FMC²
Financial Mathematics and Computation Research Cluster

Michael J. Brennan Irish Finance Working Paper Series

sfi
Fondúireacht Eolaíochta Éireann
Science Foundation Ireland

UCD DUBLIN

NUI MAYNOOTH

DCU
Dublin City University

Bubbles Talk: Narrative Augmented Bubble Prediction

Yuting Chen* Don Bredin† Valerio Potì ‡

March 31, 2023

Abstract

Financial bubble theories emphasize the importance of behavioral mechanisms centered around investor beliefs, which can be potentially gleaned from prevailing narratives, that reflect investors' psychological states and link them to economic events. By summarizing market narratives into meaningful and economically relevant features, guided by bubble theories, we offer a novel approach to bubble prediction. We then test whether the variation of narratives and bubble measures are related on a predictive basis, as bubble theories imply. Our findings reveal that most of our narrative features exhibit statistically significant predictive power for bubble measures, and that the narrative-augmented models outperform non-augmented benchmarks in out-of-sample tests. These results offer new insights into the understanding of bubbles and lay the foundation for using narratives to develop early warning systems (EWS) for bubble formation and deflation, and for investigating the causal relationship between narratives and economic events.

Keywords: narrative economics, asset bubbles, natural language processing
JEL Classification: E71, C53, E37

*College of Business, University College Dublin, Dublin, Ireland; Rennes School of Business, Rennes, France. Email: yuting.chen@ucd.ie

†College of Business, University College Dublin, Dublin, Ireland. Email: don.bredin@ucd.ie

‡Corresponding author: Valerio Potì, College of Business, University College Dublin, Blackrock, Dublin, Ireland. Email: valerio.poti@ucd.ie. The authors gratefully acknowledge the support of Science Foundation Ireland under grant number 16/SPP/3347 and of H2020 COST Action CA19130, "Fintech and Artificial Intelligence in Finance - Towards a transparent financial industry".

1 Introduction

An asset bubble is typically referred to as the set of circumstances when the price of an asset greatly exceeds the value that can be justified based on rational (or even just reasonable) estimates of future cash flows (Flood and Hodrick (1990), De Long et al. (1990), Abreu and Brunnermeier (2003) and Aliber and Kindleberger (2015)). The bursting of asset price bubbles can negatively impact the financial system and potentially lead to systemic financial crises (as seen in Brunnermeier et al. (2020) and Aliber and Kindleberger (2015)). It is, therefore, important to monitor and predict bubbles in financial markets. Empirical real-time bubble detection and forecasting, however, are challenging.

This study explores the potential usefulness of narratives in forecasting the formation and evolution of bubbles, as well as to understand how market narratives and bubbles interact. Guided by bubble theories, we identify crucial narrative features that can be empirically estimated using natural language processing (NLP) tools aided by standard econometric identification arguments. Here, by estimation of the narrative features we mean their empirical measurement, albeit with sampling error. In a nutshell, we apply the NLP techniques to financial market news in order to construct a set of time series representing changes in investor perception. Via predictive regression analysis, we show that these narrative features can provide additional predictive power for both ex-ante and ex-post bubble measurements, beyond common financial and economic indicators, and that the signs of these predictive relations are consistent with implications of the bubble theories that we use to identify the features. Our findings highlight the significance of narratives as a valuable but previously neglected source of information in bubble prediction, contributing to the existing literature.

Theoretical work in the bubble literature suggest that bubbles can occur through various mechanisms such as self-fulfilling expectations (Flood and Hodrick, 1990), mispricing of fundamentals (Froot and Obstfeld, 1991), heterogeneous beliefs accompanied by limits to arbitrage (Scheinkman and Xiong, 2003), asymmetric information with disagreement on bubble timing (Abreu and Brunnermeier, 2003), extrapolation (Barberis et al., 2018) and fades (Shiller, 2015). A shared characteristic among the models is the emphasis on investor beliefs. Nonetheless, typical approaches to empirically identifying the presence of bubbles only utilize price data and observable financial or economic indicators. By using information contained in market narratives, distilled from media news, we complement this data with hitherto neglected information on investor beliefs.

