The Review of Financial Studies

War Discourse and Disaster Premium: 160 Years of Evidence from the Stock Market

David Hirshleifer
Marshall School of Business, University of Southern California, United States

Dat Mai
MKT MediaStats LLC, United States

Kuntara Pukthuanthong
Robert J. Trulaske, Sr. College of Business, University of Missouri-Columbia, United States

In addition to summarizing their paper, replicate parts of their analysis with data from Wikipedia pageviews and Google Trends data until April 2025.

You can retreive data on the S&P 500 and other predictor variables from https://sites.google.com/view/agoyal145

Using a semisupervised topic model on 7 million *New York Times* articles spanning 160 years, we test whether topics of media discourse predict future stock market excess returns to test rational and behavioral hypotheses about market valuation of disaster risk. Media discourse data address the challenge of sample size even when disasters are rare. Our methodology avoids look-ahead bias and addresses semantic shifts. Our discourse topics positively predicts market excess returns, with *War* having an out-of-sample R^2 of 1.35%. We call this effect the *war return premium*. The war return premium has increased in more recent time periods. (*JEL* G10, G11, G12, G14, G17, G41, N00)

Received: May 2, 2023; Editorial decision: June 2, 2024
Editor: Juhani Linnainmaa
Authors have furnished an Internet Appendix, which is available on the Oxford University Press Web site next to the link to the final published paper online.

We thank Juhani Linnainmaa (editor) and two anonymous referees for helpful comments and suggestions that substantially improved the paper. We thank Fred Bereskin, Vineer Bhansali, Murray Frank (discussant), John Howe, Tim Loughran (discussant), Michael O'Doherty, Stephen Owen (discussant), Seth Pruitt (discussant), Richard Roll, Jeffrey Stark (discussant), Yinan Su (discussant), and Russ Wermers (discussant) and the seminar and conference participants at the Bank of Portugal, the University of Cologne, the University of Missouri-Columbia, the University of Missouri-St. Louis, Missouri State University, the Doctoral Tutorial at the European Finance Association (2021), the PhD Student Symposium at the University of Texas at Austin (2021), the Chicago Quantitative Alliance (CQA) Annual Academic Competition (2021), the Northern Finance Association Meeting (2021), the Financial Management Association Meeting (2021), the Southern Finance Association Meeting (2021), the Financial Research Association Meeting (2021), the American Finance Association PhD Student Poster Session (2022), the Midwest Finance Association Meeting (2022), the Young Scholars Finance Consortium (2022), the Boulder Summer Conference on Consumer Financial Decision Making (2022), the Chicago Quantitative Alliance (CQA) Annual Meeting (2023), and LongTail Alpha LLC for helpful comments. We acknowledge computing support from the Research Computing Support Services (RCSS) at the University of Missouri-Columbia. The views expressed herein are solely of the author. Supplementary data can be found on *The Review of Financial Studies* web site. Send correspondence to Kuntara Pukthuanthong, pukthuanthongk@missouri.edu.

The Review of Financial Studies 00 (2024) 1–50

https://doi.org/10.1093/rfs/hhae081 Advance Access publication 23 November 2024

Several rational and behavioral theories suggest that the prospect of major disaster may help explain major asset pricing puzzles. In some rational models, there is a premium for disaster risk, which can help address the equity premium puzzle (Barro 2006, 2009). A further implication is that higher disaster risk will predict higher future stock market excess returns. Two behavioral hypotheses offer a similar implication. The first is attentional and belief-based: that investors overestimate the probability of rare disasters owing to the high salience of extreme outcomes. This possibility is supported by evidence that people overestimate the probabilities of rare events (Fischhoff, Slovic, and Lichtenstein 1977; Snowberg and Wolfers 2010). The second is preference-based: that investors overweight low probabilities, as in the expected value function of cumulative prospect theory (Tversky and Kahneman 1992).

The rarity of major disasters is a well-known obstacle to empirically testing the relationship between their occurrence and future stock returns. This rarity limits statistical power.

In this paper, we circumvent the obstacle of small sample size by using data on investors' attention to rare disaster risks derived from news. This provides a much larger sample of changing perceptions of disaster probabilities over 160 years. Shifts in media topic coverage over time can potentially capture investor assessments of future prospects, including the risk of rare disasters.

The novelty of our approach to capturing perceptions of disaster risk is twofold. First, ours is the first study to compare the effects of disaster- and non-disaster-related topics of discourse systematically for the pricing of the aggregate stock market. Second, we apply a novel approach, "seeded Latent Dirichlet Allocation" (Lu et al. 2011, henceforth sLDA), to extract topics of popular discourse over time. This method has several key advantages over existing empirical applications of language models to asset pricing.

On the first point, for non-disaster-related topics, we consider the 12 "narratives" (topics) discussed in Shiller (2017, 2019). This allows us to obtain interpretable findings, as opposed to using a purely statistical procedure to extract topics. We make necessary adjustments to these topics to effectively implement sLDA. The disaster-related topics we consider are war and pandemics.[1]

We provide measures of media discourse coverage of different topics. For brevity, we call these media discourse measures "topics." We find that non-disaster-focused topics predict stock market excess returns in-sample and have limited out-of-sample predictive power. The *Pandemics* topic has limited and inconsistent predictive power even in-sample. In contrast, *War* has substantial predictive power both in- and out-of-sample. Our study is the first to show that

[1] Both can have massive human and economic costs and highly uncertain outcomes, as exemplified by the COVID-19 pandemic. Oleg Itskhoki, the winner of the 2022 John Bates Clark Medal, suggested in an interview with *Bloomberg* on August 2, 2022, that existing wars at that time presented an even greater economic risk than COVID-19 (Miller 2022).

War as a discourse topic is more powerful than non-disaster-focused topics in predicting excess returns.

On the second point, our use of sLDA topic modeling is a prevalent dimension-reduction strategy in the machine learning and natural language processing literature. It is used to summarize information from large amounts of text into a limited set of topics. The topic model we use, sLDA, is a recent extension of the canonical unsupervised LDA model (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004).[2] The key benefits are efficiency, interpretability, forward-looking prediction, and the ability to address semantic change over time. To understand these benefits, it is necessary to describe how sLDA works.

Under traditional unsupervised LDA, the model arbitrarily gathers common phrases into topics based on word frequencies. In contrast, under sLDA, the creation of topics allows control over the content to be extracted. sLDA fits our research goal of testing the consequences of disaster-focused and non-disaster-focused topics in media discussions. In this approach, we feed the model with the seed words associated with each topic and let the algorithm choose the phrases that often appear with these seed words.[3,4]

Our semisupervised topic model performs two key tasks: (1) it classifies market attention throughout 7 million articles published in the 160 years history of *The New York Times* into several disaster-focused and non-disaster-focused themes, and (2) it traces the evolution of media attention to these themes. These provide new quantitative measures of the market attention to topics of public discourse. Specifically, the model estimates the fraction of an article's text devoted to each topic. Aggregating over articles, these proportions measure the amount of news coverage each topic receives.

Central to sLDA is the identification of words that co-occur with the seed words. Our topic weights quantify the market's interest in each topic based on the frequency of terms that co-occur with it. We gather seed words from well-known media and publications, such as *Nature* for disaster-focused topics and from Shiller's book *Narrative Economics* for other topics. As we study 2 disaster-focused topics, *War* and *Pandemic*, and 12 non-disaster-focused topics discussed in Shiller (2019), we have 14 seeded topics in total. We add one unseeded topic as a catch-all for everything not captured by the seeded ones.

---

[2] The burgeoning popularity of LDA in computer science and other social science fields is evident. For surveys, see Steyvers and Griffiths (2007), Blei (2012), and Boyd-Graber, Hu, and Mimno (2017).

[3] Recent papers in natural language processing, such as Lu et al. (2011), Jagarlamudi, Daumé III, and Udupa (2012), Eshima, Imai, and Sasaki (2024), and Watanabe and Zhou (2020), have documented the advantages of a (semi)supervised LDA model over the unsupervised one. Among other desirable features, a guided LDA model ensures the interpretability of topics, avoids the need to label extracted topics ex post to interpret them, and, as a more constrained approach, helps alleviate possible overfitting.

[4] A third group of topic models is fully supervised methods (see, e.g., Mcauliffe and Blei (2007); Ramage et al. (2009)). These supervised models extract topics predictive of document tags or labels, so labeled documents are needed to train them. These models are not suitable for us because instead of predicting article titles, we want to use topics from news to predict market returns.

We acknowledge two key challenges in estimating the predictive power of topics of discourse: possible look-ahead bias and semantic changes over time. These problem are intertwined. The meanings of words and phrases shift extensively over time since our sample spans 160 years. An empirical approach that pools the entire sample to identify word lists and estimate the model parameters could be invalidated by such semantic shifts.

To address this issue, our analysis regularly updates the word list that constitutes topic weights. Although the list of seed words remains unchanged, the model is reestimated monthly using data from the most recent 10 years of news articles (including the current month). Therefore, the words clustered within any given topic reflect those popularly used at that time and vary monthly depending on semantic shifts.

This ability to reflect semantic shifts over time is a key advantage of sLDA. In contrast, monthly rolling estimation is impossible under unsupervised LDA because it is not designed to yield consistent thematic content across estimations. This makes it impossible to address these two challenges.

Specifically, our estimation process recalculates the topic weight on a monthly rolling forward basis. The output of this process is an article-level weight vector, where each element represents the proportion of content (or attention) devoted to a specific topic within that article. We then compute the economy-wide monthly time series of topic weights for each topic from the article-level topic weights and use this aggregate time series to test for return predictability.

To quantify perceptions of rare disaster risk, we extract disaster-related discussion from news media text. An alternative to our approach is to condition on macroeconomic variables that are hypothesized to capture risk premia (Cochrane 1996). Empirically, however, macroeconomic variables perform poorly, with variation that is too low to match asset returns. A benefit of our approach, as compared to macroeconomic variables, is that news is available at a high frequency to track shifts in market attention to the prospect of disaster.[5]

Past literature has provided evidence that news about rare disasters affects investors' expectations and the equity return premium (Rietz (1988), Barro (2006), and Julliard and Ghosh (2012)). Nevertheless, the rarity of extreme disasters suggests that it is valuable to have long time period tests that draw on the *NYT* over 160 years.[6] This can provide insight about whether effects are robust and consistent over time (Schwert 1990).

[5] See the literature on rare disaster risks using contemporaneous price data (Ferguson 2006; Le Bris 2012; Oosterlinck and Landon-Lane 2006) to understand stock returns and bond returns during wartime. Le Bris (2012) uses an event study approach to infer the effects of disasters. Berkman, Jacobsen, and Lee (2011) use a count of crisis events to proxy for rare disaster risk.

[6] We use the first 10 years of news data to train our sLDA model to obtain the first month's topic weights and continue rolling the window forward. Our market index is unavailable from the Global Financial Data until 1871; thus, our training period starts in 1861, 10 years after the *NYT*'s inception in 1851. Our main results are based on data from the *NYT*; we also verify with tests based on the *Wall Street Journal (WSJ)*. The *NYT* and the *WSJ* are the two largest national media outlets, and thereby can reflect the attention and perception of a large general audience.

We find that the topics extracted from the *NYT* provide strong power to predict stock market excess returns. To examine this issue, we construct a discourse topic index from all 14 topics via the two-step partial least squares (PLS) approach of Kelly and Pruitt (2013, 2015). PLS is a method of extracting a single return predictor from text information from the *NYT* jointly across topics.

We find that the monthly predictive regression of market returns on the PLS index yields a slope of 4.65% and an R^2 of 0.58% over the whole test sample of 149 years from 1871 to 2019 and a slope of 9.61% and an R^2 of 3.22% over the past 20 years.[7] For the subperiods (1871–1949, 1950–2019), the PLS index remains a significant predictor at the 10% and 1% level, respectively. The predictive power of the PLS index is not subsumed by common macroeconomic, sentiment, and uncertainty variables introduced in the literature.

Among these topics, we find that *War* is the strongest market predictor. The PLS index heavily loads on *War* with a correlation of 82%, indicating a strong similarity. Not surprisingly, it has somewhat similar predictive power (even though, as seen in Table 2, the PLS index also loads nontrivially on other topic indexes). In particular, the predictive power of *War* for the equity return premium increases over time. Over the test period of 149 years, a one-standard-deviation increase in *War* predicts a 3.80% increase in annualized excess returns in the next month, and the monthly in-sample R^2 is 0.39%. In comparison, over the past 20 years, the respective numbers are 9.83% and 3.39%. *War* is significant for both subperiods (1871–1949 and 1950–2019).

This is economically substantial in comparison with the average annualized monthly excess stock market return over the same period, 6.44%. As another benchmark, the average R^2 of the 40 well-known predictors is only 0.73% in-sample and -1.01% out-of-sample (see Goyal, Welch, and Zafirov (2024), who discuss the weak out-of-sample performance of most stock return predictors). The R^2 of *War* indicates that its predictive power is economically substantial. Over the test period, *War* and the PLS index also significantly predict market returns up to a horizon of 36 months ahead. Consistent with the time-varying disaster risk model and with behavioral theories of disaster risk, we find that mean market excess returns increase with market attention to rare disasters.

We conduct standard out-of-sample tests to investigate whether discourse topics create value for real-time investors. With expanding window estimation, *War* outperforms all individual economic predictors studied in this paper in terms of out-of-sample R^2 (R^2_{OS}), which compares the forecasting power of a predictor against the historical mean return used as a forecast. The out-of-sample R^2 of *War* is strong (0.17%), with strongest return predictability in the last 20 years (1.35%).

[7] The first 10 years from 1861 to 1870 are used to compute the first set of topic weights.

A possible caveat to these conclusions is that our tests include 14 different topics, any one of which could have turned out to be the best. Furthermore, the iterative process of seed word selection, sLDA model estimation, and predictive regression leads to alternative possible specifications that differ in out-of-sample R^2_{OS} and other metrics. To address these Multiple Hypothesis Testing (MHT) concerns, we employ two methods. First is a placebo bootstrapping test using seed words that are unlikely to predict returns to assess the likelihood that a seed word is predictive by chance. Second, we apply the Bonferroni correction, a conservative method for adjusting for MHT. Our findings are robust; see Appendix D.3.

We perform several tests to explore the sources of the predictive power of *War* and the PLS index for stock returns. We first test for the contemporaneous relationship between innovations in *War* and returns. We would expect to see a negative contemporaneous correlation if innovations in *War* are associated with either increases in risk premia or if increases in *War* are bad news for future cash flows (especially if investors overreact to this bad news). On the other hand, we might expect a positive contemporaneous correlation if innovations in *War* contain positive information about future cash flows. Our point estimate for this correlation is negative and economically substantial but statistically insignificant. This indicates that the statistical power of this test is not sufficient to draw strong conclusions about this relationship.

To further explore the sources of the predictive power of *War* and the PLS index, we apply the approach of Campbell (1991) using VAR(1) to decompose realized returns into expected returns and unexpected returns, which are then separated into cash flow and discount rate news. We find *War* and the PLS index predict future returns by forecasting all component of returns: expected returns, cash flows news, and discount rates news. Overall, the cash flow channel appears to play a more important role.

An important question is whether the predictive power of *War* and the PLS index derives from risk or psychological bias. We find that *War* and the PLS index are *negatively* correlated with volatility indicators such as VIX and NVIX, and with subsequent realized volatility.[8]

War and the PLS index also have insignificant predictive power for most financial crisis-related variables used by Greenwood et al. (2022). On the other hand, they are associated with negative return skewness.

These findings do not support the hypothesis that war return premium is a rational risk premium for volatility. However, the skewness findings are consistent with the hypothesis that the returns reflect the risk of rare disasters. Alternatively, they are also consistent with the possibility that the market overreacts to rare disaster risk, resulting in high subsequent returns.

[8] See also Cortes, Vossmeyer, and Weidenmier (2022), who find that massive war spending during war times by the U.S. government make future corporate profits more predictable, thereby reducing future stock volatility.

