Haven't We Seen This Before? Return Predictions from 200 Years of News

AJ Chen Gerard Hoberg Miao Ben Zhang *

June 27, 2025

Abstract

We postulate that our historical record has become adequately long and informative that newly arriving economic states often resemble historical states. Building on this insight, we develop a framework to predict future economic outcomes using the average of the realized outcomes that follow highly similar historical states. Using 210 million newspaper articles from 1815 to 2021, we identify historically similar months for each focal month and construct a predictor of aggregate U.S. stock returns, "SeenItRet". SeenItRet strongly forecasts future market-wide stock returns up to two years ahead, with an annualized impact of 4–7% for a one standard deviation shift. Our framework is general and also predicts real economic outcomes, including recessions, inflation, and patenting activity. A virtue of our approach is its use of economic principles to reduce the high dimensionality of the underlying state space to an ex-ante measurable and intuitive unidimensional predictor. Our model performs better when historical states are more similar to the focal state, and it offers interpretable economic insights by highlighting the specific themes that drive its predictions.

Keywords: *Return predictability, economic state, textual analysis, news, history rhymes, volatility, treasury yield, recession, patent*

*AJ Chen is from the University of British Columbia Sauder School of Business and can be reached at aj.chen@sauder.ubc.ca. Gerard Hoberg and Miao Ben Zhang are from the University of Southern California Marshall School of Business and can be reached at hoberg@marshall.usc.edu and miao.zhang@marshall.usc.edu, respectively. We thank Gregor Schubert for helpful comments and Vishnu Shetty Belanje, Jayantraj Coimbatore Selvakumar, and Ronak Shah for excellent research assistance. We also thank everyone on the historical news team for providing insightful advice and institutional knowledge.

"History doesn't repeat itself, but it often rhymes." —Mark Twain

The U.S. and world economies have experienced over 200 years of detailed history summarized by very large amounts of text in public media and historical records. This history spans a wide-ranging set of economic scenarios, raising the tantalizing possibility that society has seen a significant fraction of all potentially important economic scenarios (more formally, economic "states"). The implications of such a possibility for policymakers, investors, corporate managers, and risk management experts are far-reaching.

We propose a simple prediction framework based on the idea that, although history does not repeat, "history does rhyme" as Mark Twain suggested. Thus, we can identify a group of past states highly similar to the current one. Importantly, we can predict the future evolution of economic outcomes based on the realization of the outcomes observed following the historical months with similar states, which we label a "SeenIt" prediction framework.[1]

For example, investors currently facing a complex state with elevated inflation, intensified trade and geopolitical tensions, a major innovation wave, cutbacks in fiscal spending, concerns about consumer confidence, and several other important developments can search history for the 25 historical months that experienced the most similar array of economic issues. They can then predict future stock market returns and other economic variables, such as treasury yields and recession likelihoods, by averaging what actually materialized for these variables following those 25 historical peer months. The more history rhymes, the better the predictions.

The foundation for this predictability rests on a basic theoretical idea that economic states follow a transition probability process, where future economic outcomes depend on current economic conditions. Hence, our "SeenIt" framework has the virtue of accounting for the fact that economic evolutions are likely a complex function of a high-

[1]We note that a growing body of literature uses advanced technologies such as natural language processing and machine learning to extract signals that predict stock returns (e.g., Bybee et al. (2023), Bybee et al. (2024), Kelly et al. (2022), Kelly et al. (2023), and Li et al. (2025)). Our *SeenIt* framework offers an entirely new approach compared to the literature that is guided by novel and fundamental economic principles. In Section 1, we discuss related work and detail our unique contribution to this literature.

dimensional set of existing economic conditions, such as those in our example above. Yet it is devoid of any assumptions regarding the functional form, since the materialized outcomes during historically similar months already integrate all state conditions into the prediction. Another virtue is that our predictive framework provides a general economic expectations operator by offering historically similar months to any focal month, which is easy to share. Any researcher can predict any variable of interest by taking a simple average of the variable's past outcomes over the historically similar months.

Our framework begins with a massive corpus of 210 million business-related newspaper articles from NewsArchives, dating back over 200 years. Using the full text of these articles, we approximate the space of U.S. economic states as a rich 648-element vector of loadings on economic themes interacted with positive, negative, or uncertain tone. Our use of rich text data overcomes the challenge of lacking detailed numerical data to describe the complex and high-dimensional features of economic states over the past 200 years. Next, we use vector similarities to identify 25 past states most similar to the current observed state. These similarities facilitate an economic prediction model based on simple averages of actual outcomes realized following the peer historical states, which we label as the "SeenIt" predictors.

Our empirical analysis reveals significant predictive success of the SeenIt framework across an array of economic variables. To keep our paper focused, our primary analysis is on market-wide stock returns, where we find economically large predictability of SeenItRet lasting up to two years. We view this analysis as core because the stock market is difficult to predict and is seen as a first principal component for many important economic outcomes. Yet, we also find similar successes predicting treasury yields, volatility, inflation, recessions, and patenting waves. Asymptotics are also excellent, as we believe future research can use this framework with ever-increasing power as history generates more data and spatial representations of the state space improve.

Our proposed framework addresses two key empirical challenges in constructing return predictors from an extraordinarily high-dimensional narrative space. The first challenge concerns the selection of "relevant" narratives that predict economic out-

comes. Li et al. (2025) demonstrate that human curation of predictive features is crucial for significantly predicting stock returns. We thus approximate "relevant" dimensions of the state space by extracting an exhaustive list of economic themes from textbooks and related sources, and interacting each theme with measures of tone (positive and negative) and uncertainty.

The second challenge concerns the extraordinarily high dimensionality of the narrative themes, which makes the prediction exercise nearly impossible if one considers all nonlinear interaction effects (without the SeenIt simplification). As an example, if one considers the realizations of each element in our 648-dimensional theme vector as being only {low, medium, high}, the possible predictors will reach 3^{648}, or approximately 10^{309}. In contrast, our "history rhymes" thesis offers a methodology to dramatically reduce the dimensionality of the prediction problem to just one, i.e., the SeenIt predictor. Intuitively, the stock market following the historically similar months has already selected the most "relevant" economic themes in the right combination and reflected the economic state. Therefore, we can simply use the average returns following the 25 historical months with the most similar economic states to predict returns that should follow the current month. As a result, our SeenIt framework addresses the empirical fitting challenge by generating a single ex-ante measurable variable for predicting any economic variable such as stock returns, treasury yields, and recession likelihoods.

Our framework also has its limitations. First, history may not repeat itself, as every new episode of time we experience may include some nuances relative to history, such as new developments in breakthrough technologies, global political landscape, and population growth. While this limitation is currently impossible to eliminate, it appears not severe enough to render our framework based on "history rhymes" useless. In Section 3, we show that our framework's empirical predictability is stronger in episodes when economic states are more familiar to those seen in history, consistent with the predictions of our history-rhymes thesis. Second, implementing our framework requires a long time series of the economic variable being predicted. In this study, we assemble historical data on stock returns and other economic variables via various sources dating back to 1800s. In particular, our SeenItRet for the stock market return is constructed

using the CRSP database and also the historical price-weighted monthly returns from Goetzmann et al. (2001) since 1815.[2] The limitation is that it is difficult to predict variables that dont have a long historical time series of past values available.

Our main result is that our key variable SeenItRet, ex-ante measured as the average return of the 25 most similar historical months, strongly predicts ex-post monthly market-wide stock returns. One standard deviation higher SeenItRet predicts roughly 0.5% higher returns in the next month. Yet, our most compelling finding is that SeenItRet's ability to predict returns is remarkably long-lasting and maintains a large economic magnitude even with longer lags. If one smooths SeenItRet over the past 12 months (still ex-ante measurable but reflecting that prices can update gradually),[3] SeenItRet's return predictability is more than 5% annually and lasts a full 12 months. It then decays gradually but remains statistically significant for up to 30 months after measurement. For equal- or value-weighted market returns, the magnitude is closer to 6.5% and 4.5% annually, although significant equal- or value-weighted predictability also lasts for roughly 30 months after measurement. We are not aware of other variables with comparable size and consistency in predicting market-wide returns. For example, the past market return (popular in the literature) only predicts returns for one month.

Our main results are robust across several specifications. First, we show that the return predictability is not driven by spurious time-series trends in the return data. In a placebo test, we construct a placebo SeenItRet by reassigning the 25 most similar historical months to be a year earlier. These reassigned months are no longer similar to the focal month but are similarly distant from it, as the average time lapse between historical similar months and the focal month is 542 months, or about 45 years.[4] Unlike our baseline SeenItRet, the placebo SeenItRet does not predict return at all. Second, our SeenItRet also passes the stringent out-of-sample (OOS) test proposed by Welch and Goyal (2008), which compares the predictability of our variable with the historical

[2]The now-standard value-weighting was not well understood and accepted back in the 19th century (see in-depth discussions in Section 2).

[3]Evidence of slow price updating is pervasive in the literature. For example, see Cohen et al. (2020) for general results and the lagged propagation literature for evidence across many economic links (e.g., Hou, 2007; Cohen and Frazzini, 2008; Menzly and Ozbas, 2010; Hoberg and Phillips, 2018; Lee et al., 2019, 2024).

[4]Our baseline SeenItRet was by construction based on historical months that are at least 5 years earlier than the focal months to avoid information entanglement.

mean model based on rolling OLS regressions. Our SeenItRet generates positive OOS R^2, suggesting that our model, which selects past returns based on historically similar months, outperforms the historical mean model that simply averages all past returns over a long history without selection. Third, the SeenIt framework also predicts excess returns, suggesting that the return predictability of SeenItRet is not driven by the predictions of risk-free rates.

Our framework also generates a natural prediction that the more history rhymes, the stronger the predictability of SeenItRet. We confirm this in the data by computing the average similarity of the 25 past peer states, named "SeenItFamilarity." Months featuring more novel and unfamiliar states will have a lower SeenItFamiliarity. By interacting SeenItRet with SeenItFamiliarity, we find that states that are more familiar in history, i.e., when history rhymes more, show stronger return predictability of SeenItRet. Our results overall support the intuitive premise that we have seen adequate variation in past states to make significant predictions about market returns.

Next, we shed light on economic interpretations by examining which interpretable economic themes are most responsible for SeenItRet's return predictability documented above. Specifically, we examine how much return predictability is lost when each economic theme is left out of the prediction model one at a time. We highlight the themes that are important for SeenItRet's return predictability in the short term (one month), medium term (six months), and long term (24 months) are different. News about momentum is the most important theme in facilitating the prediction of the short and medium-term returns. Themes that are most important to predict short-term returns tend to be more directional such as "reduce" or liquidity-related such as "margin" or "scarcity" which are arguably vague. As one extends the horizon to the medium term, the important themes shift to more definitive issues that entail a longer resolution, including deregulation, war, consumers, and socialism. The importance of war echoes the findings in Hirshleifer et al. (2024, 2025). The most important economic themes that facilitate long-term returns for SeenItRet include more intricate themes known to have a lasting impact on the economy. These include themes such as inflation, expansions, currency, and bonds.

We also report which economic themes are most important for SeenItRet's predictions during various decades. We find, for example, the 2000s include recessions, poverty, treasury, and stimulus; the 1940s load heavily on war, depression, scarcity, and rationing; and the 1870s feature fraud, antitrust, and bubbles. These interpretative results are intuitive given the events that occurred during these decades.

After examining the predictability of SeenItRet, we explore a second-moment measure from our SeenIt framework, SeenItRisk. We construct SeenItRisk as the standard deviation of the returns following the 25 historically similar months, unlike the average as was used to construct SeenItRet. We find that SeenItRisk is also a novel predictor of market-wide returns, as it positively predicts returns after controlling for SeenItRet or observed volatility of the current month. One interpretation is that investors demand a premium for investing in historically risky economic states.

We note that the SeenIt framework, while particularly compelling for understanding market-wide returns—which serve as a "catch-all" for economic information—can be broadly applied to predict any economic variable of interest, effectively serving as a new empirical expectations operator. By using the outcomes that followed the 25 most historically similar states, this approach generates predictors for a range of economic variables. For example, we find that SeenIt-based predictors for treasury yields, volatility, NBER recessions, inflation, and patent applications all strongly and positively forecast their respective outcomes. These results affirm the generalizability and utility of the SeenIt framework for various economic agents seeking robust empirical tools for forecasting a wide array of variables and managing emerging risks.

The paper proceeds as follows: Section 1 discusses the related literature and highlights our contribution to the literature. Section 2 describes our data and measurement system for identifying recurring economic states and constructing SeenIt predictors. Section 3 presents the main results of using SeenItRet to predict stock market returns and the important themes that drive this predictability. Section 4 demonstrates that our SeenIt predictors can predict broader economic variables beyond the stock market, highlighting SeenIt as a general empirical expectations operator. Section 5 concludes.

1 Related Literature

Our study is related to several strands of literature. First, it contributes to the literature on predicting stock returns using text. Prior studies in this literature have identified a remarkable set of specific narratives that predict stock returns. In a foundational study, Manela and Moreira (2017) construct a text-based measure of economic uncertainty using front-page articles of the Wall Street Journal from 1890 to 2009 and show that it predicts aggregate equity premia. Hirshleifer et al. (2025) show that war discourse in the news, in particular, captures time-varying rare disaster risk and strongly predicts aggregate stock returns using 7 million New York Times articles spanning 160 years.[5] Distinct from the above studies focusing on a given economic theme, Bybee et al. (2024) employ topic modeling on Wall Street Journal articles to systematically identify economic narratives that predict stock market returns, such as the recession narratives. Based on the topics, Bybee et al. (2023) propose a narrative factor model that explains the cross-sectional stock returns.[6]

Our study departs from previous literature by not directly drawing return predictions from narratives. Instead, we introduce a novel spatial model that measures the similarity between economic states in historical and current months, using high-dimensional narratives spanning 1815 to 2021 (this long time series is crucial to our approach, and also differentiates our study from many of the earlier works). Building on the idea that "history rhymes," we predict future returns based on stock market performance following historical months with similar economic states (SeenItRet), rather than directly relying on narrative content. Thus, our study complements prior research by proposing a largely orthogonal approach to constructing return predictors from narratives, based on essential economic principles. Our proposed SeenItRet exhibits economically significant and long-lasting predictability of market-wide stock returns, which passes stringent OOS tests. Our approach is general, as we also show strong

[5] War narratives also explain cross-sectional stock returns as shown in Hirshleifer et al. (2024).

[6] Relatedly, a large body of literature on news and asset pricing highlights that the sentiment or coverage of news predicts stock returns (see Tetlock (2007), Engelberg and Parsons (2011), Solomon et al. (2014), Huberman and Regev (2001), Peress (2008), Fang and Peress (2009), Tetlock (2010), Jeon et al. (2022), among others. For comprehensive reviews of the literature on texts and finance, see Loughran and McDonald (2016), Gentzkow et al. (2019), and Hoberg and Manela (2025).

predictability relating to risk exposures and predicting a broader range of economic variables.

We emphasize that the predictability of SeenItRet arises naturally as long as the stock market reflects the economic states perceived as important by investors. This assumption can be motivated by both rational and behavioral theories of finance. For example, the predictability of SeenItRet arises if the economic themes we model, which we broadly call economic states, reflect either investors' risk perceptions or subjective beliefs. In either case, the ex-post returns of similar past states will predict the ex-post returns of the current state.

Second, our work also complements emerging research on machine learning and return predictions (Gu et al. (2020), Adämmer and Schüssler (2020), Kelly et al. (2022), Kelly et al. (2023), Didisheim et al. (2023a), Didisheim et al. (2023b), Kelly et al. (2024), Nagel (2025), Li et al. (2025), among others). Existing studies focus on maximizing signal extraction and return predictability by exploring the interplay between and non-linearity of high-dimensional signals. We contribute to this literature by extending the approaches for synthesizing high-dimensional signals. If the stock market prices similar underlying economic states consistently, albeit the states can be quite complex, we can use SeenItRet based on historically similar months for return predictions in the current month. As we demonstrate, a simple "seen it before" framework using long history data generates highly significant predictions of market returns and a broader set of economic variables in a unified and simple one-dimensional framework. While incorporating our proposed SeenIt variables, SeenItRet and SeenItRisk in particular, into machine learning models is beyond the scope of our study, we believe our framework offers a promising direction to improve research in this area by providing informative SeenIt features (see Li et al. (2025) for an example of feature engineering) and through reduced dimensionality.