The need for a good proxy of fundamental values does lead to difficulties in detecting the presence of bubbles. Firstly, such proxies usually reflect only past information, while prices and rational bubbles are determined by the expectation of future fundamentals. Secondly, measuring fundamental values is a non-trivial task (Flood and Hodrick, 1990) with no consensus on the best approach.[1] Thirdly, the way market participants learn about fundamentals is often left unspecified (Blanchard and Watson, 1982). [2]

One of the current state-of-the-art bubble detection methods, developed by Phillips et al. (2015), applies statistical tests to identify price explosiveness, interpreted as an indicator of the presence of a bubble. This anticipative econometric technique is widely used in the literature (for example, Brunnermeier et al. (2020), Pavlidis et al. (2016) and Hu and Oxley (2018a)). By using this method, it is common to use only price series or scaled price series using proxies for the underlying fundamental values, such as price-to-rent and price-to-dividend ratios. Another method, used by Jordà et al. (2015), considers fundamental

[1] Flood and Hodrick (1990), p 87.
[2] Blanchard and Watson (1982), p 303.

values by using the price deviation from a long-run trend. The explosiveness of price-to-dividend ratio (Phillips et al., 2015) or the deviation from a long-run trend (Jordà et al., 2015) are subject to measuring errors and composite hypothesis testing issues. Finally the long-run price elevation approach used by Greenwood et al. (2019) ignores fundamental values.

The use of information beyond price and ratios can thus potentially improve the empirical detection and prediction of bubbles. Incorporating information related to investor beliefs, as reflected in media news, might provide valuable complementary information to financial variables. We utilize natural language processing techniques to extract narrative information from media news to gain insight into investor psychological states. Narratives as highlighted by Shiller (2015) can be informative of the economy and future economic events, as significant market events often occur when similar thinking prevails among large groups of people. (Shiller (2015), p 101). Media and social media can enhance and amplify this similar thinking through a "feedback loop", which is a crucial factor in the formation of stock market bubbles (Shiller (2015), p 84). As narratives can shape public opinions, influence actions and map actions into consequences (Eliaz and Spiegler, 2020), it is essential to consider their evolution alongside official macroeconomic variables in order to better detect and predict bubbles.

We employ narratives to augment bubble prediction in several ways. Firstly, we use market narrative features in predicting measures of market bubbliness, evaluating their additional predictive power over traditional financial and economic variables. Secondly, we use narrative information to predict the scale of market drops, which serves as a proxy for the severity of bubble consequences. Thirdly, we assess the out-of-sample forecasting performance of the narrative models compared to the benchmark models and alternative specifications that do not use narrative features as predictive variables. Lastly, conditional on ex-ante bubble signals, we use the attributes of narratives to predict the probability of ex-post bubbles.

To operationalize the narrative characteristics, we draw on the financial textual analysis and natural language processing literature and consider a set of interpretable measures. We summarize market narratives by (1) narrative volume, (2) narrative sentiment, (3) sentiment dispersion, (4) topic consensus, and by characterization as (5) risk narratives, (6) opportunity narratives and (7) bubble narratives. Firstly, the narratives are located by combinations of data source selection, searching query constrains, named-entity recognition and semantic similarity estimation. Secondly, the volume of narratives are proxied by the count of content units or alternative intensity measures of the narratives. Thirdly, narrative sentiments are estimated through textual sentiment measures. Fourthly, the level of topic concentration is quantified by the entropy of topic distribution. Lastly, the sentiment dispersion is proxied by the standard deviation of sentiment measures.

Our findings indicate the feasibility and usefulness of incorporating narratives in bubble forecasting, providing practical benefits to financial practitioners in monitoring bubble activity. Thus, by empirically examining the relationship between market narratives and bubbles, the present study contributes to both the narrative economics literature and the financial bubble literature.

Our results support Shiller's argument that narratives play a crucial role in driving economic events. The results of this study provide new empirical evidence for the role of narratives in the formation of bubbles, as indicated by the sign of the narrative features in the bubble prediction models. However, although we have minimized the possibility of reverse causality and endogeneity issues, it should be noted that this study does not focus on establishing causality between narratives and bubbles. Further research to examine the direction and mechanisms of this relationship are left to future studies.

The remainder of the paper is structured as follows: In Section 2, we conduct a comprehensive review of the relevant literature, with a focus on the formation mechanisms of financial bubbles and their connection to narratives, and on bubble detection methods. Section 3 details our selection of narrative features for bubble prediction. In Section 4, we present our empirical methodology. Section 5 provides the details of the data used in our analysis. The results and discussion thereof are presented in Section 6. Finally, in Section 7, we draw our conclusions and provide insights for future research.