As we have discussed, there are behavioral arguments that when *War* or the PLS index is high, the market will be underpriced (or less overpriced), and will subsequently experience high returns. We find that the correlations between *War* or the PLS index with proxies for sentiment or disagreement proxies are negative, as predicted by these arguments, but generally modest and not always significant.[9]

To test whether short sale constraints / overpricing effects (Stambaugh, Yu, and Yuan 2012) explain the war return premium, we condition on high versus low sentiment periods using the investor sentiment index of Baker and Wurgler (2006). Under this hypothesis, the predictive power of *War* and the PLS index should be stronger during high sentiment periods (see Appendix C.7). We find that the predictive power of *War* and the PLS index are stronger during low sentiment periods, which does not support the short-sale constraint / overpricing hypothesis.

Overall, the evidence from the several tests just described are supportive of some versions, but not others, of the risk and behavioral theories for the war market return premium.

Based on the idea that wars can trigger inflationary policies or defaults, we also examine whether *War* predicts bond returns. We find evidence consistent with fear of war triggering a flight to quality. Such flight could occur for either rational or behavioral reasons. We find that *War* positively predicts excess returns on mid- to long-term high-yield corporate bonds, while negatively predicting excess returns on safer investments such as short-term government and investment-grade corporate bonds. These findings suggest that investors demand higher premiums to hold riskier assets and lower premiums for safer assets when returns are more skewed to the left.

Our results are robust to accounting for the high autocorrelation of the variable *War* and the PLS index by using its innovation. We also find that the predictive power of *War* and the PLS index remains unchanged when we alter the seed words for *War*, add *Natural Disasters* topic, increase the number of seed words of *Pandemic*, and increase the number of topics.

Contributions. This paper contributes to several lines of research. First, it contributes to social finance and narrative economics in testing how topics of media discourse are related to stock market pricing (see also Section 1). As such, it builds on a literature on the relationship of news content to economic and financial outcomes.

Second, our paper contributes to the literature on rare disaster risks, which incorporates disaster probabilities and loss into the standard consumption-based model to explain the high equity premium (Barro 2006, 2009; Gabaix 2012; Wachter 2013).

[9] Our sentiment measure includes news sentiment we construct from NYT, market sentiment from Baker and Wurgler (2006), and managerial sentiment from Jiang et al. (2019). Disagreement is from Huang, Li, and Wang (2020).

Third, it contributes to the literature on using textual data to predict stock market returns (e.g., Pástor and Veronesi (2013) and Brogaard and Detzel (2015)). In contrast with the long-horizon predictability of those papers, our *War* and PLS topic index are stronger predictors of one-month returns. So our index captures a different aspect of textual discourse and risk. Our time-series tests also yield stronger predictability for stock and bond returns than the dictionary approach of Caldara and Iacoviello (2022) (see Internet Appendix E).

Finally, our paper contributes to the burgeoning literature on applying natural language processing tools to business and economic research. A growing body of research utilizes topic modeling tools to extract thematic content from texts.[10] Whereas most finance papers use the traditional unsupervised LDA model, our semisupervised LDA model allows us to extract a predefined set of topics in the news, which enhances interpretability. Adämmer and Schüssler (2020), Manela and Moreira (2017), and Bybee et al. (2024) investigate how news media text can be used to predict aggregate stock market returns. Adämmer and Schüssler (2020) use correlated topic model, a variation of the unsupervised LDA model, while Manela and Moreira (2017) employ a support vector regression model, and Bybee et al. (2024) apply unsupervised LDA. The method we apply, rolling estimation of sLDA, enables us to address semantic changes over time, and to avoid look-ahead bias in predicting returns (see Section 1).

1. Contribution and Related Research

We discuss here the relation of this paper to existing research in more depth. Our focus on predicting the aggregate market return distinguishes our paper from most existing research on news content and financial outcomes. Our finding that *War* is a powerful predictor of market returns is a uniquely distinctive feature of our paper.

To our knowledge, this is the first paper that analyzes news articles from *all* newspaper sections of *NYT* since its inception. Our sample includes nearly 7 million articles from the *NYT* and 600,000 from the *WSJ*, making it a relatively comprehensive and extensive data set. As compared with most existing studies, this data set may be more representative of the topics of actual concern to investors.

Bybee, Kelly, and Su (2023) combine disaster- and non-disaster-focused topics to form a set of factors. Their recession topic is the most important for explaining the cross section of expected stock returns. Although not the main focus of their paper, their topics seem to have little power to predict the

[10] See, for example, Dyer, Lang, and Stice-Lawrence (2017), Hansen, McMahon, and Prat (2018), Larsen and Thorsrud (2019), Choudhury et al. (2019), Brown, Crowley, and Elliott (2020), and Bybee et al. (2024).

aggregate market return (their figure 6). In contrast, our focus is on aggregate market return predictability.

Applying data from many U.S. local newspapers over a century, van Binsbergen et al. (2022) construct a measure of economic sentiment and find that it predicts future economic fundamentals, such as gross domestic product (GDP), consumption, and employment growth. Our paper differs in studying stock and bond market return predictability.

Adämmer and Schüssler (2020), Manela and Moreira (2017), and Bybee et al. (2024) extract information from news media text to predict aggregate stock market returns. Adämmer and Schüssler (2020) do so using a variation of the unsupervised LDA model to extract topics from news articles. Our approach differs in three key respects.

First, they focus on economic news in the *NYT* and *Washington Post* from 1980 to 2018, whereas our study utilizes all *NYT* sections starting from more than a century earlier. As emphasized by Lundblad (2007), it is crucial to consider long time series data to reliably test for return predictability.

Second, the approach of Adämmer and Schüssler (2020) generates outputs that are challenging to interpret. The authors choose 100 topics for their model and, based on the full sample, identify one of them (topic 20) as the most important. Based on this information, they then interpret the meaning of topic 20 in terms of its key words. Any reestimation of the model using updated data would result in a new set of extracted topics, and potentially a different interpretation. In contrast, our approach identifies investor concern with rare disasters, and in particular war, as crucial for return prediction.

Third, and relatedly, since they use an unsupervised topic model, their approach does not address semantic changes over time.[11] In contrast, our use of a semisupervised LDA model allows us to address semantic changes via monthly rolling estimation of the topic model.

The top-line predictability reported in their study is a remarkable out-of-sample R^2 of 6.52%. However, their initial training window for return prediction of 3 years is too short for a reliable out-of-sample R^2 estimate.[12]

Manela and Moreira (2017) provide evidence that news events are positively associated with forward-looking volatility and equity risk premiums. They construct news implied volatility (NVIX) from the front page of *WSJ* starting from 1890. Our paper differs in several ways. First, our return predictor topics (notably *War*) can capture perceived disaster risk, not just volatility.

Second, based on their support vector regression model training procedure, their risk measure captures only terms that appeared in the last 20 years of

[11] They train their topic model using news data from 1980 to 1995 and apply the trained model to extract topic weights from news articles from 1996 to 2018. They argue that because their sample is short, language change is not a concern (their footnote 7).

[12] An out-of-sample R^2 is computed by comparing the return forecast of a given model against the forecast using the historical mean return, which cannot be reliably estimated with only 3 years of data. Consistent with this, their out-of-sample R^2 is highly sensitive to different start dates of the evaluation sample (see their table AV).

their sample period. This potentially induces look-ahead bias. In contrast, we estimate our model on a monthly rolling basis using data from the preceding 10 years. This allows us to address semantic changes over the 160-year sample period while avoiding look-ahead bias.

Third, we obtain stronger and more robust return predictability, including out-of-sample tests. Using the data provided on the authors' website, we find that over 1900-2016 a standard deviation increase in $NVIX^2$ is associated with an increase in the annualized market excess return by 0.22% over the next month. A one-standard-deviation increase in our *War* variable is associated with an annualized market excess return of 2.7%. We find that *War* predicts stock market return from 1 month to 3 years while their predictor, $NVIX^2$, does not predict returns until 6 months ahead.[13] *War* is a powerful predictor in both in and out of sample; Manela and Moreira (2017) do not report out-of-sample R^2. Also, when we control for NVIX or $NVIX^2$, *War* is still a substantial and significant return predictor over the entire sample. We report these results in Table 6.

In independent work, Bybee et al. (2024) use traditional unsupervised LDA on news content to select 180 topics, to fit contemporaneous financial and macroeconomic variables, and to forecast macroeconomic variables and stock market returns. Our approach differs in using a semisupervised approach to focus on just 15 topics (14 seeded plus one unseeded topic) to test hypotheses about rare disaster risk. Our analysis covers a much longer sample period (all sections of the *NYT* from 1861 to 2019); Bybee et al. (2024) focus on economic news in the *WSJ* from 1984 to 2017. Crucially, our analysis avoids look-ahead bias by estimating on a rolling forward basis.[14]

2. Method

In this section, we briefly discuss the setup of the sLDA model (Lu et al. 2011) and our implementation of it to extract news topics.

2.1 The stochastic topic model

Stochastic topic models are based on the core idea that documents can be described as mixtures of topics, where each topic is associated with a probability distribution over words (Steyvers and Griffiths 2007; Blei 2012). In this approach, latent topic weights are extracted from news articles. To do so, we assume that each text document is generated by a simple stochastic process that starts with a document-specific distribution over topics (the document-topic distribution). Each word in the document is chosen first by picking a

[13] Manela and Moreira (2017) report prediction results over 1945–2009 using $NVIX^2$.

[14] Bybee et al. (2024) address look-ahead bias by performing a rolling estimation of LDA via online LDA (oLDA). However, their use of 180 topics in the online LDA scheme is optimized over the entire sample period.

topic randomly from the document-topic distribution and then drawing a word from the topic-word distribution for that topic.

The document-topic distribution for each document and topic-word distribution for each topic (the same across documents) are unobserved parameters that are estimated from the observable word frequencies in the document collection. We use standard statistical techniques to estimate the generative process, inferring the topics responsible for generating a collection of documents (Steyvers and Griffiths 2007).

The most widely used topic model is latent Dirichlet allocation (LDA) as introduced by Blei, Ng, and Jordan (2003) and further developed by Griffiths and Steyvers (2004). Under LDA, a document is generated under the hierarchical process described above. Each word in the document is selected by first randomly selecting a topic from the document-topic distribution, and then for that topic, the word is selected from the topic-word distribution. Under LDA, the document-topic distribution (a vector of probabilities over the topics) for each document is selected from a prior Dirichlet distribution (see Appendix A for details of the LDA and sLDA methodologies). The topic-word distribution is global; it does not depend on the document. It is also assumed to be drawn from a prior Dirichlet distribution. Since the topic-word distribution is a set of probabilities for drawing each possible word, the distribution for the number of instances of each word in an entire document is multinomial with these probabilities, with N being the number of words in the document.

The unknown parameters of the multinomial distributions are estimated using the frequencies of different words in the documents in the sample.[15] Specifically, we use Gibbs sampling to simulate the posterior distribution of words and documents and estimate the two hidden model parameters, namely, the document-topic distribution (τ_d) and the topic-word distribution (ω_k).[16]

In traditional unsupervised LDA, only the number of topics K is prespecified; the model clusters words into these topics based on word frequencies in a completely unsupervised manner. The model automatically extracts underlying topics. The LDA model is more likely to assign a word w to a topic k in a document d if w has been assigned to k across many different documents and k has been used multiple times in d (Steyvers and Griffiths 2007).

Since we are interested in uncovering the effects of specific and interpretable topics relating to rare disaster risk, we instead employ a recent extension of LDA called seeded LDA (sLDA) developed by Lu et al. (2011). sLDA allows users to regulate topic contents using domain knowledge by injecting seed words (prior knowledge) into the model. When a seed word is not present in

[15] An exception is that the two hyperparameters of the two prior Dirichlet distributions are taken from LDA topic modeling literature.

[16] Gibbs sampling is a sampling technique that simulates a high-dimensional distribution by sampling from lower-dimensional subsets of variables where each subset is conditioned on the value of all others. See Griffiths and Steyvers (2004) for details on the implementation of Gibbs sampling in LDA.

a text collection, it does not enter the sLDA model and has no impact on the estimation process.

2.2 Seed words

A key component of an sLDA model is the set of seed words representing the prior knowledge of each topic. As emphasized by Watanabe and Zhou (2020), a dictionary of seed words needs to be carefully chosen based on field-specific knowledge independent of word frequencies in the collection of texts used. Table 1 lists the lemmatized seed words for each topic. (Lemmatization is the removal of word endings such as *s, es, ing, ed*.) Our seed words for *War* include *conflict, tension, terrorism, terrorist, war* and seed words for *Pandemic* include *epidemic, pandemic*.[17]

Since word meanings evolve, it is important that the seed words be general fundamental concepts that have reasonably stable meanings over very long periods.[18] Our methodology allows for the fact that the meanings of other words (such as "nuclear") may evolve over time or may even be neologisms that do not exist early in the sample.[19]

The seed words for the non-disaster-focused topics are manually collected from Shiller (2019). These words are discussed extensively in Shiller (2019). We also add certain words that help define the themes of the topics. Importantly, to avoid any look-ahead bias, in selecting the seed words, we exclude any words that were only introduced recently, such as *bitcoin*, *machine learning*, or *great recession*. As shown in Table 1, we have reclassified the 9 topics from Shiller (2019) into 12 topics to facilitate our estimation. Specifically, as *Panic* and *Confidence* are opposing notions, we split them into two topics. Similarly, *Frugality versus conspicuous consumption* is split into *Frugality* and *Conspicuous consumption*. We further divide *Real estate booms and bursts* into two separate topics, namely, *Real estates booms* and

[17] In the setup of LDA, "tension(s)" tends to not be assigned to *War* in documents that talk little about war (such as articles about tension headaches), and to be assigned to *War* in documents that talk a lot about war (such as articles about international tensions).

[18] As an illustration of semantic shifts, "inflation" once referred to an increase in the money supply, but since the early twentieth century it has referred to a general increase in the prices of goods and services in an economy (Homer and Sylla 1996). The word "amortization," once referred only to the reduction of debt over time (Dictionary 1993), but in the twentieth century has come to also refer to the gradual reduction in the value of an asset (Financial Accounting Standards Board [FASB] started this definition in 1973). In the eighteenth century, the word "budget" referred to a financial statement outlining a government's anticipated expenses and revenues for the coming year. By the 1850s, the term expanded to include nongovernmental entities, eventually encompassing the financial accounts of families or individuals (https://www.merriam-webster.com/words-at-play/financial-word-origins).

[19] During the early twentieth century, the term "nuclear" was primarily employed within the realm of atomic structure and nuclear physics (see Rutherford (2012)). As the mid-twentieth century approached, the development and utilization of nuclear weapons during World War II led to an association between "nuclear" and the immense destructive force of such armaments (Rhodes (2012)). In the aftermath of World War II and throughout the Cold War era, "nuclear" was increasingly linked to the application of nuclear technology for energy production (Walker (2004)). Advancing into the late twentieth and early twenty-first centuries, the scope of "nuclear" broadened to encompass the concept of nuclear families (Cherlin (2010)).

In your analysis you should include the 14 discourse topics, but you do not need to include all seed words. I recommend adding a 15th topic related to trade wars and tariffs.

Table 1
Seed words

Narrative	Short name	Seed words
War	War	conflict, tension, terrorism, terrorist, war
Pandemic	Pandemic	epidemic, pandemic
Panic	Panic	bank failure, bank panic, bank run, crisis, depression, downturn, fear, financial panic, hard time, panic, recession
Confidence	Confidence	business confidence, consumer confidence
Savings	Saving	compassion, family morale, frugal, frugality, modesty, moral, poverty, saving
Conspicuous consumption	Consumption	american dream, conspicuous consumption, consumption, equal opportunity, equality, homeownership, luxury, patriotism, prosperity
Monetary standard	Money	bimetallism, devaluation, gold, gold standard, inflation, monetary standard, money, silver
Technology-replacing jobs	Tech	automate, computer, digital divide, electronic brain, invention, labor save, labor save machine, machine, mechanize, network, technocracy, technological unemployment, technology, unemployment
Real estate booms	Real estate boom	boom, bubble, flip, flipper, home ownership, home purchase, house boom, house bubble, land boom, land bubble, price increase, real estate boom, real estate bubble, speculation
Real estate busts	Real estate crash	bust, crash, house bust, house crash, land bust, land crash, price decrease, real estate bust, real estate crash
Stock market bubbles	Stock bubble	advance market, boom, bubble, bull, bull market, bullish, earnings per share, inflate market, margin, margin requirement, market boom, market bubble, price earn ratio, price increase, sell short, short sell, speculation, stock market boom, stock market bubble
Stock market crashes	Stock crash	bear, bear market, bearish, bust, crash, fall market, market crash, stock crash, stock market crash, stock market decline
Boycotts and evil business	Boycott	anger, boycott, community, evil business, excess profit, fair wage, moral, outrage, postpone purchase, profiteer, protest, strike, wage cut
Wage and labor unions	Wage	consumer price, cost of live, cost push, cost push inflation, high wage, increase wage, inflation, labor union, rise cost, wage, wage demand, wage lag, wage price, wage price spiral

This table lists the *lemmatized* seed words for each of the 14 discourse topics. The first column presents the full name of the topic, and the second column reports the short name used in the paper.