Third, our study is also related to the study of using natural language processing methods to capture important economic states and test hard-to-measure theoretical constructs. Baker et al. (2016) show that newspaper text searches can be used to measure economic policy uncertainty. van Binsbergen et al. (2024) leverage word em-

beddings to extract granular and forward-looking sentiment over 170 years in the U.S. Liu and Matthies (2022) use news coverage to capture investor concerns about economic growth prospects, uncovering a persistent component of consumption growth that supports the long run risk model in asset pricing. Fisher et al. (2022) construct novel measures of macroeconomic attention (MAI) from New York Times and Wall Street Journal articles which predict FOMC announcement-day excess returns and VIX declines. Caldara and Iacoviello (2022) develop a newspaper-based index of geopolitical risk (GPR) and show that it predicts a higher probability of economic disaster. A recent important work is by Bybee et al. (2024), which estimates a topic model that summarizes business news into 180 interpretable topical themes using Wall Street Journal articles from 1984 to 2017. Previous studies have also used textual analysis to measure industry classifications (Hoberg and Phillips, 2016, 2025). Like these studies, our work supports the significant role of narratives in understanding economic dynamics (Shiller, 2017, 2020). Unlike the literature, we emphasize the use of narratives to identify historical periods with similar economic states, rather than using themes to directly predict economic outcomes.

2 Data and Measurement

We develop a novel measurement system to quantify economic narratives using historical newspaper content in the U.S. over the past 200 years. This system enables us to identify recurring patterns in economic discourse throughout history. Using this system, we construct predictors of stock market returns based on the "historical rhymes" principle.

2.1 Historical News Data and Economic Themes

We collect our data from NewspaperArchive, a database of digitized historical newspapers serving as one of the world's largest newspaper archives. The database encompasses not only major national newspapers but also numerous smaller local ones, enabling us to comprehensively characterize the U.S. economic states since the 19th century and

extract diverse views across the country. We focus on business-related articles by using a selected keyword search query to extract business news from all U.S. newspaper outlets in the NewspaperArchive.[7]

In total, we obtain more than 210 million business-related news articles from 1815 to 2021. Our analysis focuses on the abstract of each article to capture the core information while keeping the computation manageable. These articles cover narratives of both national business trends and local events. Before we use the news content to characterize economic states, we perform text cleaning by removing HTML tags, non-alphanumeric characters, and normalizing spaces and lowercasing all the text. We then pre-process the text using standard natural language processing techniques, including lemmatization and the removal of common stop words. Finally, we exclude articles that are too short, i.e., fewer than ten tokens in the English language dictionary, or exposed to significant OCR errors, i.e., more than 50% of the article's tokens are not in the English language dictionary. This filter yields a final sample of 160 million high-quality articles.

To identify the semantic state spaces for each article, we construct a comprehensive list of themes by taking the union of economic terms from several sources, including economic terms organized by the Economist, glossary of economics from Wikipedia, glossary from the Federal Reserve Bank of St. Louis, glossary from Commodity Futures Trading Commission (CFTC), political terms from Baker et al. (2016), and the commodity-related terms from Corporate Finance Institute. We selected sources that have a long time span and extensive coverage, e.g., the Economist was first issued in September 1843. We also utilize ChatGPT to further complement these word lists (see the Internet Appendix IA.1 for more details). We use unigrams to capture the relevant themes and ensure the processing of our very large corpus is manageable.

We next identify significant themes by applying an occurrence filter to retain themes that appear with sufficient prominence across the news text corpus.[8] Finally, we manu-

[7]Our search query for extracting business-related news from NewspaperArchive is: *"stock market, business, businesses, economy, investor, investors, investing, household, housing, inflation, recession, real estate, commodities, unemployment, layoff, layoffs, technology, wall street, company, companies, transportation, metals, industry, energy, agriculture, gold, oil, innovation, innovations, politics, political, trade, war."* Our database of 210 million articles includes all articles containing any one of these broad terms.

[8]Specifically, we keep economic themes that occur more than five thousand times in at least one year

ally refine the list of themes to remove uninformative unigrams, resulting in 162 curated, high-quality, interpretable themes. These themes reflect economic narratives spanning over two centuries and facilitate interpretability in our subsequent analysis in Section 3. We note that our lists were curated before any analysis was conducted, and our approach thus follows the expert curation approach used in Loughran and McDonald (2011). In our case, our goal was to use a broad set of sources as noted above to capture economic terms that are likely long-lived in history and likely economically relevant. Despite that, we note that our results are also robust to using an alternative set of themes, specifically the business-related themes proposed by Bybee et al. (2024) based on recent news articles from 1984 to 2017 (see Section 3.3). Our results are also robust to applying a data-driven "importance weighting" scheme that uses each theme's co-variation with past market returns to put higher weight on more important themes (we discuss this approach in Section 3.4.3).

For each of the 162 economic themes, we generate 100 related unigram keywords using OpenAI to facilitate the detection of the theme from our news corpus.[9] Using the keywords of the themes and our newspaper text corpus, we characterize each historical month in terms of economic themes from January 1815 to December 2021. We start by computing the cosine similarity between the language used in each day's newspaper texts and each theme's keyword list. To do so, we represent each day's news texts as a vector of word frequencies, X_t, and each theme as a binary vector indicating its 100 keywords, X_k, where $k \in [1, ..., 162]$. The relevance of a theme k on a given day t can thus be represented by a cosine similarity between the daily word-frequency vector and the theme's keyword vector, i.e.,

$$\rho(t,k) = \frac{X_t \cdot X_k}{\|X_t\|\|X_k\|}. \tag{1}$$

The result is a panel of 162 cosine similarity scores — one for each theme — on each day, capturing how closely that day's news text aligns with each economic theme. Using

in our sample. Intuitively, a word that never reached five thousand occurrences in any year throughout the two centuries is unlikely to be a significant enough narrative that drives overall stock returns.

[9]We feed each theme word into a prompt, which is chosen to encourage the generation of vocabulary representative of how economic topics are typically presented in general news coverage. See details in the Internet Appendix IA.1.

a similar procedure, we also compute three additional scores that capture the positive tone, negative tone, and uncertainty of each day's news text based on established dictionaries for sentiment and uncertainty.[10] We aggregate the 162 theme cosine similarity scores and the 3 sentiment scores (positive, negative, uncertain) to the monthly level by taking the average over all of the days in each month.

We next standardize the 162 theme scores and the 3 sentiment scores so that we can identify the extent to which a given loading is "abnormally high or low" relative to its recent history. Indeed, not much would be revealed about the economic state if a variable theme that always has a high loading continues to do so. For each month, we thus construct z-scores by comparing the score for each theme to its 36-month trailing history excluding the most recent 12 months to prevent information contamination.[11] These z-scores thus indicate how unusually prominent each theme is in a given month relative to recent history, an approach used earlier by Hanley and Hoberg (2019).[12] We thus have 3 standardized "abnormal sentiment scores" for the given month and 162 standardized theme loadings.

To facilitate cross-month comparability for the 162 themes, we further normalize the 162 z-scores to have a mean of zero and a standard deviation of one in each month. We refer to these final standardized 162 theme values as *abnormal theme loadings* (ATLs), which quantify the relative narrative salience of each theme in a given month.

2.2 A Spatial Model of Economic States

We now develop a novel measurement system to quantitatively identify economic states in the U.S. over the past 200 years. We will define the information environment for each month using a high-dimensional state-space representation in which we characterize each month using a sparse 648-dimensional vector. The 648 dimensions will summarize the intensity of the 162 themes, the interaction of the 162 themes with

[10] Our sentiment keywords are from Loughran and McDonald (2016), while our uncertainty keywords are based on the 100 synonyms of the word uncertainty (see Internet Appendix IA.1).

[11] The z-score is simply the current month's cosine score minus the average of the 37 scores including itself and the 36 lagged values, all divided by the standard deviation of these 37 scores.

[12] See similar standardization procedures also in Bybee et al. (2024), Kelly et al. (2024), and Chen et al. (2024), among others.

positive sentiment, the interaction with negative sentiment, and the interaction with uncertain sentiment. The result will be 648 dimensions obtained by appending these four vectors, each of length 162.

We start by identifying the top 25 themes with the highest ATLs (z-scores) in each month. These constitute the month's most activated themes or "signature narratives." All remaining themes are treated as inactive, and their values for z-scores are set to zero. This step yields a sparse 162-dimensional vector of *activated abnormal theme loadings* (AATL), with 25 non-zero entries corresponding to the most prominent themes of the month. The number 25 is chosen to capture themes that have a prevalence roughly one standard deviation or more above the mean. This approach is consistent with the intuition that themes with lower intensity levels were not particularly relevant to the population at the time and with there being some limits to human attention.

Finally, we construct our 648-dimensional state representation by appending four vectors. The first 162 dimensions represent thematic intensity and are the raw AATL values described above. The second 162 dimensions (elements 163 to 324) are then set to be the raw AATL values multiplied by the abnormal negative sentiment score. The third 162 dimensions are the raw AATL values multiplied by the abnormal positive sentiment score, and the final 162 are similarly computed based on the abnormal uncertain sentiment score. The result is a 648-element state vector for each month, indicating thematic intensity and underlying sentiment levels. In each month, 100 (25 themes $\times$ 4 intensity or sentiment channels) of the 648 elements are activated and are non-zero, and remaining elements are zero per the above construction. We refer to this 648-dimensional monthly vector as the *signed activated abnormal theme loading* (**SAATL**) vector.

To assess the recurrence of economic states over time, we compute the cosine similarity between each month's SAATL vector and those of all prior months, excluding the most recent five years to ensure sufficient separation and avoid entanglement with overly recent states, i.e.,

$$\text{Economic State Similarity}(t, t-h) = \frac{SAATL_t \cdot SAATL_{t-h}}{\|SAATL_t\| \, \|SAATL_{t-h}\|}. \tag{2}$$

A high similarity score indicates that a focal month's narrative structure closely resembles that of a past month — i.e., "we have seen this state before." This pairwise similarity is the core input into our spatial model of the narrative economy. Each pair of months is associated with (i) the cosine similarity of their economic states and (ii) the time elapsed between them in months. As an illustration, Figure 1 plots the average cosine similarity of the top 25 most similar historical months for each focal month, and Figure 2 plots the average elapsed months from the top 25 most similar historical months to the focal month. These data enable us to analyze how repeated narrative environments relate to economic outcomes and stock market returns.

2.3 Constructing "SeenItRet" in the Stock Market

2.3.1 Historical Stock Market Returns

Constructing SeenItRet requires using stock returns from long historical periods. We obtain pre-CRSP historical returns from Goetzmann et al. (2001), who construct price-weighted monthly stock returns based on over 600 NYSE individual stocks from February 1815 to December 1925.[13] The price-weighted returns likely capture investors' perceptions of stock market movements during the 1900s. The now-standard value-weighting method was not well understood or widely accepted at that time. For instance, the first U.S. stock market index, created by Charles H. Dow, was price-weighted, as financial newspapers typically reported prices and trading volumes, but not shares outstanding. Breakthroughs occurred following the work by Fisher (1922), which argued that commodities with greater total value should be weighted more, and the introduction of Standard Statistics Company's market-cap-weighted index in 1923, the precursor to today's S&P 500 (Lo (2016)). Another advantage of the return data from Goetzmann et al. (2001) is that they account for both capital gains and dividend yields, whereas many indices do not capture dividend yields (Hartzmark and Solomon (2022)).

[13] We download the (2020 version) of the price-weighted historical returns from William Goetzmann's website at https://som.yale.edu/sites/default/files/2021-12/Price-Weighted-Index-Returns-2020-08-20.xls [last accessed on January 20, 2025].

For the post-CRSP period from January 1926 to December 2021, we construct three versions of stock market returns. The first version is the price-weighted returns (PWRet), constructed based on individual stock returns in the CRSP database, following Goetzmann et al. (2001), which ensures consistency with the historical data. The second and third versions are the equal-weighted (EWRet) and value-weighted returns (VWRet) provided by the CRSP database. Our constructed PWRet is 95% and 93% correlated with the CRSP VWRet and EWRet, respectively, during the post-CRSP period, while VWRet and EWRet are 91% correlated. By appending the historical data from Goetzmann et al. (2001) to each of the three versions, we obtain the long time series of PWRet, EWRet, and VWRet from February 1815 to December 2021.

2.3.2 Constructing SeenItRet

We construct SeenItRet for the three versions of stock market returns in two steps. First, for each month t, we obtain the top 25 historical months with the highest economic state similarities as in equation (2), denoted $\tau \in S_t$.[14] Importantly, we obtain the *next-month* returns for each of the historical months, $R_{\tau+1}$. Second, we compute SeenItRet for the current month t by averaging the next-month returns of the 25 most similar historical months:

$$SeenItRet_t = \text{Avg}_{\tau \in S_t} \left(R_{\tau+1} \right). \tag{3}$$

Based on this definition, we construct SeenItRet for each of the three versions of monthly stock market returns: PWRet, EWRet, and VWRet. Prior literature suggests that stock prices may reflect fundamental information with certain delays (e.g., Cohen et al. (2020), Cohen and Frazzini (2008), Menzly and Ozbas (2010), and Hoberg and Phillips (2018)). Therefore, we also construct a smoothed SeenItRet for each of the three versions by taking the 12-month moving average of the past monthly SeenItRet,

$$Smooth\ SeenItRet_t = \text{Avg}_{h \in (0,11)} \left(SeenItRet_{t-h} \right). \tag{4}$$

[14] Our main findings are robust to choosing the top 15 or top 35 most similar historical months, see the Internet Appendix Table IA.1.

We note that all of our SeenItRet variables are ex ante measurable and can be used to predict next month returns. Our final sample includes the monthly returns, SeenItRets, and the smoothed SeenItRets of the three versions of stock market returns from January 1825 to December 2021.[15] Panel A of Table 1 shows the summary statistics of PWRet, EWRet, and VWRet with average month returns of about 0.56% and 0.7%. Panel B shows the summary statistics of SeenItRet and smoothed SeenItRet. Despite these measures being constructed as averages of historical monthly returns, they still exhibit substantial time-series variations, with standard deviations ranging from 0.99% to 1.18% per month for SeenItRet, and from 0.59% to 0.68% for smoothed SeenItRet.

3 Stock Market Return Predictability

In this section, we present our main results using SeenItRet to predict aggregate U.S. stock market returns. Along with the main results, we provide a rich set of findings that explore the familiarity of historical states, i.e., "how intense history rhymes," the longer-term predictions of SeenItRet, and the economic interpretation of SeenItRet in each historical decade.

3.1 SeenItRet and Monthly Returns

Our main analysis examines the predictability of SeenItRet at time t on the next month's stock market return at $t+1$ by running the following time-series regression:

$$Ret_{t+1} = \beta \cdot SeenItRet_t + \epsilon_{t+1}, \tag{5}$$

where we individually examine the predictability of SeenItRet and smoothed SeenItRet, and we conduct Newey-West adjustments for standard errors with three-month lags.

Panel A of Table 2 reports the results. Column (1) shows that SeenItRet sig-

[15] We start from 1825 to allow enough historical return data to be seen by investors, as the New York Stock Exchange (NYSE) was only founded in 1792 and formalized in 1817 (see https://en.wikipedia.org/wiki/New_York_Stock_Exchange). Section 3.2 provides detailed analysis on how similarity of historical months affects SeenItRet's predictability.

nificantly predicts future price-weighted stock returns, our preferred version of stock market returns, with a t-statistic of 3.42. A one-standard-deviation increase in SeenItRet corresponds to a 0.38% increase in future monthly returns = (0.0103 × 0.364) (or 4.5% in annual terms), which is a 62% increase relative to the mean of the future monthly PWRet (i.e., 0.61%). Figure 3 illustrates that our SeenItRet tracks the actual monthly returns, with both being smoothed over the past three years for ease of visual presentation.

Column (2) shows that our smoothed SeenItRet predicts PWRet with an even higher t-statistic of 4.46 and an R^2 of 0.9%, consistent with the slow updating of stock prices, i.e., the "lazy prices" a la Cohen et al. (2020). The economic magnitude is also enhanced: A one-standard-deviation increase in smoothed SeenItRet corresponds to a 0.44% increase in future monthly returns, or a 72% increase in the average future monthly PWRet.

