and rerun our empirical asset pricing tests. Finally, in Section 4.5.5 we provide additional results validating our main results.

[7] In Dangl and Salbrechter (2023) we show that increasing values of γ are associated with higher classification accuracies. We choose the value of 1.5 to balance the trade-off between improved accuracy and reduced sample size.

4.5.1 Firm-specific Topic Exposure

We calculate firm-specific topic exposures by summing up the loadings $\bar{I}_{t,k,p}$ over the period Jan. 1996 to Dec. 2020 for the topics (a) regulatory climate risk, (b) physical climate risk and (c) sustainability. As explained before, we consider stocks of the CRSP universe with share codes 10 or 11 that have a market cap above the median market cap and we exclude Depository, Credit and Brokerage institutions (SIC sector codes: 60, 61, 62, 66, 67). We exclude these financial companies because they often appear in the news metadata as they provide analyst ratings and reports without being the actual subject of the news. Thus, by including these firms we would significantly overestimate their true risk exposures. Table 4.2 highlights the 30 firms with the highest exposures. We observe that the firms listed in (a) are predominantly energy producers or firms related to the Oil & Gas industry. Companies in (b) that are exposed to physical climate risk include also energy producers as well as insurance and food companies. This result seems intuitive, as insurance and food companies are heavily affected by natural disasters and extreme weather events. Companies shown in (c) include solar and renewable energy companies, all of which focus on sustainability in their business models.

Next, we sort firms into major industry groups based on the first two digits of the SIC codes and calculate industry exposures by adding up the exposures of all firms within an industry. Since the number of companies N_i varies considerably across industries, large industries consisting of many companies would be biased to have larger exposures. We therefore adjust the industry exposures by dividing the sum of individual firm-level exposures by $log(N_i) + 1$, using the logarithm to avoid overly penalizing large industries.

Table 4.3 highlights the industries ranked by their adjusted news exposure. We find that *Electric, Gas, And Sanitary Services*, *Coal Mining* and *Petroleum Refining And Related Industries* have the highest exposure to regulatory climate risk, as one would expect. For physical climate risk, we find that the *Insurance Carriers* industry has the second highest exposure and the *Food And Kindred Products* industry has the fourth highest exposure. The high exposures of these industries seem plausible given that both are highly exposed to the risk of damages caused by natural disasters. Furthermore, the industries

	(a) Reg. Climate Risk	(b) Phys. Climate Risk	(c) Sustainability
0	Arch Coal Inc	Allstate Corp	First Solar Inc
1	American Electric Power Co Inc	Travelers Companies Inc	Sunpower Corp
2	CNX Resources Corp	PG & E Corp	Canadian Solar Inc
3	Peabody Energy Corp	Consolidated Edison Inc	Clean Energy Fuels Corp
4	Southern Co	Entergy Corp New	Yingli Green Energy Hldg Co Ltd
5	Massey Energy Co	Centerpoint Energy Inc	Trina Solar Limited
6	James River Coal Co	American Electric Power Co Inc	Suntech Power Holdings Co Ltd
7	Cinergy Corp	Dominion Energy Inc	JA Solar Holdings Co Ltd
8	Firstenergy Corp	Chubb Ltd	Jinkosolar Holding Co Ltd
9	Alpha Natural Resources Inc	Chubb Corp	Plug Power Inc
10	Duke Energy Corp New	Hartford Financial Svcs Grp Inc	Green Plains Inc
11	TECO Energy Inc	Duke Energy Corp New	Nextera Energy Inc
12	Cummins Inc	Nextera Energy Inc	Fuelcell Energy Inc
13	Walter Energy Inc	Progress Energy Inc	Energy Conversion Devices Inc
14	NRG Energy Inc	Anadarko Petroleum Corp	Edison International
15	XCEL Energy Inc	Exelon Corp	Hanwha Q Cells Co Ltd
16	Cloud Peak Energy Inc	Eversource Energy	Renesola Ltd
17	Union Pacific Corp	Edison International	Sempra Energy
18	CSX Corp	Archer Daniels Midland Co	PG & E Corp
19	Exelon Corp	Southern Co	LDK Solar Co Ltd
20	CVR Energy	Firstenergy Corp	Sunedison Inc
21	Norfolk Southern Corp	Bunge Ltd	Tesla Inc
22	Valero Energy Corp New	Progressive Corp Oh	AES Corp
23	Public Service Enterprise Gp Inc	OGE Energy Corp	Evergreen Solar Inc
24	Dominion Energy Inc	Everest Re Group Ltd	Enphase Energy Inc
25	Navistar International Corp	Valero Energy Corp New	China Sunergy Co Ltd
26	International Coal Group Inc	APA Corp	Duke Energy Corp New
27	DTE Energy Co	Cincinnati Financial Corp	NRG Energy Inc
28	Nv Energy Inc	Murphy Oil Corp	General Electric Co
29	Pinnacle West Capital Corp	Union Pacific Corp	Ormat Technologies Inc

Table 4.2: Top 30 companies ranked by their topic exposure to (a) regulatory climate risk, (b) physical climate risk and (c) sustainability. The topic exposure is calculated over the period Jan 1996 to Dec. 2020.

most exposed to the Sustainability topic are *Electronic And Other Electrical Equipment*, *Electric Gas And Sanitary Services*, *Chemicals And Allied Products* and *Business Services*. Interestingly, we observe an overlap of some Industries (e.g. *Electric Gas And Sanitary Services*, *Oil And Gas Extraction*, *Chemicals And Allied Products*, *Transportation Equipment*) having a pronounced exposure to all three topics. One explanation of this finding is that firms within these industries exhibit a particularly high variation of firm-specific climate risk exposures.

	(a) Reg. Climate Risk	Exposure	(b) Phys. Climate Risk	Exposure	(c) Sustainability	Exposure
0	Electric, Gas, And Sanitary Services	2069.45	Electric, Gas, And Sanitary Services	1585.48	Electronic And Other Electrical Equipment And ...	3531.45
1	Coal Mining	1088.46	Insurance Carriers	891.04	Electric, Gas, And Sanitary Services	2868.70
2	Petroleum Refining And Related Industries	504.30	Oil And Gas Extraction	580.54	Chemicals And Allied Products	1497.82
3	Oil And Gas Extraction	414.17	Food And Kindred Products	348.00	Business Services	1469.73
4	Chemicals And Allied Products	392.45	Chemicals And Allied Products	308.21	Oil And Gas Extraction	1413.57
5	Transportation Equipment	365.36	Transportation By Air	277.27	Industrial And Commercial Machinery And Comput...	1071.44
6	Railroad Transportation	295.15	Petroleum Refining And Related Industries	265.14	Transportation Equipment	764.44
7	Industrial And Commercial Machinery And Comput...	246.10	Apparel And Accessory Stores	235.66	Measuring, Analyzing, And Controlling Instrume...	654.41
8	Food And Kindred Products	232.01	Electronic And Other Electrical Equipment And ...	204.78	Nonclassifiable Establishments	592.12
9	Primary Metal Industries	227.50	Industrial And Commercial Machinery And Comput...	197.51	Communications	580.67
10	Business Services	203.65	General Merchandise Stores	194.10	Engineering, Accounting, Research, Management,...	527.05
11	Water Transportation	181.21	Business Services	186.84	Food And Kindred Products	381.80
12	Communications	175.60	Insurance Agents, Brokers, And Service	181.89	Apparel And Accessory Stores	348.05
13	Insurance Carriers	174.79	Railroad Transportation	180.08	Insurance Carriers	313.66
14	Electronic And Other Electrical Equipment And ...	131.56	Transportation Equipment	165.65	Petroleum Refining And Related Industries	262.34
15	Transportation By Air	123.38	Automotive Dealers And Gasoline Service Stations	147.53	Miscellaneous Retail	227.66
16	Metal Mining	115.07	Communications	147.47	Primary Metal Industries	225.23
17	Measuring, Analyzing, And Controlling Instrume...	86.17	Eating And Drinking Places	131.91	General Merchandise Stores	200.58
18	Fabricated Metal Products, Except Machinery An...	79.72	Miscellaneous Retail	114.35	Fabricated Metal Products, Except Machinery An...	193.97
19	Automotive Dealers And Gasoline Service Stations	74.63	Pipelines, Except Natural Gas	109.67	Transportation By Air	193.16

Table 4.3: Top 20 industries with the highest exposure to (a) regulatory climate risk, (b) physical climate risk and (c) sustainability. The topic exposure is calculated over the period Jan. 1996 to Dec. 2020.

To better understand the time-series variation, we plot the firm-specific news exposures over the period Jan. 1999 to Dec. 2020 for the firms Allstate (ALL), Apple (AAPL), Exxon Mobil (XOM), Procter&Gamble (PG), Union Pacific (UNP) and Nextera Energy (NEE) in Figure 4.4 and 4.5. We smooth the monthly observations over two-year rolling windows using arithmetic means.

Figure 4.4 shows the firm-specific exposure to physical climate risk. As expected, we observe a high exposure for the insurance company Allstate (blue line). Exxon Mobil, Union Pacific and Nextera Energy also have significant exposures throughout the sample period. In contrast, the physical climate risk of Apple and Procter&Gamble is close to zero (red and green lines). In Figure 4.5 we subtract the exposure to regulatory risk from the exposure to Sustainability. Put differently, we calculate a measure of green-minus-brown exposure. Positive values indicate green firms, negative values indicate brown firms. We observe that Nextra Energy is a sustainable firm according to this measure, as it has the largest positive exposure. Also, Apple, Procter&Gamble and Allstate have positive exposures on average, but of smaller magnitudes. Union Pacific and Exxon Mobil both show strong negative exposures. We would, thus, classify them as being strongly exposed to regulatory risk.

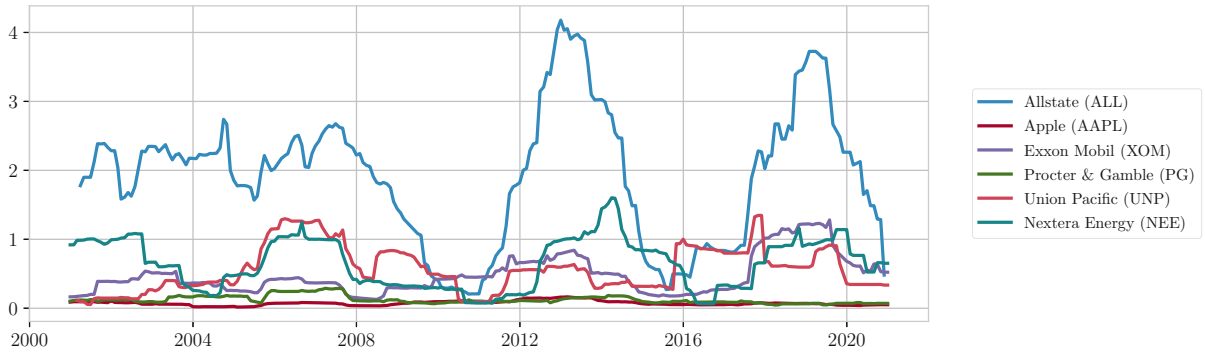


Figure 4.4: Firm-specific news indices showing the exposure of individual firms to physical climate risk over the period Jan. 1999 to Dec. 2020. We smooth the values by calculating the mean over a two-year rolling window.

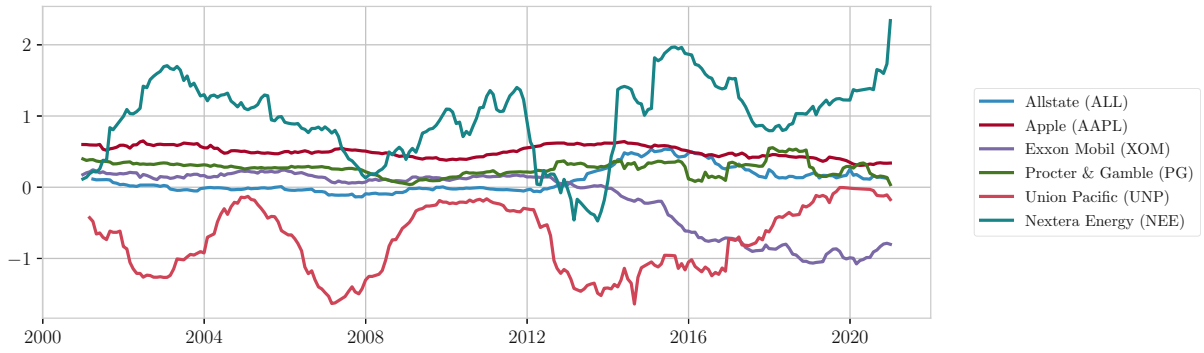


Figure 4.5: We subtract the firm-specific exposure to transition risk from the exposure to sustainability after calculating means over two-year rolling windows (green-minus-brown). The Figure shows the green-minus-brown news indices for individual firms. Positive (negative) values indicate a high exposure to sustainability (transition risk) relative to the exposure to transition risk (sustainability).

4.5.2 Topic Exposure Sorted Climate Risk Portfolios

To assess the return implications of our news-based exposures, we form zero-investment portfolios. We identify green and brown firms based on their exposures to the topics sustainability (k=1) and regulatory climate risk (k=2). To determine the constituents of this green-minus-brown (GMB) portfolio, we measure each firms' topic exposure, denoted as $\bar{E}_{t,k,p}$, over a 24-month rolling window in accordance with Equation (4.9). Since we will use these exposures also to determine portfolio weights and their distribution across firms can be highly skewed, we apply a log-transformation.

In a next step we calculate measures of relative exposure between green and brown stocks by subtracting the exposures to regulatory climate risk from the exposures to sustainability at the individual firm level ($\tilde{E}_{t,p}$). After ranking the firms from high to low we form portfolios by putting stocks in the top decile into the "Green" portfolio and stocks in the bottom decile into the "Brown" portfolio. We then weight firms, within these portfolios, relative to their topic index exposures. The portfolio weights are calculated according to Equation (4.10) with the topic exposures of top (bottom) decile stocks $\tilde{E}^{top}_{t,p}$ ($\tilde{E}^{bottom}_{t,p}$). We study monthly returns and re-balance at the end of each month.

$$\bar{E}_{t,k,p} = \log\left(\sum_{t-25}^{t-1} \bar{I}_{t,k,p}\right) \tag{4.9}$$

$$w_{t,p,green} = \frac{\tilde{E}^{top}_{t,p}}{\sum_{p} \tilde{E}^{top}_{t,p}} \qquad \text{and} \qquad w_{t,p,brown} = \frac{\tilde{E}^{bottom}_{t,p}}{\sum_{p} \tilde{E}^{bottom}_{t,p}} \tag{4.10}$$

Similar to the GMB portfolio, we also calculate a topic-exposure weighted long-only physical climate risk portfolio that includes the top decile of firms ranked by their news-based exposures to physical climate risk ($\bar{E}_{t,k=3,p}$) with weights calculated according to Equation (4.11). In addition, we construct a long-short, high-minus-low physical climate risk portfolio, denoted as PhysCR, where we go long in the topic-exposure weighted portfolio of stocks with high physical climate risk and short in all remaining stocks, i.e., all stocks with no exposure to physical climate risk over the previous 24-month rolling window period. Since these stocks are equal in regard to their topic exposure, we weight them equally.

$$w_{t,p,phys} = \frac{\bar{E}^{top}_{t,k=3,p}}{\sum_{p} \bar{E}^{top}_{t,k=3,p}} \tag{4.11}$$

Figure 4.6a shows the cumulative excess returns (in excess of the risk-free rate) of the "Green" and "Brown" portfolios, and of the portfolios with high and low physical climate risk over the period Jan. 2002 to Dec. 2020. Figure 4.6b displays the performance of the GMB and PhysCR portfolio over the same period. We observe that green stocks underperform brown stocks until 2011 (blue line). In 2011, the trend changes and from 2012 to 2014, we observe a strong outperformance of green over brown stocks. This upward trend, however, is interrupted during the four-year period 2015 to 2018. In 2015, green companies record a negative annual return, while the brown stock portfolio closes more or less at the same level as at the beginning of the year. In the years that follow, brown firms experience windfall gains due to the Trump administration's policies, leading to a further decline of the news-exposure-weighted GMB portfolio. From 2018, green stocks continue

their upward trend until the end of the observation period.

For the high-minus-low physical climate risk portfolio (PhysCR) in Figure 4.6b (brown line) we observe a steady upward trend from 2010 to 2019, indicating a stronger performance for the portfolio of stocks with exposure to physical climate risks compared to the equal weighted portfolio of stocks with low physical climate risk exposure. This is accompanied by two drawdowns: in 2009, after the great financial crisis and in 2020, after the coronavirus crash. In both cases, the drawdown is caused by a quicker recovery of the low physical climate risk portfolio relative to the high physical climate risk portfolio (see Figure 4.6a). One explanation could be that, at least during the 2020 pandemic, the demand and prices of internet firms, which tend to be firms with low physical climate risk, surged and thereby contributed to the weak relative performance of firms with high physical climate risk.

(a)

(b)

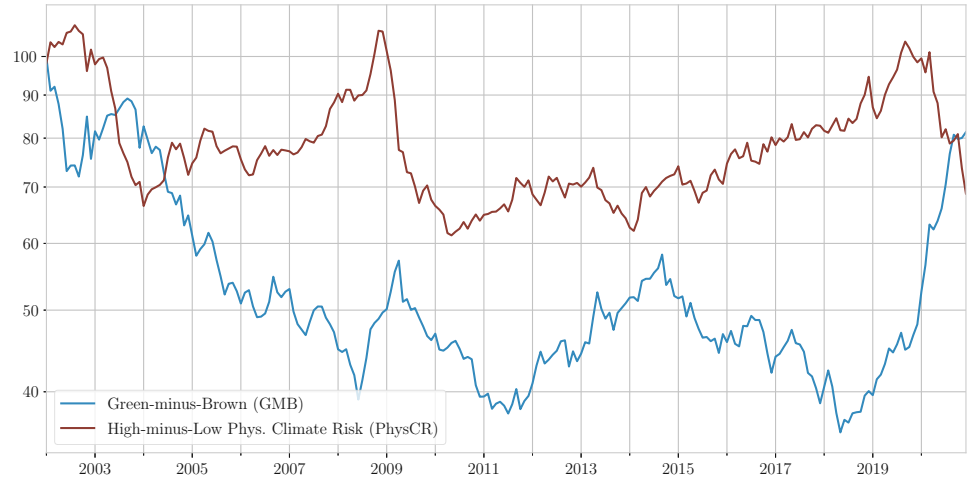


Figure 4.6: Cumulative returns of the topic-exposure weighted (a) "Green" (sustainability), "Brown" (regulatory climate risk), high- and low physical climate risk portfolio and (b) the green-minus-brown (GMB) and high-minus-low physical climate risk (PhysCR) portfolio over the period from Jan. 2002 to Dec. 2020.

Table 4.4 reports the summary statistics of the portfolio returns for the periods Jan. 2002 to Dec. 2020 (Panel A), Jan. 2002 to Dec. 2011 (Panel B), and Jan. 2012 to Dec. 2020 (Panel C). Over the full period (Panel A) the "High Physical Climate Risk Portfolio" has a similar cumulative performance as the "Green Portfolio" with values of 534.52% and 471.29% respectively. Thus, both portfolios outperform the value-weighted market portfolio with a cumulative performance of 329.25%. On a risk-adjusted basis we observe the highest Sharpe-Ratio for the "High Physical Climate Risk Portfolio" followed by the

market portfolio and the "Green Portfolio" with values of 0.64, 0.52 and 0.48, respectively. Overall, the equal-weighted "Low Physical Climate Risk Portfolio" has the highest cumulative performance over the full period (558.59%), which is consistent with existing evidence that the 1/N portfolio is a hard benchmark to beat (DeMiguel et al., 2009). Over the full period the GMB portfolio has a cumulative return of -18.67%. In Panel B, the GMB portfolio has cumulative returns of -60.45% and a CAGR of -8.93% while during the second subperiod, starting in 2012 (Panel C), the GMB portfolio has a performance of 105.65% with an CAGR of 8.34% and a Sharpe-Ratio of 0.64. Also, the "High Physical Climate Risk Portfolio" outperforms the "Low Physical Climate Risk Portfolio" over the period 2012 to 2020 in absolute and risk-adjusted performance with Sharpe-Ratios of 1.08 vs. 0.67.

The average portfolio sizes indicate a substantially higher number of constituents in the "Green" portfolio relative to the "Brown" and the "high physical climate risk" portfolios (see Table 4.4). This is the case since we only consider stocks with meaningful exposures to these topics in the rolling window aggregation of news and, therefore, exclude firms with topic exposures below a pre-defined threshold.

Portfolio	Performance (%)	CAGR (%)	SD (%)	Sharpe Ratio	Drawdown (%)	Drawdown (months)	Avg. Portf. Size	CAPM Beta	CAPM Beta (t-value)
Panel A: Full Period, Jan. 2002 to Dec. 2020									
Green (Sustainability)	471.29	9.61	19.82	0.48	-50.83	65.0	174.32	1.23	45.89
Brown (Reg. Climate Risk)	409.14	8.94	23.09	0.39	-59.65	33.0	65.52	1.25	21.97
GMB	-18.67	-1.08	13.71	-0.08	-64.07	228.0		-0.01	-0.24
High Physical Climate Risk	534.52	10.21	15.85	0.64	-50.99	46.0	54.39	0.94	32.10
Low Physical Climate Risk	558.59	10.43	21.87	0.48	-60.71	43.0	3256.02	1.30	33.15
Mkt-Rf	329.25	7.97	15.28	0.52	-51.44	54.0		1.00	inf
Panel B: Period Jan. 2002 to Dec. 2011									
Green (Sustainability)	32.99	2.92	21.63	0.13	-50.83	51.0	188.82	1.26	32.16
Brown (Reg. Climate Risk)	186.07	11.18	24.01	0.47	-59.65	33.0	57.13	1.25	17.36
GMB	-60.45	-8.93	13.86	-0.64	-62.17	120.0		0.01	0.11
High Physical Climate Risk	83.03	6.29	17.52	0.36	-50.99	46.0	49.88	1.00	26.99
Low Physical Climate Risk	112.77	7.91	23.50	0.34	-60.71	43.0	3587.71	1.33	25.06
Mkt-Rf	20.63	1.91	16.22	0.12	-51.44	51.0		1.00	inf
Panel C: Period Jan. 2012 to Dec. 2020									
Green (Sustainability)	329.57	17.58	17.51	1.00	-21.70	6.0	158.20	1.19	33.28
Brown (Reg. Climate Risk)	77.98	6.61	22.12	0.30	-52.47	29.0	74.84	1.29	14.55
GMB	105.65	8.34	13.12	0.64	-38.49	67.0		-0.10	-1.05
High Physical Climate Risk	246.67	14.81	13.76	1.08	-25.15	11.0	59.42	0.86	18.41
Low Physical Climate Risk	209.54	13.38	19.99	0.67	-34.59	25.0	2887.47	1.30	22.33
Mkt-Rf	255.85	15.15	14.01	1.08	-20.48	8.0		1.00	inf

Table 4.4: Portfolio statistics calculated over the periods Jan. 2002 to Dec. 2020 (Panel A), Jan. 2002 to Dec. 2011 (Panel B) and Jan. 2012 to Dec. 2020 (Panel C)

Table 4.5 shows the average weights of the top 30 holdings of the "Green" and "Brown" portfolio calculated over the period Jan. 2002 to Dec. 2020. Note that due to averaging

over time some firms appear in both portfolios. Firms, however, can only be in one portfolio at a specific point in time.

For the regulatory climate risk ("Brown") portfolio the largest holdings are Oil&Gas firms like Exxon Mobil and Chevron, energy suppliers such as American Electric Power and Duke Energy Corp as well as railway companies such as Union Pacific. Companies involved in the energy business, such as Consolidated Edison or PG&E, are also highly exposed to physical climate risk, along with insurance companies such as Allstate and Travelers Companies. The largest holdings of the sustainability ("Green") portfolio are technology firms like Apple, Microsoft and Alphabet as well as energy companies such as First Solar and Sunpower and telecommunication firms like AT&T and Verizon Communications.