Real estates crashes.[20] In addition to *Stock market bubbles*, we add *Stock market crashes*. In contrast, because of their similarities, we combine *Labor saving machines* and *Automation and artificial intelligence* into one topic.

In addition to the 14 topics discussed above, we include one additional "garbage collector" to absorb everything else in the news unrelated to these topics.

[20] We replace the term "bursts" with "crashes," as the phrase "real estate burst" is not common in popular usage, and the word "burst" might be taken to mean a burst of positive activity, which is not the intended meaning.

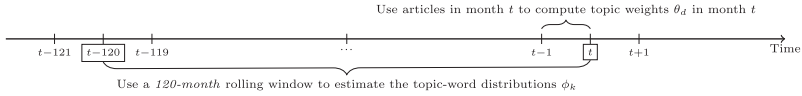

Figure 1
Estimation scheme
This figure plots the rolling estimation scheme for the sLDA model. Every month t, news articles in the previous 120 months (including month t) are used to estimate the sLDA model, and then articles in month t are used to compute topic weights in that month.

2.3 Estimation

Figure 1 illustrates the rolling estimation scheme used in the paper. At the end of each month t, we run the sLDA model using all news data over the preceding 120 months (months $t-119$ to t). We use 10 years of news data in the monthly estimation to balance the amount of news data required to estimate the model and computational costs. On average, every 10 years of historical data consists of around 460,000 articles, which should be sufficient to extract the topic weights at the time of estimation reliably. Notably, within topic models, rolling estimation is viable only under the sLDA model because the seed words that guide this approach provide consistency of thematic content over time.

We use Gibbs sampling to estimate the parameters of the model. We draw 200 drawings from the posterior distribution of z_{dv}, the realized topic for word location v in document d in the sLDA model, where we are conditioning on observed word frequencies.[21] In each drawing, we condition on the estimated values of the parameters of the model derived from previous drawing (where in the first draw, the initial estimate comes from a random number generator). In the last draw, we estimate our final value of the document-topic weights τ_d; that is, we estimate one 14×1 vector $\tau_d=[\tau_d^1,\tau_d^2,\ldots,\tau_d^{14}]$ for each news article, d, in the estimation window.

We then provide estimates of model parameters that condition on month t within the data set. We compute the global monthly weights of each topic k ($k=1,2,\ldots,14$) as the average weight of each topic across all articles in month t, weighted by the length $L(d)$ of each article:

$$\tau_t^k=\frac{\sum_{d=1}^{n_t}\tau_d^k L(d)}{\sum_{d=1}^{n_t}L(d)}, \tag{1}$$

where τ_t^k is the weight of topic k in month t, n_t is the total number of news articles in month t, and $L(d)$ is the total number of unigrams (one-word terms), bigrams (two-word terms), and trigrams (three-word terms) in article d.[22] (Equal weighting of topic weights across articles yields similar results.)

[21] In addition to the number of topics and articles, the number of samples drawn from the posterior distribution is a computational cost consideration in any topic model.

[22] An n-gram is a sequence of n words. For instance, "San Diego" is a bigram, and "A study of topics is needed" is a 6-gram.

Although 10 years of news articles are used to estimate the model each month, the final topic weights in month t are computed from the news articles of that month only. The final output of the estimation process is a time series of monthly weights for each of the 14 topics. This time series is used for our economic and financial forecasting applications.

In your analysis you can use Wikipedia Pageviews data and Google Trends data (only for an in-sample analysis as values are scaled).

3. Data

We exploit the richness of full newspaper texts using articles since the beginning of the *NYT* inception. We remove articles with limited content, such as those that contain mostly numbers, names, or lists. We then conduct standard text processing steps, as reported in detail in the Internet Appendix B and Table C.1. Our analysis is based on all articles from the *NYT* since 1861 and all *WSJ* articles since 1990. Other finance research that uses *NYT* text includes Garcia (2013) and Hillert and Ungeheuer (2019).

Articles from the first 10 years of the *NYT* since its inception are used to estimate the first monthly topic weights. We start our *NYT* sample in 1871 as the S&P 500 data are available from that year.[23] (These data were provided to us by a private company.)

Then, for each month t, we create a document term matrix containing all articles over the preceding 10 years through the current month. Each row of the matrix is an article, each column is a n-gram, and each entry is the count of that term in the article. The document-term matrix and topic-based seed words are input into the sLDA model to estimate monthly topic weights as described in the previous section. To streamline the presentation, we report the results for the *WSJ* in Internet Appendix F.

Panel A of Figure C.1 plots the time series of monthly article counts in our sample. After removing articles with limited content, since 1871, our *NYT* data has more than 6.8 million news articles with a monthly average of 3,800.[24] Before 1900, the *NYT* published about 2,000 articles a month. The number of monthly articles increased gradually after 1900, hovering between 4,000 and 6,000 until the end of the twentieth century. Amidst industry-wide struggles related to declining ad revenues and subscriber bases beginning in the 2000s, the *NYT* started scaling down its publishing capacity to around 2,000 articles a month during the 2010s.[25] However, the number of monthly articles surges back to just under 4,000 toward the end of the sample. A newspaper strike occurred from 1902 to 1903, and news articles spiked at the start of World War I.

---

[23] Bybee et al. (2024) focus on economic articles in the *WSJ* with a sample that starts in 1984. Our sample starts 100 years earlier. Manela and Moreira (2017) include the articles from the *WSJ* since 1890 but focus on the headline and title of articles on the front page, whereas we cover all articles. Baker, Bloom, and Davis (2016) and Caldara and Iacoviello (2022) study an extensive collection of newspapers but apply a dictionary approach. Caldara and Iacoviello (2022) count, each month, the number of articles discussing rising geopolitical risks. They do not consider the number of words in each article whereas our methodology does.

[24] Data are missing for September and October 1978 (because of strikes) and thus are excluded from Figure C.1.

[25] For more details, see Pew Research Center (2023).

Panel B of Figure C.1 reports the average monthly article length, which is defined as the total count of unigrams (one-word terms), bigrams (two-word terms), and trigrams (three-word terms). (Internet Appendix B provides more details on the construction of n-grams.) While Bybee et al. (2024) consider only unigrams and bigrams, we extend the analysis to trigrams as a majority of the seed words have three words. Examples include *real estate boom, stock market bubble*, and *cost push inflation*. Over 1871–2019, articles have an average length of 493 n-grams. Articles tended to have around 500 n-grams until the 1920s. After that, the number hovered just above 400 n-grams until the 1960s. Since then, article length has increased, reaching about 600 n-grams during the 2010s.

The second set of data concerns stock market outcomes. We obtain the total S&P 500 index from Global Financial Data (GFD) with monthly data from January 1871.[26] Monthly risk-free rates are downloaded from Professor Kenneth French's website. For monthly risk-free rates before 1927, we use the series from Goyal and Welch (2008).

4. Discourse Topics

We examine the contents of news topics in Section 4.1. In Section 4.2, we discuss the summary statistics of our topics and in Section 4.3, we discuss their time series.

4.1 Contents of news topics

In a semisupervised topic model, such as sLDA, the favored approach in the literature to evaluate the choice of seed words is to investigate the most common terms within a topic post-estimation to determine whether the topics feature the desired contents (Lu et al. 2011; Watanabe and Zhou 2020). Hence, to investigate the contents of the 14 extracted topics, during every monthly estimation of the sLDA model, we retain the 30 most common n-grams per topic, that is, those having the highest probabilities in that topic. Then the most important words for each topic are identified as those that have the highest frequency over time.

To visualize each topic, we create word clouds using the top words from each topic; the higher the frequency of a word in the topic, the larger the word size. We report the word clouds of six main topics (based on their weights in the PLS index discussed next) in Figure 2, and the remaining topics in Figure C.2 in the Internet Appendix.

[26] The GFD description is as follows: "The S&P 500 Total Return Index is based upon GFD calculations of total returns before 1971 […] Beginning in 1871, data are available for stock dividends for the S&P Composite Index from the Cowles Commission and from S&P itself. We used this data to calculate total returns for the S&P Composite using the S&P Composite Price Index and dividend yields through 1970, official monthly numbers from 1971 to 1987, and official daily data from 1988."

Return data on the S&P 500 and risk-free rates can be found on Amit Goyal's website (https://sites.google.com/view/agoyal145). For a description of variables refer to Goyal and Welch (2008).

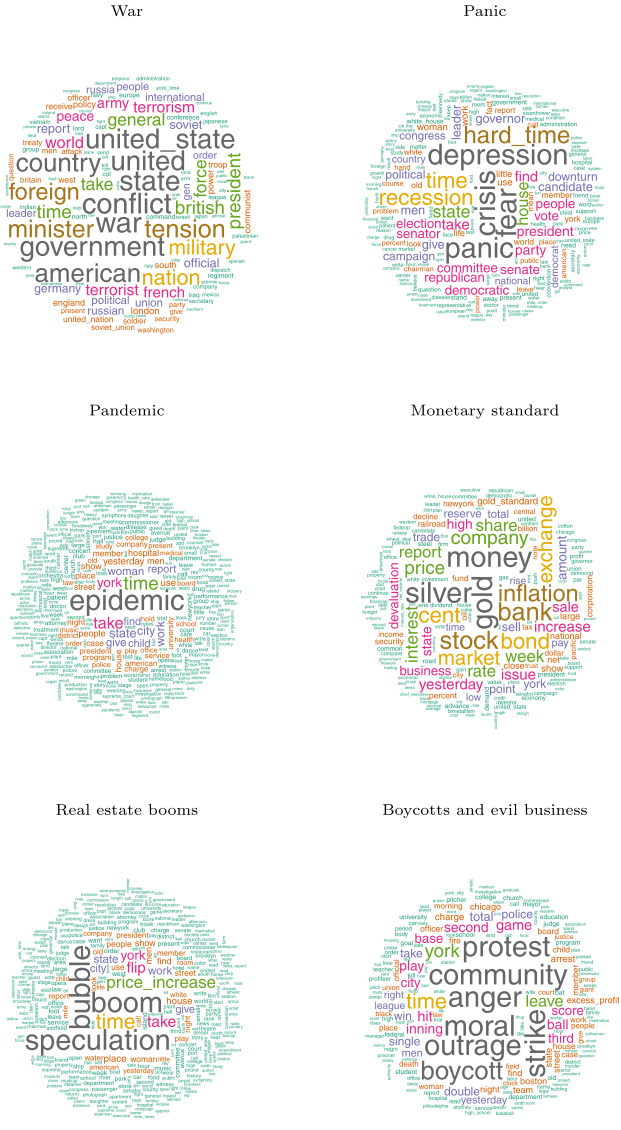

Figure 2
Discourse topic contents
This figure plots the frequencies of n-grams per topic over time. Frequencies are constructed according to the sLDA model described in Section 2, and the size of each n-gram indicates its frequency. The sample period is from January 1871 to October 2019.

Instead of creating word clouds as in Figure 2, use a simplified Topic Attention Measure based on Wikipedia data to construct a figure of which topics captured public attention over time.

As indicated by Figure 2, the sLDA model seems to perform well at extracting these topics from the *NYT* articles. For example, the most common terms for *War* extracted by the model are *conflict, war, government, tension* and for *Panic* are *panic, fear, crisis, depression, recession, hard_time*, all of which strongly overlap with the seed words. Although both *War* and *Panic* feature stress and anxiety, they capture distinct themes, and their correlation is –17% as reported in Table C.2. The top words for *Monetary* are *money, gold, silver, inflation, bank*; for *Real Estate Booms* are *bubble, boom, speculation, price increase*; and for *Boycott* are *boycott, outrage, strike, moral, anger, community, protest*. Except for *Pandemic*, the thematic contents of these extracted topics are consistent with the predefined list of seed words.

4.2 Summary statistics

We report the summary statistics for the 14 topic weights in Table 2. *War* receives the most attention on average, with a mean time-series weight of 9.7%. About 10% of the monthly *NYT* articles use one of the *War* words at least once. Table 2 shows that *War* is also the most volatile topic with a standard deviation of 3.7%, followed by *Stock Market Crash* at 3.1%.

For predicting stock market returns, we create a composite topic index by extracting and combining the signals most relevant to return prediction from all topics via the two-step PLS method, which has recently gained wide popularity in the literature (Kelly and Pruitt 2013, 2015; Huang et al. 2015; Huang, Li, and Wang 2020). As a first step, the time series of each topic weight is regressed on the time series of next-month market returns using the whole sample. Second, in each period t, the vector of topic weights is regressed on the vector of slopes obtained in the first step. The slope in the second step regression is a value of the PLS index in period t.

The second to last column of Table 2 reports the PLS loadings (the slope in the time-series regressions) for all topics. In this methodology, only the relative PLS weights of the components are meaningful. *War* receives the highest weight, and its positive loading indicates that *War* is a positive predictor of market returns. Other essential topics in the PLS index include *Real estate booms*, *Pandemic*, and *Panic*. Surprisingly, the topics receiving the smallest weights are *Stock market bubbles* and *Stock market crashes*. These facts are potentially useful for future theorizing about economic narratives and the stock market.

The last column of Table 2 reports the correlations between the 14 topics and the PLS index. As expected, the PLS index is highly correlated with *War* with a correlation coefficient of 82%.

4.3 Time series of discourse topics

Next, we examine fluctuations in topic weights over time. We plot the time series of each demeaned topic weight from January 1871 to October 2019. The results for the six main topics and the PLS index are displayed in Figure 3, and

Provide summary statistics for the Wikipedia pageview data on your topics. You can report N (number of articles included for a topic),mean, median, SD and AR(1) coefficient.

Table 2
Summary statistics

	N	Mean	SD	Q1	Median	Q3	AC(1)	PLS weights	Corr PLS
War	1784	9.71	3.73	6.56	9.64	11.92	84.84	5.22	82.09
Pandemic	1784	5.72	2.34	4.25	5.38	6.72	7.54	−2.26	−28.53
Panic	1784	8.30	2.70	6.21	7.94	10.13	37.03	2.19	15.74
Confidence	1784	5.72	2.45	3.97	5.39	6.99	7.61	0.60	−0.47
Saving	1784	5.84	2.17	4.39	5.51	6.84	29.78	−1.09	−34.12
Consumption	1784	7.36	2.85	5.46	6.72	9.23	27.62	0.88	−4.58
Money	1784	6.58	2.06	5.21	6.46	7.87	60.55	−1.77	−15.26
Tech	1784	6.61	2.52	4.99	6.57	8.07	54.58	−0.79	−8.15
Real estate boom	1784	5.95	2.52	4.20	5.60	7.22	9.31	−2.82	−32.73
Real estate crash	1784	5.57	2.16	4.23	5.41	6.49	23.16	0.47	−1.12
Stock bubble	1784	5.79	2.30	4.28	5.74	7.23	48.98	−0.40	−14.67
Stock crash	1784	7.40	3.13	5.06	6.86	9.60	26.56	0.86	−2.31
Boycott	1784	5.79	2.62	4.14	5.51	7.37	67.00	−1.47	−41.69
Wage	1784	7.90	2.59	6.05	7.60	9.50	38.53	1.07	37.33
PLS	1784	49.99	44.73	18.75	46.68	77.28	70.35		

This table presents the summary statistics for the time series of 14 monthly topic weights constructed according to the sLDA model described in Section 2. All numbers (except sample size) are expressed as percentages. The sample period is from January 1871 to October 2019.