Columns (3)-(6) show that SeenItRet and the smoothed SeenItRet based on EWRet and VWRet also strongly predict the corresponding versions of future monthly returns. Panels B and C in Figure 3 illustrate that SeenItRet also tracks the EWRet and VWRet well.

Placebo Tests Next, we conduct placebo tests to strengthen our inference that the above return predictability is driven by historical months that represent similar economic states to the current month, i.e., the "history rhymes" inference, rather than time-trend relations in the time series of stock returns.[16] To conduct the placebo tests, for each historically similar month τ in equation (3), we reassign the month to be a year earlier than the original historically similar month, i.e., $\tau - 12$, and then compute the placebo SeenItRet based on the next-month returns of the 25 reassigned months.[17]

This reassignment, which moves the historical months to be one year earlier, is unlikely to materially affect any time-trend relations between the placebo SeenItRet

[16]Note that our historically similar months, by construction, are at least five years earlier than the current month. Hence, our results are unlikely to be driven by short-term information overlap between SeenItRet and future returns.

[17]That is, $Placebo\ SeenItRet_t = Avg_{\tau \in S_t}\left(R_{\tau-11}\right)$.

and actual returns, if they exist, as the 25 historically similar months are, on average, 542 months earlier than the focal month. However, this reassignment substantially diminishes the similarity between the placebo historical months and the focal months, with the average rank of similarity declining from 13 for SeenItRet's underlying months to 333 for the placebo SeenItRet's underlying months. In other words, while the placebo SeenItRet is based on similarly distant historical returns as the original SeenItRet, their economic states are no longer similar to the focal month. As history stops rhyming, we expect the placebo SeenItRet not to predict future returns.

Panel B of Table 2 confirms our prediction. We observe that neither the placebo SeenItRet nor its smoothed version predicts future monthly returns. The coefficients are no longer significant, and the R^2s drops to essentially zero. These placebo tests support our "history rhymes" hypothesis and further rule out spurious explanations of the results.

Out-of-Sample Tests To gauge the strength of SeenItRet's predictability relative to existing benchmark predictors, we conduct the out-of-sample (OOS) test proposed by Welch and Goyal (2008). Welch and Goyal (2008) argue that few predictors can outperform the simple average of historical returns out of sample. They develop an OOS R^2, which compares the new model with the historical mean model using rolling ("out-of-sample") OLS regressions instead of full-sample OLS regressions:

$$OOS\ R^2 = 1 - \frac{\sum_{t=1}^{T-1}\left(Ret_{t+1} - \widehat{Ret}_{t+1}\right)^2}{\sum_{t=1}^{T-1}\left(Ret_{t+1} - \overline{Ret}_{t+1}\right)^2}, \tag{6}$$

where $\widehat{Ret}_{t+1}$ is the fitted value from a predictive regression of the new model estimated from the beginning of the sample through period t, and $\overline{Ret}_{t+1}$ is the benchmark historical average return estimated from the beginning of the sample through period t. In our case, this means that at each month t, investors compare the predictability of the SeenItRet model and the historical mean model for the next month's stock market return, using all information from the beginning of the sample to t. If the $OOS\ R^2$ is positive, then the SeenItRet model outperforms the benchmark as it has a lower

average mean-squared prediction error than the historical average return.

Following Campbell and Thompson (2008), we choose our full period from January 1825 to December 2021 as the sample period and the CRSP sample from January 1926 to December 2021 as the forecasting period when we compute the *OOS* R^2.

Table 3 presents the results. Two observations stand out. First, the out-of-sample R^2 is positive for all three versions of aggregate stock market returns and for both the SeenItRet and the smoothed SeenItRet models, which indicates that our SeenItRet model outperforms the historical mean model. While other return predictors have been shown to outperform the historical mean model under weak restrictions on the signs of the regression coefficients (see details in Campbell and Thompson (2008)), the improved performance of SeenItRet stands even without any restrictions. Second, the out-of-sample R^2s are highly comparable to the in-sample R^2s in magnitude. This strong predictability is not common, as Welch and Goyal (2008) highlight that only one out of seventeen existing economic predictors under examination shows such strong out-of-sample performance.

The strong out-of-sample performance of our SeenItRet model reinforces the "history rhymes" principle, which states that history predicts the future only when we focus on returns during past economic states that are similar to the current state, as is the case with our SeenItRet model. If investors simply average past returns over a long history without this selection, the predictive power on future returns is significantly lower.

Robustness Our findings are robust across several specifications. First, our results are robust when we choose the top 15 or 35 historically similar months, rather than the top 25 months, to compute SeenItRet (see the Internet Appendix Table IA.1). Second, we also find very similar results when using SeenIt "excess returns" to predict future excess returns, instead of raw returns, suggesting that our findings are not primarily driven by SeenItRet's predictability of risk-free rates (see the Internet Appendix Table IA.2 for full-sample prediction and Table IA.3 for out-of-sample prediction of excess returns). Third, SeenItRet robustly predicts returns during the post-modernization

period when ticker and telephone are introduced to the NYSE (e.g., post-1876) and the CRSP period (e.g., post-1926), suggesting that our results are not solely driven by historical periods (see the Internet Appendix Table IA.4).

3.2 SeenItRet and Historical Familiarity

It is plausible that some months have been "seen" with high familiarity in the past, while others may be relatively more novel and distinct from the historical record. We explore this heterogeneity by examining a month's similarity to historical states. Specifically, we average the economic state similarity scores (see equation (2)) of the 25 most historically similar months, which we refer to as the *SeenItFamiliarity* of the focal month.

Figure 1 plots the SeenItFamiliarity for each month from 1825 to 2021. Interestingly, the aftermath of the Great Recession during 2010 and 2011 shows substantially low familiarity with history. There can be several possible reasons for this period to feel unfamiliar, considering the unprecedented non-war-time stimulus including the large-scale American Recovery and Reinvestment Act, the back-then near-record deficits of $1.3 trillion by the government, and heightened political discussions that highlighted socialism, communism, and government intervention. Indeed, a salient observation in the news context during this period is the *combined rise* in themes featuring "deficits", "stimulus", "communism", and "socialism".[18] Targeting the exact reasons why certain periods are more novel than others is beyond the scope of our study. Instead, we focus on whether our SeenItRet predicts future returns more effectively during periods when history truly "rhymes" than during periods when history provides less guidance. Specifically, we conduct the following tests by interacting SeenItRet with $(1 - SeenItFamiliarity)$, which proxies for unfamiliarity. Our hypothesis is that Seen-

[18] Another observation in Figure 1 is that SeenItFamiliarity appears to rise in the earlier years of our sample (from 1825 to 1850), reaching a relatively more stationary trend afterwards. A driver for the rising similarity in earlier years is that our underlying newspaper data begins in 1815, resulting in fewer very similar past months for those years. However, it is also possible that investors in those early years had not observed long enough historical returns and financial news, as the NYSE was just founded at the beginning of the century. Nevertheless, we show in the Internet Appendix Table IA.5 that our findings in this section are robust to excluding the first 25 years from our sample.

ItRet should exhibit stronger predictability during periods of high SeenItFamiliarity:

$$\begin{aligned} Ret_{t+1} =& \beta \cdot SeenItRet_t + \phi \cdot (1 - SeenItFamiliarity_t) \\ &+ \gamma \cdot SeenItRet_t \times (1 - SeenItFamiliarity_t) + \epsilon_{t+1}. \end{aligned} \tag{7}$$

Table 4 confirms this prediction. We observe the coefficient for the interaction term, γ, to be significantly negative, suggesting that SeenItRet does not predict future returns as strongly during highly unfamiliar periods when very similar economic states have not been observed in the past, i.e., when "history does not rhyme." In contrast, when we linearly extrapolate SeenItFamiliarity to asymptotically approach 1, i.e., when history "repeats," the predictability of SeenItRet, as reflected by the coefficient of the standalone SeenItRet, β, becomes more than four times greater than the unconditional predictability estimated in Table 2. While it is unlikely that history repeats itself fully, this extrapolated estimate suggests that as we continue to observe more economic states in the future, SeenItFamiliarity is likely to grow, and the predictive power of SeenItRet is likely to increase over time.[19]

3.3 SeenItRet and Longer-Term Returns

After demonstrating the short-term predictability of SeenItRet on next month's stock market returns, we now examine longer-term predictability. Specifically, we run regressions as in equation (5), using monthly returns from $t + 1$ to $t + 36$ as the dependent variable, one at a time. To highlight SeenItRet's predictability of *future returns*, we also directly examine the association between SeenItRet with past monthly returns

[19]In the Internet Appendix, we examine another heterogeneity based on the average elapsed time between the historically similar months and the current month. Importantly, we ask whether investors are more likely to incorporate the information embedded in SeenItRet if the historically similar months are more recent. We present two results that suggest this is not the case. First, Table IA.6 shows that interacting SeenItRet with a Recency measure, i.e., the negative of the average raw or detrended time distance between the 25 historical months and the current month, results in small and statistically insignificant coefficients of the interaction term. Second, Table IA.7 shows that an alternative SeenItRet, constructed by restricting the historical months to be within an investor's active trading lifespan (e.g., 50 years), shows similar results to the unconstrained SeenItRet. These findings do not rule out the possibility that the predictability of SeenItRet can be explained by behavioral theories or risk-based models. Delving deeper into these more fundamental explanations remains fruitful for future research but is outside the scope of our study.

from $t-2$ to t. For ease of comparison between the predictability of future and past returns, we standardize SeenItRet to have a mean of zero and a standard deviation of one.

Table 5 shows the results. Panel A displays the predictability of the unsmoothed SeenItRet. We highlight two key findings. First, SeenItRet, constructed based on the next-month returns of historically similar months, is not associated with stock market returns in the past months or the current month, as the coefficients from returns at $t-2$ to t are all small. Hence, the strong predictability of SeenItRet for next month's returns, which we presented earlier, is likely driven by the predictor providing new information about "future" economic states. Second, the new information brought by SeenItRet appears to persistently unfold in the future months, as SeenItRet persistently predicts stock returns up to $t+24$, though the predictions are stronger for some future months and weaker for others. We observe similar findings regarding the predictions for all three versions of stock market returns.

Panel B reports the longer-term predictability of smoothed SeenItRet, which not only shows greater magnitudes than SeenItRet, but also more persistent and robust predictability in the longer run. The coefficients for the standardized smoothed SeenItRet initially increase, reaching a peak in predicting future monthly returns at $t+5$, with magnitudes of 0.47%, 0.58%, and 0.38% per month (or 5.6%, 7.0%, and 4.6% annually) for PWRet, EWRet, and VWRet, respectively. Then, the predictability gradually decays but remains significant until $t+25$. Figure 4 visualizes the longer-term predictability of both SeenItRet and smoothed SeenItRet.

As a final robustness check, we highlight that the short- and long-term return predictability of SeenItRet is robust to using alternative business news themes to construct the historically similar months. We consider Bybee et al. (2024), who develop the structure of themes that comprehensively cover modern business news since the 1980s. While we choose our own themes designed to capture economic states relevant throughout the past 200 years, we obtain the themes from www.structureofnews.com and construct an alternative SeenItRet by applying their themes to our methodology in Section 2. We select the most relevant 100 unigrams for each of the themes identified in the structure

of business news from Bybee et al. (2024). In particular, we use the topic word lists provided by these authors instead of our manually curated lists to construct alternative themes. We otherwise use the same formulation throughout. The Internet Appendix Table IA.8 shows that the smoothed SeenItRet using the alternative themes also successfully produces significant and persistent long-term return predictability, albeit with economic magnitudes that are smaller than our baseline SeenItRet. Overall, this finding highlights the robustness of our measurement system in generating return predictors based on the "history rhymes" principle.

3.4 SeenItRet Interpretations

In this section, we provide a framework for interpreting the strong return predictability we documented above.

From the literature, we expect that certain dimensions of the state space are more influential than others in driving market returns. For example, following Hirshleifer et al. (2024, 2025), we expect that the presence of war discourse to be important. As financial economists, we also expect discussions of market bubbles and unemployment to be important. Most importantly, we also expect the specific themes that are important in any period to be widely varying both in cross-section (what matters in a given month) and in time series (how quickly topics rise and fall in prevalence. Hence, many other economic themes, such as durable goods, distribution, and energy, might be important in some periods but not in others. It is ultimately an empirical question.

Our framework also offers a simple way to assess what themes are most important for the stock market from a historical perspective. We conduct a holistic assessment throughout our sample period regarding which themes are most important in generating our economically large return predictability, both in the short-term one-month horizon, and in medium and long-term horizons including 6 and 24 months.

3.4.1 Methodology for Assessing Thematic Relevance

We start by assessing which themes are most important regarding short-term return predictability. Our approach is simple: we recompute SeenItRet using all information available in the state space vector (648 dimensions) except that we leave out one theme. As each theme enters the computation four times (raw volume of the theme, and the interaction of the theme with positive, negative, and uncertain tenor), running SeenItRet with one theme left out results in a state space of 644 dimensions. We thus compute an alternative "left-one-out" SeenItRet and compare the adjusted R^2 of the baseline 648-D SeenItRet model with the 644-D SeenItRet model that leaves out a focal theme. Our regressions are based on the one-month (or 6 to 24 month) ex-post market-wide price-weighted stock return (we use PWret, our baseline return variable) as the dependent variable, and SeenItRet as the key RHS variable. We compute the fraction of adjusted R^2 that is lost when we use the given 644-D SeenItRet with one theme dropped instead of the baseline 648-D SeenItRet. A given theme is deemed to be "more important" if the fraction of adjusted R^2 lost is larger.

We thus loop through all 162 baseline themes and compute the fraction of adjusted R^2 lost for each theme when it is left out of SeenItRet's calculation. We then sort themes based on these losses from highest to lowest. We then present a list in descending order of the top 50 themes that result in the largest adjusted R^2 loss in explanatory power. The highest ranking themes are likely most important for "news watchers" to pay attention to when considering predictions for the next month.

3.4.2 Important Themes that Most Explain Market Returns

Table 6 displays the results of running the leave-one-out analysis using regressions predicting short-term next-month returns (Panel A), medium-term ex-post 6-month-ahead returns (Panel B), or long-term ex-post 24-month-ahead returns (Panel C). Our regressions are based on our main specification, in which the dependent variable is PWRet over any specified horizon, and the right-hand-side (RHS) variable is the smoothed SeenItRet.

Panel A shows that the five most important news themes for predicting next-month returns are momentum, reduce, margin, government, and scarcity. Many of these top themes have similar economic magnitudes ranging from 10% to 20% contributions to adjusted R^2. Other notable themes in the top 50 include entrepreneur, subsidy, manufacturing, deregulation, agriculture, president, equity, mortgage, and oil. This wide array of distinct economic themes illustrates the power of the SeenItRet platform's ability to dynamically incorporate information from a complex and somewhat comprehensive set of candidate state features. Generally, themes appearing on this top 50 list are quite important given there are 162 total themes. We remind readers that these results in Panel A are most effective in helping to explain shorter-term one-month ex post returns, and we note below that the nature of which themes are important varies in interesting ways as one extends the horizon up to 24 months.

Panel B of Table 6 displays the results for the 6-month-ahead return prediction horizon and illustrates that important themes change materially. While momentum remains important, the themes of deregulation, war, consumer, and oil round out the top 5. The 6-month horizon thus has only one overlapping theme with the short-term horizon of one month, namely oil. Intuitively, the results are consistent with the longer-horizon results highlighting issues that will impact society for longer periods of time, such as war, energy, consumer demand, and deregulation. The results for war echo the findings of Hirshleifer et al. (2024, 2025) on war discourse.

The results for shorter-term returns in Panel A are less definitive by comparison and are more consistent with relatively more vague themes that have less clear-cut long-term predictions, such as reduce, margin, and government. This suggests that short-term market movements are more based on knee-jerk reactions to more vague but high-level concepts (such as "reduce" indicating some form of contraction to come, or "government" indicating some non-specific changes to politics or rulemaking to come) when they first appear in the news. Yet over time, more concrete and specific themes then become clear (such as those in Panel B), and indicate the longer-term outlook more clearly. These results illustrate the main finding of our study that SeenItRet is more powerful in understanding longer-term outcomes (traditionally more difficult to

predict in asset pricing) than short-term outcomes over a single month.