	Regulatory Climate Risk (Brown) Portfolio			Physical Climate Risk Portfolio			Sustainability (Green) Portfolio		
	Weight (%)		Company Name	Weight (%)		Company Name	Weight (%)		Company Name
0	3.07	3.07	American Electric Power Co Inc	2.36	2.36	Allstate Corp	1.42	1.42	General Electric Co
1	2.71	5.78	CNX Resources Corp	1.85	4.21	Entergy Corp New	1.38	2.80	Apple Inc
2	2.44	8.22	Southern Co	1.73	5.94	Consolidated Edison Inc	1.36	4.16	Microsoft Corp
3	2.35	10.57	Arch Coal Inc	1.71	7.65	Travelers Companies Inc	1.28	5.44	Intel Corp
4	2.31	12.88	Peabody Energy Corp	1.62	9.27	Hartford Financial Svcs Grp Inc	1.19	6.64	Alphabet Inc
5	2.27	15.15	Valero Energy Corp New	1.53	10.80	Anadarko Petroleum Corp	1.18	7.82	International Business Machs Cor
6	1.91	17.06	Massey Energy Co	1.53	12.33	Centerpoint Energy Inc	1.11	8.93	Boeing Co
7	1.60	18.66	CSX Corp	1.41	13.74	APA Corp	1.08	10.01	Lockheed Martin Corp
8	1.53	20.19	Exxon Mobil Corp	1.35	15.10	Dominion Energy Inc	1.07	11.08	Cisco Systems Inc
9	1.45	21.65	Union Pacific Corp	1.29	16.38	PG & E Corp	1.02	12.10	First Solar Inc
10	1.45	23.09	Firstenergy Corp	1.22	17.61	Valero Energy Corp New	0.98	13.07	Walmart Inc
11	1.36	24.45	Cummins Inc	1.19	18.79	American Electric Power Co Inc	0.98	14.05	Qualcomm Inc
12	1.32	25.77	United States Steel Corp New	1.17	19.97	Duke Energy Corp New	0.97	15.02	Amazon Com Inc
13	1.27	27.04	Duke Energy Corp New	1.14	21.10	Nextera Energy Inc	0.95	15.96	HP Inc
14	1.26	28.30	Norfolk Southern Corp	1.12	22.23	Union Pacific Corp	0.91	16.87	Sunpower Corp
15	1.14	29.44	Chevron Corp New	1.02	23.25	Cincinnati Financial Corp	0.91	17.78	Pfizer Inc
16	1.14	30.58	Cinergy Corp	1.02	24.27	Jetblue Airways Corp	0.89	18.67	Oracle Corp
17	1.13	31.71	XCEL Energy Inc	1.02	25.29	OGE Energy Corp	0.86	19.53	Merck & Co Inc New
18	1.12	32.83	Phillips 66	1.01	26.30	Southern Co	0.84	20.38	AT & T Inc
19	1.10	33.93	NRG Energy Inc	1.01	27.30	Exxon Mobil Corp	0.83	21.21	Verizon Communications Inc
20	1.07	35.00	Marathon Petroleum Corp	0.98	28.29	CSX Corp	0.83	22.04	Tesla Inc
21	1.04	36.04	Exelon Corp	0.97	29.25	Chubb Corp	0.81	22.84	Raytheon Technologies Corp
22	1.03	37.07	Archer Daniels Midland Co	0.96	30.21	Murphy Oil Corp	0.79	23.63	Johnson & Johnson
23	1.01	38.09	AK Steel Holding Corp	0.95	31.16	Exelon Corp	0.77	24.40	Advanced Micro Devices Inc
24	1.00	39.09	Newmont Corp	0.95	32.11	Archer Daniels Midland Co	0.76	25.16	PG & E Corp
25	0.93	40.02	Du Pont EI De Nemours & Co	0.94	33.05	Conocophillips	0.75	25.91	Nextera Energy Inc
26	0.92	40.94	Hollyfrontier Corp	0.91	33.96	Tyson Foods Inc	0.73	26.65	Northrop Grumman Corp
27	0.86	41.79	Nucor Corp	0.91	34.87	Marathon Oil Corp	0.70	27.35	Procter & Gamble Co
28	0.85	42.65	Burlington Northern Santa Fe Cp	0.90	35.77	Progress Energy Inc	0.70	28.05	Sempra Energy
29	0.84	43.49	Vectren Corp	0.90	36.66	Aon Plc New	0.70	28.74	Yahoo Inc

Table 4.5: Top 30 companies of the regulatory climate risk (Brown), physical climate risk and sustainability (Green) portfolios. The weights are averages in %, calculated over the period Jan. 2002 to Dec. 2020.

Next, we calculate correlations between the climate-related risk factors, i.e., the green-minus-brown (GMB) portfolio and the high-minus-low physical climate risk portfolio (PhysCR), and the standard Fama-French risk factors, i.e., the market factor (Mkt-Rf), size factor (SMB), value factor (HML), profitability factor (RMW), the investment factor (CMA) and the momentum factor (UMD), using monthly returns over the period Jan. 2002 to Dec. 2020 (Table 4.6, Panel A) and Jan. 2012 to Dec. 2020 (Table 4.6, Panel B). In

addition, we include individual climate risk related portfolios also as long-only portfolios in the analysis.

For Panel A, we observe a negative correlation between GMB and HML of -0.30, suggesting that green stocks are rather growth stocks and brown stocks are rather value stocks. Also, green stocks exhibit weaker operating profitability than brown stocks since the correlation of GMB with RMW is -0.255. This changes for the period starting in 2012 (Panel B) where the correlation with RMW becomes slightly positive. Moreover, the correlations with CMA, SMB and HML become increasingly negative. Thus, green firms tend to be large, aggressively investing growth stocks while brown firms tend to be small, conservatively investing value stocks. Pástor et al. (2022) show that the brown nature of value stocks has a significant contribution to the poor performance of the value strategy in recent years, just as the green nature of momentum stocks explains the positive performance of the momentum strategy experienced during the most recent period.

The high physical climate risk portfolio (High Phys. CR) has positive correlation coefficients with SMB and HML, with values of 0.482 and 0.366, respectively, indicating that firms exposed to physical climate risks are rather small value stocks. The correlations with CMA, RMW and GMB are 0.10, -0.297 and -0.194. For the period staring in 2012 the correlation of the high physical climate risk portfolio with GMB becomes -0.303 which indicates that brown firms tend to be more affected by physical climate risks than green firms. Furthermore, the GMB portfolio and the high-minus-low physical climate risk portfolio (PhysCR) are slightly negatively correlated with a coefficient of -0.16.

	PhysCR	Green	Brown	High Phys. CR	Low Phys. CR	Mkt-RF	SMB	HML	CMA	RMW	UMD
Panel A: Full Period, July 2002 to July 2021											
GMB	-0.160	0.088	-0.518	-0.194	-0.057	-0.016	-0.256	-0.300	-0.061	-0.255	-0.144
PhysCR		-0.579	-0.402	-0.268	-0.716	-0.488	-0.575	-0.013	0.025	0.519	0.407
Green			0.806	0.855	0.922	0.950	0.421	0.163	-0.005	-0.485	-0.535
Brown				0.850	0.825	0.825	0.513	0.318	0.032	-0.265	-0.374
High Phys. CR					0.865	0.906	0.482	0.366	0.100	-0.297	-0.446
Low Phys. CR						0.911	0.649	0.272	0.059	-0.485	-0.535
Mkt-RF							0.396	0.232	0.000	-0.389	-0.478
SMB								0.360	0.155	-0.280	-0.198
HML									0.433	0.004	-0.343
CMA										-0.070	-0.119
RMW											0.313
Panel B: Period July 2012 to July 2021											
GMB	-0.025	-0.023	-0.612	-0.303	-0.195	-0.102	-0.454	-0.481	-0.258	0.051	0.134
PhysCR		-0.635	-0.488	-0.264	-0.748	-0.543	-0.630	0.018	0.253	0.390	0.422
Green			0.805	0.823	0.926	0.955	0.468	0.195	-0.143	-0.048	-0.566
Brown				0.831	0.848	0.816	0.640	0.439	0.039	-0.069	-0.528
High Phys. CR					0.838	0.873	0.494	0.400	0.041	0.069	-0.527
Low Phys. CR						0.908	0.697	0.265	-0.115	-0.173	-0.602
Mkt-RF							0.420	0.216	-0.140	0.024	-0.505
SMB								0.360	0.006	-0.326	-0.432
HML									0.508	0.070	-0.506
CMA										0.106	-0.192
RMW											-0.065

Table 4.6: Correlations among risk factors calculated using monthly returns over the period Jan. 2002 to Dec. 2020 (Panel A) and the period Jan. 2012 to Dec. 2020 (Panel B). "GMB" denotes the green-minus-brown portfolio and "PhysCR" denotes the high-minus-low physical climate risk portfolio.

4.5.3 Climate Risk Premia

To more systematically assess the return implications of climate-risk related news exposures, we perform Fama-MacBeth cross-sectional regressions (see, Equation (4.12)) with firm specific characteristics, i.e., the firm-specific exposures to the topics regulatory climate risk (Reg), physical climate risk (Phys) and sustainability (Sus).

In addition, we consider firm characteristics standard in the empirical asset pricing literature, namely CAPM beta, size, as measured by the log of market capitalization (*log_mktcap*), book-to-market ratio (*B2M*), operating profitability (*OP*), and investment (*INV*) as explanatory variables.[8] Equation 4.12 shows the details of the Fama-MacBeth regression. $R_{p,t} - R_{f,t}$ is the return of stock p in month t in excess of the risk free rate $R_{f,t}$. We load corporate financial data from Compustat and calculate B2M, OP and INV as described by (Fama and French, 1992, 2015). To avoid a look ahead bias we calculate all metrics as at the end of July using data from the previous fiscal year.

[8]The CAPM beta is estimated over a rolling window of 60 months of monthly return data. In case of missing values we calculate betas if at least 36 months of return data are available.

$$R_{p,t} - R_{f,t} = \delta_{0,t} + \delta_{1,t}\beta_{p,t-1} + \delta_{2,t}Size_{p,t-1} + \delta_{3,t}B2M_{p,t-1} + \delta_{4,t}OP_{p,t-1} + \delta_{5,t}INV_{p,t-1} + \delta_{6,t}Sus_{p,t} + \delta_{7,t}Reg_{p,t} + \delta_{8,t}Phys_{p,t} + \epsilon_{p,t} \tag{4.12}$$

We report the results of the Fama-MacBeth regressions in Table 4.7 showing the annualized risk premia with the corresponding t-statistics. In total, we estimate seven models, starting with Model 1 that focuses on the classic Fama and French five-factor model. We estimate Model 1 on the full CRSP universe, only excluding penny stocks, with 810.766 observations (Model 1a). The purpose of this model is to establish a first and very general basecase result using the standard controls.

We then rerun the basecase specification on the smaller sample of firms, for which we have news-based climate risk exposures (Model 1b), giving us 189.586 monthly observations. Model 1b acts as the main baseline for comparison with Model 5 that contains all climate characteristics, as well as models 6 and 7 where we further control for fixed effects using 10 sector dummies (Model 6) or 65 industry dummies (Model 7). Models 2 to 4 are univariate extensions of Model 1a where we add each climate risk related exposure individually. We omit those results for the sake of brevity (see the full Table 17 in the Appendix).

Adding the news-based climate risk related factors to the baseline model, i.e., when comparing Model 5 to Model 1b, we observe an increase in adjusted R^2 from 5.43% to 6.26% (6.01% to 7.10% for the first- and 4.78% to 5.32% for the second subperiod). In relative terms, this means an increase of 15.3% for the full period, 18.1% in the first- and 11.3% in the second subperiod. Considering Model 5, we find a significant positive risk premium of 1.50% p.a. (t-value = 2.37) for physical climate risk over the full period (Jan. 2002 to Dec. 2020). Thus, a one standard deviation increase in the exposure to physical climate risk leads to a positive risk premium of 1.50% p.a. This result is robust when controlling for fixed effects (Model 6 and 7), as the risk premium even increases to 1.75% p.a. (t-value = 2.73) with sector fixed effects and 1.94% p.a. (t-value = 3.38) with industry fixed effects. This indicates further that the risk premium for physical climate

risk is a firm-specific effect rather than a sector- or industry-specific effect.

The risk premium for regulatory climate risk is positive in the first subperiod (Model 5, Jan. 2002 to Dec. 2011) with a coefficient of 1.54% p.a. (t-value 1.61) and significantly negative in the second subperiod (Jan. 2012 to Dec. 2020) with a value of -2.56% p.a. (t-value = -2.94). As a consequence, the change in the premium over the two subperiods leads to an insignificant premium over the full period. These findings are in line with the literature. Hsu et al. (2022) finds a positive premium for firms with high toxic emissions over low emitting firms for the period 1996 to 2016 of 4.42% p.a. In contrast, Pástor et al. (2022) finds a strong outperformance of green stocks over brown stocks for the period 2013 to 2020. These contradictory results, most likely, emerge in response to the pronounced shift towards increased ESG awareness and "green investing", which caused a noticeable increase in demand for green stocks relative to brown stocks.

Interestingly, the risk premium for regulatory climate risk becomes more significant in the first subperiod (t-values = 1.61/2.38/2.34 for Model 5/6/7) and less significant in the second subperiod (t-values = -2.56/-1.66/-0.69 for Model 5/6/7) when we also control for sector and industry dummies. This indicates that in the first subperiod the regulatory climate risk premium is determined by firm-specific effects, while in the second subperiod whole industries and sectors are affected. Especially the sector *Mining* shows a sharp decline in sector average returns form the first to the second subperiod. The coefficients of exposures to sustainability always have the opposite sign compared to those of regulatory climate risk which gives further support for a shift towards "green investing". Also, we observe that the coefficients become less significant after controlling for sector and industry fixed effects in the first- (t-values = -1.24/-1.17/-0.29 for Models 5/6/7) and second subperiod (t-values = 1.39/0.95/0.67 for Models 5/6/7). Thus, again whole industries benefit, while others suffer from the transition towards more sustainable societies. Among the other explanatory variables, only *size* turns out to be significant over the entire period in case of the full model (Model 5) and after controlling for firm- and sector fixed effects (models 6 and 7). In the second subperiod, the coefficient of the market-beta becomes significantly negative once we control for fixed effects. Overall, traditional firm characteristics do not seem to play an important role, not in absolute terms and not relative to

climate-related exposures, in these regressions.

	Model 1a			Model 1b			Model 5			Model 6			Model 7		
Period	Full	1	2	Full	1	2	Full	1	2	Full	1	2	Full	1	2
Const	12.64	10.48	15.03	11.46	8.96	14.24	11.46	8.96	14.24						
	(2.46)	(1.26)	(2.63)	(2.19)	(1.06)	(2.44)	(2.19)	(1.06)	(2.44)						
Beta	-2.91	-0.61	-5.48	-0.86	0.81	-2.73	-0.64	1.18	-2.67	-0.85	1.05	-2.95	-1.40	0.55	-3.55
	(-2.16)	(-0.29)	(-3.66)	(-0.64)	(0.36)	(-2.02)	(-0.49)	(0.57)	(-1.95)	(-0.64)	(0.49)	(-2.16)	(-1.25)	(0.32)	(-2.86)
Size	7.01	5.09	9.15	5.83	3.98	7.88	6.21	4.32	8.30	6.83	4.60	9.31	7.08	4.75	9.68
	(5.24)	(2.50)	(5.76)	(3.24)	(1.47)	(3.54)	(3.21)	(1.45)	(3.63)	(3.65)	(1.60)	(4.22)	(4.14)	(1.84)	(4.73)
B2M	-2.17	-2.28	-2.05	-2.30	-0.89	-3.87	-2.28	-1.23	-3.44	-1.51	-0.88	-2.20	-1.30	-0.75	-1.92
	(-1.76)	(-1.18)	(-1.36)	(-1.45)	(-0.37)	(-1.93)	(-1.48)	(-0.54)	(-1.69)	(-1.02)	(-0.39)	(-1.19)	(-0.87)	(-0.32)	(-1.07)
OP	0.54	0.50	0.57	-0.60	-0.89	-0.27	-0.56	-0.90	-0.17	-0.53	-1.01	0.01	-0.73	-1.12	-0.29
	(0.92)	(0.76)	(0.58)	(-0.83)	(-0.94)	(-0.25)	(-0.78)	(-0.94)	(-0.16)	(-0.79)	(-1.08)	(0.01)	(-1.14)	(-1.22)	(-0.33)
INV	0.32	-0.82	1.58	0.37	-0.37	1.19	0.42	-0.23	1.14	0.22	-0.55	1.07	0.49	-0.08	1.11
	(0.54)	(-0.95)	(2.19)	(0.54)	(-0.40)	(1.19)	(0.63)	(-0.25)	(1.15)	(0.34)	(-0.66)	(1.14)	(0.81)	(-0.09)	(1.27)
Sus							-0.05	-1.38	1.44	-0.10	-1.06	0.96	0.20	-0.24	0.69
							(-0.06)	(-1.24)	(1.39)	(-0.15)	(-1.17)	(0.95)	(0.31)	(-0.29)	(0.67)
Reg							-0.40	1.54	-2.56	0.28	2.04	-1.66	0.60	1.75	-0.69
							(-0.57)	(1.61)	(-2.94)	(0.45)	(2.38)	(-2.08)	(1.04)	(2.34)	(-0.85)
Phys							1.50	2.08	0.85	1.75	2.17	1.27	1.94	2.15	1.70
							(2.37)	(2.23)	(1.04)	(2.73)	(2.34)	(1.47)	(3.38)	(2.67)	(2.08)
Fixed effects	None			None			None			Sector fixed effects			Industry fixed effects		
Months	228	120	108	228	120	108	228	120	108	228	120	108	228	120	108
Observations	810766	810766	810766	189586	189586	189586	189586	189586	189586	189586	189586	189586	189586	189586	189586
Firms	8151	8151	8151	3005	3005	3005	3005	3005	3005	3005	3005	3005	3005	3005	3005
R2 (%)	2.86	3.12	2.57	6.04	6.65	5.37	7.24	8.11	6.26	11.36	12.77	9.79	29.20	32.70	25.32
Adj. R2 (%)	2.71	2.99	2.40	5.43	6.01	4.78	6.26	7.10	5.32	9.41	10.80	7.87	22.63	26.17	18.70

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4.7: Fama-MacBeth regression performed over the periods Jan. 2002 to Dec. 2020 (Full), and the two subperiods Jan. 2002 to Dec. 2011 (1) and Jan. 2012 to Dec. 2020 (2). Model 1 is comprised of the classic risk factors that enter the Fama French five-factor model. Models 2 to 4 are univariate extensions of the Fama French 5-factor model, which we omit for the sake of brevity (see the full Table 17 in the Appendix). In Model 5 we include the climate factors and in Model 6 and 7 we control for fixed effects using sector dummies in Model 6 and industry dummies in Model 7. We consider 10 sectors (divisions) and 65 industries (with at least 1000 observations each) according to the SIC scheme. We report the annualized risk premia in percent and heteroskedasticity and autocorrelation (HAC) adjusted t-values (Newey and West (1986) standard errors with three lags). All characteristics except dummy variables are standardized for each of the n cross-sectional regressions. Also we exclude all observations with missing values in the cross sectional regression which causes the number of observations to decline relative to Model 1 as the number of firms with topic exposures is limited.

4.5.4 Climate Risk Betas

A limitation to consider when using firm-specific news exposures is the limited coverage, as not all firms are consistently mentioned in the news. To mitigate this issue and to evaluate whether the results presented before extend beyond large firms with news coverage, we follow the multi-factor framework of Fama and French (1993, 2015) and calculate beta

coefficients for each type of climate risk for all firms and not just those firms covered by the media. Specifically, we determine climate risk betas by regressing individual stock returns on the returns of the GMB and the high-minus-low physical climate risk portfolio (PhysCR) defined before. Thus, we distinguish between regulatory climate risk betas (β_{RegCR}) and physical climate risk betas (β_{PhysCR}) in the following analysis.

In order to calculate the regulatory climate risk betas, we extend the market model by our green-minus-brown (GMB) portfolio (regulatory climate risk factor) (Equation (4.13)). Positive regulatory climate risk betas indicate green firms while negative betas indicate brown firms. Similarly, we calculate physical climate risk betas according to Equation (4.14). In this case, we also control for size by including the SMB factor. This is necessary because we introduce a systematic bias towards small firms with the equal-weighted low physical climate risk portfolio (the correlation coefficient with SMB is 0.649, see Table 4.6). Positive physical climate risk betas indicate firms that are highly exposed to physical climate risks, i.e., storms, hurricanes, wildfires, droughts, etc., while firms with negative climate betas have no or minor exposure to physical climate risks.[9]

$$R_{p,t} - R_{f,t} = a_p + \beta_p(R_{M,t} - R_{f,t}) + \beta_{RegCR,p} \times GMB_t + \epsilon_t \tag{4.13}$$

$$R_{p,t} - R_{f,t} = a_p + \beta_p(R_{M,t} - R_{f,t}) + \beta_{size,p} \times SMB_t + \beta_{PhysCR,p} \times PhysCR_t + \epsilon_t \tag{4.14}$$

Characteristics of Climate Risk Beta Sorted Portfolios

Again, we form climate risk portfolios, this time however, we sort stocks based on their climate risk betas. Every month we rank firms by their regulatory climate risk beta and form a "High Regulatory Climate Risk" portfolio ("Brown" portfolio) by selecting firms with betas in the bottom third and a "Low Regulatory Climate Risk" portfolio

[9]The climate risk betas are calculated by regressing monthly excess stock returns (in excess to the risk free rate R_f) onto the returns of the factor models over a 72-month rolling window (at least 36 month of return data has to be available) with betas updated every month.

("Green" portfolio) with betas in the top third. Similarly we form a "High Physical Climate Risk" portfolio by selecting firms with physical climate risk betas in the top third and a "Low Physical Climate Risk" portfolio by selecting firms with betas in the bottom third. Figure 4.7 shows the cumulative returns of the (a) value- and (b) equal-weighted portfolios together with the market portfolio.

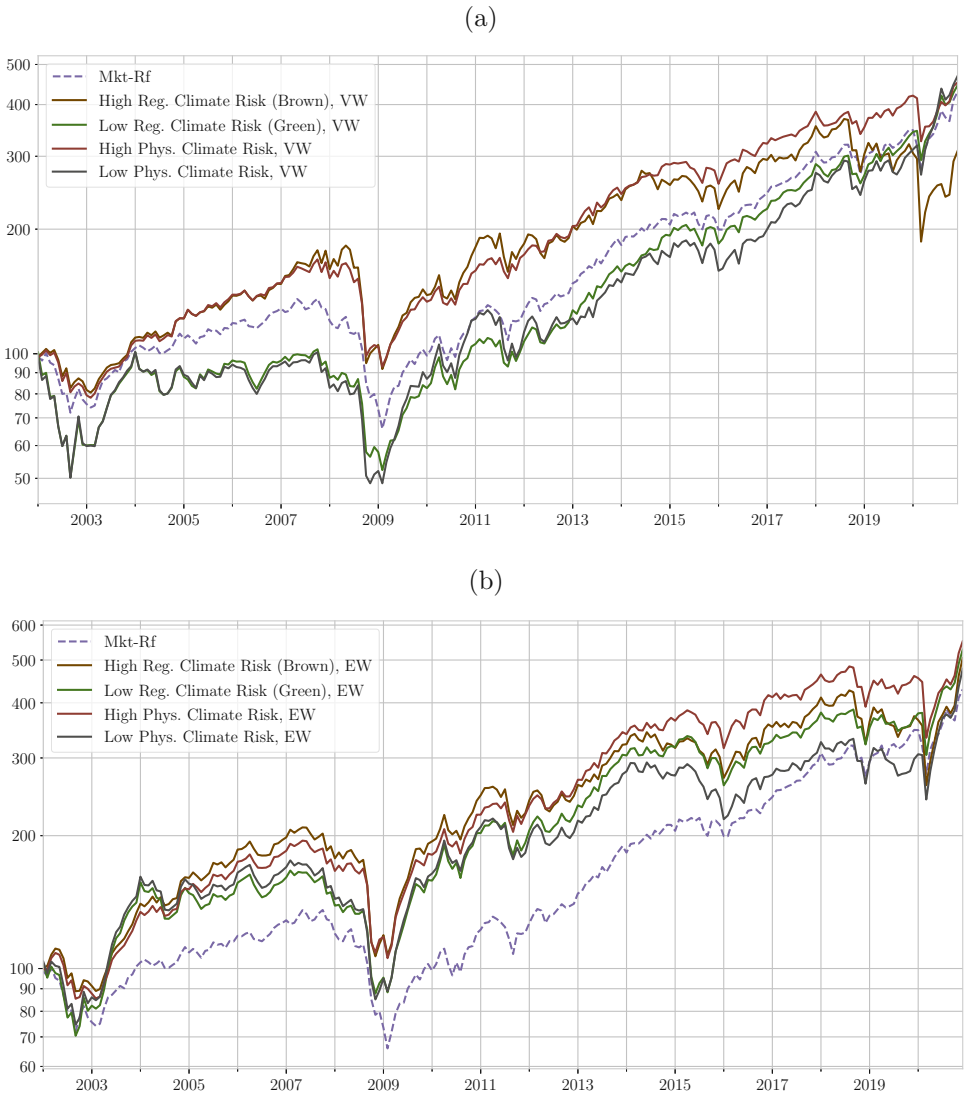


Figure 4.7: Cumulative returns of (a) value- and (b) equal-weighted climate risk beta sorted portfolios over the period Jan. 2002 to Dec. 2020. We sort stocks in a high- and low regulatory climate risk portfolio, as well as a high- and low physical climate risk portfolio.

In Table 4.8, we highlight the constituents of the beta sorted climate risk portfolios that have the largest average portfolio weight over the period from Jan. 2002 to Dec. 2020. We observe that the top holdings of the "Green" portfolio are dominated by technology,

telecommunications and software firms while the "Brown" portfolio is dominated by firms operating in the Oil&Gas sector. The top holdings of the physical climate risk portfolio also include by Oil&Gas firms like Exxon Mobil and Chevron next to insurance companies like American International Group and telecommunication firms like AT&T and Verizon Communications.[10]

Due to value weighting, portfolio weights are tilted towards large firms by construction. The cumulative weight of the 30 largest holdings is 30.47% for the "Brown" portfolio, 48.34% for the "Green" portfolio and 39.91% for the physical climate risk portfolio. Big tech stocks like Apple, Microsoft, Amazon, etc. and large Oil&Gas producers like Exxon Mobil and Chevron receive large weights due to their high market caps. Note, that two technology companies, Apple and Meta Platforms, also appear in the "Brown" portfolio, which are likely misclassifications. In Appendix C.3 we calculate beta-sorted portfolios by excluding all firms with a highly insignificant beta (t-value ≤ 1) and find that these firms vanish from the "Brown" portfolio.

We highlight the industries that are most exposed to the climate risk portfolios in Table 13 (see, Appendix C.1). The industries most exposed to regulatory climate risk are *Oil and Gas Extraction*, *Petroleum Refining and Related Industries* and *Electric, Gas, And Sanitary Services*. Among the industries less exposed to regulatory climate risk are the industries *Business Services*, *Electronic And Other Electrical Equipment* and *Industrial And Commercial Machinery And Computer Equipment*. Also, we find that the firms of the industries *Electric, Gas, And Sanitary Services*, *Chemicals And Allied Products* and *Insurance Carriers* are among the industries with the highest exposure to physical climate risk.