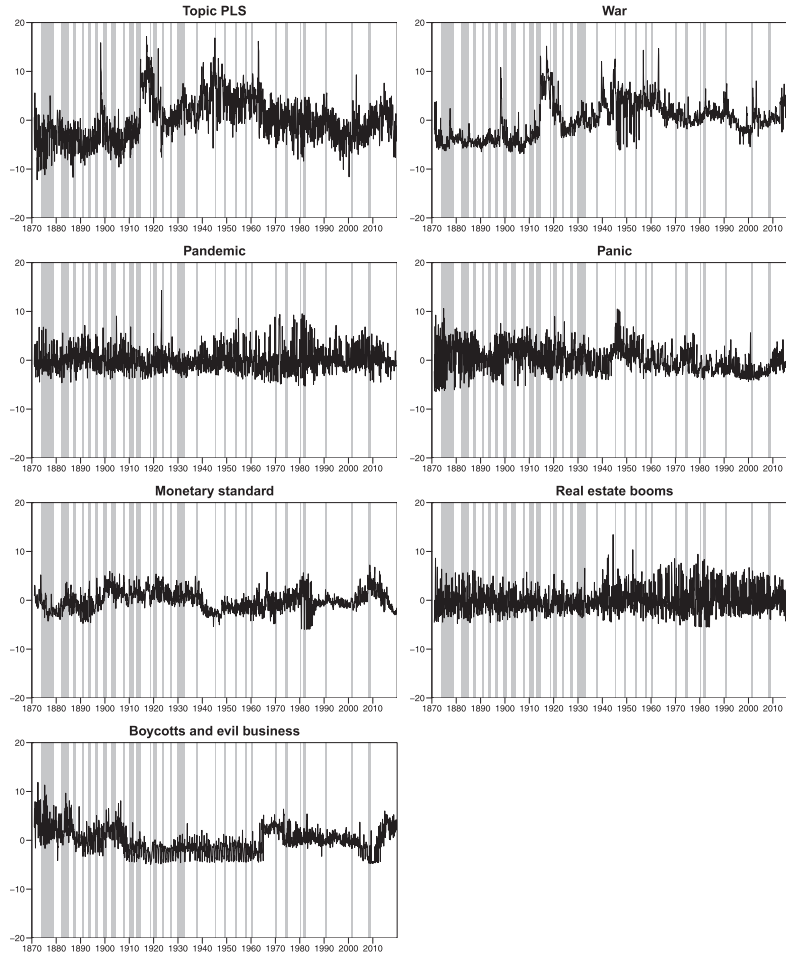

Figure 3
Time series of discourse topic weights
This figure plots the time series of monthly topic weights constructed according to the sLDA model described in Section 2. Topic weights are demeaned to improve visualization. The gray-shaded areas represent NBER-defined recessions. The sample period is from January 1871 to October 2019.

a larger plot of all topics is shown in Figure C.3 in the Internet Appendix. As can be seen from the graphs, except for *War*, the topics do not display any clear patterns. Thus, we focus our discussion on the time series of *War*.

Figure 3 describes the time series of *War*. *War* spiked in the 1870s, the Reconstruction period after the American Civil War. It also surged during the 1890s, a period that featured the Spanish-American War in 1898 and the Philippine-American War of 1899–1902. *War* rose to its highest since the start of the sample during World War I from 1917 to 1918. It remained low during

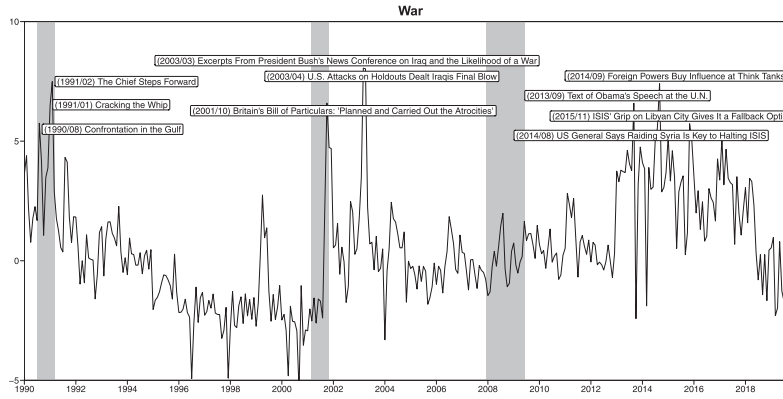

Figure 4
Articles making the biggest contribution to *War* spikes since 1990
This figure plots the 10 articles that have contributed significantly to 10 monthly heights of *War* since 1990. Topic weights are demeaned. The gray-shaded areas represent NBER-defined recessions. The sample period is from January 1990 to October 2019.

the 1920s and 1930s before surging again during World War II. *War* reached its all-time high in 1963 due to major developments of the Vietnam War. The figure also shows the graphs of other topics as well for comparison.

Figure 4 describes the time series of *War* over the last 30 years of the sample. We track the 10 articles with the most significant contributions to the 10 highest monthly scores of *War* hikes since 1990.[27] Over the last 30 years, *War* spiked in the early 1990s during the Gulf War, and surged again at the end of 2001 after the 9/11 terrorist attack. In recent years *War* has remained high, especially from 2014 to 2018. During this time, the important articles reflect the climate of the period: stories are full of international tensions, notably including the Russian annexation of Crimea and the nuclear weapons threat from North Korea.

5. Discourse Topics as Stock Market Predictors

The primary empirical question of the paper is whether rare disasters and non-disaster-focused discourse topics predict the U.S. stock market return. In Section 5.1, we consider 1 month return prediction. In Section 5.2, we consider long-horizon prediction. In the later subsections, we control for standard return predictors from past literature and examine whether discourse topics have incremental predictive power.

[27] Each month, the most influential article is the article with the highest product of article-level topic weight and article length, that is, the numerator in Equation (1). Equal weighting, ignoring the article length, can help one identify slightly different influential articles. Still, these other articles are generally thematically similar to the most influential articles reported here.

The is the standard predictive regression model you should use with your discourse topic data from Wikipedia (and Google Trends for an in-sample analysis.

5.1 Predicting 1-month-ahead returns

To investigate the return predictive power of discourse topics, we run the following standard predictive regression:

$$R^e_{t+1} = \alpha + \beta x_t + \epsilon_{t+1}, \tag{2}$$

where R^e_{t+1} is the annualized excess market return over the next month, x_t is one of the topics or the topic PLS indexes standardized to zero mean and unit variance, and β, the coefficient of interest, measures return predictability. The reported t-statistics are computed with Newey and West (1987) standard errors.

Table 3 reports the results. Over the whole 1871–2019 sample, among the 14 topics, *War* is the strongest positive predictor, with the coefficient statistically significant at the 1% level. Economically, a one-standard-deviation increase in *War* is associated with a 3.8% increase in the annualized excess return in the next month.

Please see for additional context on equity premium predictability.

In addition to the full sample analysis, we run predictive regressions over three subperiods: 1871–1949, 1950–2019, and 2000–2019. This addresses possible concerns about text quality in the earlier part of the sample. Furthermore, it is interesting to examine whether financial market behavior is different during the latest two decades, with the rise of internet usage and new communication technologies. The results during this period may be the most relevant for the future, as emphasized in Goyal and Welch (2008).

The positive association between *War* and future market returns remains in both subperiods with significance at the 5% level. Furthermore, *War* yields an impressive forecasting power over the past two decades with a coefficient of 9.8%, significant at the 1% level, and an in-sample R^2 of 3.4%.

Among the remaining economic discourse topics, *Pandemic* and *Real Estate Boom* are negative return predictors over the whole sample, both significant at the 5% level. In contrast, *Panic* is a positive predictor of market returns, significant at the 10% level. Among these nonwar topics, only *Real Estate Boom* yields meaningful predictions across all subsamples.

The last portions of Table 3 report return prediction results using the PLS method. We recursively construct the PLS indexes using only data available up to each month with an initial estimation window of 120 months.

The PLS index constructed from all 14 topics predicts returns more strongly than *War* alone. Over the total sample, a one-standard-deviation increase in the PLS index is associated with a 4.65% increase in the annualized return in the next month, with an in-sample R^2 of 0.58%. Moving from earlier to later subsamples, the PLS index displays increasingly strong predictive results, significant at the 1% level even in the later subsamples. This suggests that the combined information in all topics has predictive power for long time-series data.

We also examine the predictive power of only the topics discussed by Shiller (2019). To do so, we construct the "Shiller PLS" index by excluding *War* and *Pandemic* and report the prediction results using this index in the last row of

Your sample period would be smaller with Wikipedia Pageview data only starting in 2015 and Google Trends data becoming available in 2004.

Table 3
Predicting 1-month market returns

	1871-2019	1871-1949	1950-2019	2000-2019
War	3.80***	3.49**	4.06**	9.83***
t-stat	(3.35)	(2.02)	(2.56)	(3.43)
R^2 (%)	0.39	0.20	0.55	3.39
Pandemic	−2.61**	−3.98**	−1.55	−2.31
t-stat	(−2.06)	(−2.14)	(−0.91)	(−0.70)
R^2 (%)	0.15	0.29	−0.02	−0.21
Panic	2.21*	3.06*	2.57*	3.36
t-stat	(1.76)	(1.71)	(1.69)	(1.20)
R^2 (%)	0.09	0.13	0.15	0.02
Confidence	0.66	−0.14	1.13	0.77
t-stat	(0.55)	(−0.08)	(0.67)	(0.24)
R^2 (%)	−0.04	−0.11	−0.07	−0.40
Saving	−1.37	−1.62	−1.44	−0.98
t-stat	(−1.05)	(−0.92)	(−0.81)	(−0.37)
R^2 (%)	0.00	−0.04	−0.04	−0.39
Consumption	0.84	0.63	2.58	0.07
t-stat	(0.66)	(0.32)	(1.61)	(0.02)
R^2 (%)	−0.03	−0.10	0.15	−0.42
Money	−2.33	−1.23	−3.41*	−0.70
t-stat	(−1.64)	(−0.61)	(−1.76)	(−0.17)
R^2 (%)	0.11	−0.07	0.36	−0.40
Tech	−0.85	0.54	−3.10	−14.25***
t-stat	(−0.58)	(0.26)	(−1.62)	(−3.24)
R^2 (%)	−0.03	−0.10	0.27	7.59
Real estate boom	−3.03**	−3.51**	−3.02*	−6.57**
t-stat	(−2.50)	(−2.05)	(−1.82)	(−2.16)
R^2 (%)	0.23	0.20	0.25	1.28
Real estate crash	0.59	0.59	0.93	−2.58
t-stat	(0.48)	(0.34)	(0.55)	(−0.78)
R^2 (%)	−0.05	−0.10	−0.08	−0.16
Stock bubble	−0.47	−1.67	0.39	−4.59
t-stat	(−0.37)	(−0.96)	(0.21)	(−1.31)
R^2 (%)	−0.05	−0.04	−0.11	0.41
Stock crash	0.75	2.08	−0.23	−0.83
t-stat	(0.54)	(1.02)	(−0.14)	(−0.26)
R^2 (%)	−0.04	0.00	−0.12	−0.40
Boycott	−1.52	−2.36	−0.43	5.33
t-stat	(−1.23)	(−1.41)	(−0.23)	(1.38)
R^2 (%)	0.01	0.04	−0.11	0.70
Wage	1.12	0.70	1.87	8.62***
t-stat	(0.74)	(0.32)	(1.04)	(2.81)
R^2 (%)	−0.02	−0.09	0.02	2.51

(continued)

Replicate Table 3 with data you obtain from Wikipedia Pageviews and Google Trends (as this is in-sample). You might also want to consider normalizing (Z-score or Min-Max normalization) your topic measures for articles (as raw views on individual articles within a topic may differ significantly for a topic.

Table 3
Continued

	1871-2019	1871-1949	1950-2019	2000-2019
PLS	4.65***	3.92*	5.33***	9.61***
t-stat	(3.52)	(1.82)	(3.35)	(3.58)
R^2 (%)	0.58	0.24	1.04	3.22
Shiller PLS	2.77**	2.34	3.66**	2.76
t-stat	(2.05)	(1.10)	(2.34)	(0.94)
R^2 (%)	0.17	0.01	0.43	−0.12

This table presents the results of the following predictive regression:

$$R^e_{t+1} = \alpha + \beta x_t + \epsilon_{t+1},$$

where R^e_{t+1} is the excess market return over the next month, x_t is one of the discourse topics or the PLS indexes, and β, the coefficient of interest, measures the strength of predictability. "Shiller PLS" uses only the topics from Shiller (2019), excluding *War* and Pandemic. Returns are expressed as annualized percentages, and the independent variable is standardized to zero mean and unit variance. Reported are the OLS estimates of β and t-statistics computed with Newey and West (1987) standard errors. The sample period is from January 1871 to October 2019. *$p < .1$; **$p < .05$; ***$p < .01$.

Table 3. Accordingly, the Shiller PLS index displays similar prediction patterns as the composite PLS index, albeit with smaller magnitudes. For example, over the whole sample, the Shiller PLS has a prediction coefficient of 2.77% and an R^2 of 0.17% compared to 4.65% and 0.58%, respectively, of the composite PLS index.

Table C.4 reports the small sample bias correction for the prediction slope and t-statistics in Table 3 as proposed by Boudoukh, Israel, and Richardson (2022). The results for *War* and the PLS indexes hardly change.

Following Golez and Koudijs (2018), we compute the cumulative in-sample R^2 in predicting the next month return, reported in panel A of Figure 5. An upward trend indicates a predictor performs well during the sample period. Both *War* and the composite PLS index experience poor performances during 1910–1930 but strongly recover after that. Again, both cumulative R^2 suffer from a slight decline for a short period before 2000.

Overall, Table 3 indicates that *War* and the PLS index are strong market predictors, and that their forecasting power increases in more recent periods. The predictive power of *War* and the PLS index is most pronounced from 2000 to 2019. We conjecture that the digitization of news and the technology that accelerates the diffusion of information drive this result. This result complements that of Obaid and Pukthuanthong (2022), who find strong market predictive power in the sentiment of photos and text starting in 2010s.

5.2 Predicting long-horizon returns

You do not need to do your analysis for longer horizons.