The longer-term results (predictors of 24-month-ahead ex-post returns) reported in Panel C are also quite different from those for the short (1-month) and medium (6-month) horizons. Panel C shows that the top 5 themes are inflation, expansion, currency, bond, and competition. These themes are notable for their tendency to potentially impact economic outcomes over longer horizons. Indeed, the top theme, inflation, is known to be sticky and very difficult for regulators to control, and hence its rise as a theme portends market-wide returns that are likely to be impacted for many months to come. Similarly, expansions (the second most important) are traditionally regarded as long-lasting episodes of growth. Other notable themes include food, conditions in Europe, and bubbles.

Overall, we believe our interpretive results are intuitive across horizons, and we believe these results can further guide financial economists in understanding how to utilize deep historical context to understand how investors form beliefs about asset prices at the aggregate level (e.g., Malmendier and Nagel (2011), Greenwood and Shleifer (2014), Giglio et al. (2021), and Bybee (2023)).

3.4.3 Important Themes for Specific Decades

We complete our analysis of interpretability with an assessment of which themes were most impactful in which past decade. This analysis illustrates the diversity of state distributions that investors have lived through since the 1800s. We report results for all decades starting with the 1870s and through the 2010s.

Because ten years is not long enough to run reliable return prediction regressions as we did in the previous section, we illustrate thematic importance in each decade simply by reporting the top ten themes in each decade with the highest absolute comovement with stock returns in the given decade. We compute thematic importance for a given theme as the average of the absolute value of the theme's standardized monthly value

(ATL) multiplied by the same-month return as follows:

$$\text{Thematic importance weights}(k, T) = \frac{\sum_{t \in T} \| ATL_k \times ret_t \|}{N[T]}. \tag{8}$$

Themes with high importance thus experienced more extreme intensities when stock returns also had extreme values. Such themes are likely important in shaping how economic agents from the given period formed beliefs about stock returns. We report the ten themes in each decade with the highest thematic importance weights.[20]

Table 7 displays the results. We first report which themes have the highest overall importance weights throughout our entire sample period in the first row. The results indicate that momentum is the most important theme, followed by war, inflation, price, and fraud. These results are intuitive, given the literature's strong focus on momentum and the importance of war, as noted above.

Looking at specific decades, we note that indications of fraud, cartels, and antitrust appeared high in importance weights throughout the late 1800s. These results are intuitive, given that the Clayton Act became law by 1890, and they illustrate the importance of understanding competition to stock market traders of this time. Real economy themes, such as imports, wages, and revenue, dominated the 1880s, while a focus on households, advertising, durable goods, and oil became important in the 1900s.

Not surprisingly, the 1920s (known for rising markets and speculation) saw the rise of themes including speculators and momentum, and also a focus on more speculative concepts such as technology and a focus on the future. Yet other more traditional themes such as competition and households were also important. By the 1930s, a focus on inflation, currencies, Europe, and stimulus became important.

Not surprisingly, the 1940s saw a top focus on war driving market returns given the events of World War II. Interestingly, the war theme was not among the top ten in the 1910s during World War I. This does not mean that war was not in the media,

[20] In the Internet Appendix Table IA.9, we also report a robustness test where we rerun our return predictability tests after recomputing SeenItRet using economic state similarities (equation (2)) in which the themes are weighted using the importance weights in equation (8). The results are similar to our baseline results.

but rather, it means that stock market returns were not particularly correlated with war in the 1910s even though they were in the 1940s. The lower return-importance of war in the 1910s also indicates a limitation in our study given the intuition regarding its importance at the time. Partly this is due to reporting decade-long importance averages. We note in year-by-year analysis that war does become important in 1915 specifically, for example. We note that war again became important in the 1950s given the Korean War, and again in the 1990s given the Gulf War. Hence, the theme generally appears with high weight when we would expect.

Other important findings include the importance of oil and energy as top themes in the 1970s (where oil prices and scarcity were important), and the importance of recession themes in the 2000s as the financial crisis came about. Themes in the 2000s also include foreclosure, poverty, treasuries, default, and stimulus. These themes well represent the intuition of events at the time. By the 2010s, trade-related themes emerged as important, echoing the importance of trade imbalances, imports and exports increased during this period, something that is clearly business-relevant and also thought to be a factor in the election of Donald Trump in 2016 and again in 2024.

4 Risk and Broader Economic Predictions

Our SeenIt platform, in its full generality, motivates a novel and intuitive "empirical expectations operator." The framework thus can predict other economic quantities beyond market-wide returns. The simple idea is that many important economic quantities should evolve as functions of the underlying state space. As such, we can predict their outcomes too by assessing the average evolution of these same variables in the past months following the most highly similar past states.

Another novel feature of the SeenIt platform is that it not only offers a way to predict any economic variable of interest, but it also provides an intuitive measure of any variable's second or higher moments. Second moments can be forecasted by computing the standard deviation of a key variable's outcomes that transpired following past similar states. We explore market risk (second moment of the market return)

derived from our SeenIt platform in the next subsection, and we end our study with an analysis of a wider array of economic variables including treasury yields, volatility, recession prediction, inflation, and patenting activity. Yet, we note a limitation of this expanded analysis is that we can only conduct this analysis for variables that have a reasonably long historical time series available (a material limitation for many variables of interest).

4.1 A SeenItRisk Premium

Our first extended analysis examines the second moment of expected aggregate market returns (our baseline results are based on the first moment, "SeenItRet"). We define "SeenItRisk" as simply the standard deviation of the 25 next-month stock return observations of the 25 most similar historical months instead of the average of these returns (SeenItRet). A high SeenItRisk indicates that market-wide returns are likely to be volatile in the coming months, and thus, the systematic risk of equities is likely to be high. A wide array of finance theories would predict that stock market investors will demand a risk premium in such states, and that overall expected returns will be higher.

We test this prediction in Table 8, and we regress the ex-post next-month $t+1$ stock return on the smoothed SeenItRisk, which we note is fully measurable as of month t ensuring no look-ahead bias. As before, we consider price-weighted, equal-weighted, and value-weighted returns. Column (1) illustrates our main finding that SeenItRisk indeed positively predicts ex-post returns, indicating a risk premium for investing in market conditions which history suggests are likely to entail high systematic risk in a forward-looking sense. This is consistent with the theoretical prediction of a risk premium to holding equity. In Column (2), we show that SeenItRisk's ability to predict returns is distinct from our main variable SeenItRet, as both remain significant when included together. Yet including SeenItRet reduces the magnitude of SeenItRisk's coefficient by roughly 25%, suggesting some interrelatedness.

Finally, in Column (3), we compare SeenItRisk's predictability to market volatility measured as the standard deviation of daily stock returns in month t (which is also

ex-ante measurable as we predict returns in month $t+1$). We find that volatility negatively predicts returns, unlike SeenItRisk. Moreover, including volatility as a control strengthens SeenItRisk's coefficient by roughly 25%. SeenItRisk thus offers additional predictions that cannot be explained by traditional volatility metrics.

The remaining six columns show robustness to using equal-weighted or value-weighted returns. As was the case in our main results, our effects are strongest for equal-weighted market returns but remain robust and significant for all three return variables. These results overall reinforce the theoretical consistency of the SeenIt framework and its ability to predict higher moments, while also documenting an additional source of return predictability beyond SeenItRet.

4.2 Beyond Stock Market Predictions

We conclude our analysis by extending beyond stock market returns and exploring the efficacy of our SeenIt expectations operator in predicting a broader range of economic variables with reasonably long time series, for which our framework is applicable.

We compute SeenIt predictions for five economic variables, including **Treasury yields** of the U.S. 10-year treasury note (constant maturity) for each month from 1815 to 2021 obtained from St. Louis Fed, with the historical data prior to January 1925 extrapolated from a Business Insider publication; **Volatility** which is based on the within-month standard deviation of daily stock returns from 1885 to 2021, where daily data from 1885 to 1925 are from William Schwert's website and CRSP afterwards; **NBER Recession** indicator for each month from 1815 to 2021 with the historical data before 1855 obtained from the Wikipedia; **Inflation** as the log difference in monthly CPI from 1871 to 2021 downloaded from Robert Shiller's website; and **Patent** as the log difference in yearly number of patent applications from 1815 to 2021 downloded from the USPTO website. All variables are monthly, except for patent applications.

Similar to SeenItRet, we compute the SeenItVar for each of the five economic variables as the average of the next-month outcome of the variable in the 25 most similar historical months. For patent applications, we first compute the SeenItVar at the

monthly level, where we use the corresponding yearly patent applications for each historical month, and then average the monthly SeenItVar within each calendar year. For each economic variable, we regress its next-period $t+1$ value on its ex-ante measurable SeenItVar constructed at t. These regressions thus test the SeenIt prediction for a given economic variable using our empirical expectations operator generalized beyond stock returns. For example, we can regress our inflation measure (CPI growth) for month $t+1$ on SeenItInflation measured as the average of the 25 next-month CPI growth of the historical similar months for month t.

The results presented in Table 9 illustrate the general utility and appropriateness of the SeenIt framework as an expectations operator. We find highly significant and consistent predictability across all five variables. We believe the SeenIt framework thus offers excellent potential for future researchers and policymakers interested in predicting outcomes, perhaps especially when there are fewer econometric tools available to predict the given quantity, and when untested theory predicts the evolution of the variable through state transitions.

5 Conclusion

We propose that the historical record has become adequately long with rich narratives for researchers to predict future outcomes by identifying past times with an economic state that resembles the current scenario. If economic outcomes evolve as state transitions, a wide array of economic variables should be predictable by computing their past outcomes following the historically similar dates. We label this logic as the "history rhymes" principle and we develop a "SeenIt" methodology to generate predictors.

The idea is that a policymaker facing a world with high inflation, the possibility of a market bubble, risks of war, declining trade flows, and other concerns can predict outcomes of their interested economic variables by identifying times in the past where society faced a similar basket of state descriptors. This framework naturally requires a long sample period and high-dimensional narrative representations based on a substantial amount of text. Its simple non-parametric form also internalizes interaction

effects and non-linearities for predicting economic outcomes, as many of these state features would require policymakers to navigate complex interacting economic forces at the same time.

We consider a massive corpus of 210 million newspaper articles from Newspaper Archive, dating back to 1815, and develop an empirical model of the 648-dimensional state space using economic themes extracted from textbooks and various online resources. For each month, we identify the 25 most similar historical months and propose an empirical expectations operator as simply the average of any variable's outcome following these 25 past months, which we label as a "SeenIt" predictor. This SeenIt predictor is fully ex-ante measurable.

Market-wide stock returns are the primary focus of our empirical analysis because the stock market aggregates the impact of many economic forces on business activities. We compute a stock market predictor, "SeenItRet," and find that a one standard-deviation shift in this variable predicts annualized market-wide returns about 4-7% higher. This impact is highly significant for price-weighted, equal-weighted, and value-weighted returns and also performs well using the more stringent out-of-sample R^2 analysis proposed by Welch and Goyal (2008). We note that SeenItRet is remarkable in predicting market returns not only for the next month, but also for longer horizons. In particular, its predictability maintains nearly full strength for the next 12 months after measurement, and then gradually fades to insignificance over the subsequent months, from 13 to 30 after the date of measurement.

Our SeenIt framework is general and can be used to estimate both first and second moments for a wide array of economic variables, including those beyond stock market returns. An important restriction is that a long historical time series is required. We find that a measure of future expected risk, "SeenItRisk," also predicts market-wide returns and is distinct from our main variable SeenItRet. This result suggests a potential risk premium for investing during periods when history indicates that a wider range of future realizations is possible. We also find that the SeenIt framework generates successful predictors of other variables including treasury yields, volatility, inflation, NBER recessions, and patenting activity. We believe the framework offers

many advantages and novel features relevant to future researchers, practitioners, and policymakers alike. As we continue to observe economic states and accumulate records over time, the predictive power of our SeenIt framework is likely to strengthen.

References

Adämmer, Philipp, and Rainer A Schüssler, 2020, Forecasting the Equity Premium: Mind the News!*, *Review of Finance* 24, 1313–1355.

Baker, Scott R, Nicholas Bloom, and Steven J Davis, 2016, Measuring economic policy uncertainty, *The quarterly journal of economics* 131, 1593–1636.

Bybee, J. Leland, 2023, The ghost in the machine: Generating beliefs with large language models, *University of Chicago, Working Paper* .

Bybee, Leland, Bryan Kelly, Asaf Manela, and Dacheng Xiu, 2024, Business news and business cycles, *The Journal of Finance* 79, 3105–3147.

Bybee, Leland, Bryan Kelly, and Yinan Su, 2023, Narrative asset pricing: Interpretable systematic risk factors from news text, *The Review of Financial Studies* 36, 4759–4787.

Caldara, Dario, and Matteo Iacoviello, 2022, Measuring Geopolitical Risk, *American Economic Review* 112, 1194–1225.

Campbell, John Y., and Samuel B. Thompson, 2008, Predicting excess stock returns out of sample: Can anything beat the historical average?, *The Review of Financial Studies* 21, 1509–1531.

Chen, A. J., Gerard Hoberg, and Miao Ben Zhang, 2024, Institutional Participation in Information Production and Anomaly Returns, *USC Marshall School of Business Research Paper Sponsored by iORB* .

Cohen, Lauren, and Andrea Frazzini, 2008, Economic Links and Predictable Returns, *The Journal of Finance* 63, 1977–2011.

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen, 2020, Lazy prices, *The Journal of Finance* 75, 1371–1415.

Didisheim, Antoine, Shikun Ke, Bryan T. Kelly, and Semyon Malamud, 2023a, APT or "AIPT"? The Surprising Dominance of Large Factor Models.

Didisheim, Antoine, Shikun Ke, Bryan T. Kelly, and Semyon Malamud, 2023b, Complexity in Factor Pricing Models.

Engelberg, Joseph E., and Christopher A. Parsons, 2011, The Causal Impact of Media in Financial Markets, *The Journal of Finance* 66, 67–97.

Fang, Lily, and Joel Peress, 2009, Media coverage and the cross-section of stock returns, *The journal of finance* 64, 2023–2052.

Fisher, Adlai, Charles Martineau, and Jinfei Sheng, 2022, Macroeconomic attention and announcement risk premia, *The Review of Financial Studies* 35, 5057–5093.

Fisher, Irving, 1922, *The Making of Index Numbers: A Study of Their Varieties, Tests, and Reliability* (Houghton Mifflin).

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, 2019, Text as data, *Journal of Economic Literature* 57, 535–574.

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus, 2021, Five facts about beliefs and portfolios, *American Economic Review* 111, 1481–1522.

Goetzmann, William N., Roger G. Ibbotson, and Liang Peng, 2001, A new historical database for the NYSE 1815 to 1925: Performance and predictability 4, 1–32.

Greenwood, Robin, and Andrei Shleifer, 2014, Expectations of Returns and Expected Returns, *The Review of Financial Studies* 27, 714–746.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *The Review of Financial Studies* 33, 2223–2273.

Hanley, Kathleen Weiss, and Gerard Hoberg, 2019, Dynamic interpretation of emerging risks in the financial sector, *The Review of Financial Studies* 32, 4543–4603.

Hartzmark, Samuel M., and David H. Solomon, 2022, Reconsidering returns, *The Review of Financial Studies* 35, 343–393.

Hirshleifer, David, Dat Mai, and Kuntara Pukthuanthong, 2025, War discourse and disaster premium: 160 years of evidence from the stock market, *The Review of Financial Studies* 38, 457–506.

Hirshleifer, David A, Dat Mai, and Kuntara Pukthuanthong, 2024, War discourse and the cross-section of expected stock returns, *Journal of Finance, Forthcoming* .

Hoberg, Gerard, and Asaf Manela, 2025, The natural language of finance, *Foundations and Trends in Finance, Forthcoming* .

Hoberg, Gerard, and Gordon Phillips, 2016, Text-based network industries and endogenous product differentiation, *Journal of political economy* 124, 1423–1465.

Hoberg, Gerard, and Gordon M. Phillips, 2018, Text-based industry momentum, *Journal of Financial and Quantitative Analysis* 53, 2355–2388.

Hoberg, Gerard, and Gordon M Phillips, 2025, Scope, scale, and concentration: The 21st-century firm, *The Journal of Finance* 80, 415–466.