To alleviate the influence of the value weighting, we also calculate the top positions of equally weighted portfolios in Table 14 (see, Appendix C.2). As expected, the equally weighted portfolios are less influenced by large technology companies. The top holdings of the "Brown" portfolio belong to the industries Coal Mining and Oil&Gas Extraction, among others.

[10] According to the Wireless Infrastructure Association there are 142,100 cellular towers and 209,500 macrocell sites in operation by telecommunication firms (Wireless Infrastructure Association, 2023). This type of infrastructure is highly exposed to physical climate risks such as storms and hurricanes.

	Regulatory Climate Risk (Brown) Portfolio			Physical Climate Risk Portfolio			Sustainability (Green) Portfolio		
	Weight (%)		Company Name	Weight (%)		Company Name	Weight (%)		Company Name
0	6.81	6.81	Exxon Mobil Corp	5.81	5.81	Exxon Mobil Corp	7.43	7.43	Microsoft Corp
1	3.11	9.92	Chevron Corp New	3.07	8.88	Johnson & Johnson	4.41	11.85	Intel Corp
2	2.02	11.94	Conocophillips	2.71	11.59	Chevron Corp New	4.09	15.94	Cisco Systems Inc
3	1.27	13.21	Eog Resources Inc	2.57	14.16	Walmart Inc	3.42	19.36	General Electric Co
4	1.19	14.40	Occidental Petroleum Corp	2.30	16.46	Coca Cola Co	2.92	22.28	Apple Inc
5	1.11	15.51	Apple Inc	2.11	18.57	Berkshire Hathaway Inc Del	2.87	25.15	Oracle Corp
6	1.08	16.59	Meta Platforms Inc	2.04	20.61	AT & T Inc	1.96	27.11	Amazon Com Inc
7	0.82	17.41	American International Group Inc	2.02	22.63	Procter & Gamble Co	1.77	28.88	Alphabet Inc
8	0.81	18.22	Coca Cola Co	1.67	24.30	Pepsico Inc	1.70	30.58	Berkshire Hathaway Inc Del
9	0.76	18.99	Devon Energy Corp New	1.65	25.94	Verizon Communications Inc	1.41	31.98	Dell Inc
10	0.72	19.70	Caterpillar Inc	1.20	27.15	General Electric Co	1.21	33.20	Walmart Inc
11	0.71	20.41	Halliburton Company	1.02	28.17	Mcdonalds Corp	1.14	34.34	Johnson & Johnson
12	0.67	21.08	Johnson & Johnson	1.01	29.18	Conocophillips	1.09	35.43	Qualcomm Inc
13	0.67	21.76	Procter & Gamble Co	0.84	30.02	Pfizer Inc	1.07	36.50	International Business Machs Cor
14	0.64	22.40	Valero Energy Corp New	0.81	30.83	Philip Morris International Inc	1.06	37.57	Comcast Corp New
15	0.63	23.03	Unitedhealth Group Inc	0.81	31.64	American International Group Inc	1.03	38.60	Amgen Inc
16	0.63	23.66	Exelon Corp	0.72	32.36	Bristol Myers Squibb Co	0.98	39.57	EMC Corp Ma
17	0.63	24.29	Berkshire Hathaway Inc Del	0.67	33.03	Intel Corp	0.97	40.54	Home Depot Inc
18	0.62	24.90	Philip Morris International Inc	0.64	33.67	Home Depot Inc	0.91	41.45	HP Inc
19	0.58	25.48	APA Corp	0.63	34.30	Unitedhealth Group Inc	0.80	42.26	Time Warner Inc New
20	0.56	26.04	Anadarko Petroleum Corp	0.62	34.92	3M Co	0.76	43.01	Pepsico Inc
21	0.53	26.57	Dominion Energy Inc	0.60	35.52	United Parcel Service Inc	0.73	43.74	Procter & Gamble Co
22	0.51	27.08	Hess Corp	0.60	36.12	Union Pacific Corp	0.63	44.37	Merck & Co Inc New
23	0.51	27.59	Southern Copper Corp	0.56	36.68	Altria Group Inc	0.62	44.98	Verizon Communications Inc
24	0.51	28.10	Freeport Mcmoran Inc	0.56	37.24	Southern Co	0.59	45.58	Yahoo Inc
25	0.49	28.59	Deere & Co	0.55	37.79	Costco Wholesale Corp New	0.57	46.14	Applied Materials Inc
26	0.49	29.08	Boeing Co	0.54	38.33	Merck & Co Inc New	0.57	46.71	Corning Inc
27	0.48	29.55	NOV Inc	0.53	38.86	Dominion Energy Inc	0.55	47.26	Bristol Myers Squibb Co
28	0.47	30.02	Baker Hughes Co	0.53	39.39	Duke Energy Corp New	0.55	47.81	Target Corp
29	0.45	30.47	Marathon Oil Corp	0.53	39.91	Exelon Corp	0.53	48.34	AT & T Inc

Table 4.8: Top 30 companies of the value weighted climate risk beta sorted regulatory climate risk (brown), physical climate risk and sustainability (green) portfolios. The weights are averages in %, calculated over the period Jan. 2002 to Dec. 2020.

Climate Risk Premia using Beta Sorted Portfolios

Analogue to Section 4.5.2, we form a zero-investment portfolio that is long (short) the beta sorted "Green" ("Brown") portfolio. Figure 4.8 shows the cumulative returns of the equal- and value-weighted, beta sorted, GMB portfolios.

We observe a downward trend for the GMB portfolio until 2011. This is followed by an upward trend that lasts until the end of the sample period. Since we use value weighting which is commonly used in the literature, we compare this result with two important results from the literature: Hsu et al. (2022) find that polluters outperform non-polluting companies while Pástor et al. (2022) reports a strong outperformance of green over brown stocks. We show that the authors obtain these contrary results as they focus on different sample periods.

While Hsu et al. (2022) consider the period 1991 to 2016, Pástor et al. (2022) analyze the period Nov. 2012 to Dec. 2020 which can be characterized by the rise of sustainable finance and growing availability of ESG data. Hsu et al. (2022) argue that brown stocks have higher realized returns because investors demand higher ex ante risk premia for high-emission firms as they carry a higher risk of being effected by policy regime shifts towards

a more environmentally friendly economy. Pástor et al. (2022) also argues that brown firms have a higher risk premium and thus higher ex ante expected returns. However, the authors show that the strong performance of green stocks since 2012 is caused by unexpected windfall gains due to increased climate concerns and rising investor demand.

Given the longer time-series, compared to the literature, that we work with, we can contribute to the above discussion and provide more nuanced empirical evidence. Our findings suggest that these windfall gains surpassed the climate risk premium around 2012. The trend towards sustainable investing further contributed to an increased demand for green assets, resulting in positive returns for the GMB portfolio in 2012 and thereafter. The fact that the same pattern, albeit less pronounced, is observed in the equally weighted portfolio tells us that the outperformance of green versus brown stocks from 2011 onward is not only due to large technology stocks, but as conjectured due to a broad shift in investor demand towards sustainable businesses.

Moreover, in comparison to the topic-exposure weighted portfolio in Section 4.5.2, we do not observe a drawdown but only a sideways movement from 2015 to 2018 due to the broader coverage of the investment universe and the different weight distribution of the climate risk beta sorted portfolios. Also, note that the positive returns of the GMB portfolio are driven by an underperformance of brown stocks rather than an outperformance of green stocks relative to the market portfolio (see Figure 4.7). This also shows up in Table 4.7 (Model 5, subperiod 2) as the positive risk premium for sustainability (Sus) is smaller in absolute terms than the negative risk premium for regulatory climate risk (Reg).

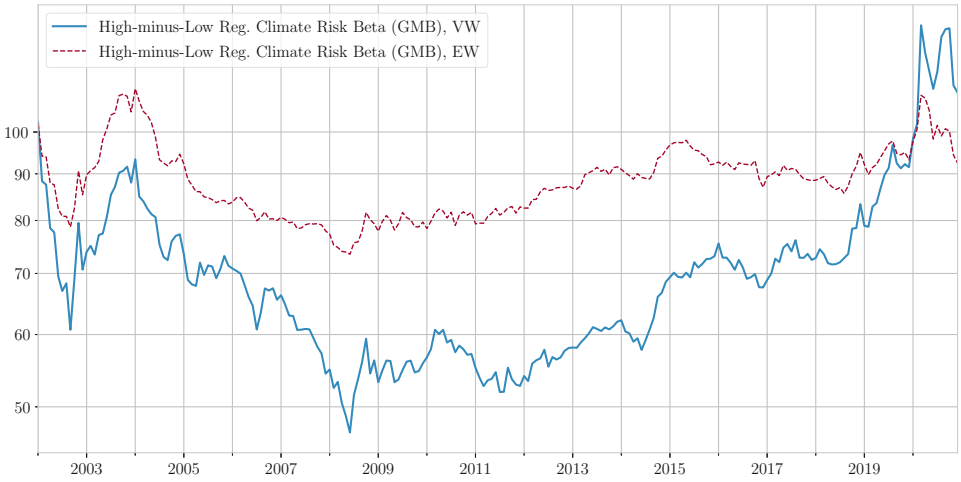


Figure 4.8: Cumulative returns of the value-weighted (VW) and equal-weighted (EW), beta sorted, green-minus-brown (GMB) portfolio over the period from Jan. 2002 to Dec. 2020.

Finally, we quantify the similarity between the news exposure sorted GMB portfolio and the climate risk beta sorted GMB portfolio by calculating the correlation between the two time-series over the period Jan. 2002 to Dec. 2020. The resulting correlation coefficient is 0.67 for quarterly returns (0.54 for monthly- and 0.64 for annual returns). The high correlation coefficient shows that the broader beta-based approach is closely related to firm-specific climate risk characteristics extracted from news.

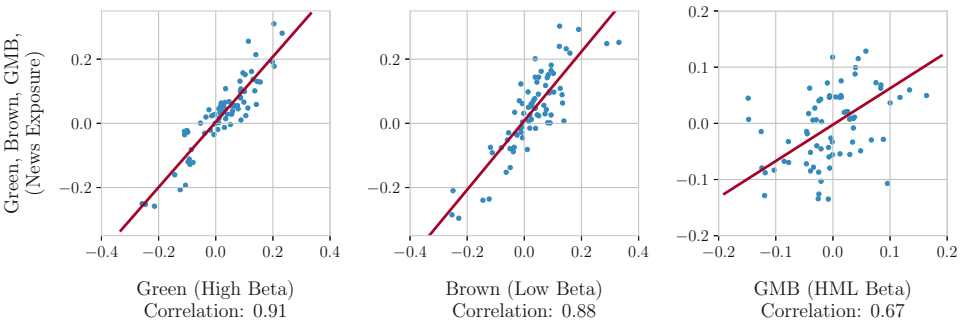


Figure 4.9: Correlation between the news exposure sorted and value-weighted beta sorted regulatory climate risk portfolios over the period Jan. 2002 to Dec. 2020 calculated on quarterly returns.

Next, we turn to physical climate risk. Figure 4.10 provides insights into the relative performance of the value- and equal-weighted high-minus-low physical climate risk portfolios. The equal-weighted portfolio shows a pattern similar to the topic-exposure weighted portfolio in Figure 4.6: an upward trend from 2011 to 2019 that ends with the start of the

pandemic in 2020. This positive performance is also in line with the significant physical climate risk premium reported in Table 4.7.

The value-weighted portfolio, on the other hand, is experiencing a sharp downturn after peaking at the end of 2008. This is due to the fact that large-cap technology firms, which experience exceptionally strong returns over the second half of the sample, are concentrated in the low physical climate risk portfolio. This exceeds the premium on physical climate risk and thus results in a negative performance.

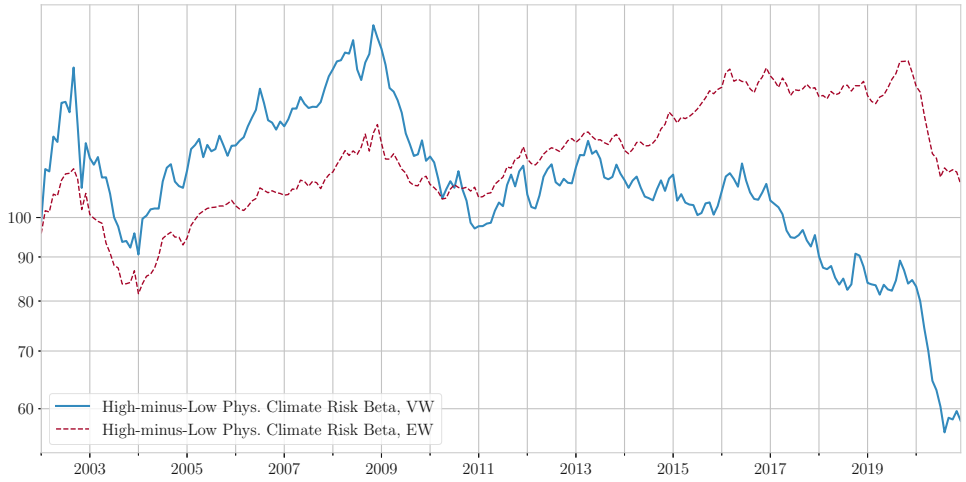


Figure 4.10: Cumulative returns of the value-weighted (VW) and equal-weighted (EW), beta sorted, high-minus-low physical climate risk beta portfolio over the period from Jan. 2002 to Dec. 2020.

Again, we quantify the similarity between the topic-exposure sorted and climate beta sorted portfolios by calculating the correlation between the two time-series over the period Jan. 2002 to Dec. 2020. The resulting correlation coefficient is 0.63 for quarterly returns (0.52 for monthly- and 0.74 for annual returns).

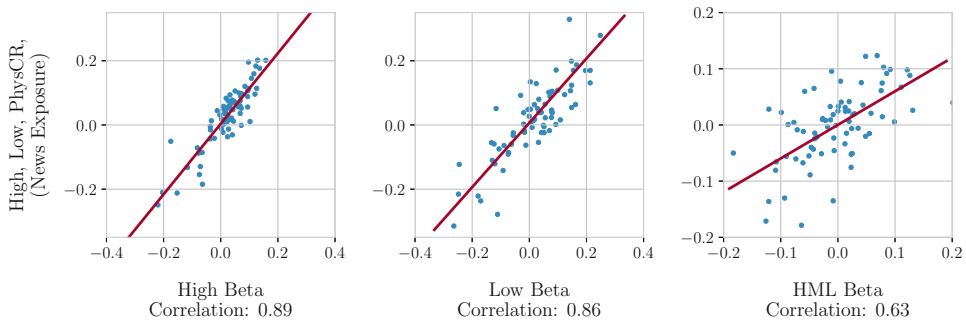


Figure 4.11: Correlation between the news exposure sorted and value-weighted beta sorted physical climate risk portfolios over the period Jan. 2002 to Dec. 2020 calculated on quarterly returns.

4.5.5 Validation

In this section, we pool complementary results that should provide additional validation to our news-based approach.

Comparison With an ESG-sorted GMB Portfolio

We further validate our methodology and results by comparing our beta-sorted GMB portfolio with the E-score-sorted GMB portfolio of Pástor et al. (2022). The authors use MSCI ESG ratings data to calculate firm-level environmental scores. Based on the calculated scores they form value-weighted portfolios by selecting the top third of firms with the highest environmental score (green portfolio) and the bottom third of firms with the lowest environmental score (brown portfolio). The correlation between our beta-sorted GMB portfolio and the GMB portfolio of Pástor et al. (2022) is 0.64, calculated on the basis of quarterly returns (0.46 for monthly- and 0.70 for annual returns) over the period Jan. 2009 to Dec. 2020 (see Figure 4.12). In addition, the correlation between the green and brown portfolios is 0.92, also calculated on the basis of quarterly returns. These high correlations indicate a high degree of similarity between the underlying portfolios. This can be observed when plotting the cumulative returns, as shown in Figure 4.13. Figure 4.13a shows the cumulative returns of the green, brown and the market portfolio, in comparison with the green and brown portfolio of Pástor et al. (2022). The cumulative returns of our "Green" and "Brown" portfolio (solid lines) align almost perfectly with the "Green" and "Brown" portfolio of Pástor et al. (2022) (dashed lines). Figure 4.13b plots the cumulative

returns of the GMB portfolios. Again, the two portfolios tend to have a small tracking error. Only after 2018 this error increases moderately, as our "Brown" portfolio slightly underperforms the E-score-sorted portfolio.

The high degree of similarity between these portfolios is surprising, as the underlying methodology for identifying green and brown companies is completely different. ESG scoring is an elaborate "bottom-up" approach that requires a thorough analysis of a company's operations, including key product/business segments, and calculations of exposures to key environmental risks. It also involves measurements of carbon intensity and emissions at the firm level. Our approach, in contrast can be seen as a "top-down" approach. Brown (green) firms have a higher chance of being mentioned in news that cover brown (green) topics. In addition, firms whose returns covary with the returns of these identified firms, as measured by the climate risk beta, are most likely also exposed to the same risks. We argue that a simple metric such as our climate risk beta, calculated in a similar way to the common risk factors available in the literature, can be used as an alternative to environmental scores to identify climate risks of individual companies.

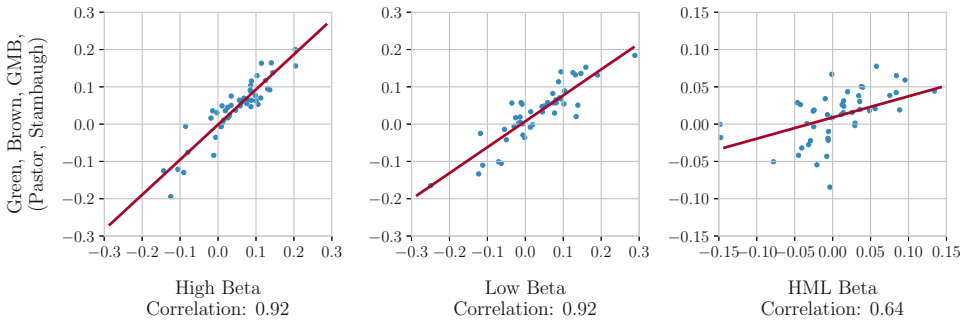


Figure 4.12: Correlation between the regulatory climate risk beta sorted green, brown and green-minus-brown (GMB) portfolio with the green, brown and GMB portfolio of Pástor et al. (2022) using quarterly returns over the period Jan. 2009 to Dec. 2020.

Exposure to Well-known Risk Factors

To assess to what extent the climate risk premia identified and discussed before are robust to existing empirical asset pricing models, we regress the returns of the value- and equal-weighted GMB portfolio (Table 4.9) and the high-minus-low physical climate risk

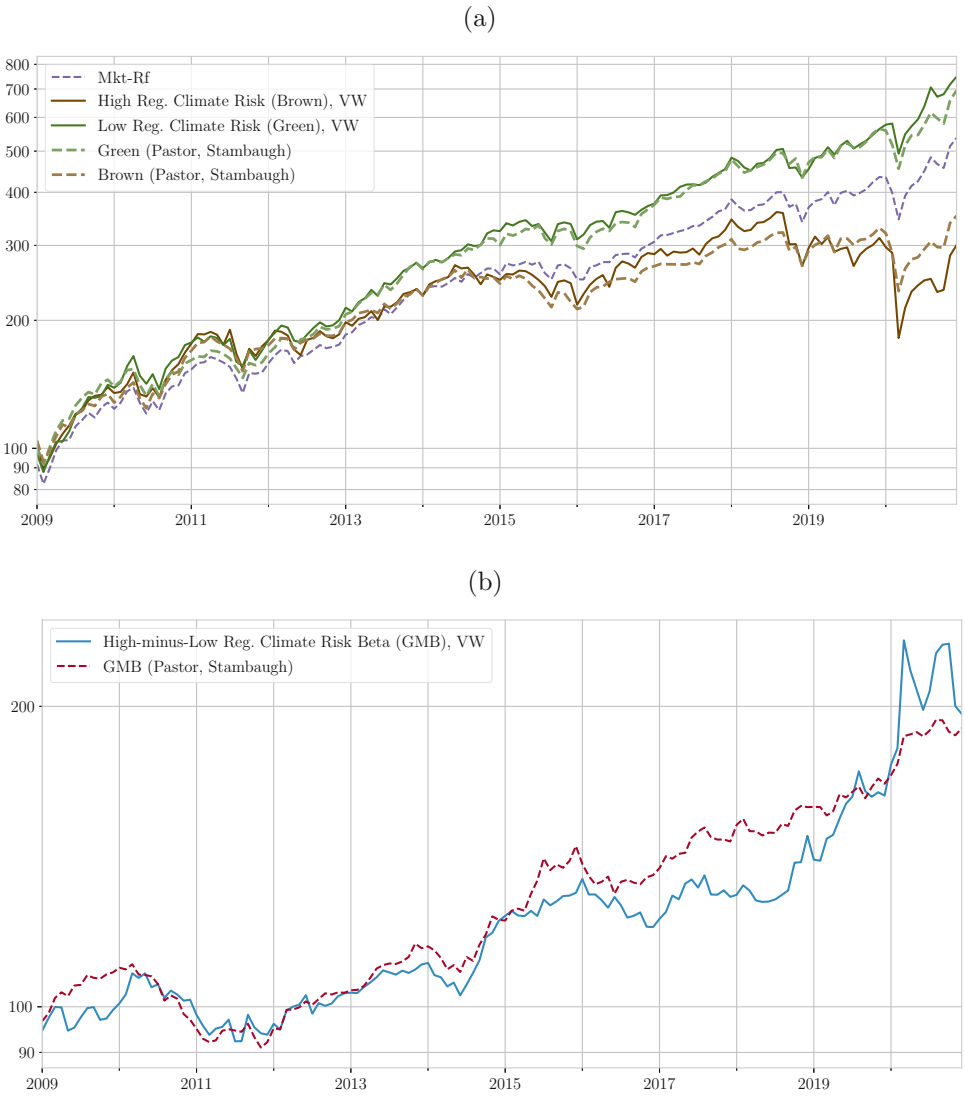


Figure 4.13: Cumulative returns of (a) the green (sustainability), brown (regulatory climate risk) portfolio and (b) the beta sorted green-minus-brown (GMB) portfolio over the period from Jan. 2009 to Dec. 2020.

portfolios (Table 4.10) on several well known risk factors documented by Fama and French (1993, 2015) and Carhart (1997). The explanatory variables are the market portfolio (Mk-Rf), size factor (SMB), value factor (HML), profitability (RMW), investment (CMA) and momentum (UMD). For each regression we show results for the full period ranging from Jan. 2002 to Dec. 2020 (Full) as well as for the two subperiods Jan. 2002 to Dec. 2011 (1) and Jan. 2012 to Dec. 2020 (2).

Most importantly, we find a positive significant alpha of 80 bps per month (9.60% p.a.) with a t-value of 2.80 for the value-weighted GMB portfolio from 2012 to 2020 (last

column of Panel A in Table 4.9). Thus, the common risk factors are not able to fully explain the strong performance of the value-weighted GMB portfolio from 2012 onwards. For the equal-weighted GMB portfolio the alpha is only 18 bps per month (2.16% p.a.). However, with a t-value of 1.10 this alpha is insignificant. Furthermore, we observe that the explained variance R^2 is smaller for the full period than for each of the two subperiods. In Panel A, the R^2 is 31.99% for the full-, 39.31% for the first- and 56.54% for the second subperiod. The low R^2 for the full period is a direct result of the regime shift in 2012. The difference between the two subperiods is also reflected in the coefficients. While for the first subperiod only the coefficient of RMW is significant, in the second subperiod Mkt-RF, SMB, HML and CMA are significant with a positive coefficient on CMA and negative coefficients on the market, SMB and HML. For the equal-weighted portfolio in Panel B we get a R^2 of 33.28% for the full-, 41.84% for the first- and 45.35% for the second subperiod. In the first subperiod, the coefficients of HML, RMW and CMA are significant. For the second subperiod Mkt-RF, SMB, HML and RMW are significant.

For the equal-weighted high-minus-low physical climate risk portfolios we find an alpha of 16 bps per month (1.92% p.a.). However, with a t-value of 0.987 it is insignificant (last column of Panel A in Table 4.10). For the value-weighted portfolio the alpha is negative and insignificant.