We have found that *War* and the PLS index predict market returns at a 1 month horizon. We now examine the long-horizon predictive power of *War* and the PLS index by running the predictive regression:

$$R^e_{t+1\to t+h} = \alpha + \beta x_t + \epsilon_{t+1\to t+h}, \tag{3}$$

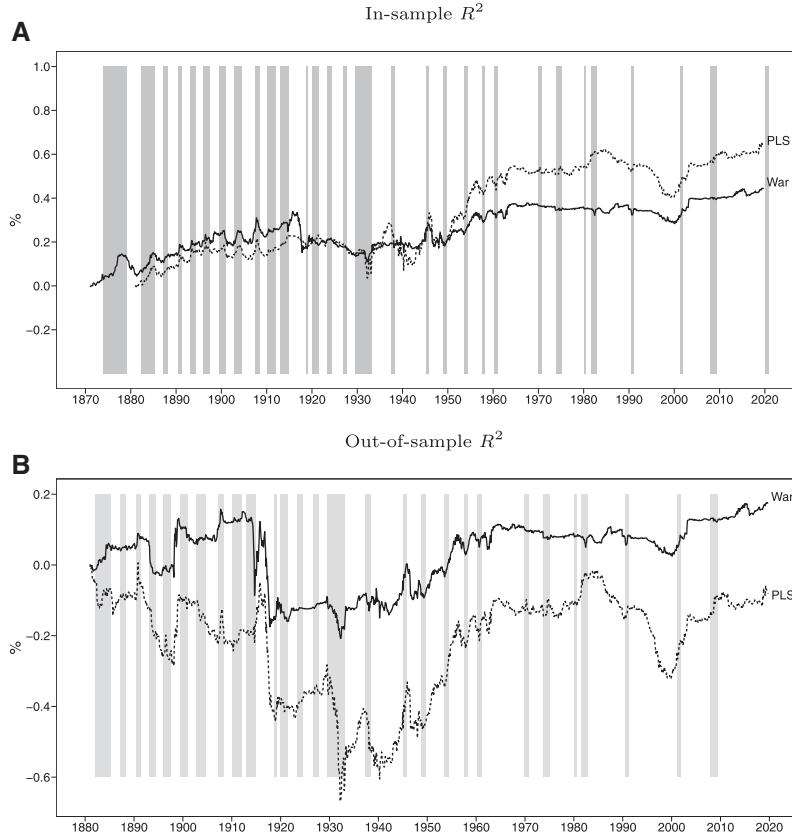

Figure 5
Cumulative R-squared in 1-month return prediction
Panel A plots the cumulative in-sample R^2 computed as

$$\left(\sum_{s=1}^{t}\left(R_s^e-\bar{R}^e\right)^2-\sum_{s=1}^{t}\left(R_s^e-\hat{R}_s^e\right)^2\right)/\sum_{s=1}^{T}\left(R_s^e-\bar{R}^e\right)^2,$$

where $\bar{R}^e$ is the sample mean of excess return and $\hat{R}_s^e$ is the fitted value from regression (2). The sample period is from January 1871 to October 2019. Panel B plots the cumulative out-of-sample R^2 computed as

$$\left(\sum_{s=1}^{t}\left(R_s^e-\bar{R}_s^e\right)^2-\sum_{s=1}^{t}\left(R_s^e-\hat{R}_s^e\right)^2\right)/\sum_{s=1}^{T}\left(R_s^e-\bar{R}_s^e\right)^2,$$

where $\bar{R}_s^e$ and $\hat{R}_s^e$ are the historical mean and predicted value, estimated based on the preceding estimation window. The evaluation period is from January 1881 to October 2019.

where $R^e_{t+1\to t+h}$ is the annualized excess market return over the following h months, x_t is either *War* or the PLS index, and β, the coefficient of interest, measures the strength of predictability. To account for potential

Table 4
Predicting long-horizon market returns

	War	*t*-stat	R^2 (%)	PLS	*t*-stat	R^2 (%)
			1871-2019			
$h = 1$	3.80***	(3.35)	0.39	4.65***	(3.52)	0.58
$h = 3$	2.87***	(2.81)	0.57	3.72***	(3.55)	0.96
$h = 6$	3.03***	(2.83)	1.36	3.28***	(3.26)	1.56
$h = 12$	2.79**	(2.48)	1.93	2.76***	(2.66)	1.86
$h = 24$	2.09**	(2.08)	1.91	2.39**	(2.39)	2.55
$h = 36$	2.28**	(2.22)	3.07	3.06***	(3.10)	5.72
			1871-1949			
$h = 1$	3.49**	(2.02)	0.20	3.92*	(1.82)	0.24
$h = 3$	2.51	(1.60)	0.26	3.00*	(1.75)	0.37
$h = 6$	2.88*	(1.77)	0.94	3.00*	(1.85)	0.95
$h = 12$	2.87*	(1.73)	1.59	2.76*	(1.74)	1.36
$h = 24$	1.90	(1.38)	1.25	2.54*	(1.73)	2.27
$h = 36$	2.11	(1.45)	2.17	3.32**	(2.31)	5.63
			1950-2019			
$h = 1$	4.06**	(2.56)	0.55	5.33***	(3.35)	1.04
$h = 3$	3.18**	(2.41)	1.06	4.40***	(3.64)	2.14
$h = 6$	2.93**	(2.24)	1.64	3.41***	(2.83)	2.26
$h = 12$	2.21	(1.50)	1.64	2.56*	(1.94)	2.23
$h = 24$	1.85	(1.24)	1.93	1.99	(1.54)	2.25
$h = 36$	1.96	(1.36)	2.85	2.49*	(1.90)	4.70
			2000-2019			
$h = 1$	9.83***	(3.43)	3.39	9.61***	(3.58)	3.22
$h = 3$	7.64***	(4.07)	6.07	5.24***	(3.12)	2.63
$h = 6$	6.08***	(3.31)	6.76	4.05***	(3.13)	2.76
$h = 12$	5.42**	(2.47)	9.68	3.10**	(2.51)	2.89
$h = 24$	4.78**	(2.28)	12.78	2.47*	(1.96)	3.11
$h = 36$	4.59***	(2.86)	17.72	3.24***	(3.43)	8.61

This table presents the results of the following predictive regression:

$$R^e_{t+1\to t+h} = \alpha + \beta x_t + \epsilon_{t+1\to t+h},$$

where $R^e_{t+1\to t+h}$ is the excess market return over the next h months, x_t is either *War* or the PLS indexes constructed from 14 topics, and β, the coefficient of interest, measures the strength of predictability. Returns are expressed as annualized percentages, and the independent variable is standardized to zero mean and unit variance. Reported are the OLS estimates of β and *t*-statistics computed with Newey and West (1987) standard errors using the corresponding h lags. The sample period is from January 1871 to October 2019. $^*p < .1$; $^{**}p < .05$; $^{***}p < .01$.

autocorrelations of the residuals in the long-horizon predictive regressions, we compute the Newey and West (1987) standard errors with corresponding h lags.

The first row of each panel in Table 4 repeats the results for $h=1$ for comparison. Table 4 reports the results for the full test period from 1871 to 2019. Over these 149 years, *War* and the PLS index significantly predict market returns up to a horizon of 36 months ahead.

In the subsample analysis, the predictive power of *War* is relatively weak during the first half of the sample period (significant at the 5% level for the 1 month horizon). Still, it is significant at the 5% level from 1- to 6-month horizons during the second subperiod (1950 to 2019). The predictive power of

the PLS index is weak during the first half of our sample period, but it becomes stronger during the second half.

The strongest effects are obtained starting from the year 2000. *War* yields impressive predictive power over the last 20 years of the sample; its in-sample adjusted R^2 ranges from 3.4% (1 month) to 18% (36 months). During this period, the PLS index yields strong results, significant at the 1% level across all forecasting periods except for the 12- and 24-month horizons. As for economic magnitudes, a one-standard-deviation increase in *War* is associated with an annualized increase of 9.8% in next month return over the 2000–2019 period. The corresponding number for the PLS index is 9.6%. The mean S&P500 annualized excess return during the same period is 5.1% suggesting the predictive power of *War* and the PLS index is economically substantial. Table C.5 reports the small sample bias correction for the prediction slope and t-statistics in Table 4 as proposed by Boudoukh, Israel, and Richardson (2022). Similar to Table C.4, the results for *War* and the PLS indexes hardly change.

5.3 Predicting 1-month-ahead returns: *War* versus economic and topic predictors

The previous two subsections show that *War* is a strong predictor of stock market returns. Next, we investigate whether *War* has predictive power beyond standard economic predictors and the remaining 13 topics studied in this paper. For economic predictors, we include the dividend-price ratio (DP), earnings-price ratio (EP), dividend payout ratio (DE), stock variance (SVAR), and Treasury-bill rate (TBL) from Goyal and Welch (2008). We include these variables since they are available for our full sample period of 1871 to 2019. We run the following bivariate regression:

$$R^e_{t+1} = \alpha + \beta War_t + \gamma z_t + \epsilon_{t+1}, \tag{4}$$

where z_t is either each of the economic or remaining topic predictors. All independent variables are standardized to zero mean and unit variance and t-statistics are computed with the Newey and West (1987) standard error.

Panel A of Table 5 reports the results for the economic predictors. In all bivariate regressions between *War* and each economic predictor, *War* remains significant at the 1% level with economic magnitude larger than those of the economic predictors. In the final column of panel A, when we run a kitchen sink regression that includes *War* and all economic predictors,[28] *War* is still significant at the 5% level.[29]

In panel B of Table 5, when *War* is tested against the remaining topic predictors, its statistical and economic significance remain intact in either bivariate or kitchen sink regressions.

[28] We exclude DE to avoid perfect collinearity because it is a linear combination of DP and EP.

[29] In Table C.6, we document that the predictive power of *War* remains intact when we control for market returns, conditional skewness, and conditional volatility.

Table 5
Predicting 1-month market returns: *War* versus economic and other topic predictors

	A. Economic predictors					
	(1)	(2)	(3)	(4)	(5)	(6)
War	3.85***	3.55***	3.75***	3.80***	3.23***	2.52**
	(3.36)	(3.08)	(2.97)	(3.36)	(2.81)	(1.98)
DP	1.54					−0.90
	(0.80)					(−0.34)
EP		2.03				3.88
		(1.29)				(1.44)
DE			−0.25			
			(−0.11)			
SVAR				−0.17		−0.28
				(−0.04)		(−0.07)
TBL					−3.09*	−4.21***
					(−1.92)	(−2.76)
R^2 (%)	0.40	0.46	0.33	0.33	0.61	0.75

(continued)

Table 5
Continued

B. Other topic predictors

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
War	3.64***	4.30***	3.87***	3.67***	4.31***	3.60***	3.78***	3.74***	3.89***	3.79***	4.11***	3.69***	3.73***	5.38**
	(3.21)	(3.74)	(3.41)	(3.13)	(3.51)	(3.16)	(3.35)	(3.30)	(3.39)	(3.34)	(3.48)	(3.08)	(3.20)	(2.29)
Pandemic	−2.38*													−1.48
	(−1.88)													(−0.89)
Panic		2.94**												3.33*
		(2.31)												(1.67)
Confidence			0.97											1.63
			(0.81)											(0.89)
Saving				−0.55										0.48
				(−0.41)										(0.28)
Consumption					1.97									2.12
					(1.44)									(1.13)
Money						−1.97								−1.82
						(−1.37)								(−1.04)
Tech							−0.79							−0.29
							(−0.55)							(−0.15)
Real estate boom								−2.97**						−2.05
								(−2.45)						(−1.24)
Real estate crash									0.98					1.33
									(0.79)					(0.80)
Stock bubble										−0.09				0.72
										(−0.07)				(0.41)
Stock crash											1.56			2.18
											(1.10)			(0.96)
Boycott												−0.33		0.11
												(−0.25)		(0.06)
Wage													0.27	0.55
													(0.18)	(0.26)
R^2 (%)	0.50	0.59	0.36	0.34	0.44	0.45	0.35	0.60	0.36	0.33	0.40	0.33	0.33	0.67

This table presents the results of the following predictive regressions:

$$R^e_{t+1} = \alpha + \beta War_t + \gamma z_t + \epsilon_{t+1},$$

where R^e_{t+1} is the excess market return over the next month and z_t is one of the economic predictors from Goyal and Welch (2008) (panel A) or the remaining topics (panel B). The last column reports the results when *War* is tested against all predictors in each panel. Returns are expressed as annualized percentages, and the independent variable is standardized to zero mean and unit variance. t-statistics are computed with Newey and West (1987) standard errors. The sample period is from January 1871 to October 2019. $^*p < .1$; $^{**}p < .05$; $^{***}p < .01$.

Overall, we find that investors' perception of war risks as captured by our *War* index is a robust predictor of stock market returns—there is a war market return premium. *War* outperforms other common economic variables and non-disaster-related discourse topics in predicting the next 1-month market returns.

5.4 Predicting 1-month-ahead returns: *War* versus other media-based uncertainty indexes

We have just seen that *War* has stronger predictive power than common economic predictors and non-disaster-focused discourse topics. However, the literature has introduced other news-based proxies for disaster risks, notably including the news implied volatility (NVIX) from Manela and Moreira (2017) and the geopolitical risks (GPR) from Caldara and Iacoviello (2022).[30] We now investigate whether our *War* contains incremental predictive power over these two measures.

In panel A of Table 6, we use *War*, NVIX^2, and GPR to predict the excess market return 1 month ahead as in Equation (4). We first run univariate regressions and then compare *War* and GPR against NVIX^2. We do not directly compare *War* with GPR because the two variables are highly correlated (correlation of 60%) over the sample period January 1900 to March 2016.[31] In contrast, *War* and NVIX^2 have a -5% correlation.

Over this period, both *War* and GPR are positive return predictors, significant at the 5% and 10% levels, respectively, where *War* yields a larger economic magnitude. In contrast, NVIX^2 does not predict the market at the 1 month horizon, consistent with the results in Manela and Moreira (2017). We observe the same results over the subsample 1950–2016. Over the most recent sample, 2000–2016, the *War* return premium dominates in terms of both economic and statistical significance.[32]

Overall, *War* produces stronger short-term predictive power for market returns than the other two media-based disaster risk measures, especially after year 2000.

5.5 Predicting 1-month-ahead returns: *War* versus crisis event counts

Berkman, Jacobsen, and Lee (2011) measure investor perceptions of disaster risks by counting the number of monthly crisis events. We now run a horse-race test of the effectiveness of this measure, which is based on actual crisis events, versus our media-based measure, which is based on textual discourse, in predicting returns.[33] We include the aggregate crisis index (*Crisis*) as this

[30] We thank the authors of these papers for making their data available.

[31] NVIX is only available until March 2016. Following their paper, we use NVIX^2 in our analyses. Using NVIX yields almost the same results.

[32] In unreported results, when we put three predictors together in the same regression, *War* drives out the significance of GPR during the period 2000–2016.

[33] The data have been updated to 2018 and available at https://sites.duke.edu/icbdata/.

Table 6
Predicting 1-month market returns: *War* versus other media-based uncertainty and crisis event count indexes

A. Other media-based uncertainty indexes

			1900-2016		
War	2.69**			2.70**	
	(1.99)			(2.00)	
$NVIX^2$		0.07		0.22	0.07
		(0.03)		(0.10)	(0.03)
GPR			2.48*		2.48*
			(1.72)		(1.72)
R^2 (%)	0.12	−0.07	0.09	0.05	0.02
			1950-2016		
War	3.96**			3.90**	
	(2.42)			(2.29)	
$NVIX^2$		1.01		0.67	0.96
		(0.32)		(0.21)	(0.31)
GPR			3.51*		3.50*
			(1.91)		(1.91)
R^2 (%)	0.50	−0.09	0.37	0.39	0.28
			2000-2016		
War	10.08***			10.19***	
	(3.11)			(3.05)	
$NVIX^2$		−0.63		−1.42	−0.07
		(−0.11)		(−0.25)	(−0.01)
GPR			7.32**		7.32**
			(2.27)		(2.32)
R^2 (%)	3.19	−0.50	1.44	2.75	0.92

(continued)

is the main variable studied in Berkman, Jacobsen, and Lee (2011) and the war count index (*CWar*), which is mostly related to our *War*. In panel B of Table 6, we show that over the 100-year period 1918–2018, both news-based (*War*) and event-based (*CWar*) war indexes predict next-month market returns, both significant at the 5% level. However, over the two subsamples 1950–2018 and 2000–2018, *War* dominates the event count indexes.[34]

We conclude that investors' perception of rare disaster risks as extracted from news media is an important predictor of stock market return, even after controlling for event-count variables.

5.6 Predicting international stock returns

Although *NYT* is an American newspaper, it covers global conflicts that have implications for stock returns in other countries. This raises the question of whether *NYT War* predicts the equity returns of other countries.

To investigate this, we use data from Global Financial Data. We collect equity index data for the United Kingdom, which dates back to 1871, as well

[34] In Internet Appendix C, we explore a larger set of real crisis events obtained from Global Financial Data. We create dummy variables to capture the occurrences of the following events: recessions, bank failures, wars, natural disasters, epidemics, and any of them. As reported in Table C.7, *War* retains its predictive power after controlling for these events.