2 Literature Review

2.1 Theories of Bubble

The literature is rich in modelling bubbles with different generation mechanisms. A rational bubble is defined as a result of some rational reasons including self-confirming expectations, misvalued fundamentals (intrinsic rational bubbles), and mispricing with exogenously determined factors (extrinsic rational bubbles) (see Flood and Hodrick (1990), Froot and Obstfeld (1991), Azariadis (1981), and Diba and Grossman (1988)). It is, however, difficult to distinguish the contribution of rational bubble to the explosive price from the contribution of unobservable fundamentals or expectations (Diba and Grossman, 1988; Dale et al., 2005). In addition, rational bubble models assume homogeneous investor beliefs and cannot explain variation in trading volumes (Barberis et al., 2018). By contrast, disagreement-based models (e.g. Harrison and Kreps (1978) and Scheinkman and Xiong (2003)) and the extrapolation model (Barberis et al., 2018) explain the bubble generating process with heterogeneous beliefs among the agents. When there are disagreements about the fundamentals, bubbles could be formed if there are arbitrage limitations (Scheinkman and Xiong, 2003). When there are disagreements about the timing of price correction after overreaction to good news, bubbles could form and last if rational arbitrageurs ride the bubble (Abreu and Brunnermeier, 2003). Another strand of literature argues that bubbles may result from "animal spirits" such as irrationally optimistic expectations, fashion, or fads (e.g. Shiller (2015, 2019)). During irrational bubbles, the relationship between fundamental values and prices breaks down and psychological factors plays a more important role (Dale et al., 2005).

2.2 Empirical Bubble Detection Methods

The current state-of-the-art "bubble" detection method developed by Phillips et al. (2015) is based on time series statistics.[3] It identifies explosiveness in asset prices based on the recursive right-tailed unit-root test.[4] With the assumption of a "greater than unit" root in the data generating process during a bubble, the identification and date stamping are achieved by the hypothesis testing for "no explosiveness". Comparing with classic unit-root tests (as that used in Diba and Grossman (1988)), the PSY method has an "ex-ante" nature in dating and advantages in dealing with "pseudo stationary behavior". As claimed by Evans (1991) and Phillips et al. (2015), prices might appear to be stationary when multiple collapsing bubbles are present.[5]

Recent studies, including Brunnermeier et al. (2020), Pavlidis et al. (2018) and Hu and Oxley (2018a), rely on this approach to identify bubble periods in various asset classes. When using this method, it is common to use price series or price series normalized by

[3] Phillips et al. (2015) refer to "explosive behaviors" when they use the term "bubbles".

[4] The identified explosive episode could be either in exuberance or collapse phases, as reported in Phillips et al. (2015), p 1066 - 1067. We provide an example in Appendix C.

[5] Evans (1991) p 922; Phillips et al. (2015), p 1051

a proxy of the fundamental value. Proxies of observable fundamental values include, for example, dividends (Phillips et al., 2015; Phillips and Shi, 2018; Brunnermeier et al., 2020) or earnings (Leone and de Medeiros, 2015; Deng et al., 2017; Hu and Oxley, 2018b) for equities and rents (Giglio et al., 2016; Engsted et al., 2016; Pavlidis et al., 2016; Brunnermeier et al., 2020) or income (Anundsen et al., 2016; Pavlidis et al., 2016) for real estates. However, the suitability of the observable proxies is questioned by Basse et al. (2021).

Alternative methods uses price elevation and price deviation. The long-run price elevation approach, as demonstrated in Greenwood et al. (2019), ignores fundamental values and indicates evidence of bubbles when a price elevation exceeds a reasonable threshold. The price deviation approach, as described by Jordà et al. (2015), determines the existence of a bubble through the examination of significant deviations from a long-run trend in the asset price.

2.3 Bubbles and Narratives

The behavioral finance literature consider factors such as "animal spirits" (Akerlof and Shiller, 2010), investor confidence (e.g. Daniel et al. (1998), Shiller (2000), and Barber and Odean (2001)), investor sentiment (e.g. Barberis et al. (1998), Berger and Turtle (2015) and Baker and Wurgler (2006))) and narratives (e.g. Shiller (2017), Ter Ellen et al. (2021), and Nyman et al. (2021)) in explaining investor behaviors and the movement of asset prices. Along with the development of natural language processing techniques, many of the behavioral variables can be proxied by textual data based variables. For instance, Tetlock (2007) applies textual sentiment analysis on market news to quantify the tone of media news; Baker et al. (2016) create an economic policy uncertainty (EPU) index based on the frequency of news articles containing words in a short word list; Nimark and Pitschner

(2019) implement Latent Dirichlet Allocation (LDA) topic model on newspaper stories to reveal news focus around major events; Chen et al. (2022) use multiple NLP techniques to study risk narratives that are perennial and went viral during the COVID-19 pandemic.

Shiller (2015, 2017, 2019) emphasises the role of narratives in the life cycle of asset bubbles and highlights that narratives are understudied in Finance and Economics. There is a growing strand of literature finding evidence of useful information from narratives that could predict economic events (e.g. Bertsch et al. (2021), Larsen et al. (2021) and Nyman et al. (2021)), but the narrative information have not been widely employed in the bubble detection literature.