Period	Full	1	2	Full	1	2	Full	1	2	Full	1	2
Panel A: Dependent Variable: VW GMB Portfolio												
Constant	0.0015	-0.0050	0.0139***	0.0007	-0.0052	0.0089***	0.0043*	0.0013	0.0082***	0.0044*	0.0008	0.0080***
	(0.5316)	(-1.2894)	(3.9576)	(0.2495)	(-1.3054)	(3.0711)	(1.7333)	(0.3763)	(2.9054)	(1.7529)	(0.2340)	(2.8041)
Mkt-RF	-0.0273	0.2641***	-0.5008***	0.0769	0.2712***	-0.3004***	-0.0569	-0.0643	-0.2401***	-0.0715	-0.1026	-0.2164***
	(-0.4230)	(3.1525)	(-6.0003)	(1.1352)	(2.9119)	(-4.1006)	(-0.9022)	(-0.6871)	(-3.1998)	(-1.0809)	(-1.0811)	(-2.7226)
SMB				-0.2525**	0.0567	-0.4960***	-0.4370***	-0.0879	-0.5767***	-0.4310***	-0.0420	-0.5542***
				(-2.1003)	(0.3279)	(-4.1363)	(-4.1121)	(-0.6008)	(-4.4300)	(-4.0402)	(-0.2858)	(-4.1794)
HML				-0.3407***	-0.1275	-0.5396***	-0.3665***	-0.0307	-0.6486***	-0.3897***	-0.0945	-0.6074***
				(-3.1102)	(-0.7798)	(-5.1626)	(-3.4444)	(-0.2096)	(-5.2550)	(-3.5107)	(-0.6329)	(-4.6170)
RMW							-1.0338***	-1.1860***	-0.3803*	-1.0110***	-1.0838***	-0.3751*
							(-7.8331)	(-6.7244)	(-1.9477)	(-7.4516)	(-5.9114)	(-1.9183)
CMA							0.4901***	0.2791	0.4393**	0.4912***	0.3078	0.4563**
							(2.8779)	(1.2518)	(2.0038)	(2.8815)	(1.3910)	(2.0721)
UMD										-0.0449	-0.1305*	0.0870
										(-0.7415)	(-1.8269)	(0.9110)
Observations	228	120	108	228	120	108	228	120	108	228	120	108
R^2	0.0008	0.0777	0.2535	0.0811	0.0826	0.5308	0.3182	0.3751	0.5618	0.3199	0.3931	0.5654
Adjusted R^2	-0.0036	0.0699	0.2465	0.0688	0.0589	0.5173	0.3029	0.3477	0.5403	0.3015	0.3608	0.5395
Residual Std. Error	0.0429	0.0428	0.0349	0.0413	0.0431	0.0280	0.0357	0.0359	0.0273	0.0358	0.0355	0.0273
F Statistic	0.1789	9.9385***	36.0030***	6.5873***	3.4819**	39.2178***	20.7260***	13.6878***	26.1537***	17.3283***	12.1967***	21.8967***
Panel B: Dependent Variable: EW GMB Portfolio												
Constant	-0.0000	-0.0016	0.0041**	-0.0005	-0.0020	0.0022	0.0014	0.0014	0.0018	0.0014	0.0014	0.0018
	(-0.0100)	(-0.7316)	(2.3072)	(-0.3118)	(-0.8953)	(1.3419)	(1.1017)	(0.7458)	(1.1598)	(1.0935)	(0.7105)	(1.1089)
Mkt-RF	-0.0100	0.1323***	-0.2303***	0.0316	0.1358***	-0.1396***	-0.0389	-0.0474	-0.1069**	-0.0370	-0.0518	-0.1006**
	(-0.2934)	(2.7922)	(-5.4429)	(0.8689)	(2.6325)	(-3.4207)	(-1.1753)	(-0.9364)	(-2.5727)	(-1.0672)	(-0.9940)	(-2.2785)
SMB				-0.0750	0.1242	-0.2666***	-0.1780***	0.0339	-0.3372***	-0.1787***	0.0392	-0.3312***
				(-1.1641)	(1.2964)	(-3.9908)	(-3.1971)	(0.4288)	(-4.6764)	(-3.1944)	(0.4860)	(-4.4953)
HML				-0.1785***	-0.2014**	-0.1622***	-0.2086***	-0.1677**	-0.1964***	-0.2057***	-0.1750**	-0.1855**
				(-3.0410)	(-2.2226)	(-2.7851)	(-3.7423)	(-2.1124)	(-2.8734)	(-3.5340)	(-2.1352)	(-2.5375)
RMW							-0.5681***	-0.6560***	-0.2852***	-0.5710***	-0.6443***	-0.2838**
							(-8.2175)	(-6.8717)	(-2.6377)	(-8.0250)	(-6.4031)	(-2.6131)
CMA							0.3297***	0.2534**	0.1787	0.3296***	0.2567**	0.1832
							(3.6965)	(2.0999)	(1.4716)	(3.6867)	(2.1137)	(1.4974)
UMD										0.0056	-0.0149	0.0231
										(0.1770)	(-0.3808)	(0.4358)
Observations	228	120	108	228	120	108	228	120	108	228	120	108
R^2	0.0004	0.0620	0.2184	0.0586	0.1042	0.4066	0.3327	0.4176	0.4525	0.3328	0.4184	0.4535
Adjusted R^2	-0.0040	0.0540	0.2111	0.0460	0.0810	0.3895	0.3177	0.3921	0.4256	0.3147	0.3875	0.4210
Residual Std. Error	0.0227	0.0242	0.0177	0.0221	0.0239	0.0156	0.0187	0.0194	0.0151	0.0188	0.0195	0.0152
F Statistic	0.0861	7.7966***	29.6255***	4.6475***	4.4968***	23.7538***	22.1349***	16.3504***	16.8572***	18.3705***	13.5473***	13.9677***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4.9: We run monthly time-series regressions over the periods Jan. 2002 to Dec. 2020 (Full) as well as the two subperiods Jan. 2002 to Dec. 2011 (1) and Jan. 2012 to Dec. 2020 (2). In Panel A, the dependent variable is the value-weighted (VW) beta sorted green-minus-brown (GMB) portfolio. In Panel B, the dependent variable is the equal-weighted (EW) beta sorted GMB portfolio. Mkt-Rf is the excess market return, SMB is the size- and HML is the value factor of Fama and French (1993). RMW and CMA are the profitability and investment factors of Fama and French (2015) and UMD is the momentum factor of Carhart (1997).

Period	Full	1	2	Full	1	2	Full	1	2	Full	1	2
Panel A: Dependent Variable: VW High-minus-low Physical Climate Risk Portfolio												
Constant	0.0012	0.0035	-0.0031	0.0026	0.0039	-0.0001	-0.0014	-0.0022	-0.0008	-0.0015	-0.0017	-0.0014
	(0.5026)	(0.9596)	(-1.0315)	(1.1377)	(1.1048)	(-0.0400)	(-0.6549)	(-0.6793)	(-0.3161)	(-0.7405)	(-0.5403)	(-0.5875)
Mkt-RF	-0.3814***	-0.4850***	-0.2135***	-0.4257***	-0.5190***	-0.2673***	-0.2818***	-0.2159**	-0.2146***	-0.2402***	-0.1819**	-0.1546**
	(-7.0858)	(-6.2562)	(-2.9784)	(-7.7372)	(-6.2738)	(-3.8591)	(-5.3859)	(-2.5810)	(-3.3405)	(-4.4391)	(-2.1418)	(-2.3479)
SMB				-0.1140	-0.1746	-0.0781	0.0282	-0.0551	0.0918	0.0111	-0.0957	0.1487
				(-1.1683)	(-1.1367)	(-0.6887)	(0.3197)	(-0.4210)	(0.8238)	(0.1275)	(-0.7281)	(1.3535)
HML				0.5113***	0.4711***	0.5553***	0.3835***	0.3649***	0.2365**	0.4495***	0.4213***	0.3408***
				(5.7498)	(3.2440)	(5.6201)	(4.3465)	(2.7813)	(2.2378)	(4.9525)	(3.1538)	(3.1264)
RMW							0.8811***	1.0635***	0.3778**	0.8157***	0.9729***	0.3911**
							(8.0522)	(6.7417)	(2.2594)	(7.3525)	(5.9317)	(2.4140)
CMA							0.1520	-0.1532	0.9185***	0.1488	-0.1786	0.9615***
							(1.0770)	(-0.7684)	(4.8927)	(1.0676)	(-0.9024)	(5.2693)
UMD										0.1282**	0.1156*	0.2202***
										(2.5902)	(1.8087)	(2.7821)
Observations	228	120	108	228	120	108	228	120	108	228	120	108
R^2	0.1818	0.2491	0.0772	0.2878	0.3119	0.2973	0.4488	0.5247	0.4614	0.4650	0.5381	0.4997
Adjusted R^2	0.1782	0.2427	0.0685	0.2783	0.2941	0.2770	0.4364	0.5039	0.4350	0.4505	0.5135	0.4700
Residual Std. Error	0.0358	0.0396	0.0300	0.0335	0.0382	0.0264	0.0296	0.0321	0.0234	0.0293	0.0317	0.0226
F Statistic	50.2088***	39.1406***	8.8710***	30.1726***	17.5301***	14.6679***	36.1496***	25.1700***	17.4756***	32.0177***	21.9382***	16.8154***
Panel B: Dependent Variable: EW High-minus-low Physical Climate Risk Portfolio												
Constant	0.0011	0.0020	0.0000	0.0019	0.0024	0.0016	-0.0002	-0.0014	0.0020	-0.0003	-0.0013	0.0016
	(0.7845)	(0.9560)	(0.0133)	(1.4226)	(1.1607)	(0.9195)	(-0.1589)	(-0.7906)	(1.2004)	(-0.2131)	(-0.7177)	(0.9867)
Mkt-RF	-0.0676**	-0.0679	-0.0610	-0.0798**	-0.0749	-0.0767*	-0.0022	0.1206**	-0.1114**	0.0147	0.1298***	-0.0744*
	(-2.1160)	(-1.4952)	(-1.3448)	(-2.4310)	(-1.5339)	(-1.7241)	(-0.0708)	(2.6079)	(-2.5863)	(0.4627)	(2.7349)	(-1.6756)
SMB				-0.1245**	-0.1396	-0.1135	-0.0314	-0.0555	0.0301	-0.0383	-0.0665	0.0652
				(-2.1396)	(-1.5403)	(-1.5582)	(-0.6123)	(-0.7673)	(0.4036)	(-0.7483)	(-0.9053)	(0.8803)
HML				0.2945***	0.2454***	0.3386***	0.2712***	0.1888**	0.3105***	0.2979***	0.2041***	0.3748***
				(5.5515)	(2.8626)	(5.3327)	(5.2835)	(2.6029)	(4.3836)	(5.5975)	(2.7334)	(5.0997)
RMW							0.5428***	0.6911***	0.5057***	0.5164***	0.6666***	0.5139***
							(8.5279)	(7.9236)	(4.5130)	(7.9393)	(7.2701)	(4.7055)
CMA							-0.1172	-0.1618	-0.0351	-0.1185	-0.1687	-0.0086
							(-1.4272)	(-1.4671)	(-0.2793)	(-1.4500)	(-1.5242)	(-0.0703)
UMD										0.0518*	0.0314	0.1357**
										(1.7862)	(0.8779)	(2.5431)
Observations	228	120	108	228	120	108	228	120	108	228	120	108
R^2	0.0194	0.0186	0.0168	0.1390	0.0880	0.2279	0.3662	0.4471	0.3565	0.3752	0.4509	0.3952
Adjusted R^2	0.0151	0.0103	0.0075	0.1275	0.0644	0.2056	0.3519	0.4229	0.3249	0.3582	0.4217	0.3593
Residual Std. Error	0.0212	0.0232	0.0190	0.0200	0.0226	0.0170	0.0172	0.0177	0.0157	0.0172	0.0177	0.0153
F Statistic	4.4775**	2.2357	1.8085	12.0559***	3.7325**	10.2325***	25.6513***	18.4394***	11.2995***	22.1188***	15.4637***	10.9989***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4.10: We run monthly time-series regressions over the periods Jan. 2002 to Dec. 2020 (Full) as well as the two subperiods Jan. 2002 to Dec. 2011 (1) and Jan. 2012 to Dec. 2020 (2). In Panel A, the dependent variable is the value-weighted (VW) beta sorted high-minus-low physical climate risk portfolio. In Panel B, the dependent variable is the equal-weighted (EW) beta sorted high-minus-low physical climate risk portfolio. Mkt-Rf is the excess market return, SMB is the size- and HML is the value factor of Fama and French (1993). RMW and CMA are the profitability and investment factors of Fama and French (2015) and UMD is the momentum factor of Carhart (1997).

Climate Risk Factors

In Section 4.5.5 we show that the Fama French factor models are not able to explain the outperformance of the green-minus-brown (GMB) portfolio from 2012 to 2020. Consequently, we now test whether the climate risk factor portfolios, the regulatory climate risk beta sorted GMB portfolio as well as the high-minus-low physical climate risk beta portfolio (PhysCR) are able to improve the explainability of average returns beyond the well-known factor models of Fama and French (1993, 2015). Similar to Fama and French (2015), we test how well the different factor models explain monthly excess returns ($R_i - R_f$) of 25 Size-B2M, 25 Size-OP, 25 Size-Inv portfolios as well as 10 sector and 72 industry portfolios

using the linear regression:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_i \tilde{R}_{P,t} + \tilde{\eta}_{i,t} \qquad \forall_i = 1, ..., N \tag{4.15}$$

estimated over a period of $t = 1, ..., T$ month. $\tilde{R}_{P,t} = (\tilde{R}_{1,t}, \tilde{R}_{2,t}, ..., \tilde{R}_{L,t})'$ denotes the return vector of the L factor portfolios that enter the market-model (Mkt-Rf), the three-factor model (FF3), the five-factor model (FF5) and an extension of each by our GMB and PhysCR portfolios. The ideal factor model that is able to fully explain expected returns has intercepts that are indistinguishable from zero (Fama and French, 2015). Consequently, we test whether the alphas for each set of 25, 10 or 72 time-series regressions are jointly zero by applying the GRS-test (Gibbons et al., 1989) under the null hypothesis:

$$H_0 : \alpha_i = 0 \qquad \forall_i = 1, ..., N \tag{4.16}$$

The GRS test statisitic W is calculated as follows:

$$\tilde{W} \equiv \frac{T(T-N-L)}{N(T-L-1)} \frac{\alpha \hat{\Sigma}^{-1} \hat{\alpha}'}{1 + \bar{R}_P \tilde{\Omega}^{-1} \bar{R}_P'} \quad \sim \quad F_{N,T-N-L} \tag{4.17}$$

with the factor return matrix $\tilde{R}_P = (\tilde{R}_{P,1}, ..., \tilde{R}_{P,T})'$, the variance-covariance matrix Ω and the variance-covariance matrix of the disturbances $\hat{\Sigma}$:

$$\Omega = \frac{1}{T} (\tilde{R}_P - \mathbf{1}\bar{R}_P)' (\tilde{R}_P - \mathbf{1}\bar{R}_P) \tag{4.18}$$

$$\hat{\Sigma} = \frac{\eta' \eta}{T - L - 1} \tag{4.19}$$

$\bar{R}_P \equiv \frac{1}{T}\sum_{t=1}^{T} \tilde{R}_{P,t}$ is a $1 \times L$ vector of factor means, $\mathbf{1}$ is a T-dimensional column vector filled with ones and $\eta \in \mathbb{R}^{T \times N}$ is the residual matrix $(\eta_1, \eta_2, ..., \eta_N)$. Finally the p-value of the GRS-test is calculated as:

$$p\text{-}value = 1 - F(\tilde{W},\ N,\ T\text{–}N\text{–}L) \tag{4.20}$$

with the cumulative distribution function F evaluated at $\tilde{W}$. The results of the GRS-test are shown in Table 4.12 and 4.11 for regressions performed over (1) the full period from Jan. 2002 to Dec. 2020 and (2) from Jan. 2012 to Dec. 2020. In addition to the GRS test statistic and it's p-value we also report the average absolute value of the regression intercepts $\|\alpha\| = \frac{1}{N}\sum_{i=1}^{N} \|\alpha_i\|$ and the associated t-value which we calculate as follows: We take the unbiased estimator for the residual variance $\hat{\sigma}_i^2 = diag(\hat{\Sigma})_i$ and the fist element d_0 of $d = diag((\tilde{R}_P^{'}\,\tilde{R}_P)^{-1})$ to calculate z_i (see Equation (4.21)). Then we calculate the average t-value according to Equation (4.22).

$$z_i = \frac{\alpha_i}{\hat{\sigma}_i\sqrt{d_0}} \quad \sim \quad t_{T-L-1} \tag{4.21}$$

$$t\text{-}value = \frac{1}{N}\sum_{i=1}^{N} \|z_i\| \tag{4.22}$$

We find that adding the GMB factor to the Fama-French 5-factor model (FF5) leads to a better model, as we observe a reduction in the GRS statistic in all cases (see Table 4.12 and 4.11). Consider the case of 72 industry portfolios in Panel E over the period (2): Extending the Fama French 5-factor model (FF5) by the GMB portfolio results in a reduction of the GRS statistic from 2.44 to 1.38 and in an increase of the associated p-value from 0.0027 to 0.158. Thus, while we reject the H_0 - all intercepts are jointly zero

- for the five-factor model, we cannot reject H_0 after adding GMB. This is also true for the three-factor- (FF3) and the market model (Mkt-Rf).

For PhysCR, the effect is not as consistent: we observe a small reduction in the GRS statistic in some cases, but also a worse result in others. An improvement can be observed for sector portfolios (Panel D) where the test statistic is further reduced if we add both, the GMB and the PhysCR factor. In period (2), the FF5 model has a GRS statistic of 1.999 with is lowered to 1.247 after including either the GMB portfolio. When we extend the FF5 model by both, GMB and PhysCR, we obtain a GRS statistic of 1.239. This is also the case for the three-factor- (FF3) and the market model (Mkt-Rf).

These results show, that the inclusion of climate risk factors substantially improves the explainability of variations in asset returns, especially since 2012. In Panel A to C, for 25 Size-B2M/OP/INV sorted portfolios we also observe similar results.

	(1) Full Period: Jan. 2002 to Dec. 2020				(2) Period: Jan. 2012 to Dec. 2020			
	GRS	$\|\alpha\|$ (%)	p-val. (GRS)	t-value	GRS	$\|\alpha\|$ (%)	p-val. (GRS)	t-value
Panel D: 10 Sector portfolios								
Mkt-Rf	1.1352	0.2192	0.3371	1.0506	2.3902	0.4562	0.014	1.3785
Mkt-Rf+GMB	1.1266	0.1918	0.3434	0.9766	1.0451	0.2424	0.4124	0.8245
Mkt-Rf+PhysCR	1.3818	0.2393	0.1901	1.1641	2.2471	0.4454	0.021	1.4069
Mkt-Rf+GMB+PhysCR	1.2722	0.2013	0.2476	1.0058	1.0346	0.245	0.4209	0.8713
FF3	0.9397	0.151	0.4976	0.7955	1.7691	0.352	0.0768	1.2351
FF3+GMB	0.9616	0.1459	0.4781	0.7886	1.1006	0.2411	0.3699	0.8851
FF3+PhysCR	1.1089	0.1799	0.3566	0.8941	1.7503	0.3519	0.0809	1.2654
FF3+GMB+PhysCR	0.9685	0.1398	0.472	0.7344	1.0809	0.2407	0.3849	0.9326
FF5	1.0931	0.2133	0.3687	0.9579	1.9986	0.3557	0.0421	1.2942
FF5+GMB	0.8411	0.1563	0.5896	0.7923	1.2467	0.2462	0.2726	0.9521
FF5+PhysCR	1.0543	0.2001	0.3993	0.9363	1.9906	0.3581	0.0431	1.3125
FF5+GMB+PhysCR	0.8772	0.1592	0.5554	0.8126	1.2389	0.2431	0.2774	0.9692
Panel E: 72 Industry portfolios								
Mkt-Rf	2.4765	0.2833	0.0	0.8488	2.6385	0.5034	0.0008	1.1386
Mkt-Rf+GMB	2.5458	0.2707	0.0	0.836	1.1576	0.2958	0.3183	0.677
Mkt-Rf+PhysCR	2.5259	0.2891	0.0	0.8826	2.3521	0.4804	0.0028	1.1102
Mkt-Rf+GMB+PhysCR	2.4044	0.2662	0.0	0.8381	1.2113	0.2951	0.2681	0.6873
FF3	2.5421	0.2653	0.0	0.8014	2.2481	0.3565	0.0046	0.8228
FF3+GMB	2.6513	0.263	0.0	0.8129	1.398	0.2974	0.1398	0.7049
FF3+PhysCR	2.3929	0.2638	0.0	0.7965	2.1963	0.3567	0.0062	0.8369
FF3+GMB+PhysCR	2.2247	0.2443	0.0	0.7596	1.4253	0.294	0.1292	0.7091
FF5	2.2793	0.2967	0.0	0.8678	2.4352	0.3505	0.0027	0.8265
FF5+GMB	2.1729	0.2759	0.0	0.8258	1.3754	0.3023	0.158	0.7283
FF5+PhysCR	2.2959	0.2973	0.0	0.8808	2.4314	0.3546	0.0031	0.8464
FF5+GMB+PhysCR	2.1677	0.2717	0.0	0.8181	1.4034	0.2969	0.1462	0.7259

Table 4.11: Continuation of Table 4.12. Panel D shows the GRS statistics for 10 sector portfolios and Panel E for 72 industry portfolios according to the SIC classification scheme. Using the GRS-test we test whether the alphas for each set of 10 or 72 time-series regressions are jointly zero. The table shows the GRS test statistic and it's p-value together with the average absolute value of the intercepts $\|\alpha\|$ and it's associated t-value.

	(1) Full Period: Jan. 2002 to Dec. 2020				(2) Period: Jan. 2012 to Dec. 2020			
	GRS	$\|\alpha\|$ (%)	p-val. (GRS)	t-value	GRS	$\|\alpha\|$ (%)	p-val. (GRS)	t-value
Panel A: 25 Size-B2M portfolios								
Mkt-Rf	1.554	0.193	0.0517	1.1592	1.322	0.4444	0.1743	1.7047
Mkt-Rf+GMB	1.508	0.1662	0.0646	1.0171	0.9839	0.1177	0.4971	0.4895
Mkt-Rf+PhysCR	1.6163	0.1948	0.0379	1.2152	1.245	0.3885	0.2286	1.5357
Mkt-Rf+GMB+PhysCR	1.5764	0.1692	0.0464	1.0869	0.9718	0.1176	0.5123	0.5038
FF3	1.7537	0.1135	0.0186	0.9971	1.2062	0.134	0.2607	0.8241
FF3+GMB	1.7326	0.1109	0.0209	0.9796	1.0169	0.1175	0.4571	0.7132
FF3+PhysCR	1.7503	0.1162	0.019	1.0459	1.1912	0.1321	0.2741	0.8202
FF3+GMB+PhysCR	1.6849	0.1099	0.0268	0.9934	1.0058	0.1174	0.4707	0.7198
FF5	1.362	0.0997	0.1261	0.9291	1.3062	0.1385	0.1863	0.8749
FF5+GMB	1.2433	0.0914	0.2059	0.8523	1.0939	0.1171	0.3701	0.7349
FF5+PhysCR	1.3502	0.0965	0.1328	0.9033	1.2866	0.1358	0.2001	0.869
FF5+GMB+PhysCR	1.2508	0.0932	0.2	0.8795	1.08	0.1171	0.3854	0.7413
Panel B: 25 Size-OP portfolios								
Mkt-Rf	0.9363	0.1331	0.5551	0.8323	1.3392	0.3734	0.1637	1.4663
Mkt-Rf+GMB	1.0581	0.1242	0.3947	0.8228	0.8839	0.0885	0.6248	0.4085
Mkt-Rf+PhysCR	0.9693	0.1331	0.51	0.8381	1.282	0.3378	0.2013	1.3197
Mkt-Rf+GMB+PhysCR	1.0178	0.1174	0.4457	0.7959	0.8809	0.0884	0.6286	0.4114
FF3	1.2149	0.0974	0.2294	0.8782	1.4065	0.1177	0.1285	0.7121
FF3+GMB	1.2561	0.1033	0.1956	0.9666	1.0942	0.0891	0.369	0.5637
FF3+PhysCR	1.1848	0.0903	0.2568	0.8333	1.4285	0.1188	0.1187	0.7189
FF3+GMB+PhysCR	1.1515	0.094	0.2896	0.8931	1.1084	0.0891	0.3544	0.5652
FF5	0.984	0.0831	0.4903	0.7965	1.4002	0.117	0.1326	0.7622
FF5+GMB	0.9009	0.0789	0.6041	0.7784	1.1044	0.0871	0.3589	0.584
FF5+PhysCR	0.9974	0.0835	0.4724	0.803	1.4318	0.1192	0.1182	0.7776
FF5+GMB+PhysCR	0.893	0.0778	0.6149	0.7703	1.1248	0.0867	0.3382	0.5833
Panel C: 25 Size-Inv portfolios								
Mkt-Rf	1.7634	0.1301	0.0176	0.8195	1.1976	0.3624	0.2672	1.4911
Mkt-Rf+GMB	1.7836	0.1184	0.0158	0.7598	1.0721	0.1032	0.3926	0.4901
Mkt-Rf+PhysCR	1.7527	0.1306	0.0187	0.831	1.1333	0.3204	0.3278	1.3054
Mkt-Rf+GMB+PhysCR	1.7787	0.1131	0.0163	0.7618	1.0633	0.1034	0.4027	0.5042
FF3	2.2432	0.1046	0.0011	1.0338	1.3354	0.1161	0.167	0.7832
FF3+GMB	2.2194	0.1092	0.0013	1.0883	1.2211	0.1021	0.2486	0.686
FF3+PhysCR	2.2059	0.0982	0.0014	0.9771	1.3408	0.1192	0.1643	0.8153
FF3+GMB+PhysCR	2.1245	0.1031	0.0023	1.0324	1.2139	0.1022	0.255	0.6982
FF5	1.7219	0.0793	0.0221	0.8469	1.6176	0.1276	0.0565	0.917
FF5+GMB	1.5889	0.0795	0.0438	0.8465	1.3443	0.1037	0.1632	0.7396
FF5+PhysCR	1.7063	0.0796	0.0241	0.8521	1.6409	0.1312	0.0518	0.9469
FF5+GMB+PhysCR	1.5928	0.0789	0.043	0.843	1.3533	0.103	0.1586	0.7422

Table 4.12: Summary statistics for GRS (Gibbons et al., 1989) tests of the market-, three-factor-, five-factor model and an extension of each by our regulatory- and physical climate risk factors (GMB and PhysCR) over (1) the full period from Jan. 2002 to Dec. 2020 (228 month) and (2) the period Jan. 2012 to Dec. 2020 (108 month). We test the ability of the various factor models to explain monthly excess returns on 25 Size-B2M portfolios (Panel A), 25 Size-OP portfolios (Panel B), 25 Size-Inv portfolios (Panel C). Using the GRS-test we test whether the alphas for each set of 25 time-series regressions are jointly zero. The table shows the GRS test statistic and it's p-value together with the average absolute value of the intercepts $\|\alpha\|$ and it's associated t-value.

Do Climate Betas Predict Future News Flow?

If the climate beta correctly identifies green and brown firms we would assume that companies with a high (low) regulatory climate risk beta, i.e., green (brown) companies, are linked to future news articles of these topic. Similarly, we would assume that firms with a high physical climate risk beta are also affected by future realizations of this climate risk.