Table 6
Continued

B. Crisis event count indexes

		1918-2018		
War	3.09**			2.66*
	(2.08)			(1.75)
Crisis		1.60		−0.36
		(0.99)		(−0.20)
CWar			3.75**	3.52*
			(1.99)	(1.69)
R^2 (%)	0.15	−0.02	0.26	0.27
		1950-2018		
War	4.12**			4.28***
	(2.51)			(2.58)
Crisis		1.01		0.11
		(0.66)		(0.07)
CWar			1.58	1.93
			(0.97)	(1.13)
R^2 (%)	0.57	−0.08	−0.02	0.48
		2000-2018		
War	10.23***			10.05***
	(3.30)			(2.97)
Crisis		−2.59		−2.00
		(−0.92)		(−0.64)
CWar			−3.42	−0.37
			(−1.17)	(−0.11)
R^2 (%)	3.64	−0.18	0.01	2.96

This table presents the results of the following bivariate predictive regressions:

$$R^e_{t+1} = \alpha + \beta War_t + \gamma z_t + \epsilon_{t+1},$$

where R^e_{t+1} is the excess market return over the next month and x_t is *War*. In panel A, z_t is either $NVIX^2$ from Manela and Moreira (2017), or geopolitical risk (GPR) from Caldara and Iacoviello (2022); in panel B, z_t is either Crisis (monthly count of real-world crisis events), or CWar (monthly count of real-world war events) from Berkman, Jacobsen, and Lee (2011). Returns are expressed as annualized percentages, and the independent variables are standardized to zero mean and unit variance. t-statistics are computed with Newey and West (1987) standard errors. The whole sample in Panel A is from January 1900 to March 2016 and in Panel B is from January 1918 to December 2018. $^*p < .1$; $^{**}p < .05$; $^{***}p < .01$.

as the MSCI World Index from 1969, the Dow Jones World Index (excluding the United States) from 1992, and the FTA World Index (excluding the United States) from 1919.[35]

Table 7 reports the results. The most significant results are observed for the United Kingdom from 1871 to 2019, although the effects are largely driven by the last two decades. The result from the MSCI World Index is also significant at the 1% level from 2000 to 2019; it is insignificant during 1969 to 2019. Similarly, *War* also positively predicts the returns on Dow Jones and FTA World Indexes over the past two decades.

Overall, our findings indicate that for global equities there is a strong war return premium, especially over the past 20 years.

[35] MSCI is long-term monthly historical data that covers the same countries and can be found in the GFD World Price and Return Indices. The FTA or the Financial Times/Standard and Poor's World Dollar Index is calculated jointly by the *Financial Times* and Standard and Poor's.

Table 7
Predicting international stock returns

	β	t-stat	R^2 (%)
A. UK FTSE Index			
1871-2019	2.78***	(2.97)	0.29
1950-2019	3.74*	(1.96)	0.26
2000-2019	7.16***	(2.73)	2.02
B. MSCI World Index			
1969-2019	2.00	(0.92)	–0.01
2000-2019	8.62***	(2.86)	2.37
C. Dow Jones World Index (excluding US)			
1992-2019	4.57	(1.57)	0.37
2000-2019	7.15**	(2.05)	1.14
D. FTA World Index (excluding US)			
1919-2019	0.22	(0.16)	–0.08
1950-2019	1.98	(1.17)	0.02
2000-2019	6.51*	(1.76)	0.71

This table presents the results of the following predictive regression:

$$R^e_{t+1} = \alpha + \beta x_t + \epsilon_{t+1},$$

where R^e_{t+1} is the excess market return over the next month for a given stock index, x_t is *War*, and β, the coefficient of interest, measures the strength of predictability. Returns are expressed as annualized percentages, and the independent variable is standardized to zero mean and unit variance. Adjusted R^2 is expressed as a percentage, and t-statistics are computed with Newey and West (1987) standard errors. $^*p < .1$; $^{**}p < .05$; $^{***}p < .01$.

5.7 Robustness checks: sLDA

We perform a battery of robustness checks and present the results in Internet Appendix D and E. The number and type of topics and seed words are key inputs to our technique. To examine robustness of these inputs, we first examine the strategy of using a very large number of topics. In such a case, the weights of seeded topics can be approximated by the frequency of the seed words in the corpus. Hence, we investigate this case by constructing topic weights as the counts of seed words scaled by the article length and present the results in Internet Appendix E. Frequency-based topic weights still yield results consistent with the sLDA ones, but their out-of-sample performance is weaker. We also consider variations in the choices of the numbers of topics and seed words, and find that the results are robust. See Internet Appendix D.1.

5.8 Robustness checks: Empirical design

As mentioned earlier, a possible caveat to the conclusion that *War* is an important return predictor is that our tests include 14 different topics, any one of which could have turned out to be the best. Furthermore, our approach involves different possible specification choices in for the seed words and the number of topics in the sLDA model, which then influences the independent variables in the resultant predictive regressions.

Gentzkow, Kelly, and Taddy (2019) discuss the potential arbitrariness of the number of topics selected for study. In principle, a possible solution

is to optimize the number of topics. However, doing so using data for the entire sample period would introduce look-ahead bias, which we seek to avoid. Additionally, extracting the number of topics each month is not suitable because the topic weights would vary based on the varying number of topics for a given month, so the topic weights would vary for reasons other than shifts in market attention to a given topic.

To address these multiple hypothesis testing concerns, we conduct a placebo/bootstrapping test using seed words that are not anticipated to predict returns. In our baseline model, we use five seed words for *War*; in our bootstrap, we evaluate the likelihood of randomly discovering any set of five seed words that predicts returns as effectively as *War*.

We first create a list of 1,000 words that have no relation to stock returns. We manually check them to ensure they are unrelated to economic/financial disasters. Then, during each iteration, we randomly draw five words from this list and store the return prediction statistics using the frequency of these random words in the *NYT* data. We repeat this step 10,000 times and create the empirical distribution of these prediction statistics. For this bootstrap, we only count the frequency of the random words and compare the prediction results to those using the count of our *War* seed words as reported in Internet Appendix E. We choose the simple count method because the simple frequency count of *War* seed words is a good approximation of our sLDA weight and it is computationally infeasible to run the bootstrap with sLDA.[36]

Our findings for the bootstrap are reported in Table D.6. Accordingly, the probability of matching the prediction results of *War* seed words is 0.53% over the 1871–2019 sample and 0.06% over the 1950–2019 sample. Details of our bootstrap procedure are reported in Internet Appendix D.3.

A more conservative method to account for MHT is to use the Bonferroni correction. If there are m individual hypotheses and the desired significance level is α, then each hypothesis should be tested at the α/m level. Our baseline model reported in Table 3 has 14 individual hypotheses (i.e., 14 topics). If the designed α is 0.01, we should test each hypothesis at the 0.00071 (= 0.01/14) level. For *War*, the t-statistic is 3.35 over 1871–2019 (Table 3), which corresponds to a one-sided p-value of .0004 and a two-sided p-value of .0008. Hence, after the Bonferroni correction, *War* is still significant at the 1% level in the one-sided test (which is reasonable because, in theory, we expect *War* to be a positive predictor) and marginally significant at the 1% level in the two-sided test.

As mentioned in the previous section, as an extreme robustness check, we estimate the sLDA model having only one seeded topic for *War* with one seed word "war" and 50 unseeded topics. We find the prediction results for

[36] One run of sLDA over 160 years of *NYT* data fully parallelized on 80 computational nodes requires at least 1 day.

War remain robust (see Table D.5). Our robustness checks confirm that the predictive performance of *War* is not an artifact of our estimation specification.

5.9 Robustness checks: *War* innovation

We also consider robustness with respect to autocorrelation. This is especially relevant for *War*, which has an autocorrelation of 0.85. To address this, we also perform tests using innovation as a predictor. We measure innovation as residual from an AR(1) process (using ARMA(1,1) or AR(2) yields similar results). The results are consistent for both in-sample and out-of-sample. Over the whole sample, the in- and out-of-sample predictability of *War*'s innovation is comparable to that of *War*'s level. Over the past 20 years, the predictive power of *War*'s innovation is weaker than *War*. However, it is still significant with t-statistic of 1.90. As for the PLS index, the innovation of the PLS is stronger economically and statistically than *War* in-sample in all periods. However, as with *War*, the PLS index is weaker out-of-sample during the past 20 years. See Table D.7.

6. Out-of-Sample Analysis

In addition to your in-sample analysis, please conduct an out-of-sample analysis with Wikipedia data (be careful with Google Trends data for an OOS analysis.

The predictability results in Section 5 are obtained by pooling within the 150-year sample. Such tests are subject to look-ahead bias, as with past studies that perform in-sample predictability tests or that use in-sample information to construct return predictors. To address this concern, we now perform an out-of-sample analysis, as is required to offer real-time economic value to investors (Goyal and Welch 2008). We conduct two standard out-of-sample tests to investigate whether discourse topics can help investors make better investment decisions: out-of-sample R^2 and certainty equivalent return (CER) gains.

Following Campbell and Thompson (2008), we compute the following well-known out-of-sample R^2 statistic:

$$R^2_{OS} = 1 - \frac{\sum_{t=p}^{T-1}\left(R^e_{t+1} - \hat{R}^e_{t+1}\right)^2}{\sum_{t=p}^{T-1}\left(R^e_{t+1} - \bar{R}^e_{t+1}\right)^2}, \tag{5}$$

where R^e_{t+1} is the realized excess market return, $\hat{R}^e_{t+1} = \hat{f}_t(x_t)$ is the predicted excess return with $\hat{f}_t(x_t)$ being a function of the predictors recursively estimated using only the training window, $\bar{R}^e_{t+1}$ is the historical mean excess return computed over the training window, and p is the size of the initial training window. We employ an expanding estimation window to incorporate all available information into formulating future forecasts and begin the evaluation period in January 1881 (10 years from the sample's start).

We benchmark the out-of-sample results of the 14 topics against the 6 predictors from Goyal and Welch (2008), including dividend-price ratio,

dividend yield, earnings-price ratio, dividend payout ratio, stock variance, and Treasury-bill rate, all of which are available from 1871.

We use two approaches to recursively estimate the function $f_t(x_t)$. First, we specify $f_t(x_t)$ as a linear function of the 14 topics and 6 economic predictors. Second, we specify $f_t(x_t)$ as a function of all 14 topics or all 6 economic predictors estimated via PLS as described in Section 4. Also, recall that our topic weights are extracted monthly using data over the past 10 years, so there is no look-ahead bias in the out-of-sample analysis.

When a predictor outperforms the historical mean benchmark in forecasting future returns, it produces a lower mean squared forecast error (MSFE) than the historical mean. Thus, the R^2_{OS} will be greater than zero. To test the significance of R^2_{OS}, we report the Clark and West (2007) MSFE-adjusted statistic.

Panel A of Table 8 reports the results from ordinary least squares (OLS) regressions using individual predictors. Among the six economic predictors, only the Treasury bill produces a positive and significant R^2_{OS} over the whole evaluation period, yet the magnitude is tiny at 0.07%. Meanwhile, among the 14 topics, during 1881–2019, *War*, *Pandemic*, and *Real Estate Boom* yield a significant R^2_{OS} (0.17%, 0.08%, and 0.19%, respectively). Except for *Pandemic*, *War* and *Real Estate Boom* continue to deliver out-of-sample (henceforth, OOS) predictive power over the past 20 years with magnitudes much larger than the whole-sample results, at 1.35% and 1.13%, respectively. Consistent with the in-sample results in Section 5, *War* displays strong out-of-sample predictive power in recent periods.

Panel B of Table 8 combines the signals of individual predictors via PLS. Combining all six economic predictors via PLS produces negative R^2's across all sample periods in the top row ("Economic"). In the second row ("All Topics"), combining all 14 topics via PLS yields a negative R^2_{OS} over the whole sample. However, in the two most recent subsamples, the topic PLS method delivers strong predictive power, producing R^2_{OS}'s of 0.95% over 1950–2019 and 2.23% over 2000–2019, both significant at the 1% level. The predictive power of the topic PLS provides an economically substantial superior performance compared to the R^2_{OS} of 1.7% from Gómez-Cram (2022) using macroeconomic indicators over the last 20 years. In the last row of panel B ("Shiller topics"), we use only the 12 topics from Shiller (2019) in the PLS estimation, which yields negative R^2_{OS}'s in all samples.

Panel B of Figure 5 plots the cumulative out-of-sample R^2 for *War* and the PLS method using all 14 topics. An upward trend indicates good performance during that period. Consistent with Table 8, *War* and the PLS method do not perform well during the first half of the sample, especially from 1910 to 1930, in which both display a steep downward slope in the cumulative R^2_{OS}. From 1930 to 1990, both *War* and the PLS method feature steadily upward trends, with the PLS method having a much steeper slope. However, both encounter a decline during the 1990s before having a turnaround during the last two decades of the sample.

Create a table similar to Table 8 with the OOS-R^2 for your predictors (only simple univariate predictive regression models). You can compare your R^2 with the R^2 of DP, DY, EP, DE, SVAR and TBL for the same time period.

Table 8
Out-of-sample R^2

	1881-2019	1881-1949	1950-2019	2000-2019
A. OLS				
Dividend-price ratio (DP)	−0.60	−0.81	−0.25	0.05
Dividend yield (DY)	−0.48	−0.39	−0.64	0.04
Earnings-price ratio (EP)	−0.14	−0.07	−0.26	−0.35
Dividend payout ratio (DE)	−0.83	−1.12	−0.33	−1.06
Stock variance (SVAR)	−1.68	−2.18	−0.79	−0.86
Treasury-bill rate (TBL)	0.07**	−0.05	0.26**	0.45
War	0.17***	−0.10**	0.65***	1.35***
Pandemic	0.08*	0.27**	−0.23	0.18
Panic	0.04	0.11	−0.06	0.28
Confidence	−0.09	−0.11	−0.07	−0.09
Saving	−0.03	−0.05	−0.00	0.02
Consumption	−0.10	−0.14	−0.03	−0.01
Money	0.01	−0.18	0.33*	−0.19
Tech	−0.45	−0.69	−0.01	0.12
Real estate boom	0.19**	0.21*	0.14*	1.13**
Real estate crash	−0.11	−0.15	−0.03	−0.15
Stock bubble	−0.12	−0.05	−0.23	−0.27
Stock crash	−0.10	−0.03	−0.23	−0.07
Boycott	−0.03	0.02	−0.11	−1.24
Wage	−0.16	−0.27	0.03	0.32**
B. PLS				
Economic	−0.84	−1.11	−0.38	−0.55
All topics	−0.08***	−0.67	0.95***	2.23***
Shiller topics	−0.88	−1.03	−0.62	−0.17

This table reports the out-of-sample R^2 (R^2_{OS}) statistic (Campbell and Thompson 2008) in predicting the monthly excess market return using economic predictors or discourse topics. Panels A reports the out-of-sample R^2 for each predictor using the simple univariate OLS regression. Panel B reports the out-of-sample R^2 by applying partial least square (PLS): "Economic" means using all economic variables including DP, DY, EP, DE, SVAR, and TBL; "All topics" means using all 14 discourse topics; and "Shiller topics" means using 12 topics from Shiller (2019) without *War* and *Pandemic*. All out-of-sample forecasts are estimated recursively using the data available in the expanding estimation window. All numbers are expressed as percentages. The evaluation period begins in January 1881, and the whole sample is from January 1871 to October 2019. *$p < .1$; **$p < .05$; ***$p < .01$ (based on the Clark and West (2007) MSFE-adjusted statistic).

Overall, we find that discourse topics outperform standard return predictors in out-of-sample prediction, especially during recent decades. These findings corroborate the in-sample results of earlier sections that the predictive power of discourse topics is stronger in recent periods.

We also examine the economic value of news topics from an asset allocation perspective. We compute the certainty equivalent return (CER) gain and Sharpe ratio for a mean-variance investor who optimally allocates her portfolio between the stock market and the risk-free asset using out-of-sample return forecasts. Consistent with the R^2_{OS} results, we find that discourse topics, especially *War*, offer economic gains to real-time investors. We present the results in Internet Appendix C.8.

7. Mechanisms

We now consider possible explanations for the predictive power of *War* and the PLS index.

7.1 Rational and behavioral channels

Next, we perform tests to determine whether our results are explained by behavioral mispricing or rational pricing of disaster risk. First, we examine simple *contemporaneous* correlations between *War* or the PLS index and proxies for risk or sentiment (see Table C.3). If *War* and the PLS index are proxies for risk, we expect them to be positively correlated with ex ante measures of market volatility and/or negatively correlated with skewness (negative skewness proxying for rare disaster risk). Alternatively, if *War* and the PLS index are proxies for behavioral sources of undervaluation (or lower overvaluation), they may be contemporaneously correlated with behavioral proxies for misvaluation such as sentiment, disagreement, and trading volume.