Hou, Kewei, 2007, Industry information diffusion and the lead-lag effect in stock returns, *The review of financial studies* 20, 1113–1138.

Huberman, Gur, and Tomer Regev, 2001, Contagious speculation and a cure for cancer: A nonevent that made stock prices soar, *The Journal of Finance* 56, 387–396.

Jeon, Yoontae, Thomas H. McCurdy, and Xiaofei Zhao, 2022, News as sources of jumps in stock returns: Evidence from 21 million news articles for 9000 companies, *Journal of Financial Economics* 145, 1–17.

Kelly, Bryan, Semyon Malamud, and Kangying Zhou, 2024, The Virtue of Complexity in Return Prediction, *The Journal of Finance* 79, 459–503.

Kelly, Bryan, Dacheng Xiu, et al., 2023, Financial machine learning, *Foundations and Trends® in Finance* 13, 205–363.

Kelly, Bryan T., Semyon Malamud, and Kangying Zhou, 2022, The Virtue of Complexity Everywhere.

Lee, Charles MC, Terrence Tianshuo Shi, Stephen Teng Sun, and Ran Zhang, 2024, Production complementarity and information transmission across industries, *Journal of Financial Economics* 155, 103812.

Lee, Charles MC, Stephen Teng Sun, Rongfei Wang, and Ran Zhang, 2019, Technological links and predictable returns, *Journal of Financial Economics* 132, 76–96.

Li, Bin, Alberto Rossi, S Yan, and Lingling Zheng, 2025, Machine learning from a "universe" of signals: The role of feature engineering, *Journal of Financial Economics, Forthcoming* .

Liu, Yukun, and Ben Matthies, 2022, Long-run risk: Is it there?, *The Journal of Finance* 77, 1587–1633.

Lo, Andrew W., 2016, What is an index? .

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

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

Malmendier, Ulrike, and Stefan Nagel, 2011, Depression babies: Do macroeconomic experiences affect risk taking?, *The Quarterly Journal of Economics* 126, 373–416.

Manela, Asaf, and Alan Moreira, 2017, News implied volatility and disaster concerns, *Journal of Financial Economics* 123, 137–162.

Menzly, Lior, and Oguzhan Ozbas, 2010, Market Segmentation and Cross-predictability of Returns, *The Journal of Finance* 65, 1555–1580.

Nagel, Stefan, 2025, Seemingly Virtuous Complexity in Return PredictionResearch, *Chicago Booth Working Paper* .

Peress, Joel, 2008, Media coverage and investors' attention to earnings announcements, *Available at SSRN 1106475* .

Shiller, Robert J., 2017, Narrative Economics, *American Economic Review* 107, 967–1004.

Shiller, Robert J, 2020, Narrative economics: How stories go viral and drive major economic events .

Solomon, David H, Eugene Soltes, and Denis Sosyura, 2014, Winners in the spotlight: Media coverage of fund holdings as a driver of flows, *Journal of Financial Economics* 113, 53–72.

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

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

van Binsbergen, Jules H, Svetlana Bryzgalova, Mayukh Mukhopadhyay, and Varun Sharma, 2024, (almost) 200 years of news-based economic sentiment, *NBER Working Paper* .

Welch, Ivo, and Amit Goyal, 2008, A Comprehensive Look at The Empirical Performance of Equity Premium Prediction, *The Review of Financial Studies* 21, 1455–1508.

Figure 1: Average News Similarity of "SeenIt" Historical Months

This figure plots the average text-similarity of the 25 most similar historical months and the current month (*SeenItFamiliarity*) from January 1825 to December 2021.

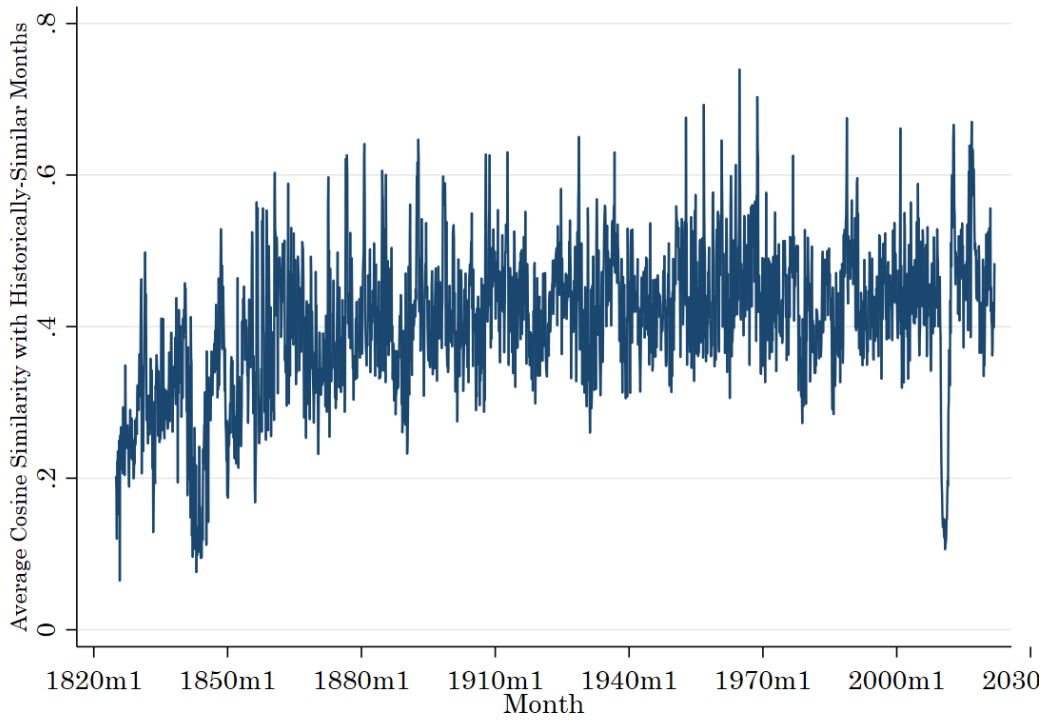
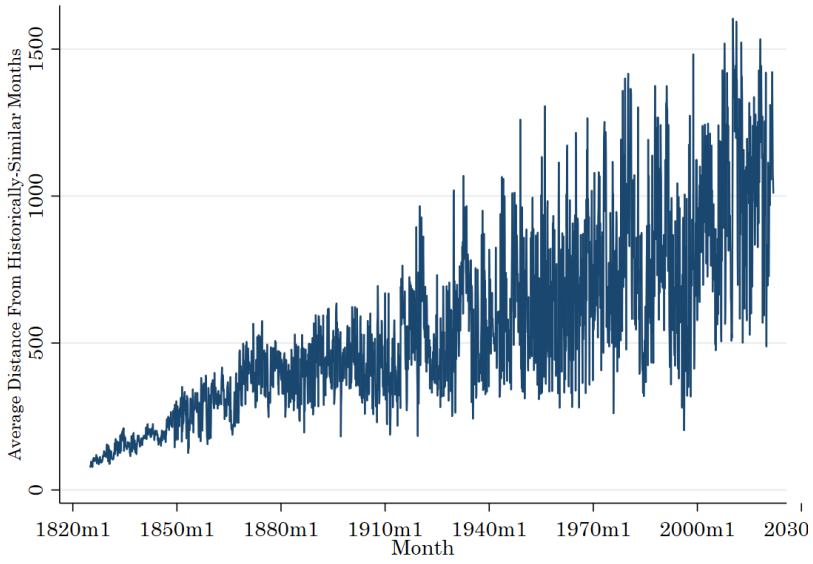


Figure 2: Average Time Distance from "SeenIt" Historical Months

This figure plots the average time length from the 25 most similar historical months to the current month from 1825 to 2021. In Panel B, we detrend this distance by first regressing the raw months on a linear month variable of the current month and then taking the regression residuals as the detrended distance. Finally, I added the sample mean of the original distance (542 months) to the detrended distance to shift the level.

Panel A: Raw Distance



Panel B: Detrended Distance

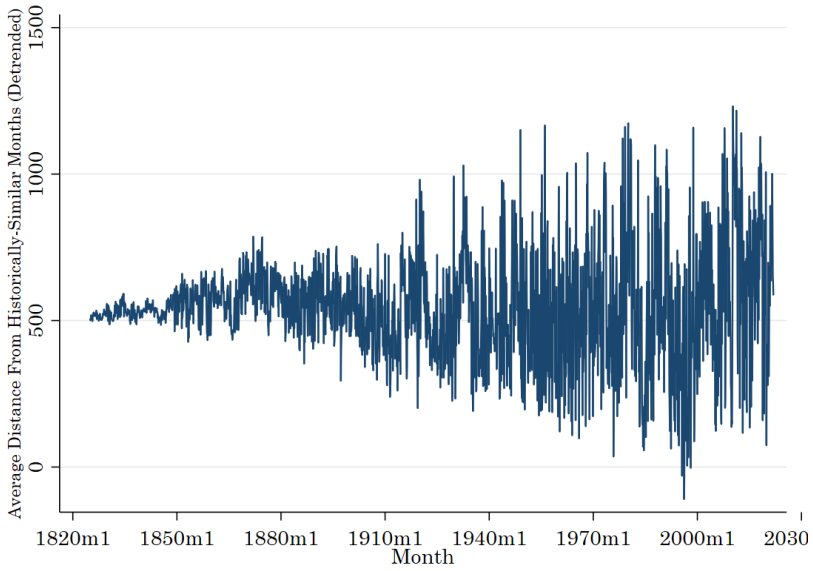


Figure 3: Stock Market Return and SeenItRet

This figure plots the 36-month backward moving average of monthly stock market return and monthly *SeenItRet*. *SeenItRet* for each month is the average of the monthly returns of the 25 most similar historical months based on the past month's news (see Section 2). *PWRet*, *EWRet*, and *VWRet* are price-weighted, equal-weighted, and value-weighted monthly stock returns and their corresponding *SeenItRet*s, respectively. See Table 1 for more details.

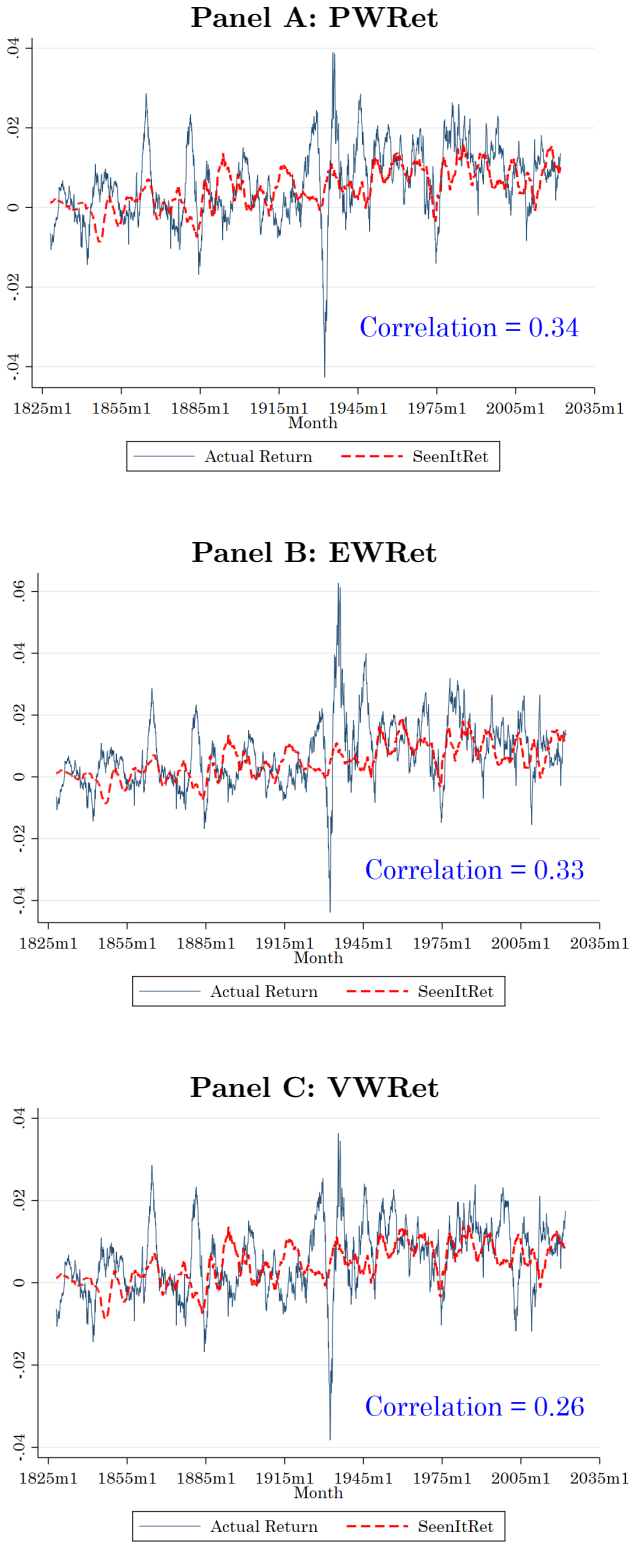
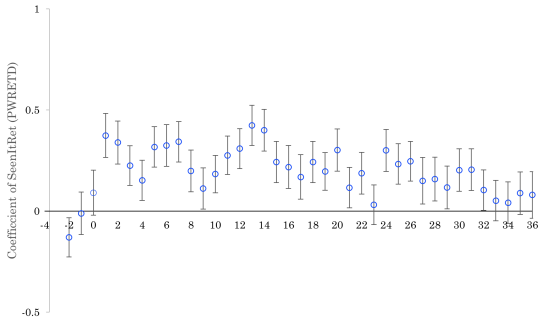


Figure 4: Long-Term Return Predictability of SeenItRet

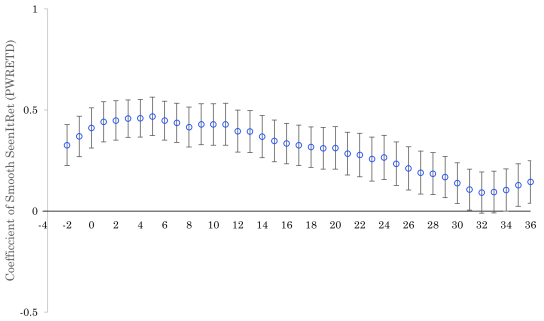
This figure plots the regression of monthly stock returns on *SeenItRet*, measured based on the average of next-month returns over 25 historical months with the highest news-based similarities to month t. PEWRet, EWRet, and VWRet are price-weighted, equal-weighted, and value-weighted stock returns, respectively (see Table 1). In the figures on the left side, each circle represents the regression coefficient of monthly stock returns at $t+i$ on SeenItRet measured at t. In the figures on the right side, each circle represents the regression coefficient of monthly stock returns at $t+i$ on the past-12-month Smooth SeenItRet at t. SeenItRet and Smooth SeenItRet are both standardized, and the dependent variables are in percentages. The vertical bars represent Newey-West adjusted standard errors with a 3-month lag. The sample period is from January 1825 to December 2021.