To test if this is the case we regress the future (log-transformed) topic exposure onto the climate betas. Therefore we measure the firm-specific exposures to the sustainability topic $\bar{E}_{k=1,p}$ and the exposures to the regulatory climate risk topic $\bar{E}_{k=2,p}$ over 24 month from t to $t+24$ (see Equation (4.23)). As the topic exposure is a highly skewed variable – with several firms having high exposure while many have very low to no exposure – we log-transform the data as before. For regulatory climate risk, we regress the difference in the exposures on the sustainability- and the regulatory climate risk topic ($\bar{E}_{t,1,p} - \bar{E}_{t,2,p}$) on the regulatory climate risk beta (Equation (4.24)). Similarly, for physical climate risk, we regress the firm-specific exposures to the physical climate risk topic $\bar{E}_{k=3,p}$ on the physical climate risk beta (Equation (4.25)).

As the climate betas $\beta_{RegCR,t,p}$ and $\beta_{PhysCR,t,p}$ are re-estimated on a monthly frequency, we also run the regression in a monthly interval from Jan. 2002 to Dec. 2018. The resulting coefficient b of the regression and it's t-values are shown in Figure 4.14. We observe that the regression coefficients are almost always positive, indicating a positive relationship between the climate beta and future news content. The average value of b (t-value) is 0.073 (3.63) for regulatory climate risk and 0.168 (5.79) for physical climate risk. This implies that firms with a positive climate risk beta are more likely affected by the consequences of climate change in the future.

$$\bar{E}_{t,k,p} = log\left(1 + \sum_{t}^{t+24} \bar{I}_{t,k,p}\right) \tag{4.23}$$

$$\bar{E}_{t,1,p} - \bar{E}_{t,2,p} = a_t + b_t \times \beta_{RegCR,t,p} + \epsilon_{t,p} \tag{4.24}$$

$$\bar{E}_{t,3,p} = a_t + b_t \times \beta_{PhysCR,t,p} + \epsilon_{t,p} \tag{4.25}$$

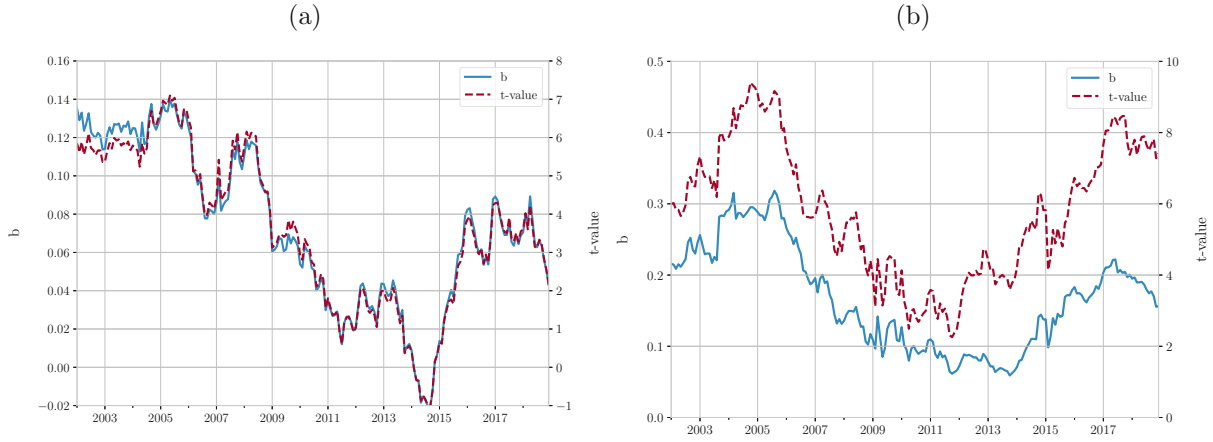


Figure 4.14: Coefficient b of the regressions shown in Formula 4.24 and 4.25 plotted together with the t-value over the period Jan. 2002 to Dec. 2018 for (a) the regulatory climate risk beta (β_{RegCR}) and (b) the physical climate risk beta (β_{PhysCR})

4.6 Conclusion

In this study, we propose a fully data-driven methodology to estimate firm-specific climate risk from public news. By utilizing a comprehensive dataset of almost 5 million U.S. news articles, we gain extensive support in the data to estimate the physical and regulatory climate risks for a wide range of U.S. stocks.

Our first main empirical finding is that we are the first to document a significant and economically sizable positive risk premium of 1.5% p.a. for physical climate risk over the period 2002 to 2020. This result is also robust to sector- and industry fixed effects indicating the risk exposure is at the individual firm-level.

Our second main result contributes to the ongoing discussion in the literature about the risk premium associated with regulatory climate risk. A portfolio that is long "green" stocks (low regulatory risk and good sustainability performance) and short "brown" stocks (high regulatory risk) reveals a regime shift occurring around 2012. The regulatory risk premium is positive from 2002 to 2012 (1.54% p.a.), but switches sign in the subsequent period from 2012 to 2020 becoming significantly negative with a point estimate of -2.56%. Thus, we contribute to the ongoing controversy in the literature about the sign of the regulatory climate risk premium as we are able to document this regime shift in a consistent

framework. This is due to the use of news data that allow us to estimate firm-specific climate risk exposures back to 2002, while traditional data sources such as ESG datasets only start in the 2010s.

Methodologically, we apply a novel machine learning technique to identify topic clusters in unstructured text, called Guided Topic Modeling. We furthermore extend the firm-specific news-based climate risk estimates to a universe of 9000 U.S. equities by calculating physical- and regulatory climate risk betas. When forming climate risk portfolios as before, but sorting stocks by their climate risk betas, we observe very similar patterns in the return series indicating that climate risk betas are useful and informative proxies for individual firm's exposures to regulatory and physical climate risks.

On a related note, a comparison between our climate beta sorted GMB portfolio and the ESG-sorted GMB portfolio of Pástor et al. (2022) shows a surprisingly high similarity in realized excess returns, yielding a correlation coefficient of 0.64. This adds validity to the proposed methodology and our results. It also suggests that news-based proxies, representing a top-down approach that only requires news as input, might be feasible, cost-effective alternative measures of company-specific climate risks compared to bottom-up ESG-scores with extensive data requirements.

Bibliography

Angelov, D. (2020). Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470.

Ardia, D., Bluteau, K., Boudt, K., and Inghelbrecht, K. (2020). Climate change concerns and the performance of green versus brown stocks. Available at SSRN 3717722.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. The quarterly journal of economics, 131(4):1593–1636.

Berkman, H., Jona, J., and Soderstrom, N. S. (2021). Firm-specific climate risk and market valuation. Available at SSRN 2775552.

Bolton, P. and Kacperczyk, M. (2021). Global pricing of carbon-transition risk. Technical report, National Bureau of Economic Research.

Carhart, M. M. (1997). On persistence in mutual fund performance. The Journal of finance, 52(1):57–82.

Choi, D., Gao, Z., and Jiang, W. (2020). Attention to global warming. The Review of Financial Studies, 33(3):1112–1145.

Dangl, T. and Salbrechter, S. (2023). Guided topic modeling with word2vec. Available at SSRN.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? The review of Financial studies, 22(5):1915–1953.

Engle, R. F., Giglio, S., Kelly, B., Lee, H., and Stroebel, J. (2020). Hedging climate change news. The Review of Financial Studies, 33(3):1184–1216.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. the Journal of Finance, 47(2):427–465.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of financial economics, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. Journal of financial economics, 116(1):1–22.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3):535–74.

Gibbons, M. R., Ross, S. A., and Shanken, J. (1989). A test of the efficiency of a given portfolio. Econometrica: Journal of the Econometric Society, pages 1121–1152.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794.

Hayes, P. J. and Weinstein, S. P. (1990). Construe/tis: A system for content-based indexing of a database of news stories. In IAAI, volume 90, pages 49–64.

Hsu, P.-H., Li, K., and TSOU, C.-Y. (2022). The pollution premium. The Journal of Finance.

King, G., Lam, P., and Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. American Journal of Political Science, 61(4):971–988.

Kölbel, J. F., Leippold, M., Rillaerts, J., and Wang, Q. (2020). Ask bert: How regulatory disclosure of transition and physical climate risks affects the cds term structure. Swiss Finance Institute Research Paper, (21-19).

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Newey, W. K. and West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix.

Pástor, L., Stambaugh, R. F., and Taylor, L. A. (2022). Dissecting green returns. Journal of Financial Economics, 146(2):403–424.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Sautner, Z., van Lent, L., Vilkov, G., and Zhang, R. (2023a). Firm-level climate change exposure. Journal of Finance. forthcoming.

Sautner, Z., Van Lent, L., Vilkov, G., and Zhang, R. (2023b). Pricing climate change exposure. Management Science.

Seltzer, L. H., Starks, L., and Zhu, Q. (2022). Climate regulatory risk and corporate bonds. Technical report, National Bureau of Economic Research.

Sia, S., Dalmia, A., and Mielke, S. J. (2020). Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! arXiv preprint arXiv:2004.14914.

Wireless Infrastructure Association (2023). Wireless infrastructure by the numbers.

Appendices

C.1 Industry Exposure of Climate Risk Beta Sorted Portfolios

	W (%)		(a) Reg. Climate Risk	W (%)		(b) Phys. Climate Risk	W (%)		(c) Sustainability
0	13.08	13.08	Oil And Gas Extraction	9.03	9.03	Electric, Gas, And Sanitary Services	19.81	19.81	Business Services
1	10.30	23.38	Petroleum Refining And Related Industries	8.90	17.93	Chemicals And Allied Products	14.55	34.36	Electronic And Other Electrical Equipment And ...
2	7.86	31.23	Electric, Gas, And Sanitary Services	7.97	25.90	Insurance Carriers	13.18	47.54	Industrial And Commercial Machinery And Comput...
3	6.15	37.38	Insurance Carriers	7.63	33.53	Petroleum Refining And Related Industries	8.97	56.51	Chemicals And Allied Products
4	5.84	43.22	Chemicals And Allied Products	6.77	40.30	Food And Kindred Products	4.74	61.25	Communications
5	5.60	48.82	Industrial And Commercial Machinery And Comput...	4.62	44.92	Measuring, Analyzing, And Controlling Instrume...	3.91	65.17	Measuring, Analyzing, And Controlling Instrume...
6	4.73	53.54	Automotive Dealers And Gasoline Service Stations	4.55	49.46	Oil And Gas Extraction	3.56	68.73	Insurance Carriers
7	4.45	58.00	Business Services	4.12	53.59	Automotive Dealers And Gasoline Service Stations	3.32	72.05	General Merchandise Stores
8	3.75	61.74	Transportation Equipment	4.04	57.63	General Merchandise Stores	2.86	74.91	Food And Kindred Products
9	3.18	64.93	Measuring, Analyzing, And Controlling Instrume...	3.74	61.37	Industrial And Commercial Machinery And Comput...	2.14	77.05	Building Materials, Hardware, Garden Supply, A...
10	3.04	67.96	Food And Kindred Products	3.30	64.67	Business Services	1.82	78.87	Transportation Equipment
11	2.14	70.10	Electronic And Other Electrical Equipment And ...	3.07	67.74	Communications	1.54	80.41	Engineering, Accounting, Research, Management,...
12	2.13	72.23	Tobacco Products	2.79	70.53	Tobacco Products	1.54	81.95	Apparel And Accessory Stores
13	1.97	74.20	Railroad Transportation	2.69	73.22	Transportation Equipment	1.35	83.29	Miscellaneous Retail
14	1.65	75.84	Nonclassifiable Establishments	2.33	75.55	Engineering, Accounting, Research, Management,...	1.26	84.56	Tobacco Products
15	1.61	77.45	Metal Mining	1.88	77.43	Railroad Transportation	1.10	85.66	Home Furniture, Furnishings, And Equipment Stores
16	1.42	78.87	Primary Metal Industries	1.86	79.29	Transportation By Air	1.00	86.66	Hotels, Rooming Houses, Camps, And Other Lodgi...
17	1.41	80.28	Wholesale Trade-durable Goods	1.77	81.07	Electronic And Other Electrical Equipment And ...	1.00	87.66	Transportation By Air
18	1.27	81.55	Communications	1.66	82.73	Eating And Drinking Places	0.98	88.64	Eating And Drinking Places
19	1.26	82.80	Fabricated Metal Products, Except Machinery An...	1.27	84.00	Building Materials, Hardware, Garden Supply, A...	0.97	89.61	Stone, Clay, Glass, And Concrete Products

Table 13: Top 20 industries of the climate risk beta sorted (a) regulatory climate risk, (b) physical climate risk and (c) sustainability portfolio calculated over the period Jan. 2002 to Dec. 2020. Industry exposures are calculated by summing up the monthly portfolio weights of each company over the period 2002 to 2020 and aggregating the weights at the industry level. To adjust for different industry sizes, in terms of industry firm count, we normalize the aggregate industry exposure by $(1 + log(industry\ size))$. We use the logarithm to avoid overly penalizing large industries.

C.2 Equal-Weighted Portfolios

	Regulatory Climate Risk (Brown) Portfolio			Physical Climate Risk Portfolio			Sustainability (Green) Portfolio		
	Weight (%)		Company Name	Weight (%)		Company Name	Weight (%)		Company Name
0	0.08	0.08	Seacor Holdings Inc	0.08	0.08	PNM Resources Inc	0.08	0.08	Best Buy Company Inc
1	0.08	0.16	Carbo Ceramics Inc	0.08	0.16	MDU Resources Group Inc	0.08	0.16	Cisco Systems Inc
2	0.08	0.25	Talos Energy Inc	0.08	0.25	UGI Corp New	0.08	0.25	Superconductor Technologies Inc
3	0.08	0.33	Universal Stnless & Aly Prods In	0.08	0.33	Firstenergy Corp	0.08	0.33	Intel Corp
4	0.08	0.41	Carrizo Oil & Gas Inc	0.08	0.41	Pricesmart Inc	0.08	0.40	Powerfleet Inc
5	0.08	0.49	Willbros Group Inc Del	0.08	0.49	Selective Insurance Group Inc	0.08	0.48	Alpha Pro Tech Ltd
6	0.08	0.57	Commercial Metals Co	0.08	0.58	J & J Snack Foods Corp	0.08	0.56	Cumulus Media Inc
7	0.08	0.66	Ion Geophysical Corp	0.08	0.66	Chevron Corp New	0.08	0.64	Repligen Corp
8	0.08	0.74	Team Inc	0.08	0.74	Idacorp Inc	0.08	0.71	Extreme Networks Inc
9	0.08	0.82	Cleveland Cliffs Inc New	0.08	0.82	Northwest Natural Holding Co	0.08	0.79	Innodata Inc
10	0.08	0.90	TRC Companies Inc	0.08	0.90	WGL Holdings Inc	0.08	0.86	Option Care Health Inc
11	0.08	0.98	Bristow Group Inc	0.08	0.99	Brady Corp	0.07	0.94	Broadvision Inc
12	0.08	1.06	United States Steel Corp New	0.08	1.07	Southwest Gas Holdings Inc	0.07	1.01	Sandisk Corp
13	0.08	1.14	Murphy Oil Corp	0.08	1.15	Black Hills Corp	0.07	1.09	Tivo Inc
14	0.08	1.22	PDC Energy Inc	0.08	1.23	RLI Corp	0.07	1.16	Windtree Therapeutics Inc
15	0.08	1.30	Casella Waste Systems Inc	0.08	1.31	Atmos Energy Corp	0.07	1.23	Mattson Technology Inc
16	0.08	1.38	Helix Energy Solutions Group Inc	0.08	1.40	Southern Co	0.07	1.30	Dot Hill Systems Corp
17	0.08	1.46	Gibraltar Industries Inc	0.08	1.48	Cincinnati Financial Corp	0.07	1.38	Amkor Technology Inc
18	0.08	1.53	SM Energy Co	0.08	1.56	Dominion Energy Inc	0.07	1.45	Anika Therapeutics Inc
19	0.08	1.61	Forward Air Corp	0.08	1.64	XCEL Energy Inc	0.07	1.52	EMC Corp Ma
20	0.08	1.69	Carpenter Technology Corp	0.08	1.73	Avista Corp	0.07	1.59	Empire Resorts Inc
21	0.08	1.77	AK Steel Holding Corp	0.08	1.81	Consolidated Edison Inc	0.07	1.66	Englobal Corp
22	0.08	1.84	Tetra Technologies Inc	0.08	1.89	DTE Energy Co	0.07	1.73	Oracle Corp
23	0.08	1.92	PHI Inc	0.08	1.97	Duke Energy Corp New	0.07	1.80	Cincinnati Bell Inc New
24	0.08	2.00	Unit Corp	0.08	2.05	Exxon Mobil Corp	0.07	1.87	Illumina Inc
25	0.08	2.07	Unifi Inc	0.08	2.14	Entergy Corp New	0.07	1.94	1 800 Flowers Com Inc
26	0.08	2.15	Energen Corp	0.08	2.22	Kelly Services Inc	0.07	2.01	Regeneron Pharmaceuticals Inc
27	0.08	2.22	Schnitzer Steel Industries Inc	0.08	2.30	Vectren Corp	0.07	2.08	Myriad Genetics Inc
28	0.08	2.30	Eog Resources Inc	0.08	2.38	Eversource Energy	0.07	2.15	CVD Equipment Corp
29	0.08	2.37	Hardinge Inc	0.08	2.46	American Financial Group Inc New	0.07	2.22	Lam Resh Corp

Table 14: Top holdings of the equal-weighted climate risk beta sorted regulatory climate risk (brown), physical climate risk and sustainability (green) portfolios. The weights are averages in %, calculated over the period Jan. 2002 to Dec. 2020.

C.3 Selective Climate Risk Beta Sorted Portfolios

We perform the same steps as described in Section 4.5.4 with the only difference that we now exclude all beta estimates with low statistical significance (t-value < 1). Table 15 highlights the top holdings of the value-weighted portfolios. It can be observed that firms like Apple and Meta are not present in the "Brown" portfolio anymore, as it was the case in Table 4.8. This also affects the cumulative portfolio returns shown in Figure 15. In particular, the "Brown" portfolio moves more or less sideways in the second half of the period, widening the gap between the "Green" and "Brown" portfolios.

	Regulatory Climate Risk (Brown) Portfolio			Physical Climate Risk Portfolio			Sustainability (Green) Portfolio		
	Weight (%)		Company Name	Weight (%)		Company Name	Weight (%)		Company Name
0	4.22	4.22	Exxon Mobil Corp	4.66	4.66	Exxon Mobil Corp	7.91	7.91	Intel Corp
1	2.08	6.30	Conocophillips	3.72	8.38	Chevron Corp New	7.33	15.24	Cisco Systems Inc
2	1.88	8.18	Occidental Petroleum Corp	2.59	10.96	Verizon Communications Inc	4.80	20.04	Microsoft Corp
3	1.74	9.92	Chevron Corp New	2.48	13.44	Walmart Inc	3.21	23.24	Apple Inc
4	1.47	11.39	Devon Energy Corp New	1.74	15.18	AT & T Inc	2.79	26.03	Home Depot Inc
5	1.27	12.66	Valero Energy Corp New	1.59	16.78	Coca Cola Co	2.51	28.54	Oracle Corp
6	1.21	13.86	APA Corp	1.40	18.17	Johnson & Johnson	2.40	30.95	Alphabet Inc
7	1.18	15.04	Eog Resources Inc	1.25	19.42	Procter & Gamble Co	2.39	33.33	Dell Inc
8	1.17	16.22	Halliburton Company	1.22	20.64	Nextera Energy Inc	2.03	35.36	Amazon Com Inc
9	1.17	17.38	Freeport Mcmoran Inc	1.19	21.83	Pepsico Inc	1.66	37.03	EMC Corp Ma
10	1.16	18.55	Anadarko Petroleum Corp	1.16	22.99	Southern Co	1.53	38.56	Yahoo Inc
11	1.00	19.54	NOV Inc	1.15	24.14	Duke Energy Corp New	1.47	40.03	Ford Motor Co Del
12	0.92	20.47	Berkshire Hathaway Inc Del	1.10	25.24	Dominion Energy Inc	1.39	41.42	Best Buy Company Inc
13	0.92	21.38	Southern Copper Corp	1.08	26.32	Berkshire Hathaway Inc Del	1.06	42.49	Corning Inc
14	0.88	22.26	Baker Hughes Co	0.98	27.29	Pfizer Inc	0.74	43.22	Johnson & Johnson
15	0.87	23.13	Coca Cola Co	0.94	28.24	Exelon Corp	0.73	43.96	Procter & Gamble Co
16	0.82	23.94	Hess Corp	0.94	29.18	Conocophillips	0.72	44.68	Juniper Networks Inc
17	0.76	24.70	American Airlines Group Inc	0.94	30.11	International Business Machs Cor	0.66	45.33	Nextel Communications Inc
18	0.76	25.46	Newmont Corp	0.86	30.97	Altria Group Inc	0.63	45.96	Las Vegas Sands Corp
19	0.75	26.21	Mosaic Company New	0.82	31.79	United Parcel Service Inc	0.62	46.58	Berkshire Hathaway Inc Del
20	0.75	26.96	Procter & Gamble Co	0.77	32.56	General Electric Co	0.56	47.14	Sun Microsystems Inc
21	0.73	27.69	Marathon Oil Corp	0.76	33.32	Allstate Corp	0.56	47.70	General Electric Co
22	0.71	28.39	Murphy Oil Corp	0.73	34.05	American Electric Power Co Inc	0.54	48.24	Gilead Sciences Inc
23	0.58	28.98	Chesapeake Energy Corp	0.73	34.79	Mcdonalds Corp	0.53	48.76	Pepsico Inc
24	0.58	29.56	Concho Resources Inc	0.73	35.52	Travelers Companies Inc	0.48	49.25	Qualcomm Inc
25	0.57	30.13	Williams Cos	0.65	36.17	Qualcomm Inc	0.47	49.71	Merck & Co Inc New
26	0.57	30.70	Noble Energy Inc	0.64	36.81	Boeing Co	0.46	50.18	Broadcom Corp
27	0.56	31.26	Exelon Corp	0.58	37.39	Caterpillar Inc	0.46	50.64	Lucent Technologies Inc
28	0.56	31.82	Continental Resources Inc	0.58	37.97	Waste Management Inc Del	0.46	51.10	International Business Machs Cor
29	0.54	32.36	Marathon Petroleum Corp	0.53	38.51	Philip Morris International Inc	0.42	51.52	Sandisk Corp

Table 15: Top 30 companies of the value-weighted climate risk beta sorted portfolios. We exclude all firms with beta estimates of low statistical significance (t-value < 1). The weights are averages in %, calculated over the period Jan. 2002 to Dec. 2020.

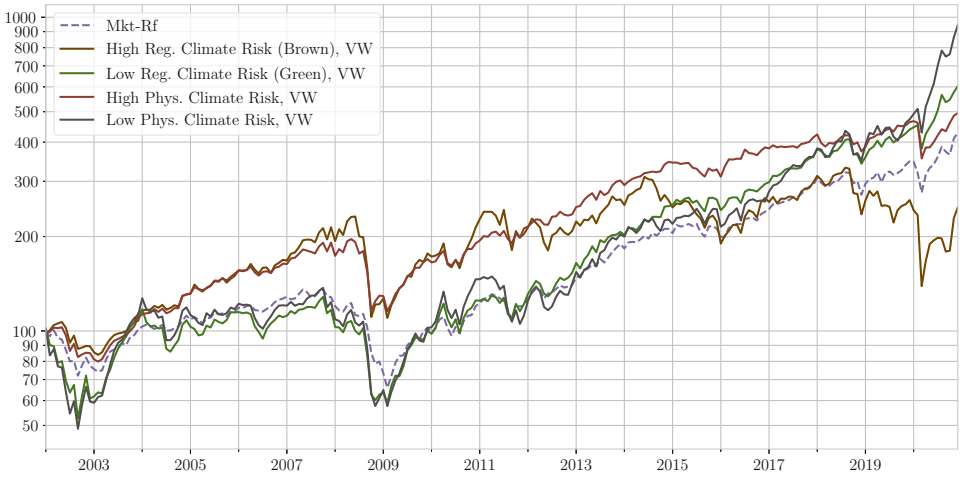


Figure 15: Cumulative returns of the value-weighted climate risk beta sorted portfolio over the period Jan. 2002 to Dec. 2020. We exclude all firms with beta estimates of low statistical significance (t-value < 1) and sort stocks in a high- and low regulatory climate risk portfolio, as well as a high- and low physical climate risk portfolio.

C.4 Climate Risk Beta Distributions

In Figure 16 we plot the cross-sectional distribution of the regulatory climate risk beta (left) and the physical climate risk beta (right) at three points in time: Jan. 2002, Jan. 2012 and Jan. 2020 and observe that the distribution of the physical climate risk beta is strongly skewed to the left in all cases (see Table 16). Also, we observe that only 28.8% of firms have a positive regulatory climate risk beta in 2020.

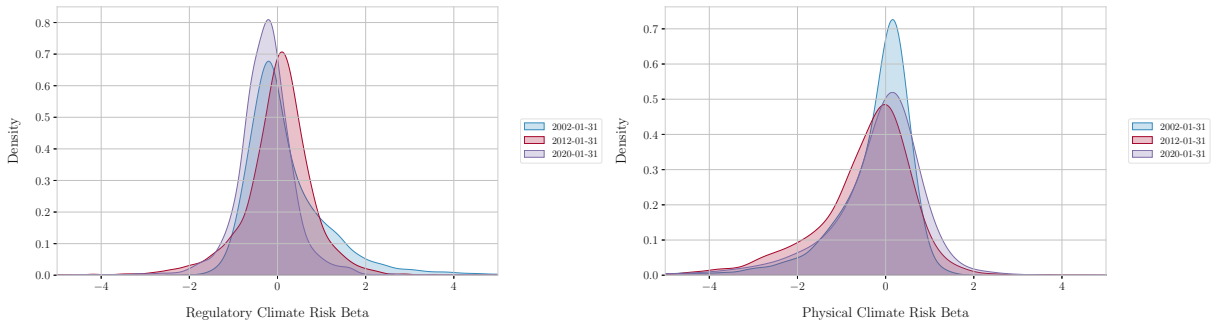


Figure 16: Distribution of the regulatory climate risk beta (left) and the physical climate risk beta (right) at three points in time: Jan. 2002, Jan. 2012 and Jan. 2020.