The most basic risk-based hypothesis is that *War* and the PLS index are capturing a risk premium for volatility. Contrary to this hypothesis, we find that *War* and the PLS index are *negatively* contemporaneously correlated with ex ante volatility indices, such as VIX and NVIX.

According to the sentiment hypothesis, high *War* will be contemporaneously associated with low sentiment, and therefore with high future returns. According to the disagreement model of Miller (1977), greater disagreement, together with constraints on short-selling, is a source of overvaluation. An implication of this is that, all else equal, when disagreement is lower, future returns will on average be higher. Gervais, Kaniel, and Mingelgrin (2001) suggest that owing to shifts in investor attention, trading volume predicts future returns (an effect that they document for individual stocks).

We find that the correlations between *War* or the PLS index and sentiment or disagreement are negative, as predicted, but generally modest and not always significant. For sentiment, there is a significant correlation of *War* with the managerial sentiment index of Jiang et al. (2019), which is based on corporate filings and earnings conference call transcripts, although this is limited to the 14-year sample period (January 2003 to December 2014) of their study. The correlation with news sentiment is negative and significant at the 10% level. We find modest negative correlations between *War* or the PLS index with investor disagreement (-9% and -8%, respectively), both of which are significant at the 5% level. These correlations are consistent with the sentiment and disagreement versions of the behavioral explanation for the *War* return premium. Additionally, we find no significant correlation between *War* or the PLS index and trading volume.

Second, to further test for risk premium or misvaluation effects, we estimate the correlations between *War* or the PLS index and ex post realizations of return volatility and skewness. A higher probability of war, as a catastrophic event, should increase return volatility and decrease skewness (i.e., increase the probability of extreme negative events), and should, according to rational asset pricing theory, cause investors to require a higher compensation for risk. Details for how we compute monthly volatility and skewness are provided in Internet Appendix C.2.

Table 9
Predicting volatility and skewness

A. Realized volatility (1927-2019)

	β	t-stat	R^2 (%)
War	−0.37**	(−2.23)	60.24
Pandemic	0.06	(0.32)	60.13
Panic	−0.26*	(−1.69)	60.19
Confidence	0.02	(0.11)	60.13
Saving	−0.09	(−0.48)	60.13
Consumption	0.01	(0.05)	60.13
Money	0.36*	(1.87)	60.23
Tech	0.03	(0.13)	60.13
Real estate boom	−0.17	(−1.05)	60.15
Real estate crash	0.11	(0.65)	60.14
Stock bubble	0.08	(0.41)	60.13
Stock crash	0.38*	(1.72)	60.25
Boycott	−0.30	(−1.63)	60.20
Wage	0.28	(1.21)	60.19
PLS	−0.16	(−1.07)	60.15

B. Implied volatility (1990-2019)

	β	t-stat	R^2 (%)
War	−0.35***	(−2.60)	81.55
Pandemic	0.18	(0.98)	81.39
Panic	−0.13	(−1.15)	81.37
Confidence	−0.13	(−1.05)	81.37
Saving	0.04	(0.28)	81.34
Consumption	0.20	(1.40)	81.41
Money	0.13	(0.74)	81.36
Tech	0.71**	(2.55)	82.13
Real estate boom	0.05	(0.37)	81.34
Real estate crash	0.00	(0.02)	81.33
Stock bubble	0.17	(1.01)	81.39
Stock crash	0.30	(1.60)	81.49
Boycott	−0.49**	(−2.42)	81.72
Wage	−0.45***	(−2.71)	81.69
PLS	−0.23*	(−1.85)	81.43

(continued)

We find that *War* and the PLS index are *negatively* correlated with future implied and realized return volatility (see Table 9). As *War* increases by one standard deviation, the subsequent realized volatility and subsequent implied volatility decrease by 0.37% and 0.35% in the next month, respectively. This effect is economically marginal relative to the mean realized and implied volatility of 9.65% and 19.19% over the sample period, respectively. The negative correlation of *War* with subsequent return volatility provides further evidence against the hypothesis that the *War* return premium is a premium for volatility.

Descriptively, compared to the other 13 discourse measures, *War* is the most economically and statistically significant for predicting subsequent realized volatility. For subsequent implied volatility, *War* is significant at the 1% level and economically is outranked only by *Tech*, *Boycott*, and *Wage*. The correlation of PLS with future volatility is more modest.

Table 9
Continued

C. Negative skewness (1927-2019)

	β	t-stat	R^2 (%)
War	−3.00*	(−1.78)	0.38
Pandemic	0.75	(0.69)	−0.06
Panic	−3.67**	(−2.54)	0.61
Confidence	−1.15	(−0.86)	−0.02
Saving	2.45	(1.47)	0.22
Consumption	−0.02	(−0.02)	−0.09
Money	2.70*	(1.89)	0.29
Tech	−0.74	(−0.46)	−0.06
Real estate boom	−0.18	(−0.13)	−0.09
Real estate crash	−2.73*	(−1.95)	0.30
Stock bubble	2.15	(1.49)	0.15
Stock crash	2.95**	(2.22)	0.36
Boycott	2.80**	(2.07)	0.32
Wage	−0.38	(−0.29)	−0.08
PLS	−4.46***	(−3.05)	0.95

This table presents results of the following predictive regression:

$$y_{t+1} = \alpha + \beta x_t + \delta' W_t + \epsilon_{t+1},$$

where y_{t+1} is either realized market volatility (panel A), implied volatility (panel B), or negative skewness (panel C) over the next month, x_t is one of the 14 topics, and W_t is a set of controls. Realized and implied volatility is in annualized percentages and independent variables are standardized to unit variance and zero mean. Realized volatility is the square root of the sum of squared daily market returns, rescaled to annual values. When σ_{t+1} is realized volatility, W_t includes two lags of realized volatility and two lags of negative market returns. When σ_{t+1} is implied volatility (VIX), W_t includes two lags of VIX, two lags of realized volatility, and two negative market returns. t-statistics are computed with Newey and West (1987) standard errors with 6 lags and R^2 is in percentages. The sample period for realized volatility and negative skewness is 1927-2019 and for implied volatility is 1990-2019. $^*p < .1$; $^{**}p < .05$; $^{***}p < .01$.

War and the PLS index are also negatively correlated with realized skewness. A one-standard-deviation increase in *War* is associated with a decrease in skewness of 3%, which is economically substantial relative to the mean skewness of –27.57% during the sample period. Compared to the other discourse topics, although the statistical significance of the negative prediction of skewness by *War* is only at the 10% level, its economic effect is substantial, following only *Panic*. For the PLS index, a one-standard-deviation increase is associated with a decrease in skewness of 4.46%. The negative correlation between *War* and subsequent skewness is consistent with the hypothesis that the *War* return premium is a premium for disaster risk, or that the market overweights disaster risk. However, *War* and the PLS index have insignificant predictive power for most of the financial crisis-related variables used by Greenwood et al. (2022).[37]

Third, mispricing might be high during high sentiment periods owing to limits to arbitrage and short-sale constraints on overpriced stocks (Stambaugh, Yu, and Yuan 2012). Under the short-sale constraints / overpricing version of

[37] We perform a regression to predict their crash dummies, including the onset of the financial crisis, bank equity crash indicator, bank failure indicator, as well as a panic indicator from Baron, Verner, and Xiong (2021), on *War*. We also include the onset of financial crisis according to Jordà, Schularick, and Taylor (2017), and the onset of financial crisis according to Reinhart and Rogoff (2011).

the behavioral explanation, the predictive power of *War* and the PLS index will be more pronounced during high sentiment periods. To test this, we partition in- and out-of-sample prediction R^2 into high and low sentiment periods. We report the results of these tests for our discourse topics in Table C.10 in Internet Appendix C.7. We find that *War* and the PLS index are stronger predictors of market returns during low sentiment periods, which does not support the short-sale constraint / overpricing hypothesis.

Overall, the evidence supports some versions, but not all, of the risk and behavioral theories for the war market return premium. Consistent with behavioral explanations, *War* is negatively correlated with indicators of overpricing, though the magnitudes are modest. There is no evidence that this premium is a compensation for volatility risk. The predictive power of *War* is stronger during low sentiment periods, inconsistent with the short-sale constraint / overpricing version of the behavioral hypothesis. The evidence that *War* predicts negative skewness is consistent either with a risk premium for rare disaster risk or market overweighting of such risk.

Finally, we also examine the predictive power of *War* on bond markets and report the results in Internet Appendix C.9. We find that *War* is a negative predictor of the returns on relatively safe fixed income assets (short-term government bonds and investment grade corporate bonds) and a positive predictor of the returns of relatively risky fixed income assets (long-term high-yield corporate bonds). This is consistent with the theoretical predictions of Gabaix (2012). More generally, this is potentially consistent with *War*, and the more left-tilted skewness associated with high *War*, triggering a flight to quality.[38] If behavioral investors were to overreact to *War* (overestimating the danger), this will yield identical implications about flight to quality. Overall, our bond results are consistent with flight to quality, either for rational or behavioral reasons.

7.2 War innovations and contemporaneous returns

If innovations in *War* and the PLS index are associated with increases in risk premiums, other things equal we expect to see a negative contemporaneous correlation of innovations with returns. This is a discount rate channel. A cash flow channel would further contribute toward negative contemporaneous correlation if innovations in *War* and the PLS index are associated with bad news about future cash flows (and especially if there is overreaction to this bad news).

However, there is also a possible force in the opposite direction. Other things equal, we expect a positive contemporaneous correlation if innovations in *War* and the PLS index are associated with good news about future cash flows.

[38] Recall that that higher *War* is associated with lower volatility, which would not induce a flight to quality, but with lower skewness, which could induce a flight to quality.

Table 10
Contemporaneous relation between *War* innovation and stock returns

	1871-2019	1871-1949	1950-2019	2000-2019
A. Contemporaneous correlation				
First difference				
Correlation	–3.41	–3.93	–2.64	–7.09
t-stat	–1.44	–1.21	–0.76	–1.09
AR(1) residual				
Correlation	–1.58	–2.25	–1.13	–3.18
t-stat	–0.67	–0.69	–0.33	–0.49
B. Contemporaneous regression				
First difference				
β	–1.95	–2.47	–1.31	–3.59
t-stat	–1.61	–1.31	–0.94	–1.16
AR(1) residual				
β	–0.90	–1.42	–0.56	–1.61
1	–0.72	–0.71	–0.40	–0.52

This table reports the contemporaneous relation between *War* and excess market returns. Panel A reports the contemporaneous correlations and panel B reports the contemporaneous regression of *War* innovation on excess returns. *War* innovation is measured as either first difference of *War* or residual from an AR(1) process. Returns are expressed as annualized percentages, and the independent variable is standardized to zero mean and unit variance. *t*-statistics are computed with Newey and West (1987) standard errors. The sample period is from January 1871 to October 2019. $*p < .1$; $**p < .05$; $***p < .01$.

To address this issue empirically, we perform tests using innovations in *War*, defined as either monthly changes in *War* or as innovations from an AR(1) process for *War*.[39] The results of contemporaneous regression of returns on innovations in *War* are provided in Table 10. The coefficients are negative but statistically insignificant. The closest to being significant is the regression with first difference over the whole sample ($t = -1.61$), which is marginally insignificant at the 10% level.

However, the point estimate is economically substantial. It implies that a one-standard-deviation greater *War* innovation corresponds to a 1.95% lower annualized returns.[40] So we cannot conclude that the true relationship is weak. The test has inadequate power to assess whether there is a substantial relationship.

7.3 *War* and cash flow versus discount rate news

We have just argued that on conceptual grounds the contemporaneous correlation could be either positive or negative (owing to possible opposing effects of discount rate and cash flow news), which can resolve the apparent puzzle. Furthermore, our evidence is not inconsistent with a substantial negative correlation; in a simple contemporaneous regression, the data simply

[39] Because of limited space, we do not report the result on the PLS index here. It is qualitatively similar to the result of *War* and it is available upon request.

[40] The average annualized excess stock market return over the analyzed period is 6.44%. This effect is about 30% as large.

Table 11
Predicting expected and unexpected returns, and cash flow and discount rate news

	r_{t+1}	$E_t r_{t+1}$	CF_{t+1}	DR_{t+1}
R, DP	3.14***	0.25*	3.33***	0.43**
	(2.80)	(1.85)	(2.66)	(2.32)
R, DP, EP	3.14***	0.74***	2.98**	0.58***
	(2.80)	(5.07)	(2.37)	(2.65)
R, DP, SVAR	3.14***	0.24*	3.33***	0.42**
	(2.80)	(1.78)	(2.66)	(2.25)
R, DP, TBL	3.14***	0.24*	3.32***	0.43**
	(2.80)	(1.81)	(2.66)	(2.28)
R, DP, EP, SVAR, TBL	3.14***	0.84***	2.87**	0.57**
	(2.80)	(5.63)	(2.24)	(2.36)

This table reports results from the following univariate predictive regressions:

$$r_{t+1} = \alpha + \beta Var_t + \epsilon_{t+1},$$

$$\mathbb{E}_t r_{t+1} = \alpha^E + \beta^E Var_t + \epsilon^E_{t+1},$$

$$CF_{t+1} = \alpha^{CF} + \beta^{CF} Var_t + \epsilon^{CF}_{t+1}, \quad \text{and}$$

$$DR_{t+1} = \alpha^{DR} + \beta^{DR} Var_t + \epsilon^{DR}_{t+1},$$

where r_{t+1} is log market return, $\mathbb{E}_t r_{t+1}$ is expected return, CF_{t+1} is cash flow news, DR_{t+1} is discount rate news, and $r_{t+1} = \mathbb{E}_t r_{t+1} + CF_{t+1} + DR_{t+1}$. $\mathbb{E}_t r_{t+1}$, CF_{t+1}, and DR_{t+1} are estimated via a VAR(1) model consisting of log return, log dividend-price (DP), log earnings-price (EP), Treasury bill (TBL), and stock variance (SVAR). Each row reports results for a different combination of the return predictors. Returns are expressed as annualized percentages, and the independent variable is standardized to zero mean and unit variance. t-statistics are computed with Newey and West (1987) standard errors. The sample period is from January 1871 to October 2019. *$p < .1$; **$p < .05$; ***$p < .01$.

do not have enough power to draw a strong conclusion about whether or not that is the case.

To further address this issue, we further examine the discount rate versus cash flow channels by employing the return decomposition framework of Campbell (1991). This models realized returns as the sum of three components: expected returns, news about future discount rates, and news about future cash flows.

We estimate this decomposition using a VAR model and then examine whether *War* and the PLS index have predictive power for each component. As shown in Table 11 and Table C.15, *War* and the PLS index positively predict all three components. So they are associated with both cash flow news and discount rate news. Specifically, *War* and the PLS index can potentially predict expected returns or unexpected returns which consist of cash flow news and discount rate news.

We follow Rapach, Ringgenberg, and Zhou (2016) in applying the Campbell (1991) return decomposition. Accordingly, we can decompose realized log market return into three components as follows:

$$r_{t+1} = \underbrace{\mathbb{E}_t r_{t+1}}_{\text{expected return}} + \underbrace{CF_{t+1} - DR_{t+1}}_{\text{unexpected returns}}, \tag{6}$$

where CF_{t+1} and DR_{t+1} are the cash flow and discount rate news or revisions, constituting the unexpected component of realized log return. To estimate the three components of realized log return, we use a VAR(1) model:

$$Y_{t+1} = AY_t + U_{t+1}, \tag{7}$$

where Y_{t+1} is an n-vector of consisting of log return r_{t+1}, log dividend-price $d_t - p_t$, and other return predictors, A is an $n \times n$ matrix of VAR slope coefficients, and U_{t+1} is an n-vector of zero-mean innovations.