Panel A: Long-Term Prediction of PWRet

(1) Predicting using SeenItRet

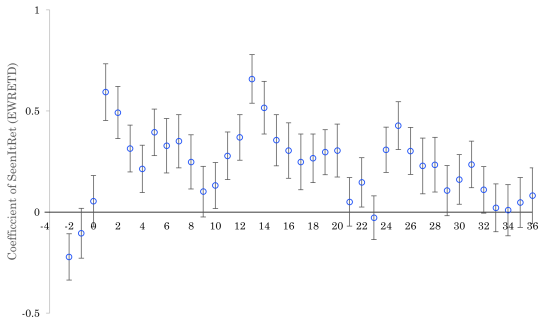


(2) Predicting using Smooth SeenItRet

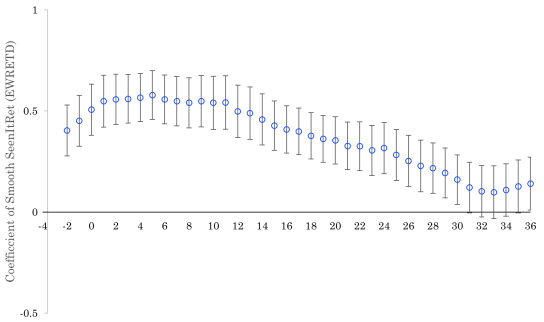


Panel B: Long-Term Prediction of EWRet

(1) Predicting using SeenItRet

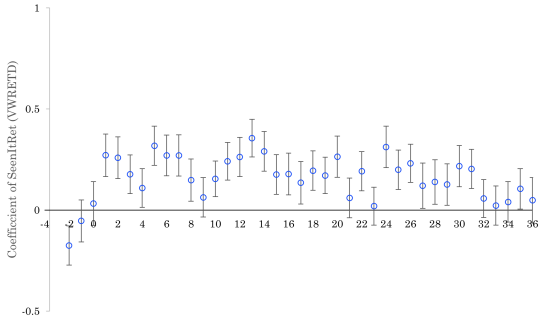


(2) Predicting using Smooth SeenItRet



Panel C: Long-Term Prediction of VWRet

(1) Predicting using SeenItRet



(2) Predicting using Smooth SeenItRet

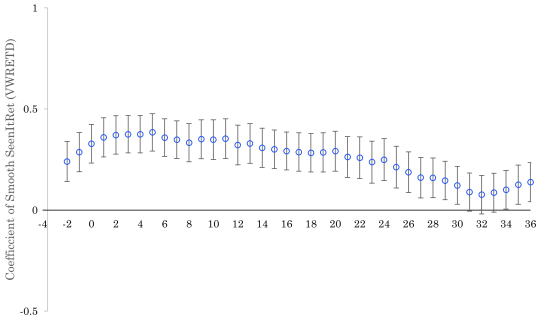


Table 1
Summary Statistics

This table presents the summary statistics for our dependent and independent variables. All dependent variables are as of next month (or next year for Patent). In Panel B, the variables labeled as *SeenItRet_* and *SeenItRisk_* are the mean and standard deviation of the corresponding next-month returns from the 25 most similar historical months to the current month (see Section 2). *Smooth SeenItRet_* are the 12-month moving averages of the corresponding SeenItRets. *SeenItFamiliarity* is the average of the Cosine similarity of the 25 most similar historical months to the current month. *SeenitVar_* is defined similarly to *SeenItRet_* for economic variables other than stock returns. *PWRet*, *EWRet*, and *VWRet* are price-weighted, equal-weighted, and value-weighted monthly stock returns, respectively. We calculate PWRet using the CRSP database and obtain EWRet and VWRet directly from the database. For periods prior to CRSP (January 1926), we impute all three return indices using the price-weighted stock return index (2020 version) downloaded from William Goetzmann's website. *Treas.Yield* is the U.S. 10-year treasury note yield (constant maturity) from St. Louis Fed, with the historical data prior to January 1925 extrapolated from the Business Insider publication. *Volatility* is the within-month standard deviation of daily stock returns, where daily data from 1885 to 1925 are from William Schwert's website and CRSP afterwards. *Recession* is the monthly NBER recession indicator with the historical data before 1855 obtained from Wikipedia. *Inflation* is the log difference in monthly CPI from 1871 to 2021 downloaded from Robert Shiller's website. *Patent* is the log difference in yearly number of patent applications from the USPTO website. Our final sample is from January 1825 to December 2021.

Variable	Mean	SD	P10	Median	P90	# Obs.
Panel A: Dependent Variables						
PWRet	0.0061	0.0474	-0.0426	0.0062	0.0558	2,364
EWRet	0.0070	0.0572	-0.0474	0.0058	0.0626	2,364
VWRet	0.0056	0.0465	-0.0429	0.0055	0.0539	2,364
Treas.Yield	0.0456	0.0207	0.0248	0.0423	0.0716	2,364
Volatility	0.0084	0.0055	0.0041	0.0069	0.0141	1,554
Recession	0.3524	0.4778	0.0000	0.0000	1.0000	2,364
Inflation	0.0020	0.0097	-0.0084	0.0017	0.0113	1,717
Patent	0.0344	0.1001	-0.0555	0.0241	0.1339	174
Panel B: Independent Variables						
SeenItRet_PWRet	0.0047	0.0103	-0.0074	0.0035	0.0186	2,364
SeenItRet_EWRet	0.0052	0.0118	-0.0078	0.0034	0.0205	2,364
SeenItRet_VWRet	0.0045	0.0099	-0.0071	0.0035	0.0180	2,364
Smooth SeenItRet_PWRet	0.0047	0.0061	-0.0029	0.0044	0.0130	2,364
Smooth SeenItRet_EWRet	0.0051	0.0068	-0.0029	0.0043	0.0147	2,364
Smooth SeenItRet_VWRet	0.0045	0.0059	-0.0029	0.0043	0.0124	2,364
SeenItRisk_PWRet	0.0364	0.0143	0.0180	0.0352	0.0550	2,364
SeenItRisk_EWRet	0.0396	0.0198	0.0180	0.0370	0.0609	2,364
SeenItRisk_VWRet	0.0358	0.0137	0.0180	0.0346	0.0536	2,364
SeenItFamiliarity	0.4079	0.0934	0.2925	0.4144	0.5166	2,364
SeenItVar_Treas.Yield	0.0464	0.0082	0.0364	0.0470	0.0550	2,364
SeenItVar_Volatility	0.0079	0.0019	0.0061	0.0074	0.0102	1,554
SeenItVar_Recess	0.5059	0.2263	0.2400	0.4800	0.8000	2,364
SeenItVar_Inflation	0.0004	0.0053	-0.0061	0.0009	0.0059	1,717
SeenItVar_Patent	0.0389	0.0380	0.0035	0.0299	0.0911	174

Table 2
Return Predictability of SeenItRet

This table shows our main prediction regression of SeenItRet. In Panel A, the dependent variable is next month's return, and the independent variable, *SeenItRet*, is the simple average of the next month's return of the 25 most similar (based on news text) historical months of the focal month. *Smooth SeenItRet* is the moving average of SeenItRet over the past 12 months (including the current month). In Panel B, we report the placebo test by regressing next-month stock return on a placebo SeenItRet, which moves each of the 25 historically similar months' future returns one year backward. The placebo Smooth SeenItRet is the 12-month moving average of the placebo SeenItRet. All three stock return indices (PWRet, EWRet, and VWRet) prior to CRSP (January 1926) use the price-weighted stock return index (2020 version) downloaded from William Goetzmann's website. t-statistics based on Newey-West adjusted standard errors with 3-month legs are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from January 1825 to December 2021.

Panel A: Main Test of SeenItRet

	PWRet		EWRet		VWRet	
	(1)	(2)	(3)	(4)	(5)	(6)
SeenItRet	0.364***		0.504***		0.274***	
	(3.420)		(4.240)		(2.583)	
Smooth SeenItRet		0.717***		0.802***		0.611***
		(4.457)		(4.295)		(3.705)
Observations	2364	2364	2364	2364	2364	2364
R^2	0.006	0.009	0.011	0.009	0.003	0.006

Panel B: Placebo Test by Perturbing SeenItRet

	PWRet		EWRet		VWRet	
	(1)	(2)	(3)	(4)	(5)	(6)
SeenItRet	0.121		0.171		0.100	
(Placebo)	(1.105)		(1.462)		(0.891)	
Smooth SeenItRet		0.249		0.165		0.286
(Placebo)		(1.398)		(0.858)		(1.642)
Observations	2364	2364	2364	2364	2364	2364
R^2	0.001	0.001	0.001	0.000	0.000	0.001

Table 3
Out of Sample Test Compared with the Historical Mean Model

This table presents statistics on forecast errors in-sample (IS) and out-of-sample (OOS) based on Welch and Goyal (2008) for monthly aggregate stock return forecasts using the corresponding *SeenItRet* and *Smooth SeenItRet* described in Table 2. *Sample Begin* denotes the beginning of the full sample from January 1825 to December 2021, while *Forecast Begin* denotes the beginning of the forecast sample for OOS tests from January 1926 to December 2021 (Campbell and Thompson (2008)). *In-Sample t-stat.* and *In-Sample* R^2 are the t-statistics and R^2 of regressing future monthly returns on predictors in the full sample, respectively, where the t-statistics are based on the Newey-West adjusted standard errors with 3-month lags. *Out-of-Sample* R^2 is the OOS R^2 constructed following Welch and Goyal (2008) which compares the predictability of the SeenItRet model with the historical mean model in the forecast period (see Section 3 for details).

	Sample Begin	Forecast Begin	In-Sample t-stat.	In-Sample R^2	Out-of-Sample R^2
			Panel A. PWRet		
SeenItRet	1825m1	1926m1	3.420	0.60%	0.75%
Smooth SeenItRet	1825m1	1926m1	4.457	0.90%	1.00%
			Panel B. EWRet		
SeenItRet	1825m1	1926m1	4.240	1.10%	1.20%
Smooth SeenItRet	1825m1	1926m1	4.295	0.90%	0.94%
			Panel C. VWRet		
SeenItRet	1825m1	1926m1	2.583	0.30%	0.33%
Smooth SeenItRet	1825m1	1926m1	3.705	0.60%	0.59%

Table 4
Interaction of SeenItRet with SeenItFamiliarity

This table shows the results of regressing next month's monthly return on the interaction of *SeenItRet* and a familiarity measure about the similarities of the historical 25 months, *SeenItFamiliarity*. *SeenItRet* is the simple average of the next month's return of the 25 most similar (based on news text) historical months of the focal month. *SeenItFamiliarity* is the average of the similarities of the 25 most similar historical months. See details of specifications in Table 2. t-statistics based on Newey-West adjusted standard errors with 3-month legs are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from January 1825 to December 2021.

	(1) PWRet	(2) EWRet	(3) VWRet
SeenItRet	1.642***	1.888***	1.463**
	(2.800)	(2.633)	(2.461)
(1-SeenItFamiliarity)	-0.023**	-0.017	-0.023**
	(-2.269)	(-1.436)	(-2.263)
SeenItRet × (1-SeenItFamiliarity)	-2.429**	-2.568**	-2.259**
	(-2.398)	(-2.077)	(-2.169)
Observations	2364	2364	2364
R^2	0.012	0.014	0.008

Table 5
Long-Term Return Predictability of SeenItRet

This table reports the regression of monthly stock returns from $t-2$ to $t+36$ on *SeenItRet* measured at t in Panel A, on the past-12-month Smooth SeenItRet measured at t without and with controlling for time trend (month) in Panels B and C, respectively. See details of specifications in Table 2, which focuses on predicting short-term returns at $t+1$. The *SeenItRet* is standardized, and the dependent variables are in percentages. Newey-West adjusted standard errors with 3-month legs are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from January 1825 to December 2021. See Figure 4 for a visualization of the coefficients.

Panel A: Predictability of SeenItRet

	PWRet		EWRet		VWRet	
	Coeff.	s.e.	Coeff.	s.e.	Coeff.	s.e.
t-2	−0.13	(0.10)	−0.22*	(0.12)	−0.18*	(0.10)
t-1	−0.01	(0.11)	−0.11	(0.12)	−0.05	(0.10)
t	0.09	(0.11)	0.05	(0.13)	0.03	(0.11)
t+1	0.37***	(0.11)	0.59***	(0.14)	0.27***	(0.10)
t+2	0.34***	(0.11)	0.49***	(0.13)	0.26**	(0.10)
t+3	0.23**	(0.10)	0.31***	(0.12)	0.18*	(0.10)
t+4	0.15	(0.10)	0.21*	(0.12)	0.11	(0.10)
t+5	0.32***	(0.10)	0.39***	(0.12)	0.32***	(0.10)
t+6	0.32***	(0.10)	0.33**	(0.14)	0.27***	(0.10)
t+7	0.34***	(0.10)	0.35***	(0.13)	0.27***	(0.10)
t+8	0.20*	(0.10)	0.25*	(0.13)	0.15	(0.10)
t+9	0.11	(0.10)	0.10	(0.13)	0.06	(0.10)
t+10	0.18**	(0.09)	0.13	(0.11)	0.15*	(0.09)
t+11	0.28***	(0.10)	0.28**	(0.12)	0.24***	(0.09)
t+12	0.31***	(0.10)	0.37***	(0.11)	0.26***	(0.10)
t+13	0.42***	(0.10)	0.66***	(0.12)	0.36***	(0.09)
t+14	0.40***	(0.10)	0.52***	(0.13)	0.29***	(0.10)
t+15	0.24**	(0.10)	0.35***	(0.13)	0.18*	(0.10)
t+16	0.22**	(0.11)	0.30**	(0.14)	0.18*	(0.10)
t+17	0.17	(0.11)	0.25*	(0.14)	0.13	(0.11)
t+18	0.24**	(0.10)	0.27**	(0.12)	0.19**	(0.10)
t+19	0.20**	(0.09)	0.30***	(0.11)	0.17*	(0.09)
t+20	0.30***	(0.10)	0.30**	(0.13)	0.26***	(0.10)
t+21	0.12	(0.10)	0.05	(0.12)	0.06	(0.10)
t+22	0.19*	(0.10)	0.15	(0.12)	0.19**	(0.10)
t+23	0.03	(0.10)	−0.03	(0.11)	0.02	(0.09)
t+24	0.30***	(0.10)	0.31***	(0.11)	0.31***	(0.10)
t+25	0.23**	(0.10)	0.43***	(0.12)	0.20**	(0.10)
t+26	0.25**	(0.10)	0.30***	(0.12)	0.23**	(0.09)
t+27	0.15	(0.11)	0.23*	(0.14)	0.12	(0.11)
t+28	0.16	(0.11)	0.23*	(0.14)	0.14	(0.11)
t+29	0.12	(0.11)	0.11	(0.12)	0.13	(0.10)
t+30	0.20*	(0.11)	0.16	(0.12)	0.22**	(0.10)
t+31	0.20**	(0.10)	0.24**	(0.11)	0.20**	(0.10)
t+32	0.10	(0.10)	0.11	(0.11)	0.06	(0.09)
t+33	0.05	(0.10)	0.02	(0.12)	0.02	(0.10)
t+34	0.04	(0.10)	0.01	(0.13)	0.04	(0.10)
t+35	0.09	(0.10)	0.05	(0.12)	0.11	(0.10)
t+36	0.08	(0.11)	0.08	(0.14)	0.05	(0.11)

Table 5
Long-Term Return Predictability of SeenItRet —*Continued*

Panel B: Predictability of Smooth SeenItRet						
	PWRet		EWRet		VWRet	
	Coeff.	s.e.	Coeff.	s.e.	Coeff.	s.e.
t-2	0.33***	(0.10)	0.40***	(0.13)	0.24**	(0.10)
t-1	0.37***	(0.10)	0.45***	(0.13)	0.29***	(0.10)
t	0.41***	(0.10)	0.51***	(0.13)	0.33***	(0.10)
t+1	0.44***	(0.10)	0.55***	(0.13)	0.36***	(0.10)
t+2	0.45***	(0.10)	0.56***	(0.12)	0.37***	(0.10)
t+3	0.46***	(0.09)	0.56***	(0.12)	0.37***	(0.09)
t+4	0.46***	(0.09)	0.57***	(0.12)	0.37***	(0.09)
t+5	0.47***	(0.09)	0.58***	(0.12)	0.38***	(0.09)
t+6	0.45***	(0.10)	0.56***	(0.12)	0.36***	(0.09)
t+7	0.44***	(0.10)	0.55***	(0.12)	0.35***	(0.09)
t+8	0.41***	(0.10)	0.54***	(0.12)	0.33***	(0.09)
t+9	0.43***	(0.10)	0.55***	(0.13)	0.35***	(0.10)
t+10	0.43***	(0.10)	0.54***	(0.13)	0.35***	(0.10)
t+11	0.43***	(0.10)	0.54***	(0.13)	0.35***	(0.10)
t+12	0.39***	(0.10)	0.50***	(0.13)	0.32***	(0.10)
t+13	0.39***	(0.10)	0.49***	(0.13)	0.33***	(0.10)
t+14	0.37***	(0.10)	0.46***	(0.13)	0.31***	(0.10)
t+15	0.35***	(0.10)	0.43***	(0.12)	0.30***	(0.10)
t+16	0.33***	(0.10)	0.41***	(0.12)	0.29***	(0.09)
t+17	0.33***	(0.10)	0.40***	(0.11)	0.29***	(0.09)
t+18	0.32***	(0.10)	0.38***	(0.11)	0.28***	(0.10)
t+19	0.31***	(0.10)	0.36***	(0.12)	0.29***	(0.10)
t+20	0.31***	(0.10)	0.35***	(0.12)	0.29***	(0.10)
t+21	0.28***	(0.11)	0.33***	(0.12)	0.26***	(0.10)
t+22	0.28**	(0.11)	0.32***	(0.12)	0.26**	(0.10)
t+23	0.26**	(0.11)	0.31**	(0.12)	0.24**	(0.10)
t+24	0.27**	(0.11)	0.32**	(0.13)	0.25**	(0.10)
t+25	0.23**	(0.11)	0.28**	(0.13)	0.21**	(0.10)
t+26	0.21**	(0.11)	0.25**	(0.13)	0.19*	(0.10)
t+27	0.19*	(0.11)	0.23*	(0.13)	0.16	(0.10)
t+28	0.18*	(0.10)	0.22*	(0.12)	0.16	(0.10)
t+29	0.17*	(0.10)	0.19	(0.12)	0.15	(0.10)
t+30	0.14	(0.10)	0.16	(0.12)	0.12	(0.09)
t+31	0.11	(0.10)	0.12	(0.12)	0.09	(0.09)
t+32	0.09	(0.10)	0.10	(0.13)	0.08	(0.09)
t+33	0.09	(0.10)	0.10	(0.13)	0.09	(0.10)
t+34	0.10	(0.10)	0.11	(0.13)	0.10	(0.10)
t+35	0.13	(0.11)	0.13	(0.13)	0.13	(0.10)
t+36	0.14	(0.11)	0.14	(0.13)	0.14	(0.10)