	Median	Mean	Std. dev.	Skewness	Kurtosis	% Pos
Reg. Climate Risk Beta						
2002-01-31	-0.039	0.216	0.990	2.253	9.732	0.470
2012-01-31	0.073	0.012	0.853	-2.391	38.762	0.555
2020-01-31	-0.254	-0.268	0.684	-0.365	26.695	0.288
Phys. Climate Risk Beta						
2002-01-31	-0.011	-0.264	0.964	-2.751	16.202	0.492
2012-01-31	-0.259	-0.493	1.234	-2.095	14.222	0.371
2020-01-31	0.004	-0.225	1.544	-18.932	735.129	0.502

Table 16: Descriptive statistics of the distributions shown in Figure 16.

In Figure 17 we plot the distribution of regulatory climate risk betas across different industries. We show the distribution for the industries *Electric, Gas, And Sanitary Services*, *Oil And Gas Extraction*, *Petroleum Refining And Related Industries* and *Business Services*. The climate risk betas of the industry *Business Services* is clearly skewed to the right, i.e., towards green firms while the betas of the industry *Oil And Gas Extraction* is skewed to the left, i.e., brown firms. The overlap between the distributions indicate a strong variation of firm-specific climate risks within industries.

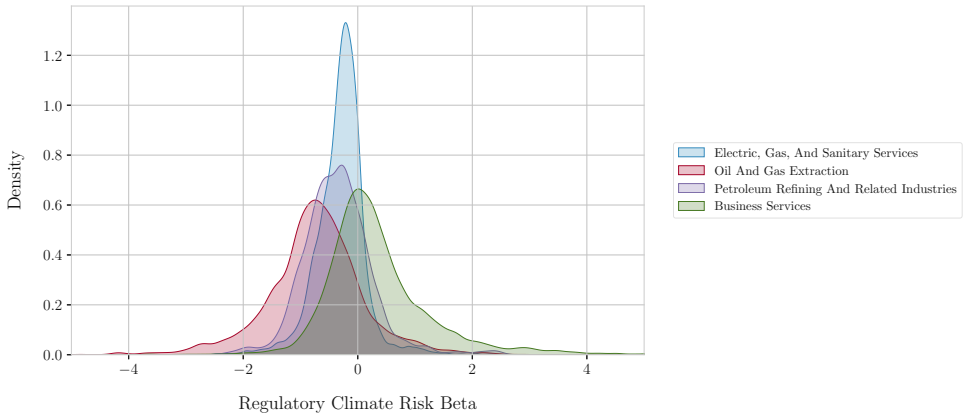


Figure 17: Distribution of the regulatory climate risk betas across different industries.

Figure 18 and 19 visualize the distribution of regulatory climate risk betas within selected industries on an annual basis via boxplots. We observe on average positive betas for firms in the industries *Business Services* and *Communications* and negative betas for firms in the industries *Oil And Gas Extraction* and *Electric, Gas, And Sanitary Services*.

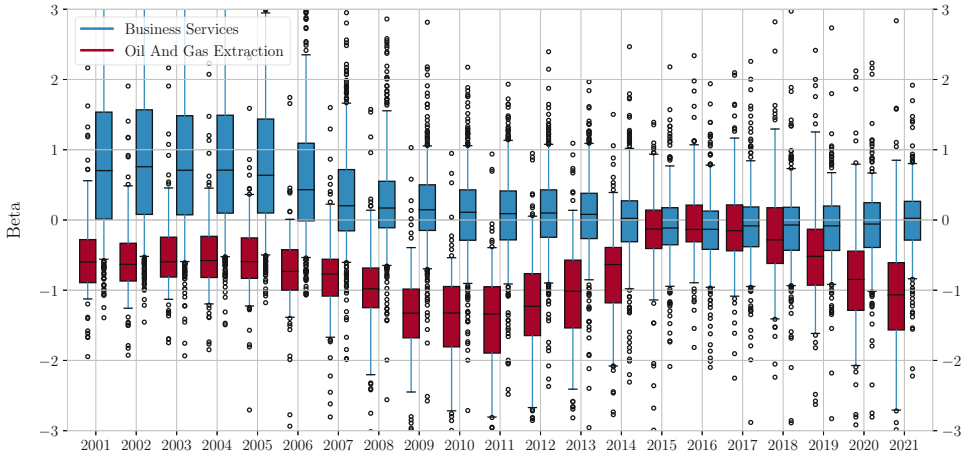


Figure 18: Annual boxplots highlighting the distribution of the regulatory climate risk beta for firms in the industries *Oil And Gas Extraction* and *Business Services*. We resample the monthly betas to an annual frequency by taking the mean.

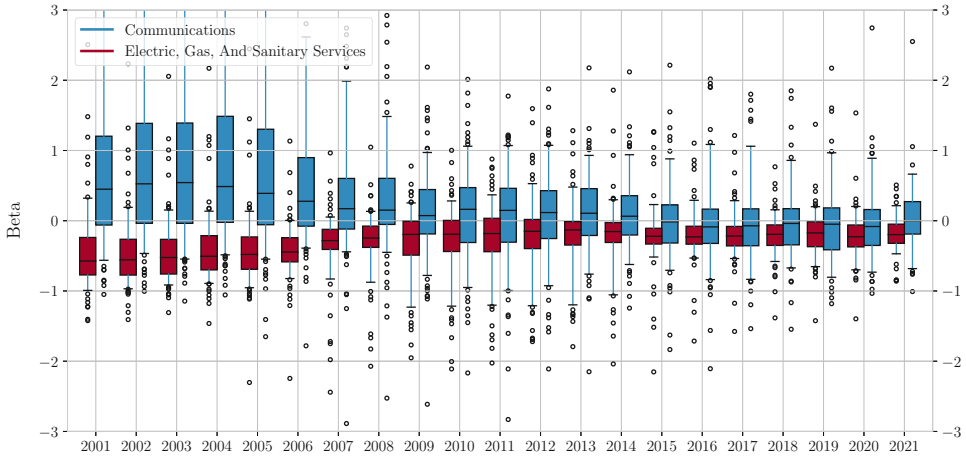


Figure 19: Annual boxplots highlighting the distribution of the regulatory climate risk beta for firms in the industries *Communications* and *Electric, Gas, And Sanitary Services*. We resample the monthly betas to an annual frequency by taking the mean.

	Model 1a			Model 1b			Model 2			Model 3			Model 4			Model 5			Model 6			Model 7		
Period	Full	1	2	Full	1	2	Full	1	2	Full	1	2	Full	1	2	Full	1	2	Full	1	2	Full	1	2
Const	12.64	10.48	15.03	11.46	8.96	14.24	12.54	10.52	14.79	12.62	10.73	14.72	11.32	8.71	14.23	11.46	8.96	14.24						
	(2.46)	(1.26)	(2.63)	(2.19)	(1.06)	(2.44)	(2.26)	(1.17)	(2.43)	(2.33)	(1.23)	(2.45)	(2.15)	(1.02)	(2.43)	(2.19)	(1.06)	(2.44)						
Beta	-2.91	-0.61	-5.48	-0.86	0.81	-2.73	-2.48	-0.59	-4.58	-1.66	0.21	-3.73	-0.95	0.92	-3.03	-0.64	1.18	-2.67	-0.85	1.05	-2.95	-1.40	0.55	-3.55
	(-2.16)	(-0.29)	(-3.66)	(-0.64)	(0.36)	(-2.02)	(-1.90)	(-0.29)	(-3.15)	(-1.27)	(0.10)	(-2.53)	(-0.73)	(0.43)	(-2.24)	(-0.49)	(0.57)	(-1.95)	(-0.64)	(0.49)	(-2.16)	(-1.25)	(0.32)	(-2.86)
Size	7.01	5.09	9.15	5.83	3.98	7.88	7.61	5.20	10.28	6.50	4.27	8.97	6.88	5.59	8.31	6.21	4.32	8.30	6.83	4.60	9.31	7.08	4.75	9.68
	(5.24)	(2.50)	(5.76)	(3.24)	(1.47)	(3.54)	(4.48)	(2.02)	(5.10)	(3.81)	(1.65)	(4.39)	(3.80)	(2.00)	(3.85)	(3.21)	(1.45)	(3.63)	(3.65)	(1.60)	(4.22)	(4.14)	(1.84)	(4.73)
B2M	-2.17	-2.28	-2.05	-2.30	-0.89	-3.87	-1.94	-1.05	-2.93	-2.06	-1.21	-3.01	-2.43	-1.15	-3.86	-2.28	-1.23	-3.44	-1.51	-0.88	-2.20	-1.30	-0.75	-1.92
	(-1.76)	(-1.18)	(-1.36)	(-1.45)	(-0.37)	(-1.93)	(-1.31)	(-0.44)	(-1.83)	(-1.44)	(-0.55)	(-1.68)	(-1.55)	(-0.47)	(-2.07)	(-1.48)	(-0.54)	(-1.69)	(-1.02)	(-0.39)	(-1.19)	(-0.87)	(-0.32)	(-1.07)
OP	0.54	0.50	0.57	-0.60	-0.89	-0.27	0.26	0.69	-0.23	-0.02	0.02	-0.07	0.21	0.15	0.29	-0.56	-0.90	-0.17	-0.53	-1.01	0.01	-0.73	-1.12	-0.29
	(0.92)	(0.76)	(0.58)	(-0.83)	(-0.94)	(-0.25)	(0.36)	(0.79)	(-0.20)	(-0.03)	(0.02)	(-0.06)	(0.32)	(0.16)	(0.29)	(-0.78)	(-0.94)	(-0.16)	(-0.79)	(-1.08)	(0.01)	(-1.14)	(-1.22)	(-0.33)
INV	0.32	-0.82	1.58	0.37	-0.37	1.19	0.40	-0.72	1.64	0.52	-0.47	1.62	0.30	-0.51	1.20	0.42	-0.23	1.14	0.22	-0.55	1.07	0.49	-0.08	1.11
	(0.54)	(-0.95)	(2.19)	(0.54)	(-0.40)	(1.19)	(0.67)	(-0.87)	(2.09)	(0.78)	(-0.52)	(1.72)	(0.47)	(-0.57)	(1.33)	(0.63)	(-0.25)	(1.15)	(0.34)	(-0.66)	(1.14)	(0.81)	(-0.09)	(1.27)
Sus							0.91	0.57	1.28							-0.05	-1.38	1.44	-0.10	-1.06	0.96	0.20	-0.24	0.69
							(1.58)	(0.64)	(1.84)							(-0.06)	(-1.24)	(1.39)	(-0.15)	(-1.17)	(0.95)	(0.31)	(-0.29)	(0.67)
Reg										-0.15	1.47	-1.96				-0.40	1.54	-2.56	0.28	2.04	-1.66	0.60	1.75	-0.69
										(-0.21)	(1.34)	(-2.59)				(-0.57)	(1.61)	(-2.94)	(0.45)	(2.38)	(-2.08)	(1.04)	(2.34)	(-0.85)
Phys													1.60	2.26	0.88	1.50	2.08	0.85	1.75	2.17	1.27	1.94	2.15	1.70
													(2.14)	(1.91)	(1.02)	(2.37)	(2.23)	(1.04)	(2.73)	(2.34)	(1.47)	(3.38)	(2.67)	(2.08)
Fixed effects	None			None			None			None			None			None			Sector fixed effects			Industry fixed effects		
Months	228	120	108	228	120	108	228	120	108	228	120	108	228	120	108	228	120	108	228	120	108	228	120	108
Observations	810766	810766	810766	189586	189586	189586	449319	449319	449319	331025	331025	331025	241756	241756	241756	189586	189586	189586	189586	189586	189586	189586	189586	189586
Firms	8151	8151	8151	3005	3005	3005	5621	5621	5621	4927	4927	4927	3660	3660	3660	3005	3005	3005	3005	3005	3005	3005	3005	3005
R2 (%)	2.86	3.12	2.57	6.04	6.65	5.37	4.03	4.54	3.46	4.53	4.73	4.32	5.54	6.31	4.67	7.24	8.11	6.26	11.36	12.77	9.79	29.20	32.70	25.32
Adj. R2 (%)	2.71	2.99	2.40	5.43	6.01	4.78	3.72	4.23	3.15	4.10	4.32	3.87	4.95	5.70	4.11	6.26	7.10	5.32	9.41	10.80	7.87	22.63	26.17	18.70

Note: *p<0.1; **p<0.05; ***p<0.01

Table 17: Fama-MacBeth regression performed over the periods Jan. 2002 to Dec. 2020 (Full), and the two subperiods Jan. 2002 to Dec. 2011 (1) and Jan. 2012 to Dec. 2020 (2). Model 1 is comprised of the classic risk factors that enter the Fama French five-factor model. Models 2 to 4 extend this model by one of the climate factors and Model 5 contains all climate factors. With Model 6 and 7 we control for fixed effects using sector dummies in Model 6 and industry dummies in Model 7. We consider 10 sectors (divisions) and 65 industries (with at least 1000 observations each) according to the SIC scheme. We report the annualized risk premia in percent and heteroskedasticity and autocorrelation (HAC) adjusted t-values (Newey and West (1986) standard errors with three lags). All characteristics except dummy variables are standardized for each of the n cross-sectional regressions. Also we exclude all observations with missing values in the cross sectional regression which causes the number of observations to decline relative to Model 1 as the number of firms with topic exposures is limited.

5 Guided Topic Modeling with Word2Vec: A Technical Note

Guided Topic Modeling with Word2Vec: A Technical Note

Thomas Dangl Stefan Salbrechter

September 19, 2023

We propose GTM (Guided Topic Modeling), an algorithm that enables the fast and flexible generation of comprehensive topic clusters from (a pair of) seed words. The unsupervised algorithm performs clustering in the word-embedding space while offering the possibility to adjust the characteristics of the topic clusters via several hyperparameters. Applications for this methodology are information retrieval, classification and the calculation of various topic indices from news feeds.

5.1 Introduction

We introduce a new algorithm, termed Guided Topic Modeling (GTM) with Word2Vec, that automatically generates topic word clusters, i.e., weighted keyword lists for an almost unlimited number of unique topics. Using the information encoded in vector representations of words that are learned by unsupervised pre-training on large text corpora, the only input required are (a pair of) seed words that are representative of a topic of interest. The iterative algorithm retrieves the most related words from the vector space to generate comprehensive topic clusters. This has the advantage that no further training dataset or expert knowledge is required as all the necessary information is already encoded in the word embeddings. Still, the algorithm is flexible and adjustable such that one can control the characteristics of the desired topic mappings. Internally, the algorithm does not simply add up the words closest to the seed words but also adjusts its topic center accordingly, such that it converges towards an optimal center. In this way, we can extract additional information from the list of topic words – a similarity parameter (weight) that is higher for words closer to the topic center, i.e., important topic words, and lower for words that are more distant form the topic center, i.e., less important words.

Applications of the GTM methodology include information retrieval and classification of text documents. A frequent task is to find all documents from a large dataset that describe a certain topic, person, concept, literature, sentiment, or event (King et al., 2017). This is usually done by keyword matching, i.e., finding all documents that contain words included in manually defined keyword lists. However, the manual creation of well specified keyword lists is a "near-impossible" task for humans (Hayes and Weinstein, 1990). Furthermore, King et al. (2017) show that the selection of an incomplete keywords list can result in a severe selection bias. The authors therefore propose an iterative, semi-automated algorithm that generates keyword suggestions. The researcher's task is then to manually select the most representative words. To implement this algorithm, a researcher, however, is required to have access to a sufficiently large dataset containing documents properly tagged with the desired concept of interest. Rinke et al. (2022) argue that a possible lack of researchers' domain-knowledge could affect their ability to select

the proper keywords from the suggested candidates. Therefore, the authors propose to include the expertise of external experts to obtain keywords through surveys. The domain expertise is then combined with an LDA (Latent Dirichlet Allocation) (see Blei et al., 2003) model, an unsupervised topic modeling algorithm, to discover relevant topics and to classify documents into these topics. However, conducting a survey for each topic is time-consuming and requires a considerable amount of effort.

GTM, in contrast, is fast and flexible as various topics can be generated in seconds. Borrowing from the idea of transfer learning (Devlin et al., 2018), word embeddings are trained once and serve as a foundation for the GTM algorithm to generate topic word clusters, i.e., keyword lists, that can be applied on several tasks and datasets without the need of further fine-tuning. The topic word clusters are generated by performing clustering in the embedding space. The GTM algorithm takes as input a list of two (or more) seed words, each associated with a weight parameter that defines the importance of each seed word in the topic cluster. Initially, the seed words span up a plane, or hyper-space in the case of more than two seed words, in the vector space. The iterative clustering algorithm then calculates projections of all words onto this plane, adds the closest word to the topic, re-fits the plane and starts all over by projecting words onto the new located plane. Thus, the topic center is not defined by the seed words but the algorithm iteratively finds a optimal topic center (i.e. the final location of the plane / hyper space). By using a weighted combination of seed words, the topics can be tilted towards specific characteristics. Depending on the settings, the clustering algorithm stops when a certain topic size is reached or when the distance to the next nearest word exceeds a predefined threshold.

Our methodology is also related to topic modeling approaches that aim to find latent topics in the data. The most commonly used topic model is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), beside more recent developments like Top2Vec (Angelov, 2020) and BERTopic (Grootendorst, 2022). These algorithms find a certain number of topics in the data, but since the user does not provide any guidance, the algorithm may detect topics that do not necessarily align with the researcher's interests. A researcher may be interested in the exposure of news articles or social media posts to specific topics like

inflation, climate risk, investor uncertainty or to any other topic. The classic clustering algorithm however does not guarantee that these topics will be identified. Our approach therefore aims to give researchers a tool to easily create topic word clusters for any topic of interest.

This allows applications such as the generation of topic indices, i.e., the calculation of topic loadings on text documents. Those topic indices can be applied to a variety of tasks in finance and economics. As an example, topic indices can be used to predict macroeconomic variables like unemployment rates, consumer price indices or house prices (Cong et al., 2019). Let's say a researcher wants to know how several topics, e.g., unemployment, uncertainty or bankruptcy, etc., are exposed in the news over time. Then, for each of those topics, one can generate a topic word cluster and calculate exposures. These exposures represent the sum of weighted counts of overlapping words among the generated topic words and the words in the news documents. Furthermore, it can be used to construct sentiment indices[1] or to measure individual firms' climate risk (Dangl et al., 2023).

The remainder of this paper is composed as follows: In Section 5.2 we provide a formal description of the clustering algorithm. In Section 5.2.1 we describe how we reduce the computational cost by applying efficient similarity search and in Section 5.2.2 we provide details about the hyperparameters. Section 5.2.3 discusses the Word2Vec algorithm which we utilize to obtain vector representations of words and in Section 5.2.4 we describe our methodology to obtain polar word embeddings, i.e., word embeddings that include sentiment information of words. In Section 5.3.1 we present a case study to quantify the capabilities of GTM to perform classification and in Section 5.3.2, we present a case study to show how the GTM algorithm is used to learn company-specific climate risks from public news.

5.2 Clustering Algorithm

The clustering algorithm intends to find optimal topic clusters for a given set of seed words. The algorithm takes as input two (or more) seed words whose vector representations

[1] In order to generate topics that distinguish between words with positive or negative polarity, we add an additional sentiment dimension to the word embeddings (see Section 5.2.4).

$\boldsymbol{x_i} \in \mathbb{R}^p$ span the initial projection plane $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_k}] \in \mathbb{R}^{p \times k}$ with $k \geq 2$ in the p-dimensional embedding space. To further adjust the orientation of the initial plane $\boldsymbol{X}$, l negative seed words $\boldsymbol{\eta_i}$ with $\boldsymbol{N} = [\boldsymbol{\eta_1}, ..., \boldsymbol{\eta_l}] \in \mathbb{R}^{p \times l}$ can be defined if needed. The seed vectors are scaled by the weight vectors $\boldsymbol{g_x} \in \mathbb{R}^k$, $\forall g \in \boldsymbol{g_x} : g > 0$ and $\boldsymbol{g_\eta} \in \mathbb{R}^l$, $\forall g \in \boldsymbol{g_\eta} : g < 0$. If more than one negative seed word is defined, we compute the average of the negative seed word vectors $\bar{\boldsymbol{\eta}} = \boldsymbol{N}\, diag(\boldsymbol{g_\eta})\, \boldsymbol{G}$, to account for the combined influence of the negative seed words. $\boldsymbol{G} \in \mathbb{R}^{l \times k}$ is a matrix filled with $1/l$. With negative seed words defined, the initial projection plane $\boldsymbol{X}$ is transformed into $\boldsymbol{X'}$ according to Equation 5.1.[2] If only one negative seed word is defined, then $\bar{\boldsymbol{\eta}}$ equals $\boldsymbol{\eta_1} \times g_\eta$ and if no negative seed words are defined, then $\boldsymbol{X'} = \boldsymbol{X}$

$$\boldsymbol{X'} = \boldsymbol{X}\, diag(\boldsymbol{g_x}) + \bar{\boldsymbol{\eta}} \tag{5.1}$$

The column vectors of $\boldsymbol{X'}$ are then transformed into vectors of unit length. A geometric interpretation is provided in Figure 5.1. After the initialization of $\boldsymbol{X'}$ all m word vectors $\boldsymbol{w}$ contained in the vocabulary $\boldsymbol{V} \in \mathbb{R}^{p \times m}$ are projected on $\boldsymbol{X'}$. The projection coefficients $\boldsymbol{b_j}$ with $\boldsymbol{B} = [\boldsymbol{b_1}, ..., \boldsymbol{b_m}] \in \mathbb{R}^{k \times m}$ are calculated by Equation (5.2).

$$\boldsymbol{B} = (\boldsymbol{X}^T\, \boldsymbol{X})^{-1}\, \boldsymbol{X}^T\, \boldsymbol{V} \tag{5.2}$$

$$\boldsymbol{B_{adj}} = diag(\boldsymbol{g_x})\, \boldsymbol{B} \tag{5.3}$$

The seed word weight vector $\boldsymbol{g_x}$ is used in Equation (5.3) to adjust the projection coefficients. Using Equations (5.4) to (5.6), the projection angles $\boldsymbol{\alpha'} = [\alpha'_1, ..., \alpha'_m] \in \mathbb{R}^m$ can be calculated with Equation (5.7). Finally, the word with the smallest projection angle α'_j is added to the topic.

[2] We stretch the column vectors of $\boldsymbol{X}$ with the weight vector $\boldsymbol{g_x}$ ($\boldsymbol{X}\, diag(\boldsymbol{g_x})$) to incorporate the seed word weights in the transformation of $\boldsymbol{X}$ to $\boldsymbol{X'}$. Thus, after the transformation, the $\boldsymbol{X'}$ column vectors remain closer to the $\boldsymbol{X}$ column vectors the larger the associated weight are. $\boldsymbol{X}\, diag(\boldsymbol{g_x})$ is of dimension $(p \times k) * (k \times k) = (p \times k)$. $\boldsymbol{N}\, diag(\boldsymbol{g_\eta})\, \boldsymbol{G}$ is of dimension $(p \times l) * (l \times l) * (l \times k) = (p \times k)$.

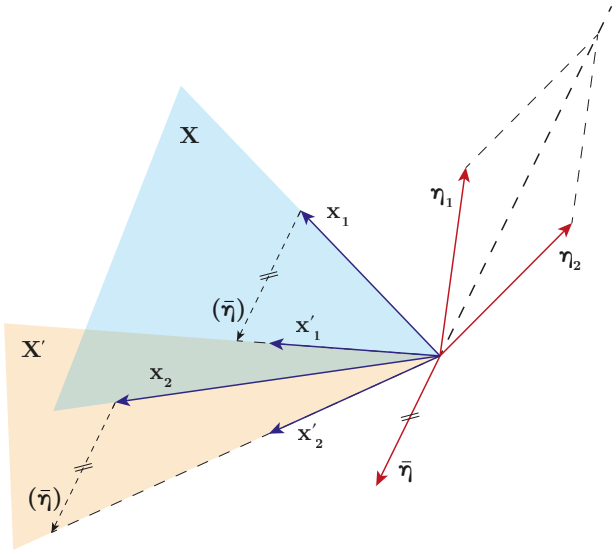


Figure 5.1: The initial Plane $\boldsymbol{X}$ that is spanned by the seed vectors $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ is transformed to $\boldsymbol{X'}$ with the negative seed vectors $\boldsymbol{\eta_1}$ and $\boldsymbol{\eta_2}$ according to Equation (5.1).

$$\hat{\boldsymbol{V}} = \boldsymbol{X}\,\boldsymbol{B} \tag{5.4}$$

$$\boldsymbol{V}^{\perp} = \boldsymbol{V} - \hat{\boldsymbol{V}} \tag{5.5}$$

$$\hat{\boldsymbol{V}}_{\boldsymbol{adj}} = \boldsymbol{X}\,\boldsymbol{B}_{\boldsymbol{adj}} \tag{5.6}$$

$$\boldsymbol{\alpha'} = \arctan\left(\frac{\sqrt{\langle \boldsymbol{V}^{\perp \boldsymbol{T}}, \boldsymbol{V}^{\perp}\rangle}}{\sqrt{\langle \hat{\boldsymbol{V}}^{T}_{\boldsymbol{adj}}, \hat{\boldsymbol{V}}_{\boldsymbol{adj}}\rangle}}\right) \tag{5.7}$$

Due to the scaling of the projection coefficients with the weight vector $\boldsymbol{g_x}$, we are able to favor the selection of words that are closer to the seed words that have higher weights assigned. As shown in Figure 5.2, a word vector $\boldsymbol{w}$ is projected on $\boldsymbol{X'}$ with the projection coefficients b_1 and b_2. By scaling the projection coefficients to $b_{adj,1}$ and $b_{adj,2}$, $\hat{\boldsymbol{w}}$ is transformed to $\hat{\boldsymbol{w}}_{\boldsymbol{adj}}$, $\boldsymbol{w}$ is transformed to $\boldsymbol{w}_{\boldsymbol{adj}}$ and the projection angle α becomes α'. Since $\boldsymbol{w}^{\perp}$ remains unchanged, stretching the projection coefficients means that $\alpha' < \alpha$ and α' becomes smaller for words where a large weight is multiplied with a large projection coefficient.