Let e_1 be a n-vector with one as its first element and zeros for the remaining elements, the three components of return are given by

$$\begin{aligned}
&\mathbb{E}_t r_{t+1} = e_1' AY_t, \\
&DR_{t+1} = (\mathbb{E}_{t+1} - \mathbb{E}_t) \sum_{j=1}^{\infty} \rho^j r_{t+1+j} = e_1' \rho A (I - \rho A)^{-1} U_{t+1}, \quad \text{and} \\
&CF_{t+1} = (\mathbb{E}_{t+1} - \mathbb{E}_t) \sum_{j=0}^{\infty} \rho^j \Delta d_{t+1+j} = r_{t+1} - \mathbb{E}_t r_{t+1} + DR_{t+1},
\end{aligned} \tag{8}$$

where ρ is the log-linearization normalizing constant equal 0.96. As it is clear from equation (8), we use the VAR model to directly estimate the expected return component and DR news and extract the CF news as the residual.

To estimate the VAR slope parameter A, we include in the Y_{t+1} vector log return, log dividend-price ratio (DP), and other return predictors available over 1871–2019, including log earnings-price ratio (EP), Treasury bill (TBL), and stock variance (SVAR). We estimate different VAR specifications using different predictors but always include log dividend-price ratio because this predictor is important to properly estimate the CF and DR components (Engsted, Pedersen, and Tanggaard 2012).

After estimating the three components of return, to see which component *War* can predict, we regress each component onto *War* in univariate predictive regressions:

$$\begin{aligned}
r_{t+1} &= \alpha + \beta War_t + \epsilon_{t+1}, \\
\mathbb{E}_t r_{t+1} &= \alpha^E + \beta^E War_t + \epsilon_{t+1}^E, \\
CF_{t+1} &= \alpha^{CF} + \beta^{CF} War_t + \epsilon_{t+1}^{CF}, \quad \text{and} \\
DR_{t+1} &= \alpha^{DR} + \beta^{DR} War_t + \epsilon_{t+1}^{DR}.
\end{aligned} \tag{9}$$

By the properties of OLS, $\beta = \beta^E + \beta^{CF} - \beta^{DR}$. We report these β estimates in Table 11. We find that across all specifications of the VAR model, *War* predicts both expected and unexpected returns. *War* yields the strongest predictive power for expected returns when we include all return predictors as reported in the last row of Table 11. *War* also predicts both the CF news and DR

news components of unexpected returns, significant at at least the 5% level across all specifications. The result for the PLS index is similar and reported in Table C.15.

Among the three components of returns, *War* has the largest comovement with cash flow news. So the ability of War to anticipate cash flow news is the most economically important source of *War*'s predictive power for stock returns.[41] The association of War with high future cash flows suggests that, other things equal, increases in War are good news for future cash flows. In other words, the cash flow effect is positive.

8. Conclusion

We test the hypothesis that rare disaster risk is priced (or mispriced) by extracting market attention to rare disasters from news media. This helps overcome the challenge of scant data on realized disasters. It also has the advantage (from a behavioral perspective) of focusing on investor attention to and perceptions of disaster, which may differ from objective risks. We provide the most comprehensive analysis for empirically testing for pricing effects of disaster risk. In addition to two topics covering rare disaster risks (*War* and *Pandemic*), we also examine 12 non-disaster-focused narratives from Shiller (2019).

We employ an advanced natural language processing tool called sLDA to extract discourse topics from nearly 7 million *New York Times* articles over the past 160 years. We create a list of topic-based seed words to input into the sLDA model to guide the topic extraction process. We employ a rolling estimation scheme to include only historical news data at every estimation time; thus, our measure avoids look-ahead bias and addresses changes in semantic usage over time.

Among the discourse topics considered, the most important is *War*, which encompasses various themes related to the danger of armed conflict. We find that both *War* and an index constructed from all topics (our PLS index) are strong positive predictors of the stock market return up to a horizon of up to 36 months. We find that the *war market return premium* increases through the sample period and that the predictive power of discourse topics holds at both the market and portfolio levels. The war market return premium remains even when our war proxy is extracted from a different media outlet, the *WSJ*.

The war market return premium is consistent with models of rare disaster risk, and with the behavioral hypothesis that investors overweight the prospect of rare disasters. Barro (2009) finds that the probability of rare disasters can

[41] That *War* positively predicts future cash flow is consistent with the findings in Cortes, Vossmeyer, and Weidenmier (2022) in a more specialized sample. Using hand-collected data on military expenditures, Cortes, Vossmeyer, and Weidenmier (2022) find that U.S. excessive military spending in war times have positive spill-over effects on future corporate earnings.

explain the high equity premium. During times when the probability of a rare disaster is higher, the equity premium should be higher, which is consistent with our finding that *War* is associated with higher subsequent stock returns. Alternatively, if investors overweight rare risks (either owing to overestimation of probability owing to salience or the overweighting of low probability events in the cumulative prospect theory utility function), we again expect higher war media discourse to be associated with high future returns.

Our results also confirm the prediction of Gabaix (2012) (or of a behavioral setting where agents overweight rare disasters). We find that *War* is associated with higher excess returns on mid- to long-term high-yield corporate bonds. In contrast, *War* negatively predicts excess returns on safer investment instruments such as short-term government bonds and investment-grade corporate bonds.

The approach proposed in this paper suggests several further directions. First, in a study that also uses the sLDA approach, Hirshleifer, Mai and Pukthuanthong (forthcoming) find that a factor that is based on our *War* variable explains the cross section of stock returns across a wide range of testing assets, and that leading benchmark factors as well as other media-based uncertainty measures do not subsume its explanatory power.

Second, with sufficient computing power the model could be estimated at a daily frequency, both to increase statistical power for testing the economic hypotheses about war risk and returns, and to see whether there is return predictability at shorter horizons.

Third, a good approximation to our method can be used for research and applications even when computing resources are limited by using a simple frequency count of *War* seed words. This is a reasonable approximation of our sLDA weight and produces consistent prediction results (as shown in Internet Appendix E).

Lastly, the estimation scheme can be modified to suit other countries by using local newspapers and translating the seed words into foreign languages. This would permit testing in countries that are potentially more heavily exposed to war risk, and to less developed stock markets in which mispricing is more prevalent.

Code Availability: The replication code is available in the Harvard Dataverse at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/7PPQ1I.

References

Adämmer, P., and R. A. Schüssler. 2020. Forecasting the equity premium: Mind the news! *Review of Finance* 24:1313–55.

Baker, M., and J. Wurgler. 2006. Investor sentiment and the cross-section of stock returns. *Journal of Finance* 61:1645–80.

Baker, S. R., N. Bloom, and S. J. Davis. 2016. Measuring economic policy uncertainty. *Quarterly Journal of Economics* 131:1593–36.

Baron, M., E. Verner, and W. Xiong. 2021. Banking crises without panics. *The Quarterly Journal of Economics* 136:51–113.

Barro, R. J. 2006. Rare disasters and asset markets in the twentieth century. *Quarterly Journal of Economics* 121:823–66.

———. 2009. Rare disasters, asset prices, and welfare costs. *American Economic Review* 99:243–64.

Berkman, H., B. Jacobsen, and J. B. Lee. 2011. Time-varying rare disaster risk and stock returns. *Journal of Financial Economics* 101:313–32.

Blei, D. M. 2012. Probabilistic topic models. *Communications of the ACM* 55:77–84.

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

Boudoukh, J., R. Israel, and M. Richardson. 2022. Biases in long-horizon predictive regressions. *Journal of Financial Economics* 145:937–69.

Boyd-Graber, J., Y. Hu, and D. Mimno. 2017. Applications of topic models. *Foundations and Trends in Information Retrieval* 11:143–296.

Brogaard, J., and A. Detzel. 2015. The asset-pricing implications of government economic policy uncertainty. *Management Science* 61:3–18.

Brown, N. C., R. M. Crowley, and W. B. Elliott. 2020. What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research* 58:237–91.

Bybee, L., B. Kelly, A. Manela, and D. Xiu. 2024. Business news and business cycles. *Journal of Finance* 79:3105–47.

Bybee, L., B. T. Kelly, and Y. Su. 2023. Narrative asset pricing: Interpretable systematic risk factors from news text. *Review of Financial Studies* 36:4759–87.

Caldara, D., and M. Iacoviello. 2022. Measuring geopolitical risk. *American Economic Review* 112:1194–225.

Campbell, J. Y. 1991. A variance decomposition for stock returns. *The Economic Journal* 101:157–79.

Campbell, J. Y., and S. B. Thompson. 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies* 21:1509–31.

Cherlin, A. J. 2010. *The marriage-go-round: The state of marriage and the family in America today*. New York: Vintage.

Choudhury, P., D. Wang, N. A. Carlson, and T. Khanna. 2019. Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles. *Strategic Management Journal* 40:1705–32.

Clark, T. E., and K. D. West. 2007. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics* 138:291–311.

Cochrane, J. H. 1996. A cross-sectional test of an investment-based asset pricing model. *Journal of Political Economy* 104:572–621.

Cortes, G. S., A. Vossmeyer, and M. D. Weidenmier. 2022. Stock volatility and the war puzzle. Working Paper, Working paper, University of Florida.

Dictionary, O. E. 1993. Oxford English dictionary. *Simpson, JA & Weiner, ESC.–1989* .

Dyer, T., M. Lang, and L. Stice-Lawrence. 2017. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics* 64:221–45.

Engsted, T., T. Q. Pedersen, and C. Tanggaard. 2012. Pitfalls in VAR based return decompositions: A clarification. *Journal of Banking & Finance* 36:1255–65.

Eshima, S., K. Imai, and T. Sasaki. 2024. Keyword assisted topic models. *American Journal of Political Science* 68:730–50.

Ferguson, N. 2006. Political risk and the international bond market between the 1848 Revolution and the outbreak of the First World War. *Economic History Review* 59:70–112.

Fischhoff, B., P. Slovic, and S. Lichtenstein. 1977. Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance* 3:552–.

Gabaix, X. 2012. Variable rare disasters: An exactly solved framework for ten puzzles in macro-finance. *Quarterly Journal of Economics* 127:645–700.

Garcia, D. 2013. Sentiment during recessions. *Journal of Finance* 68:1267–300.

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

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

Golez, B., and P. Koudijs. 2018. Four centuries of return predictability. *Journal of Financial Economics* 127:248–63.

Gómez-Cram, R. 2022. Late to recessions: Stocks and the business cycle. *Journal of Finance* 77:923–66.

Goyal, A., and I. Welch. 2008. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21:1455–508.

Goyal, A., I. Welch, and A. Zafirov. 2024. A comprehensive 2022 look at the empirical performance of equity premium prediction. *Review of Financial Studies* 37:3490–557.

Greenwood, R., S. G. Hanson, A. Shleifer, and J. A. Sørensen. 2022. Predictable financial crises. *Journal of Finance* 77:863–921.

Griffiths, T. L., and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101:5228–35.

Hansen, S., M. McMahon, and A. Prat. 2018. Transparency and deliberation within the FOMC: A computational linguistics approach. *Quarterly Journal of Economics* 133:801–70.

Hillert, A., and M. Ungeheuer. 2019. The value of visibility. *SSRN* .

Hirshleifer, D., D. Mai, and K. Pukthuanthong. Forthcoming. War discourse and the cross section of expected stock Returns. *Journal of Finance*.

Homer, S., and R. E. Sylla. 1996. *A history of interest rates*. Piscataway, NJ: Rutgers University Press.

Huang, D., F. Jiang, J. Tu, and G. Zhou. 2015. Investor sentiment aligned: A powerful predictor of stock returns. *Review of Financial Studies* 28:791–837.

Huang, D., J. Li, and L. Wang. 2020. Are disagreements agreeable? Evidence from information aggregation. *Journal of Financial Economics* 141:83–101.

Jagarlamudi, J., H. Daumé III, and R. Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 204–13. Association for Computational Linguistics.

Jiang, F., J. Lee, X. Martin, and G. Zhou. 2019. Manager sentiment and stock returns. *Journal of Financial Economics* 132:126–49.

Jordà, Ò., M. Schularick, and A. M. Taylor. 2017. Macrofinancial history and the new business cycle facts. *NBER Macroeconomics Annual* 31:213–63.

Julliard, C., and A. Ghosh. 2012. Can rare events explain the equity premium puzzle? *Review of Financial Studies* 25:3037–76.

Kelly, B., and S. Pruitt. 2013. Market expectations in the cross-section of present values. *Journal of Finance* 68:1721–56.

———. 2015. The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics* 186:294–316.

Larsen, V. H., and L. A. Thorsrud. 2019. The value of news for economic developments. *Journal of Econometrics* 210:203–18.

Le Bris, D. 2012. Wars, inflation and stock market returns in France, 1870–1945. *Financial History Review* 19:337–61.

Lu, B., M. Ott, C. Cardie, and B. K. Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th International Conference on Data Mining Workshops*, 81–8. Piscataway, NJ: IEEE.

Lundblad, C. 2007. The Risk Return Tradeoff in the Long Run: 1836–2003. *Journal of Financial Economics* 85:123–50.

Manela, A., and A. Moreira. 2017. News implied volatility and disaster concerns. *Journal of Financial Economics* 123:137–62.

Mcauliffe, J., and D. Blei. 2007. Supervised topic models. *Advances in Neural Information Processing Systems* 20:121–28.

Miller, E. M. 1977. Risk, uncertainty, and divergence of opinion. *Journal of Finance* 32:1151–68.

Miller, R. 2022. Clark Medal winner Oleg Itskhoki says war is a bigger economic risk than Covid. *Bloomberg*, August 2. https://www.bloomberg.com/news/articles/2022-08-02/clark-medal-winner-oleg-itskhoki-says-war-is-a-bigger-economic-risk-than-covid

Newey, W., and K. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55:703–8.

Obaid, K., and K. Pukthuanthong. 2022. A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news. *Journal of Financial Economics* 144:273–97.

Oosterlinck, K., and J. S. Landon-Lane. 2006. Hope Springs Eternal–French Bondholders and the Soviet Repudiation (1915–1919). *Review of Finance* 10:507–35.

Pástor, L., and P. Veronesi. 2013. Political uncertainty and risk premia. *Journal of Financial Economics* 110:520–45.

Ramage, D., D. Hall, R. Nallapati, and C. D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 248–56.

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

Reinhart, C. M., and K. S. Rogoff. 2011. From financial crash to debt crisis. *American Economic Review* 101:1676–706.

Rhodes, R. 2012. *The making of the atomic bomb*. New York: Simon and Schuster.

Rietz, T. A. 1988. The equity risk premium: A solution? *Journal of Monetary Economics* 22:117–31.

Rutherford, E. 2012. The scattering of α and β particles by matter and the structure of the atom. *Philosophical Magazine* 92:379–98.

Schwert, G. W. 1990. Stock market volatility. *Financial Analysts Journal* 46:23–34.

Shiller, R. J. 2017. Narrative economics. *American Economic Review* 107:967–1004.

———. 2019. *Narrative economics: How stories go viral and drive major economic events*. Princeton, NJ: Princeton University Press.

Snowberg, E., and J. Wolfers. 2010. Explaining the favorite–long shot bias: Is it risk-love or misperceptions? *Journal of Political Economy* 118:723–46.

Stambaugh, R. F., J. Yu, and Y. Yuan. 2012. The short of it: Investor sentiment and anomalies. *Journal of Financial Economics* 104:288–302.

Steyvers, M., and T. Griffiths. 2007. Probabilistic topic models. In *Handbook of Latent Semantic Analysis*, 439–60. New York: Routledge.

Tversky, A., and D. Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5:297–323.

van Binsbergen, J. H., S. Bryzgalova, M. Mukhopadhyay, and V. Sharma. 2022. (Almost) 200 years of news-based economic sentiment. *Working Paper, University of Pennsylvania.* .

Wachter, J. A. 2013. Can time-varying risk of rare disasters explain aggregate stock market volatility? *Journal of Finance* 68:987–1035.

Walker, J. S. 2004. *Three mile island: A nuclear crisis in historical perspective*. Berkeley: University of California Press.

Watanabe, K., and Y. Zhou. 2020. Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches. *Social Science Computer Review 40:346–66.* .