Table 6
Top Themes Explaining Future Stock Returns (Various Horizons)

This table reports, for each individual economic theme, the fraction of R^2 that is lost when each theme is individually removed from the state space used to assess similarity of past states (compared to the baseline where all states are included). The R^2 examined is the one obtained from simple time series regressions predicting the next-month market-wide stock return (Panel A). We also consider next 6 month returns in Panel B and next 24 month returns in Panel C. In all cases, SeenItRet is the RHS variable (and we consider versions of SeenItRet for each theme based on recomputing SeenItRet after leaving out the given theme as noted above). In particular, we identify the R^2 when all themes are included in the model used to compute SeenItRet as in our baseline, and assess how much R^2 is lost when each theme is respectively removed from the state space (including its interactions with tone and uncertainty). We report only the Top 50 most important themes in each Panel as those that result in the most losses in adjusted R^2 when left out.

Panel A: Top 50 Explanatory Power Themes (1-Month Returns)

Top 50 Individual Terms	momentum (19.6%), reduce (19.6%), margin (19.1%), government (15.0%), scarcity (14.7%), entrepreneur (14.7%), subsidy (14.0%), wheat (14.0%), domestic (13.2%), efficiency (12.9%), embargo (12.2%), legislation (11.7%), manufacturing (11.6%), boycott (10.7%), index (10.7%), deregulation (10.3%), agriculture (10.3%), regulation (10.1%), trademark (10.0%), president (10.0%), future (9.9%), equity (9.6%), output (9.5%), land (9.4%), mortgage (9.3%), cost (9.1%), bust (9.0%), oil (8.4%), copyright (8.3%), auction (8.3%), yield (8.3%), boom (8.2%), monopoly (8.0%), recovery (7.9%), advertising (7.9%), socialism (7.8%), investing (7.7%), retirement (7.5%), congress (7.4%), medicaid (7.4%), household (7.4%), economy (7.1%), productivity (6.8%), surplus (6.5%), innovation (6.3%), earnings (6.1%), taxation (6.0%), euro (6.0%), gold (5.5%), distribution (5.3%)

Panel B: Top 50 Explanatory Power Themes (6-Month Returns)

Top 50 Individual Terms	momentum (18.2%), deregulation (11.2%), war (10.6%), consumer (10.2%), oil (10.0%), scarcity (8.8%), substitute (7.8%), socialism (7.5%), reduce (7.3%), charity (7.2%), medicaid (7.1%), congress (7.0%), uncertainty (6.5%), domestic (6.4%), aid (6.3%), future (6.2%), efficiency (5.9%), competition (5.0%), household (5.0%), retail (4.7%), land (4.5%), recovery (4.5%), currency (4.5%), nationalization (4.4%), offshore (4.4%), mortgage (4.2%), copyright (4.1%), index (3.7%), corruption (3.7%), cartel (3.4%), government (3.2%), cost (3.1%), embargo (2.8%), energy (2.5%), pricing (2.5%), investing (2.4%), distribution (2.2%), surplus (2.1%), treasury (2.1%), security (2.0%), commodity (1.9%), innovation (1.9%), contract (1.9%), subsidy (1.7%), advertising (1.6%), legislation (1.5%), auction (1.5%), insurance (1.5%), manufacturing (1.5%), dividend (1.2%)

Panel C: Top 50 Explanatory Power Themes (24-Month Returns)

Top 50 Individual Terms	inflation (7.1%), expansion (5.9%), currency (5.1%), bond (5.1%), competition (4.9%), wheat (4.5%), land (4.3%), nationalization (4.1%), euro (3.7%), bubble (3.5%), domestic (3.5%), congress (3.4%), energy (3.2%), legislation (3.2%), treasury (2.8%), earnings (2.7%), government (2.7%), household (2.5%), luxury (2.5%), socialism (2.5%), margin (2.4%), reduce (2.3%), war (2.3%), deficit (2.2%), antitrust (2.1%), gold (1.9%), tariff (1.9%), finance (1.8%), money (1.8%), medicaid (1.8%), recession (1.8%), swap (1.7%), momentum (1.7%), agriculture (1.4%), medicare (1.4%), nominal (1.2%), demand (1.2%), dividend (1.1%), rally (1.0%), efficiency (0.7%), uncertainty (0.6%), cost (0.5%), manufacturing (0.4%), policy (0.3%), equity (0.3%), boom (-0.1%), yield (-0.2%), deregulation (-0.5%), investing (-0.6%), boycott (-0.6%)

Table 7
Top Themes by Decade

This table reports, both overall and for each decade, the individual themes with the highest likely impact on monthly returns for each decade. We compute impact for a given theme is the average of the absolute value of the theme's standardized monthly value multiplied by the same-month return. Themes with high impact have extreme values when returns have extreme values and are likely important in how economic agents from the given period think about stock returns. This calculation is computed as an average overall (Full Sample row one below) and separately for each decade as noted.

Sample	Highest Impact Newspaper Themes
Full Sample	(1) momentum, (2) war, (3) inflation, (4) price, (5) fraud, (6) reduce, (7) rationing, (8) welfare, (9) nominal, (10) household, (11) medicaid, (12) swap, (13) currency, (14) expansion, (15) embargo, (16) devaluation, (17) depression, (18) future, (19) tariff, (20) treasury, (21) energy, (22) poverty, (23) stimulus, (24) margin, (25) recovery
1870s	(1) fraud, (2) antitrust, (3) bubble, (4) aid, (5) taxation, (6) recovery, (7) household, (8) spending, (9) bond, (10) transaction
1880s	(1) import, (2) real, (3) wage, (4) revenue, (5) antitrust, (6) innovation, (7) deregulation, (8) trust, (9) loan, (10) bust
1890s	(1) fraud, (2) cartel, (3) aid, (4) welfare, (5) liability, (6) bank, (7) emergency, (8) poverty, (9) uncertainty, (10) medicare
1900s	(1) household, (2) welfare, (3) oil, (4) manufacturing, (5) technology, (6) facility, (7) advertising, (8) durable, (9) medicare, (10) agriculture
1910s	(1) welfare, (2) medicare, (3) manufacturing, (4) contract, (5) labor, (6) productivity, (7) efficiency, (8) entrepreneur, (9) oil, (10) ngo
1920s	(1) domestic, (2) competition, (3) household, (4) embargo, (5) future, (6) technology, (7) momentum, (8) facility, (9) speculator, (10) devaluation
1930s	(1) swap, (2) inflation, (3) europe, (4) currency, (5) nominal, (6) subsidy, (7) margin, (8) tariff, (9) embargo, (10) stimulus
1940s	(1) war, (2) depression, (3) scarcity, (4) rationing, (5) household, (6) unemployment, (7) aid, (8) labor, (9) contract, (10) real
1950s	(1) boom, (2) war, (3) risk, (4) future, (5) corruption, (6) depression, (7) momentum, (8) fraud, (9) efficiency, (10) expansion
1960s	(1) future, (2) population, (3) efficiency, (4) productivity, (5) risk, (6) agriculture, (7) rationing, (8) momentum, (9) advertising, (10) depression
1970s	(1) energy, (2) consumption, (3) population, (4) oil, (5) reduce, (6) advertising, (7) surplus, (8) retail, (9) medicaid, (10) commodity
1980s	(1) reduce, (2) price, (3) growth, (4) enterprise, (5) trader, (6) recession, (7) investing, (8) gold, (9) retail, (10) margin
1990s	(1) depression, (2) embargo, (3) war, (4) future, (5) poverty, (6) momentum, (7) security, (8) uncertainty, (9) durable, (10) resistance
2000s	(1) recession, (2) poverty, (3) treasury, (4) stimulus, (5) economics, (6) economy, (7) foreclosure, (8) default, (9) luxury, (10) expansion
2010s	(1) import, (2) swap, (3) export, (4) distribution, (5) population, (6) trade, (7) substitute, (8) auction, (9) barter, (10) real

Table 8
Return Predictability of SeenItRisk

The dependent variable is next month's return. The independent variable, *SeenItRisk* is the standard deviation of the monthly returns of the 25 most similar historical months, and *SeenItRet* is the simple average of the next month's return of the 25 most similar (based on news text) historical months of the focal month, and *Volatility* is the current month's daily value-weighted return volatility. All three stock return indices (PWRet, EWRet, and VWRet) prior to CRSP (1926 Jan) use the price-weighted stock return index (2020 version) downloaded from William Goetzmann's website. *t*-statistics based on Newey-West adjusted standard errors with 3-month legs are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from January 1825 to December 2021, where the Volatility measure becomes available from February 1885.

	PWRet			EWRet			VWRet		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
SeenItRisk	0.358***	0.214**	0.417**	0.352***	0.200**	0.346**	0.306***	0.185**	0.298*
	(4.132)	(2.336)	(2.501)	(4.432)	(2.135)	(2.477)	(3.541)	(2.037)	(1.767)
SeenItRet		0.559***			0.561**			0.480***	
		(3.245)			(2.546)			(2.730)	
Volatility			-2.593***			-2.475***			-2.458***
			(-4.848)			(-3.154)			(-4.753)
Observations	2364	2364	1644	2364	2364	1644	2364	2364	1644
R^2	0.006	0.010	0.081	0.008	0.011	0.049	0.004	0.007	0.075

Table 9
Predictability of Other Economic Variables

This table shows the results of regressing next month's or next year's economic variables on *SeenItVar* measure at this month or year. *SeenItVar* is the simple average of the next month's or next year's economic variable of the 25 most similar (based on news text) historical months of the focal month. *Treas.Yield* is the U.S. 10-year treasury note yield (constant maturity) from St. Louis Fed, with the historical data prior to January 1925 extrapolated from the Business Insider publication. *Volatility* is the within-month standard deviation of daily stock returns, where daily data from 1885 to 1925 are from William Schwert's website and CRSP afterwards. *Recession* is the monthly NBER recession indicator with the historical data before 1855 obtained from Wikipedia. *Inflation* is the log difference in monthly CPI from 1871 to 2021 downloaded from Robert Shiller's website. *Patent* is the log difference in yearly number of patent applications from the USPTO website. t-statistics based on Newey-West adjusted standard errors with 3 lags are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from 1825 to 2021, which may vary depending on the availability of the respective measure.

	Panel A: Monthly				**Panel B: Yearly**
	Treas.Yield	Volatility	Recession	Inflation	Patent
	(1)	(2)	(3)	(4)	(5)
SeenItVar	0.267***	0.547***	0.675***	0.254***	0.703**
	(3.654)	(3.859)	(10.618)	(3.602)	(2.131)
Observations	2364	1554	2364	1717	174
R^2	0.011	0.038	0.102	0.019	0.071

Internet Appendix for

"Haven't We Seen This Before? Return Predictions from 200 Years of News"

AJ Chen Gerard Hoberg Miao Ben Zhang

Table IA.1
Robustness: SeenItRet Based on Alternative Peer Months

This table shows that our main prediction regression of SeenItRet in Table 2 is robust to alternative choices of the most similar historical months. The dependent variable is next month's return. Panel A defines *SeenItRet* as the simple average of the next month's return of the top 15 (instead of 25 in our baseline analysis) most similar historical months of the focal month. Panel B defines *SeenItRet* using the top 35 most similar historical months. *Smooth SeenItRet* is the moving average of SeenItRet over the past 12 months (including the current month). All three stock return indices (PWRet, EWRet, and VWRet) prior to CRSP (January 1926) use the price-weighted stock return index (2020 version) downloaded from William Goetzmann's website. t-statistics based on Newey-West adjusted standard errors with 3-month legs are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from January 1825 to December 2021.

Panel A: SeenItRet Based on the Top 15 Peer Months

	PWRet		EWRet		VWRet	
	(1)	(2)	(3)	(4)	(5)	(6)
SeenItRet	0.334***		0.428***		0.269***	
	(3.919)		(4.605)		(3.157)	
Smooth SeenItRet		0.631***		0.718***		0.541***
		(4.211)		(4.124)		(3.524)
Observations	2364	2364	2364	2364	2364	2364
R^2	0.007	0.008	0.011	0.009	0.005	0.006

Panel B: SeenItRet Based on the Top 35 Peer Months

	PWRet		EWRet		VWRet	
	(1)	(2)	(3)	(4)	(5)	(6)
SeenItRet	0.420***		0.565***		0.308**	
	(3.345)		(4.161)		(2.483)	
Smooth SeenItRet		0.798***		0.887***		0.671***
		(4.739)		(4.683)		(3.961)
Observations	2364	2364	2364	2364	2364	2364
R^2	0.006	0.009	0.010	0.010	0.003	0.006

Table IA.2
Robustness: Excess Return Predictability of SeenItRet

This table shows our prediction regression of SeenItRet on excess stock returns. In Panel A, the dependent variable is next month's excess return, and the independent variable, *SeenItRet*, is the simple average of the next month's excess return of the 25 most similar (based on news text) historical months of the focal month. *Smooth SeenItRet* is the moving average of SeenItRet over the past 12 months (including the current month). In Panel B, we report the placebo test by regressing next-month stock return on a placebo SeenItRet, which moves each of the 25 historically similar months' future excess returns one year backward. The placebo Smooth SeenItRet is the moving average of the placebo SeenItRet over the past 12 months (including the current month). *Excess returns* are monthly returns minus the risk-free rate constructed following Welch and Goyal (2008), where the risk-free rate is the T-bill rate from January 1920 to December 2021, the Commercial Paper Rate of New York from January 1857 to December 1919, the Great Britain Open Market Rate from January 1824 to December 1856, and zero before January 1824. All three stock return indices (PWRet, EWRet, and VWRet) prior to CRSP (January 1926) use the price-weighted stock return index (2020 version) downloaded from William Goetzmann's website. t-statistics based on Newey-West adjusted standard errors with 3-month legs are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from January 1825 to December 2021.