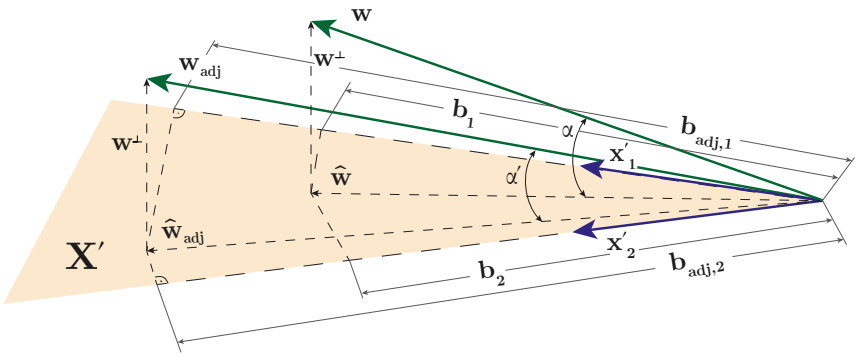


Figure 5.2: Weighting of the projection coefficients.

With the first word added to the topic, the plane $\boldsymbol{X'} \in \mathbb{R}^{p \times k}$ is re-fitted by minimizing the residual sum of squares (RSS) between the s=k+1 word vectors included in the cluster $\boldsymbol{C} \in \mathbb{R}^{(p \times s)}$. The new added word receives a weight of 1 resulting in the updated weight vector $\boldsymbol{g_x} = [\boldsymbol{g_{x1}}, ..., \boldsymbol{g_{xk}}, 1]$.

To solve the minimization problem, that is formulated in Equations (5.8) to (5.10), we use the iterative conjugate gradient method (CS). The orthogonal projection vector of the new added word $\boldsymbol{w}^{\perp}$ is multiplied with $\mathbf{1} \in \mathbb{R}^{\mathbf{k}}$, a k-dimensional row vector filled with ones, which results in $\boldsymbol{W}^{\perp} = [\boldsymbol{w}^{\perp}, ..., \boldsymbol{w}^{\perp}] \in \mathbb{R}^{p \times k}$. To minimize RSS, $\boldsymbol{X'}$ is transformed according to Equation (5.10) with $\boldsymbol{a} \in \mathbb{R}^{k}$ being the parameter vector that is initialized with $\mathbf{0}$. The CS algorithm iteratively adjusts $\boldsymbol{a}$ in order to minimize the residual sum of squares. After the algorithm converges, we update the plane $\boldsymbol{X'}$ by setting $\boldsymbol{X'} = \boldsymbol{X'_{new}}$ and the algorithm continues with the next iteration by projecting all words that are not jet included in topic C onto $\boldsymbol{X'}$. Finally, the algorithm stops when the pre-specified topic size is reached.

$$\boldsymbol{RSS} = \mathbf{1}^T \left(\boldsymbol{I} - \boldsymbol{H}(a)\right) \boldsymbol{C}\, diag(\boldsymbol{g_x})\, \boldsymbol{C}^T \left(\boldsymbol{I} - \boldsymbol{H}(a)\right)^T \mathbf{1} \tag{5.8}$$

with:

$$\boldsymbol{H}(a) = \boldsymbol{X'_{new}} \left(\boldsymbol{X'^T_{new}}\, \boldsymbol{X'_{new}}\right)^{-1} \boldsymbol{X'^T_{new}} \tag{5.9}$$

$$\boldsymbol{X'_{new}} = \boldsymbol{X'} + \boldsymbol{W}^{\perp}\, diag(\boldsymbol{a}) \tag{5.10}$$

5.2.1 Efficient Similarity Search

The calculation of the projection coefficients, as given by Equation (5.2), becomes more computationally intensive as the vocabulary size m increases. In order to reduce the computational demand we perform a pre-selection of K words with $K \ll m$. We therefore use the Python library Faiss (Johnson et al., 2019) that allows for fast and efficient approximate similarity search of dense vectors. The Faiss algorithm performs a L2 distance search in the embedding space. To increase the speed we do not perform an exhaustive but rather an approximate search. This is done by segmenting the dataset into multiple cells and assigning vectors to these cells. The query vectors are the word vectors of the positive and negative seed words. The Faiss algorithm then performs a similarity search with all datapoints that fall in the same and surrounding cells as the seed vectors. The output is a reduced vocabulary of the size K, i.e., the vocabulary contains K words that are most similar to the seed words.

Figure 5.3 visualizes the projection of all words of the vocabulary V onto the plane $\boldsymbol{X'}$ after the generation of a topic cluster (red dots). The blue dots indicate the word vectors that are identified with Faiss, the grey dots indicate all remaining words. It can be observed that the preselection works well, since no words not identified by Faiss are near the red dots, i.e., the topic word cluster.

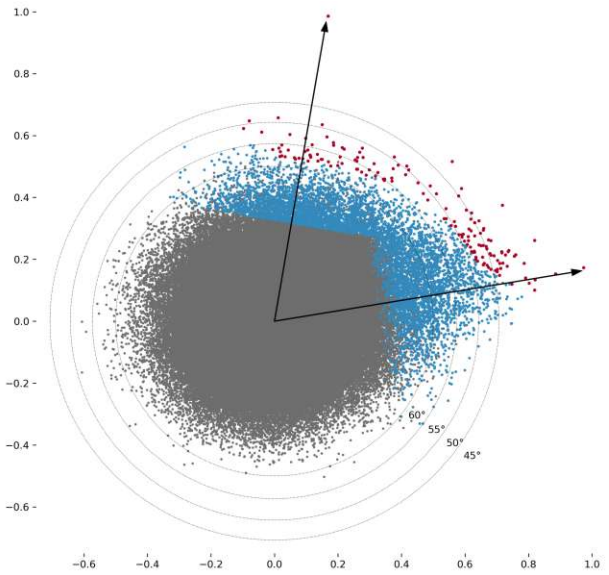


Figure 5.3: This plot visualizes the projection of all words of the vocabulary on the plane $\boldsymbol{X'}$. The red dots indicate the words that are included in the topic cluster $\boldsymbol{C}$, the blue dots indicate all K=3000 pre-selected words found by approcimate similarity search with Faiss. The grey dots indicate the projections of all remaining words of the vocabulary $\boldsymbol{V}$ that were not part of the clustering procedure. The two arrows show the vectors $\boldsymbol{x'_1}$ and $\boldsymbol{x'_2}$ that span the plane $\boldsymbol{X'}$.

5.2.2 Hyperparameters

The clustering algorithm takes several hyperparameters as input, which allow for tuning the cluster properties as well as the speed of the algorithm. The first three hyperparameters *nbrobe*, *nlist* and K adjust the properties of the similarity search with Faiss. Therby, *nlist* determines the number of cells, *nbrobe* specifies the number of surrounding cells that are taken into account for similarity search and K defines the number of similar words selected with Faiss. The values of these three hyperparameters represent a trade-off between accuracy and speed, with higher values implying higher accuracy.

The hyperparameters *cluster_size* and *alpha_max* control the topic-cluster size. The first, *cluster_size* is used to generate word clusters of fixed-size, i.e., the clustering algorithm stops when *cluster_size* is reached. Furthermore, *alpha_max* defines the maximum angle α' that is accepted. Thus, the iterative clustering algorithm stops when α', the projection angle of the closest word to $\boldsymbol{X'}$, exceeds the threshold *alpha_max*. Reasonable values of *alpha_max* are in the range of 0.5 to 1.5. The next hyperparameter *update_freq* specifies whether the plane should be re-fitted after each word added to the topic, or whether it

should be readjusted only after the number of words defined by *update_freq* respectively.

Furthermore, seed words are initialized with the weight vector $\boldsymbol{g_x}$. The default option is to set $\boldsymbol{g_x} = [1.0, 1.0]$ for the case of two seed words. Increasing the weight of one seed word tilts the topic word cloud more toward that seed word. This effect is shown in Figure 5.4 by using the two seed words "iphone" and "steve_jobs". The generated topic word cloud is shown in the left panels and the word projections onto the final plane $\boldsymbol{X'}$ in the right panels. The two arrows represent the two column vectors of the final plane $\boldsymbol{X'}$ which point close to the initial seed words (green dots). Since the location of the plane $\boldsymbol{X'}$ is adjusted during the iterative clustering procedure they usually don't align with the initial seed words. Increasing the weight of "iphone" from 1.0 to 2.0, as shown in the second row of Figure 5.4, tilts the topic closer to the concept of "iphone". This effect can also be observed in the right panel, since the red dots, i.e., the topic words, are all concentrated on the right side, close to the seed word "iphone". If we instead increase the weight of "steve_jobs" from 1.0 to 2.0, the topic is tilted towards the concept if CEOs and the red dots in the right panel are concentrated at the top of the plot, close to the seed word "steve_jobs".

In addition, the hyperparameter named *gravity* allows to control the location of the final cluster center. This is achieved by multiplying the weight vector $\boldsymbol{g_x}$ by $(1+gravity)$ in each iteration. Subsequently, the weight vector is extended by 1, i.e., the weight of a newly added word ($[\boldsymbol{g_{x1}}, ..., \boldsymbol{g_{xk}}, 1]$). The effect of the *gravity* parameter is to give precedence to words added early to the topic cluster, including the initial seed words, over those added later. It acts as a dragging force, pulling the topic center closer towards earlier added words. If the value of *gravity* is large, the topic center stays close to the initial plane $\boldsymbol{X'}$ spanned by the seed words. When the value of *gravity* is close to zero, the center of the cluster deviates more easily to denser locations (see Figure 5.5).

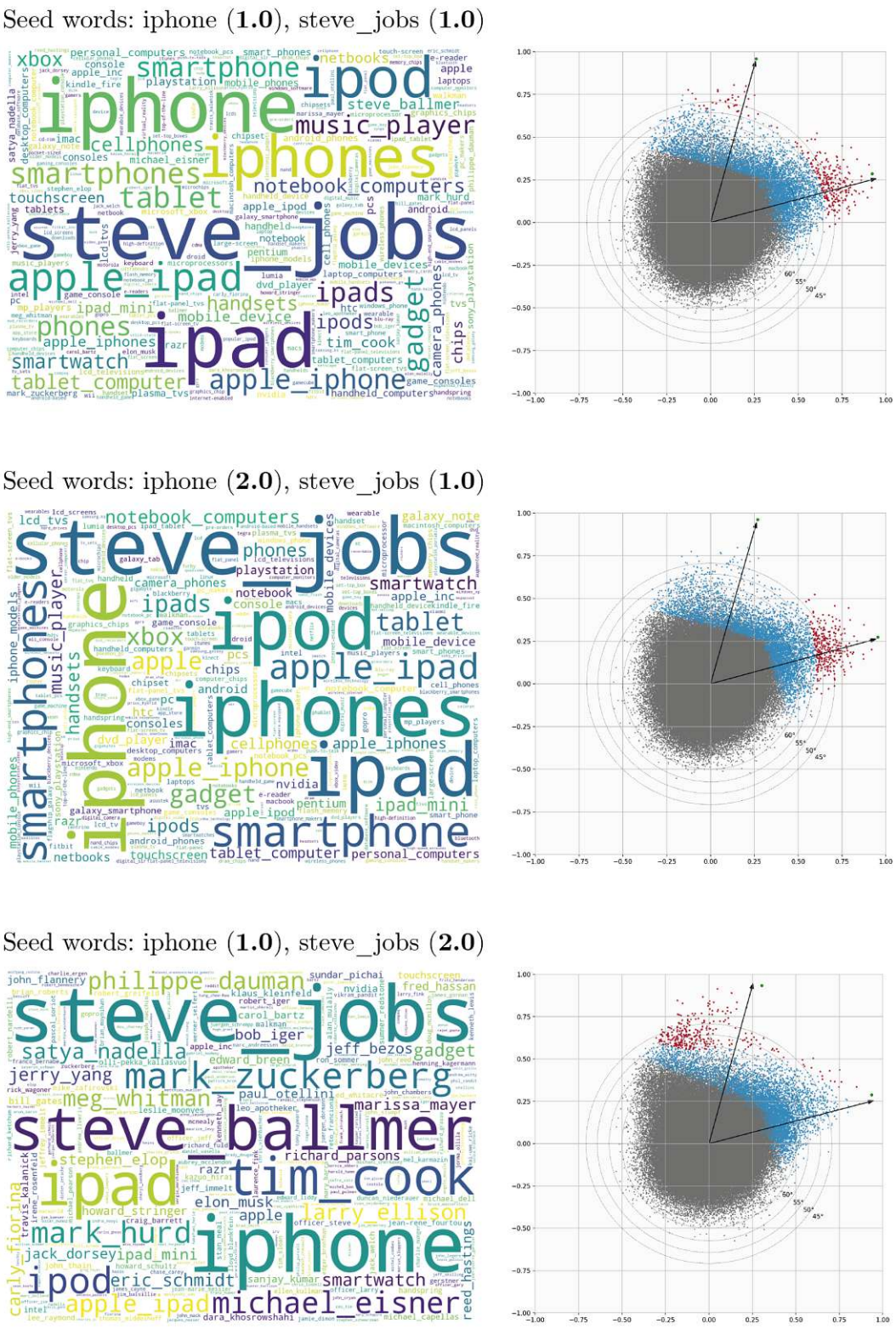
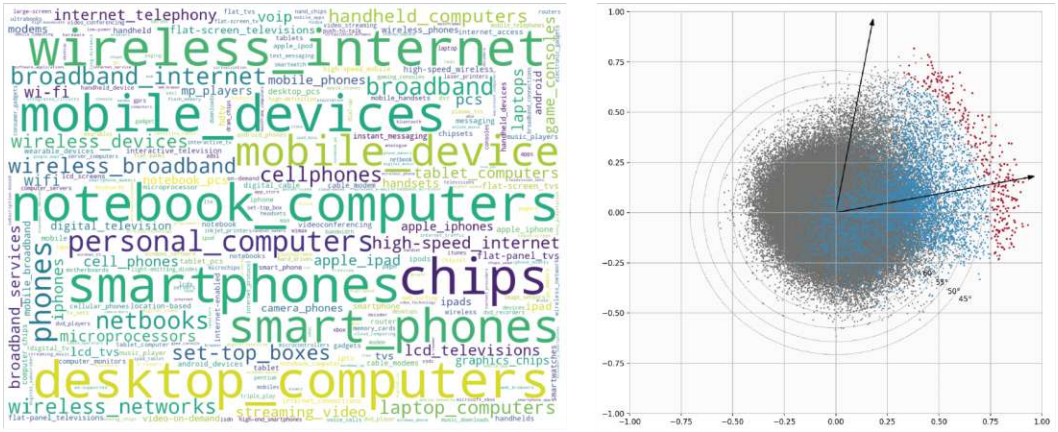


Figure 5.4: Word clouds generated with the seed words "iphone" and "steve_jobs" and the weights $\boldsymbol{g_x} = [1, 1]$ in the first-, $\boldsymbol{g_x} = [2, 1]$ in the second- and $\boldsymbol{g_x} = [1, 2]$ in the third row. The plots in right column show the projections of all words onto the final plane $\boldsymbol{X'}$. The red dots indicate all words that are included in the topic word cluster. Hyperparameters: cluster_size=300, gravity=0.15, K=3000 and update_freq=1.

Seed words: iphone (1.0), steve_jobs (1.0), gravity = 0



Seed words: iphone (1.0), steve_jobs (1.0), gravity = 0.5

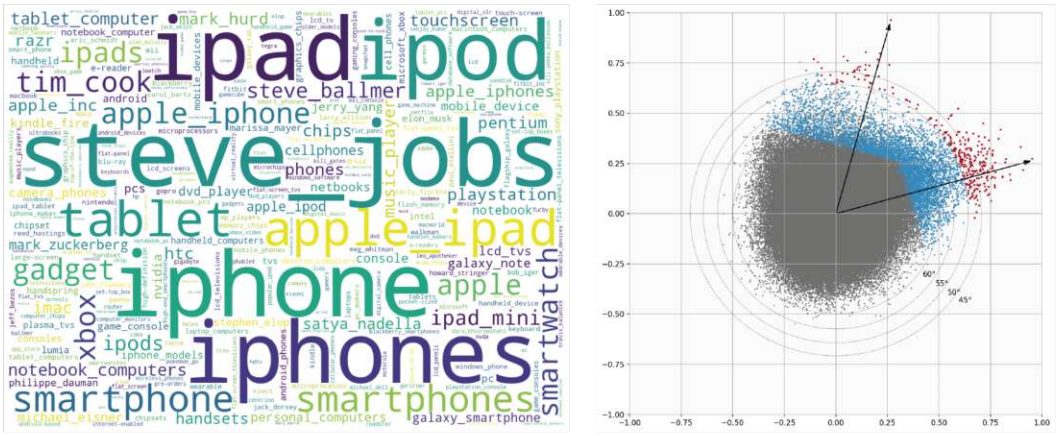


Figure 5.5: Word clouds generated with the seed words "iphone" and "steve_jobs" and the weights $\boldsymbol{g_x} = [1, 1]$ to show the effect of the hyperparameter *gravity* on the topic word cluster. In the first row we set *gravity=0* which results in a topic that is centered around the concept of computer/internet. The seed words, however are not at it's center. In contrast, if we set *gravity=0.5* (second row), the topic cluster stays centered around the seed words. The remaining hyperparameters are: cluster_size=300, K=3000 and update_freq=1.

5.2.3 Vector Representations of Words

We obtain word embeddings from a Word2Vec model (Mikolov et al., 2013a). This model can either be trained on a specific dataset or downloaded in a pre-trained version. If a large text corpus is available, training Word2Vec on this specific corpus may result in more accurate results than using a pre-trained model. However, if only a small corpus of text is available, it is probably better to use a pre-trained model.

We train Word2Vec (Mikolov et al., 2013a) on 10 million Thomson Reuters news articles (2.5 billion words) covering the period from 1996 to 2017. Before training Word2Vec we detect multi-word expressions (bigrams) in the text data by applying the Phrases model available in the Python package Gensim (Řehůřek and Sojka, 2010). This model calculates a score for every phrase as described in (Mikolov et al., 2013b). Any multi-word expression that receives a score above a certain threshold is considered a valid bigram and is included in the vocabulary. Then we train Word2Vec with the continuous bag-of-words (CBOW) algorithm using the Python library Gensim and the following hyperparameters: *vector_size*=64, *window*=18 and *negative*=10. Experiments with different hyperparameters show that for the sake of topic modeling, rather small embedding sizes in the range of 32 to 128 generate more diversified and less fragmented topics. The embedding size of 64 turned out to be a sweet spot. The window size determines the maximum distance between a target word and a context word. During training the window size varies according to a uniform distribution in the range from 1 to 18. The parameter *negative* determines the number of negative samples, i.e., the number of false training samples that are randomly drawn from the vocabulary. We train the model over 100 epochs and augment the obtained word embeddings with polarity dimensions as described in Section 5.2.4.

The Word2Vec model is available in two versions: Skip-Gram and CBOW. While in literature, the Skip-Gram model is often preferred over the CBOW model (Mikolov et al., 2013b), we find that CBOW is better suited for the task of topic clustering. Compare the topic word clouds in Figure 5.6 generated with (a) CBOW and (b) Skip-Gram. All other parameters being equal, the CBOW algorithm better generalizes the topic *"family"* as the most important words characterizing this topic, *"mother"*, *"father"*, *"son"*, *"daughter"* appear big in the word cloud, i.e., close to the topic center. The Skip-Gram model in contrast selects the rather specific bigrams *"his mother"*, *"her son"*, *"her parents"*, *"her mother"* and *"father"* as the most important words representing the topic *"family"*.

(a) CBOW (b) Skip-Gram

Figure 5.6: Comparison of topic clusters generated with GTM using Word2Vec word embeddings generated with (a) the CBOW algorithm and (b) the Skip-Gram algorithm. We generate the topic "family" using the seed words "mother" and "father". The CBOW model is able to produce a cohesive topic cluster with the most representative words (mother, father, daughter, son, husband, wife, etc.) located close to the topic center, as indicated by the large font size of these words. The Skip-Gram model in contrast, produces less cohesive and less balanced topics as rather specific bigrams like her_son, his_mother, her_sister, etc. are located close to the topic center.

Figure 5.7 provides a comparison of a topic word cluster generated using (a) our Word2Vec model trained on Thomson Reuters news and (b) a Word2Vec model trained on 100 billion google news with 300 dimensional vectors. We see no improvement in the quality of the topic cluster by using 300-dimensional embeddings. Quite the opposite is true, we find the 64-dim Thomson Reuters embeddings produces more balanced topic clusters.

(a) Thomson Reuters 64-dim (b) Google 300-dim

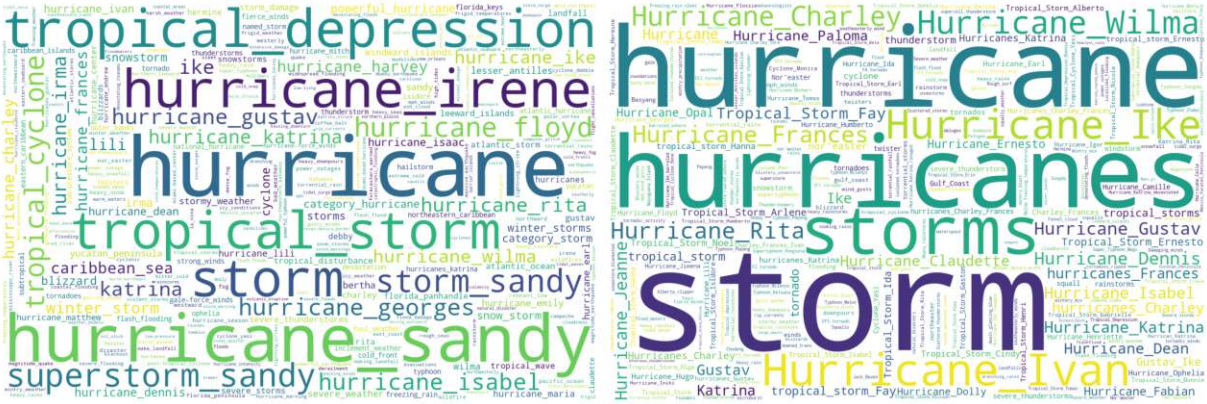


Figure 5.7: Comparison of topic clusters generated with GTM using (a) 300-dimensional Skip-Gram word embeddings trained on a 100 billion word google news dataset[3](Mikolov et al., 2013b) and (b) 64-dimensional CBOW word embeddings trained on 2.5 billion Thomson Reuters news. We use the seed words "storm" and "hurricane", both with a weight of 1 and the hyperparameters cluster_size=300, gravity=0.15, K=5000 and update_freq=1.

5.2.4 Polar Word Embeddings

Although Guided Topic Modeling (GTM) is applicable to various domains, we have developed this methodology targeting the financial domain. Especially in this domain, word polarity matters a lot. Thus, we are interested in a clear separation of words with positive and negative tone. Due to the underlying mechanisms of the Word2Vec algorithm this separation is not achieved by default. In this Section we describe a fully data driven method to obtain polar word embeddings. We therefore use the the direct feedback of the stock market in response to news releases.

We consider a total of 2.85 million firm-tagged US news articles with exact timestamp published from 1996 to 2021 and 17,432 US companies with corresponding return time series obtained from CRSP. Our goal is to obtain a score for each word in the vocabulary that indicates whether a word in more likely associated with positive or negative stock returns. For each stock we calculate weekly, daily as well as daytime and overnight returns. Daytime and overnight returns are calculated by decomposing the daily returns into the daytime period where the stock market is open (9.30am to 4:00pm) and into the overnight

[3]https://code.google.com/archive/p/word2vec/

period where the stock market is closed (4:00pm to 9:30am). Then we merge news articles for each stock and interval to obtain weekly, daily, daytime and overnight news documents. Thereafter we transform the weekly, daily, daytime and overnight returns into z-scores calculated over a rolling window of 127 trading days. With this data we aim to calculate a polarity score for each word. This is achieved by measuring their relative frequencies of co-occurrence with positive and negative z-scores across weekly, daily and intraday intervals. We apply a tf-idf vectorizer on the non-headline texts, i.e., the body of the news documents. The features of the vectorizer are limited to 150,000 words. Next we categorize the vectorized news documents into positive (negative) news documents for associated return observations with a z-score above (below) a pre-defined threshold. We set the threshold to +2.56/-2.56 for daytime and overnight intervals and +1.96/-1.96 for daily and weekly intervals. Then we sum up the tf-idf scores of all words for both, the positive and negative news documents. Thus, for each word w we get a positive pos_w and a negative score neg_w which we further use to calculate the polarity of each word.