Panel A: SeenItRet and Excess Returns

	PWRet		EWRet		VWRet	
	(1)	(2)	(3)	(4)	(5)	(6)
SeenItRet	0.343***		0.484***		0.255**	
	(3.249)		(4.122)		(2.421)	
Smooth SeenItRet		0.704***		0.787***		0.589***
		(4.304)		(4.164)		(3.523)
Observations	2364	2364	2364	2364	2364	2364
R^2	0.005	0.008	0.010	0.008	0.003	0.005

Panel B: Placebo Test by Perturbing SeenItRet for Excess Returns

	PWRet		EWRet		VWRet	
	(1)	(2)	(3)	(4)	(5)	(6)
SeenItRet	0.092		0.142		0.074	
(Placebo)	(0.823)		(1.193)		(0.648)	
Smooth SeenItRet		0.166		0.085		0.208
(Placebo)		(0.870)		(0.407)		(1.096)
Observations	2364	2364	2364	2364	2364	2364
R^2	0.000	0.000	0.001	0.000	0.000	0.001

Table IA.3
Robustness: Out of Sample Test of Predicting Excess Returns

This table presents statistics on forecast errors in-sample (IS) and out-of-sample (OOS) based on Welch and Goyal (2008) for monthly excess stock return forecasts using the corresponding *SeenItRet* and *Smooth SeenItRet* constructed using historical excess returns (see Section 3). The dependent variable, *Excess Returns*, is monthly returns minus the risk-free rate constructed following Welch and Goyal (2008), where the risk-free rate is the T-bill rate from January 1920 to December 2021, the Commercial Paper Rate of New York from January 1857 to December 1919, the Great Britain Open Market Rate from January 1824 to December 1856, and zero before January 1824. *Sample Begin* denotes the beginning of the full sample from January 1825 to December 2021, while *Forecast Begin* denotes the beginning of the forecast sample for OOS tests from January 1926 to December 2021 (Campbell and Thompson (2008)). *In-Sample t-stat.* and *In-Sample* R^2 are the t-statistics and R^2 of regressing future monthly excess returns on predictors in the full sample, respectively, where the t-statistics are based on the Newey-West adjusted standard errors with 3-month lags. *Out-of-Sample* R^2 is the OOS R^2 constructed following Welch and Goyal (2008) which compares the predictability of the SeenItRet model with the historical mean model in the forecast period.

	Sample Begin	Forecast Begin	In-Sample t-stat.	In-Sample R^2	Out-of-Sample R^2
			Panel A. PWRet		
SeenItRet	1825m1	1926m1	3.249	0.50%	0.66%
Smooth SeenItRet	1825m1	1926m1	4.304	0.80%	0.91%
			Panel B. EWRet		
SeenItRet	1825m1	1926m1	4.122	1.00%	1.10%
Smooth SeenItRet	1825m1	1926m1	4.164	0.80%	0.87%
			Panel C. VWRet		
SeenItRet	1825m1	1926m1	2.421	0.30%	0.27%
Smooth SeenItRet	1825m1	1926m1	3.523	0.50%	0.51%

Table IA.4
Robustness: Return Predictability of SeenItRet (Sub-Periods)

This table shows our main prediction regression of SeenItRet in two sub-periods of post-1876 and post-1926. The post-1876 period features a modernized financial system after several important reforms of the NYSE, including the introduction of the stock ticker in 1867 and the installation of telephones in 1878, which revolutionized market communications by enabling the quick transmission of market information across the United States. The post-1926 period refers to the time when the CRSP database became available. We end both subsamples at 2007, as the period following the 2008-09 financial crisis is known to be novel and unfamiliar to historical periods (see Figure 1 and Section 3.2), when SeenItRet is likely not to predict returns well. *SeenItRet*, is the simple average of the next month's return of the 25 most similar (based on news text) historical months of the focal month. All three stock return indices (PWRet, EWRet, and VWRet) prior to CRSP (January 1926) use the price-weighted stock return index (2020 version) downloaded from William Goetzmann's website. t-statistics based on Newey-West adjusted standard errors with 3-month legs are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

	Post-1876 Period			Post-1926 Period		
	PWRet	EWRet	VWRet	PWRet	EWRet	VWRet
	(1)	(2)	(3)	(4)	(5)	(6)
SeenItRet	0.414***	0.604***	0.310***	0.516***	0.719***	0.376**
	(3.419)	(4.420)	(2.587)	(3.073)	(4.032)	(2.228)
Observations	1584	1584	1584	984	984	984
R^2	0.008	0.015	0.004	0.010	0.017	0.005

Table IA.5
Interaction of SeenItRet with SeenItFamiliarity Ex. Early Years

This table shows the robustness of Table 4 in a sample that excludes the first 25 years in our original sample. It represents the results of regressing next month's monthly return on the interaction of *SeenItRet* and a familiarity measure about the similarities of the historical 25 months, *SeenItFamiliarity*. *SeenItRet* is the simple average of the next month's return of the 25 most similar (based on news text) historical months of the focal month. *SeenItFamiliarity* is the average of the similarities of the 25 most similar historical months. See details of specifications in Table 2. *t*-statistics based on Newey-West adjusted standard errors with 3-month legs are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from January 1850 to December 2021.

	(1) PWRet	(2) EWRet	(3) VWRet
SeenItRet	1.621**	1.891**	1.477**
	(2.458)	(2.362)	(2.208)
(1-SeenItFamiliarity)	-0.023*	-0.014	-0.022
	(-1.646)	(-0.839)	(-1.582)
SeenItRet × (1-SeenItFamiliarity)	-2.422**	-2.592*	-2.319*
	(-2.097)	(-1.855)	(-1.952)
Observations	2064	2064	2064
R^2	0.010	0.013	0.007

Table IA.6
Interaction of SeenItRet with Recency

This table shows the results of regressing next month's monthly return on the interaction of *SeenItRet*, which is the simple average of the next month's return of the 25 most similar historical months of the focal month, and *Recency*, which is the negative of the average (detrended) time distance by month in Panel A (in Panel B) from the 25 historical months and the focal month. Figure 2 plots the average time distance and detrended time distance for SeenItRet. Both measures of Recency are further standardized to have a mean of zero and a standard deviation of one. See details of specifications in Table 2. t-statistics based on Newey-West adjusted standard errors with 3-month legs are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from January 1825 to December 2021.

Panel A: Interacting SeenItRet with Recency

	PWRet	EWRet	VWRet
	(1)	(2)	(3)
SeenItRet	0.322***	0.451***	0.255***
	(3.246)	(4.115)	(2.594)
Recency	-0.003***	-0.002*	-0.003***
	(-2.785)	(-1.673)	(-2.780)
SeenItRet × Recency	0.038	-0.037	0.103
	(0.332)	(-0.311)	(0.870)
Observations	2364	2364	2364
R^2	0.009	0.012	0.006

Panel B: Interacting SeenItRet with Detrended Recency

	PWRet	EWRet	VWRet
	(1)	(2)	(3)
SeenItRet	0.361***	0.495***	0.271***
	(3.469)	(4.149)	(2.631)
Recency	-0.000	0.001	0.000
	(-0.078)	(0.415)	(0.158)
SeenItRet × Recency	-0.034	-0.086	-0.020
	(-0.325)	(-0.823)	(-0.187)
Observations	2364	2364	2364
R^2	0.006	0.011	0.003

Table IA.7
Return Predictability of Recent SeenItRet

This table shows the results of regressing next month's monthly return on an alternative version of SeenItRet based on averaging the next month's return of the 25 most similar (based on news text) historical months *within 50 years* of the focal month (*Recent SeenItRet*). *Smooth Recent SeenItRet* is the moving average of Recent SeenItRet over the past 12 months (including the current month). See details of specifications in Table 2. t-statistics based on Newey-West adjusted standard errors with 3-month legs are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from January 1825 to December 2021.

	PWRet		EWRet		VWRet	
	(1)	(2)	(3)	(4)	(5)	(6)
Recent SeenItRet	0.347***		0.420***		0.270***	
	(3.777)		(4.391)		(3.077)	
Smooth Recent SeenItRet		0.603***		0.596***		0.525***
		(4.831)		(4.439)		(4.003)
Observations	2364	2364	2364	2364	2364	2364
R^2	0.007	0.009	0.010	0.008	0.004	0.006

Table IA.8
Robustness: Return Predictability of Alternative SeenItRet

This table reports the robustness of Table 5 by regressing monthly stock returns from $t-2$ to $t+36$ on an alternative version of the *Smooth SeenItRet* measured at t, which is constructed using the business news themes from Bybee et al. (2024) instead of our chosen themes in Section 2. The alternative Smooth SeenItRet is standardized, and the dependent variables are in percentages. Newey-West adjusted standard errors with 3-month legs are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from January 1825 to December 2021.

	PWRet		EWRet		VWRet	
	Coeff.	s.e.	Coeff.	s.e.	Coeff.	s.e.
t-2	0.27***	(0.10)	0.38***	(0.13)	0.22**	(0.11)
t-1	0.29***	(0.10)	0.40***	(0.13)	0.24**	(0.10)
t	0.33***	(0.10)	0.42***	(0.12)	0.28***	(0.10)
t+1	0.35***	(0.10)	0.44***	(0.12)	0.30***	(0.10)
t+2	0.36***	(0.10)	0.45***	(0.12)	0.31***	(0.10)
t+3	0.38***	(0.10)	0.45***	(0.12)	0.32***	(0.10)
t+4	0.39***	(0.10)	0.46***	(0.12)	0.33***	(0.10)
t+5	0.40***	(0.10)	0.48***	(0.13)	0.34***	(0.10)
t+6	0.39***	(0.10)	0.45***	(0.12)	0.32***	(0.10)
t+7	0.38***	(0.10)	0.45***	(0.12)	0.32***	(0.10)
t+8	0.36***	(0.10)	0.43***	(0.12)	0.31***	(0.10)
t+9	0.34***	(0.10)	0.42***	(0.12)	0.30***	(0.10)
t+10	0.34***	(0.10)	0.42***	(0.12)	0.30***	(0.10)
t+11	0.35***	(0.10)	0.43***	(0.13)	0.31***	(0.10)
t+12	0.34***	(0.10)	0.42***	(0.13)	0.31***	(0.10)
t+13	0.32***	(0.10)	0.39***	(0.12)	0.29***	(0.10)
t+14	0.28***	(0.10)	0.35***	(0.12)	0.26***	(0.10)
t+15	0.24**	(0.10)	0.31**	(0.12)	0.22**	(0.10)
t+16	0.23**	(0.10)	0.29**	(0.12)	0.21**	(0.10)
t+17	0.22**	(0.10)	0.29**	(0.12)	0.20**	(0.10)
t+18	0.24**	(0.10)	0.31**	(0.12)	0.21**	(0.10)
t+19	0.25**	(0.10)	0.33***	(0.12)	0.23**	(0.10)
t+20	0.26**	(0.10)	0.36***	(0.12)	0.24**	(0.10)
t+21	0.28***	(0.11)	0.37***	(0.13)	0.26**	(0.10)
t+22	0.27**	(0.11)	0.36***	(0.13)	0.25**	(0.10)
t+23	0.26**	(0.11)	0.33**	(0.14)	0.24**	(0.10)
t+24	0.25**	(0.11)	0.31**	(0.14)	0.23**	(0.11)
t+25	0.25**	(0.11)	0.31**	(0.14)	0.23**	(0.11)
t+26	0.26**	(0.11)	0.31**	(0.13)	0.24**	(0.10)
t+27	0.28***	(0.11)	0.33***	(0.13)	0.26**	(0.10)
t+28	0.28***	(0.10)	0.34***	(0.12)	0.26***	(0.10)
t+29	0.25**	(0.10)	0.30**	(0.12)	0.24**	(0.10)
t+30	0.24**	(0.10)	0.27**	(0.12)	0.23**	(0.09)
t+31	0.21**	(0.10)	0.22*	(0.12)	0.20**	(0.09)
t+32	0.19*	(0.10)	0.17	(0.12)	0.18*	(0.09)
t+33	0.17*	(0.10)	0.13	(0.12)	0.17*	(0.09)
t+34	0.15	(0.10)	0.11	(0.12)	0.16*	(0.09)
t+35	0.14	(0.10)	0.10	(0.12)	0.15*	(0.09)
t+36	0.16	(0.09)	0.11	(0.12)	0.17*	(0.09)

Table IA.9
Robustness: Return Predictability of Theme-Weighted SeenItRet

This table shows the robustness of Table 2 by showing the return predictability based on an alternative SeenItRet. The dependent variable is next month's return. *SeenItRet (WT)* is the simple average of the next month's return of the 25 most similar historical months of the focal month, where the similarity is constructed by weighting each theme's relevance to the stock market as discussed in Section 3.4.3, rather than based on unweighted themes as in the baseline measure in equation (2). *Smooth SeenItRet (WT)* is the 12-month moving average of the alternative SeenItRet. All three stock return indices (PWRet, EWRet, and VWRet) prior to CRSP (January 1926) use the price-weighted stock return index (2020 version) downloaded from William Goetzmann's website. t-statistics based on Newey-West adjusted standard errors with 3-month legs are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. The sample period is from January 1825 to December 2021.

	PWRet		EWRet		VWRet	
	(1)	(2)	(3)	(4)	(5)	(6)
SeenItRet (WT)	0.318***		0.459***		0.247**	
	(3.104)		(3.984)		(2.466)	
Smooth SeenItRet (WT)		0.729***		0.813***		0.605***
		(4.933)		(4.802)		(4.123)
Observations	2364	2364	2364	2364	2364	2364
R^2	0.005	0.009	0.009	0.010	0.003	0.006

IA.1 List of Economic Themes

A key component of our measurement framework is the set of economic themes used to characterize the narrative content in newspaper articles. These themes serve as the conceptual building blocks for the monthly narrative state vectors in our economic spatial space. In this section, we outline the process by which these themes were constructed, curated, and refined.

We begin with a broad candidate set of economic topics drawn from multiple high-quality textual sources. These include economic terms organized by The Economist, the glossary of economics from Wikipedia, the glossary from the Federal Reserve Bank of St. Louis, the glossary from the Commodity Futures Trading Commission (CFTC), political terms from Baker et al. (2016) and commodity-related terms from the Corporate Finance Institute. The selection of textual sources covers encyclopedic content, professional media, and institutional reports, ensuring that the themes reflect both economic concepts and the broader set of terms used by policymakers, industry participants, and journalists. In addition, to be comprehensive, we asked ChatGPT for "200 distinct economically related terms that have been most frequently used over the years".

To ensure interpretability and thematic coherence, we apply a two-stage filtering and review process. First, themes that appear rarely in historical data are excluded. Specifically, we keep economic themes that occur more than five thousand times in at least one year in our sample. Intuitively, a word that never reached five thousand occurrences in any year throughout the two centuries is unlikely to be a significant enough narrative that drives overall stock returns. Second, financial researchers with domain expertise in finance and economics review the remaining themes for clarity and relevance. This review ensures that the final list reflects economically meaningful concepts and avoids redundancy.

The result is a curated set of 162 economic themes. They provide a conceptually grounded and empirically tractable framework for capturing the evolving content of economic and business news narratives over time:

{*advertising, agriculture, aid, antitrust, auction, automation, bank, bankruptcy, barter, bond, boom, borrower, boycott, broker, bubble, budget, bullion, bust, cartel, charity, coin, collateral, commodity, communism, competition, congress, consumer, consumption, contract, copyright, corruption, cost, credit, currency, debt, default, deficit, demand, deposit, depression, deregulation, devaluation, distribution, dividend, domestic, durable, earnings, economics, economy, efficiency, embargo, emergency, energy, enterprise, entrepreneur, equity, euro, expansion, expectation, export, facility, finance, foreclosure, fraud, future, gold, government, growth, household, import, income, index, inflation, infrastructure, innovation, institution, insurance, interest, inventory, investing, labor, land, lawsuit, lease, legislation, lender, liability, liquidation, loan, luxury, manufacturing, margin, medicaid, medicare, modelling, momentum, money, monopoly, mortgage, nationalization, ngo, nominal, offshore, oil, output, patent, peril, policy, population, portfolio, poverty, premium, president, price, pricing, productivity, profit, rally, rationing, real, recession, recovery, reduce, regulation, resistance, retail, retirement, revenue, risk, salary, scarcity, security, socialism, speculation, speculator, spending, stimulus, stock, subsidy, substitute, supply, surplus, swap, tariff, tax, taxation, technology, trade, trademark, trader, transaction, treasury, trust, uncertainty, unemployment, wage, war, wealth, welfare, wheat, wholesale, yield*}

For each of the 162 economic themes, we generate 100 related unigram keywords using OpenAI to facilitate the detection of the theme from our news corpus. We feed each theme word into the following prompt, which is chosen to encourage the generation of vocabulary representative of how economic topics are typically presented in general news coverage: *prompt = (f"Generate a ranked list of 200 unique single-word unigrams related to the topic of 'uni' that might appear in newspaper articles. Please provide each word on a separate line without any numbering. Exclude any bigrams or phrases. Please order them by relevance.").* We keep the first 100 unigrams for each theme. To conserve space in the draft, we provide the full list of 162 economic themes, along with their associated the 100 synonym unigrams per theme, at this link.