Therefore, we first calculate the positive and negative rate (Equation (5.11)) and frequency (Equation (5.12)) of each word. Then we scale those values by calculating the cumulative distribution functions and the harmonic mean from *pos_rate_cdf* and *pos_freq_cdf* ($hmean\_pos_w$) as well from *neg_rate_cdf* and *neg_freq_cdf* ($hmean\_neg_w$). Finally, the polarity is calculated by Equation (5.13). We use the harmonic mean since the difference in magnitude of *pos_rate* and pos_freq is large, as pos_freq is a very small number. If the arithmetic mean were used instead, the influence of pos_freq would not be reflected adequately.

$$pos\_rate = \frac{pos_w}{pos_w + neg_w} \quad \text{and} \quad neg\_rate = 1 - pos\_rate \tag{5.11}$$

$$pos\_freq = \frac{pos_w}{\sum_w pos_w} \quad \text{and} \quad neg\_freq = \frac{neg_w}{\sum_w neg_w} \tag{5.12}$$

$$polarity_w = hmean\_pos_w - hmean\_neg_w \tag{5.13}$$

In Figure 5.8 we highlight the words that are identified as positive and negative using intraday-, daily- and weekly intervals. The font-size of the words is proportional to the harmonic means $hmean\_pos_w$ and $hmean\_neg_w$. We see a clear separation into positive words (first row) and negative words (second row) indicating that this approach is well suited to learn about the polarity of individual words.

Next, we need to incorporate the polarity information into the pre-trained word vectors. We achieve this by adding additional "polarity dimensions" to the pre-trained word embedding vectors. We consider 64-dimensional word embeddings, trained as described in Section 5.3.2. To keep the dimensionality constant, we perform a principal component analysis (PCA) on the embedding vectors and keep the first 61 principal components. Thus, we exclude the three principal components that explain the least variance in the data. We replace these three principal components with the polarity scores of the intraday (daytime/overnight), daily and weekly intervals. Thus, we extend each embedding vector by three polarity dimensions to finally obtain 64-dimensional polar embedding vectors. In a last step we scale the vectors to unit length.

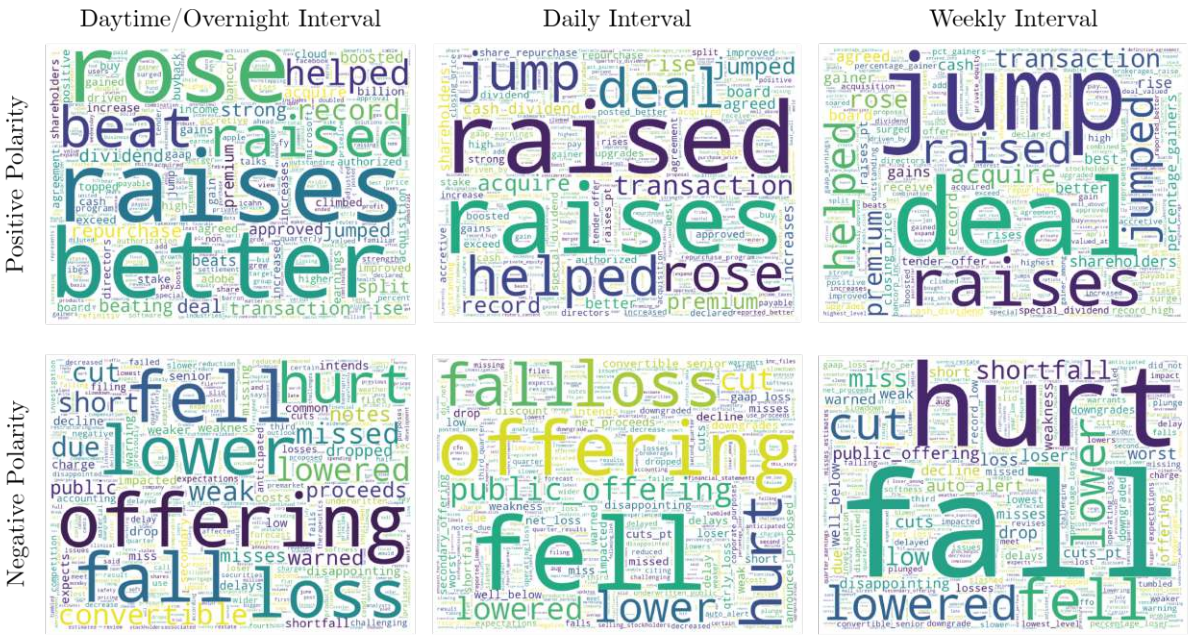


Figure 5.8: Words identified as positive (first row) and negative (second row) using the direct feedback of the stock market. We also observe a subtle variation in word polarities computed over different intervals. For example, the word "deal" obtains a higher positive score when considering weekly intervals (first row, third column) compared to intraday intervals (first row, first column). Similarly, the word "hurt" is associated with a higher negative score for the weekly interval (second row, third column) compared to the intraday interval (second row, first column).

To show how the guided topic modeling (GTM) procedure benefits from using polar word embeddings we generate a positive topic using the seed words "rise" and "surge" and a negative topic using the seed words "fall" and "lower", each associated with a weight of 1. Figure 5.9 shows the resulting topic clusters when using word embeddings without the polarity dimension (left column) and when using word embeddings with the additional polarity dimensions (right column). The word clouds without polarity fail to separate positive and negative words. The positive topic contains many negative words like fall, sharp_drop, sharp_fall, drop, slide, etc. The large font-size of these words further indicate closeness to the topic center. Similarly, the negative topic contains many positive words like higher, rise, slightly_higher, jump, etc. In contrast, by using polar word embeddings we observe a clear separation into words with positive and negative tone.

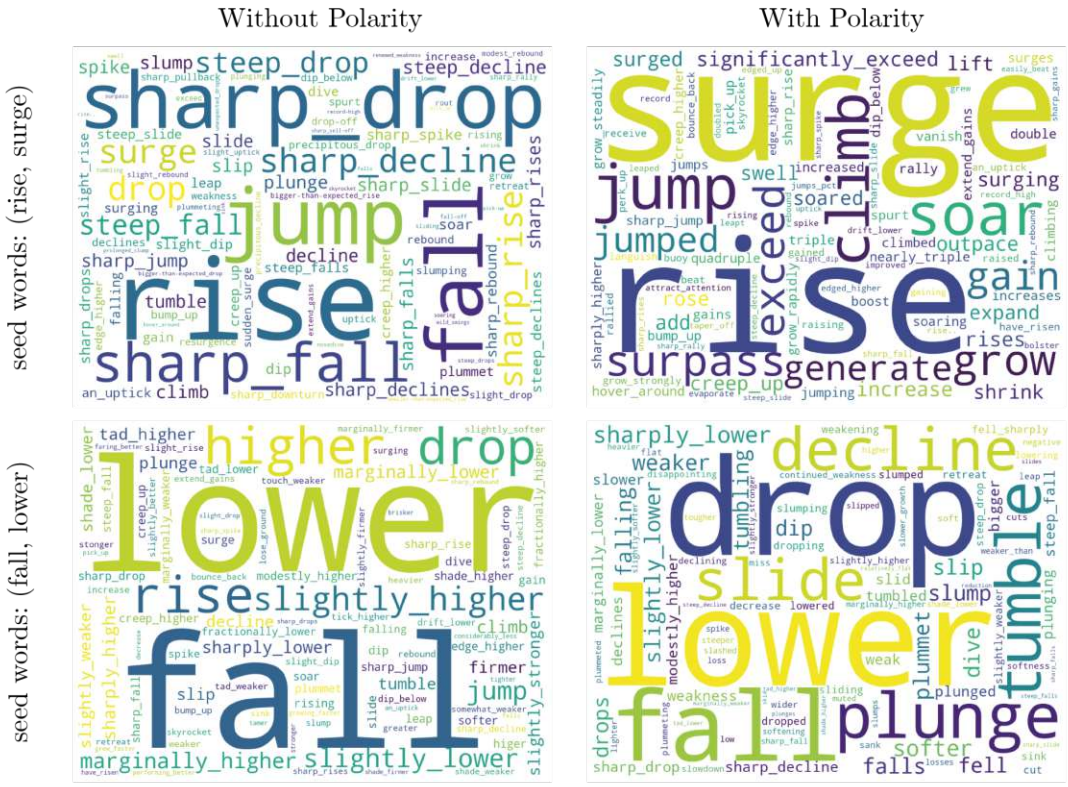


Figure 5.9: Comparison of topic word clusters generated with GTM using standard word-embeddings obtained from the Word2Vec algorithm (left column) with the enhanced polar word embeddings that account for the polarity of individual words. The first row shows positive topics generated with the seed words "rise", "surge" and the second row shows negative topics generated using the seed words "fall", "lower". Without polarity (first column) we see that positive and negative words are mixed and we get no clear separation into a positive and negative topic word cluster. When using polar word embeddings instead (right column), we obtain a clear separation into topics of positive and negative polarity.

5.3 Case Studies

5.3.1 Classification

In this case study, we apply the GTM algorithm on a classification task to quantify its capabilities. We therefore use the BBC News dataset (Greene and Cunningham, 2006) that consists of news of the five categories "tech", "business", "sport", "entertainment" and "politics". The dataset is split into a training dataset with 1225 observations and a test dataset with 1000 observations in total. The benefit of using the GTM method is that it is builds on a pre-trained model, thus no further training is required. In order to apply the method on a classification task the researcher has to define seed words for each of the

five topics. As these topics are less specific but more general, we create two topic-word clusters for each category, using different seed words to capture a wide range of terms for each category. The seed words are shown in Table 5.1 and the generated topic word clusters are visualized in Figure 5.10.

	Positive Seed Words (Weight)			Negative Seed Words (Weight)
Class 0: Technology				
Subtopic 1	software (1.0)	computer (1.0)		
Subtopic 2	website (1.0)	internet (1.0)		
Class 1: Business				
Subtopic 1	firm (1.0)	company (1.0)	stock-market (1.0)	
Subtopic 2	financial (1.0)	chief_executive (1.0)	investors (1.0)	
Class 2: Sport				
Subtopic 1	sports (1.0)	football (1.0)	athlete (1.0)	concerts (-0.5)
Subtopic 2	tournament (1.0)	winner (1.0)		concerts (-0.5)
Class 3: Entertainment				
Subtopic 1	entertainment (1.0)	movie (1.0)	theater (1.0)	
Subtopic 2	singer (1.0)	hollywood_star (1.0)		businessman (-0.5)
Class 4: Politics				
Subtopic 1	government (1.0)	party (1.0)		
Subtopic 2	minister (1.0)	election (1.0)		

Table 5.1: Seed words used in GTM to generate topic word clusters. The GTM algorithm takes as input a list of two or more seed word with positive weight. To further guide the topic in a desired direction, we also define negative seed words if needed.

With the keyword lists, i.e., topic word clouds at hand we now need to calculate loadings of the topics on the BBC news articles. Therefore we transform news articles contained in the training and test dataset into a bag-of-words (BoW) representation while considering a total of 184,968 words that are embedded in the Word2Vec model. Thus, by transforming the news articles into BoW we obtain a sparse matrix of size n_obs x 184,968 with the elements representing the count of each word in a news article. To avoid that single words that appear repeatedly in an article influence the topic loading too much, we clip the word counts at a maximum value of 2. Next, we select only the columns, i.e., words that are contained in topic i which results in a much smaller matrix $\mathbf{T_i}$ of size n_obs x topic_size with topic sizes set to 300, 500 or 800. The loadings of topic i on all news articles are then calculated with the matrix multiplication $\mathbf{T_i w_i}$ with $\mathbf{w_i}$ being a vector of word weights for topic i.

Figure 5.10: Word clouds of the topics "tech", "business", "sport", "entertainment" and "politics" generated with GTM using the seed words shown in Table 5.1 and the hyperparameters: cluster_size=800, gravity=0.15, K=5000 and update_freq=1.

With the loadings calculated for all news articles and subtopics, we calculate the final

topic loadings as the average loading on subtopic 1 and 2. An item is then assigned to the class with the highest topic loading. By doing so we obtain the classification metrics shown in Table 5.2 for the test dataset. We highlight the results for using subtopic sizes of 300, 500 and 800 words. We obtain an accuracy of 90% when using subtopic sizes of 300 and 500 words and 91% for subtopic sizes of 800 words. As the model requires no training we can also perform an out-of-sample classification on the training dataset and observe a accuracy of 90% for all topic sizes.

Instead of simply assigning an item to the class with the highest loading, we introduce the threshold parameter γ, which controls how much the class with the highest loading must exceed the class with the second highest loading for an item to be assigned to the highest loading class. If the threshold is not exceeded the item is assigned to no class. For example, if the second highest loading of an item is 10, and $\gamma = 1.5$, then the highest loading class hat to exceed a loading of 15 (= 1.5 x 10), otherwise the item is assigned to no class. In Figure 5.11 we plot the accuracy over increasing values of γ as well as the declining number of observations with not missing predictions as γ becomes more restrictive. We observe a steep increase in accuracy from 91% to 97% when we set $\gamma = 2$, which is, however, accompanied with a 20% reduction of valid classifications. In Table 5.3 we highlight the classification results if we do not consider the word weights obtained by GTM but treat each word with a weight of 1. We observe a reduction in accuracy from 90% to 88% when using cluster sizes of 300 words, while the accuracy remains unchanged for larger cluster sizes. Thus, using the word weights obtained by GTM gives an improvement, although the gains are rather small.

Next, we compare the classification metrics to three models, namely Support Vector Machines (SVM), Neural Network (NN) and DistilBERT. These are all supervised models which require training on the training dataset. We use the python library scikit-learn to implement these models. For SVM and the NN we transform the text data into a BoW representation by using the tf-idf vectorizer. Also we exclude all words with a frequency above the 99.5% and below the 0.5% threshold which results is a total of 4893 features. For SVM and NN we keep the default settings of scikit-learn. The NN has one hidden layer with 100 neurons. Also, we load the pre-trained DistilBERT (Sanh et al., 2019) model

from Huggingface (Wolf et al., 2020) which we fine-tune over 5 epochs. The classification results are reported in Table 5.4. The accuracies on the test data range from 96% obtained by the NN to 98% obtained by DistilBERT. It is no great surprise that the supervised models outperform classification by GTM since GTM was never trained on this dataset. Still, the accuracy of GTM exceeds 90% and by using the threshold parameter γ the accuracies are close to those obtained by the supervised models.

Consequently, we argue that GTM is not the right choice to use when a labelled dataset is available to train supervised models. Also it may be challenged by unsupervised topic modeling techniques like LDA for datasets that contain a small number of distinct topics. We see the primary use case for GTM when researchers are dealing with very large, comprehensive and unlabelled datasets and are interested to retrieve or classify documents related to specific topics, persons or concepts. With GTM, a researcher can specify particular topics with a small number of seed words and retrieve or classify documents from large corpora in an efficient manner.

	GTM 300			GTM 500			GTM 800			
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	support
tech	0.90	0.87	0.88	0.91	0.86	0.89	0.90	0.89	0.90	189
business	0.87	0.85	0.86	0.90	0.83	0.86	0.93	0.83	0.88	224
sport	0.94	0.96	0.95	0.95	0.98	0.96	0.95	0.98	0.97	236
entertainment	0.88	0.94	0.91	0.89	0.97	0.93	0.88	0.96	0.92	176
politics	0.88	0.85	0.86	0.86	0.88	0.87	0.87	0.88	0.87	175
Accuracy		0.90			0.90			0.91		1000

Table 5.2: Classification metrics precision, recall, F1-score and accuracy obtained by applying GTM with word clouds of size 300, 500 and 800 on the BBC News (Greene and Cunningham, 2006) test dataset. (weighted)

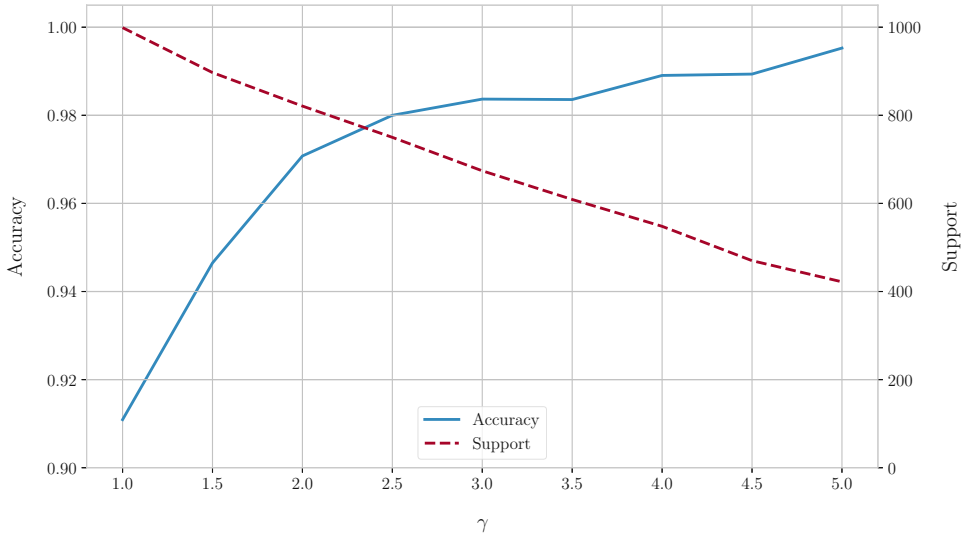


Figure 5.11: Accuracy of GTM 800 for increasing values of γ.

	GTM 300			GTM 500			GTM 800			
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	support
tech	0.88	0.89	0.89	0.91	0.88	0.89	0.90	0.91	0.90	189
business	0.81	0.87	0.84	0.88	0.85	0.86	0.89	0.84	0.86	224
sport	0.95	0.94	0.94	0.95	0.98	0.96	0.95	0.98	0.97	236
entertainment	0.88	0.93	0.91	0.89	0.97	0.93	0.90	0.96	0.93	176
politics	0.91	0.77	0.83	0.88	0.83	0.86	0.88	0.84	0.86	175
Accuracy		0.88			0.90			0.91		1000

Table 5.3: Classification metrics precision, recall, F1-score and accuracy obtained by applying GTM with word clouds of size 300, 500 and 800 on the BBC News (Greene and Cunningham, 2006) test dataset. (boolean)

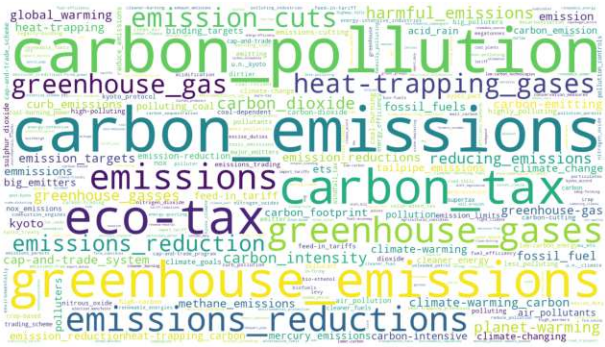
	SVM (BoW)			NN (BoW)			DistilBERT			
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	support
tech	0.97	0.96	0.96	0.95	0.95	0.95	0.96	0.99	0.98	189
business	0.93	0.96	0.95	0.93	0.95	0.94	0.96	0.96	0.96	224
sport	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	236
entertainment	0.99	0.96	0.97	0.97	0.95	0.96	0.99	0.98	0.99	176
politics	0.95	0.96	0.95	0.95	0.96	0.95	0.98	0.95	0.97	175
Accuracy		0.97			0.96			0.98		1000

Table 5.4: Classification metrics precision, recall, F1-score and accuracy obtained by applying the supervised benchmark models Support Vector Machines (SVM), neural network (NN) and DistilBERT on BBC News (Greene and Cunningham, 2006).

5.3.2 Firm-specific Climate Risk Estimated from Public News

This case study is based on Dangl et al. (2023). Here, GTM is applied to generate topic word clusters that cover various forms of climate risks and opportunities with the goal to identify individual firms' climate risk using textual analysis. We use the Word2Vec model trained on Thomson Reuters news as described in Section 5.2.3. Figure 5.12 highlights two topics that are generated using these embeddings with the GTM algorithm.

(1) Reg. climate risk



(2) Sustainability

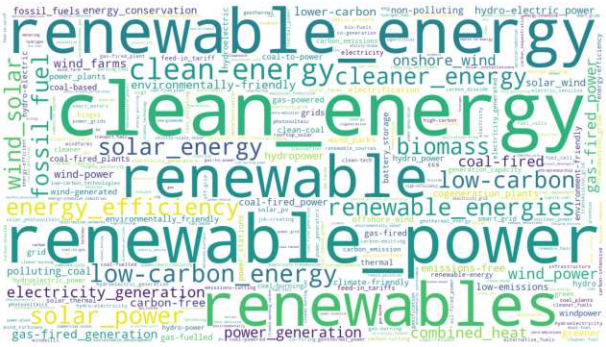


Figure 5.12: Topic word cloud (1) is generated with the seed words "eco-tax" and "carbon_tax", topic word cloud (2) is generated with the seed words "renewable_energy" and "clean_energy".

With the topic word clouds at hand we calculate firm specific topic indices that indicate how strong a specific topic is exposed in the news over time. We follow a similar procedure as described in Section 5.3.1: First, we determine the overlapping words between news articles and topic clusters and count the occurrences of each topic word in a news article. The counts are then multiplied by the weight assigned to each topic word and summed to obtain a loading score. Thus, words closer to the topic center contribute more to the calculated loading than more distant words. Second, we adjust the loadings to account for article length, word frequency and news frequency. Third, we link the news articles with individual firms to obtain firm specific topic indices as shown in Figure 5.13. The graphs show the topic indices for regulatory climate risk and sustainability for the firms Exxon Mobil (a) and Amazon (b). It can be observed that news about Exxon Mobil are more centered around the reg. climate risk topic from 2014 onwards while the news about Amazon are more exposed to the topic sustainability.

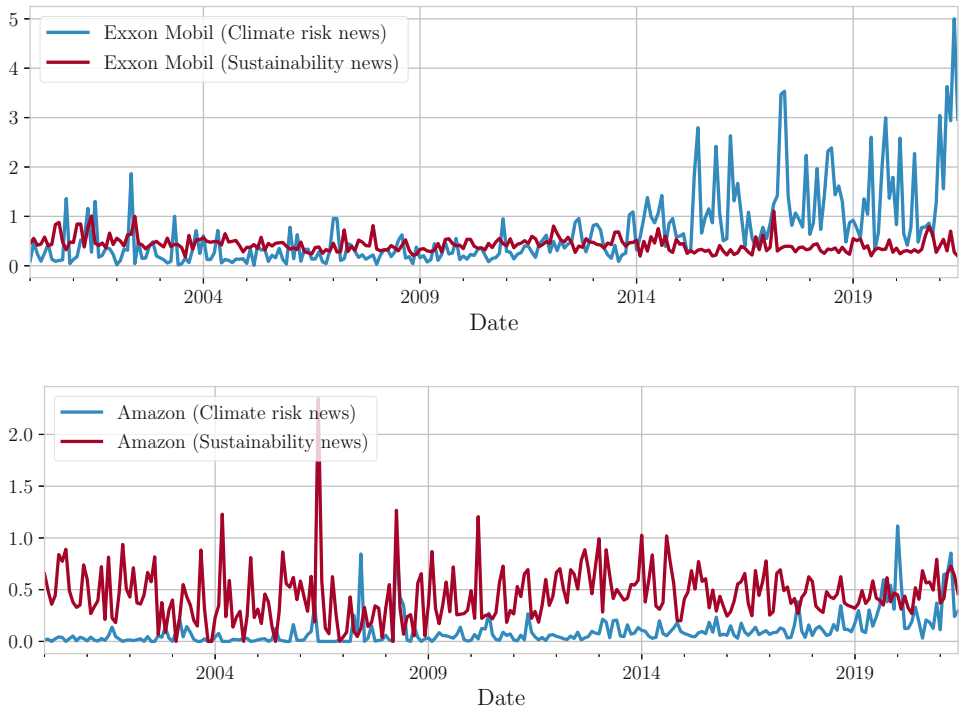


Figure 5.13: Company-specific regulatory climate risk- and sustainability topic indices for the companies (a) Exxon Mobil and (b) Amazon.

5.4 Conclusion

We introduce GTM - Guided Topic Modeling with Word2Vec, which enables the fast and flexible generation of an almost unlimited number of unique topic clusters by defining a small number of seed words. The algorithm comes with several hyperparameters that allow the researcher to adjust the characteristics of the topic clusters and direct them in the desired direction. We show that for the purpose of clustering, smaller embedding sizes are preferable. We also find that the CBOW algorithm produces more coherent topic clusters than Skip-Gram. Furthermore, we extend the word embeddings to incorporate information of word polarity by using the feedback of the stock market. Applications for GTM include information retrieval and classification. If a researcher is confronted with a large unlabeled dataset, GTM can be used to create specific topic clusters, i.e., keyword lists that serve as input for keyword searches within large text corpora. This methodology can also be applied to perform classification of text documents in cases where no training

labels are available or when computational resources are limited which makes the use of a large language model unsuitable. Also, the methodology can be applied to transform high dimensional text data into a low number of dimensions, i.e., the exposure of news to some specific topics of interest. For example, this can be used to generate indices of uncertainty, sentiment, or climate risk as shown by Dangl et al. (2023).

Bibliography

Angelov, D. (2020). Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022.

Cong, L. W., Liang, T., and Zhang, X. (2019). Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information. Interpretable, and Data-driven Approach to Analyzing Unstructured Information (September 1, 2019).

Dangl, T., Halling, M., and Salbrechter, S. (2023). Firm-specific climate risk estimated from public news. Available at SSRN.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Greene, D. and Cunningham, P. (2006). Practical solutions to the problem of diagonal dominance in kernel document clustering. In Proceedings of the 23rd international conference on Machine learning, pages 377–384.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794.

Hayes, P. J. and Weinstein, S. P. (1990). Construe/tis: A system for content-based indexing of a database of news stories. In IAAI, volume 90, pages 49–64.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3):535–547.

King, G., Lam, P., and Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. American Journal of Political Science, 61(4):971–988.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Rinke, E. M., Dobbrick, T., Löb, C., Zirn, C., and Wessler, H. (2022). Expert-informed topic models for document set discovery. Communication Methods and Measures, 16(1):39–58.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.