3 Narrative Features for Bubble Prediction from Purpose-oriented Regularization of Texual Data

The usefulness of narratives in bubble prediction is implied by the causal relationships between fundamentals/events, narratives and bubbliness, which was posit as in Figure [1]. The outcome of concern is the future bubbliness. Suggested by theories, bubbles are driven by investor beliefs, which are high dimensional and not easily observable in real-time. Investor beliefs are shaped and influenced by fundamentals, events and narratives. Fundamentals and events influence investor beliefs directly by changing the information set and investor's Bayesian updating of their beliefs, or indirectly via narratives by evoking psychological reactions or biased perceptions. Investors process their given information and communicate their beliefs via narratives, leading to a bi-directional causal relationship between narratives and investor beliefs. Taking out the non-observable node of investor beliefs, we have a standard three-node directed acyclical graph. Fundamentals/events act as confounders to the causal relationship between narratives and future bubbliness.

[Insert Figure [1] about here]

Bubble theories consider certain properties and patterns of investor beliefs as the drivers of bubbles. For the purpose of bubble prediction, given the bi-directional causal relationship between narratives and investor beliefs, we leverage the predictive information carried by observable narratives features that are channelled from the unobservable investor beliefs.

Narratives convey rich information in an enormous number of dimensions, including but not limited to general tones or emotions, aspect based sentiments, topics, events, expectations, facts or misinformation, and economic or causal inferences (see Shiller (2019) and Eliaz and Spiegler (2020)).[6] With NLP techniques, many informative narrative dimensions can be captured by proper proxies. Too much information, however, makes it difficult to be useful in any particular economic question. To optimize the information universe and avoid overfitting, one needs a purpose-oriented regularization. One promising way is to filter important narratives and narrative features with economic theories. In this paper, we select prominent (not all-inclusive) narratives and interpretable "narrative features" that are of high relevance to bubbles. Among the variable candidates, those can be properly operationalized by current technologies are selected for the empirical analysis. We rely on the textual analysis literature to build the "feasibility filter".

3.1 Opportunity Narratives and Positive Sentiment

One popular strand of the literature claims that many bubbles are driven by "animal spirits" instead of rational quantitative expectations of future cash flows.[7] Aliber and Kindleberger

[6] Shiller (2019), p 65; Eliaz and Spiegler (2020), p 3787.

[7] Akerlof and Shiller (2010) defines "animal spirits" as a modern economic term referring to our peculiar relationship with ambiguity or uncertainty, beyond the original definition of "a basic mental energy and life force", p 3-4

(2015) describe the psychological phenomenon of investors in many bubbles as "mania" and claim that there are similar patterns in those manias. They associate the mania phase of a bubble with a sense of "we never had it so good" and "making money never seemed easier." A "follow-the-leader" process drives investors away from rational behavior. [8] Similarly, Shiller (2015) describes the economic phenomenon resulting from the psychological epidemic of investor enthusiasm as "irrational exuberance". He defines a speculative bubble as "a situation in which news of price increases spurs investor enthusiasm, which spreads, goes viral and brings in more speculative investors". [9] From both seminal references, a large and increasing volume of "investment opportunity" narratives should be observed during the emerging phase of a bubble. Such narratives should go viral and deliver continuous and exaggerating optimistic, confident and exciting sentiments (Shiller, 2019). [10]

3.2 Opinion Disagreement and Topic Homogeneity

Harrison and Kreps (1978) propose that the relevant notion of intrinsic value is decided by aggregate investor assessments and attribute speculation to heterogeneous beliefs. Scheinkman and Xiong (2003) and Hong et al. (2006) show that a speculative bubble can arise when investors have heterogeneous beliefs due to overconfidencce. The price is upwardly biased because of short-sales constraints (Miller, 1977; Scheinkman and Xiong, 2003) and the resale option effect (Harrison and Kreps, 1978). Hong and Stein (2007) then argue that excess media coverage fuels investor disagreement and may help explain both the dramatic trading volume and the elevated prices. However, Yu (2011) finds a negative relationship between market disagreement and the ex-post expected market return, and Kim et al. (2014) find

[8] Aliber and Kindleberger (2015), p 11, p 29.
[9] Shiller (2015), p 2.
[10] Shiller (2019), p 9, p 21, p 41, p 226, p 228.

a negative relationship between market disagreement and future returns, only during high-sentiment periods. Ma et al. (2022) find a positive relationship between investors' belief dispersion and trading volume but confirm a negative relationship between belief dispersion and future returns. Given that bubble inflation is associated with positive returns, the empirical findings seem to be inconsistent with disagreement-based bubble theories.

In addition to the opinion disagreement, topic homogeneity is also considered to be an important feature around economic events like bubbles. Shiller (2019) suggests that dominant narratives lead to many economic events. For example, the "new era" narratives were associated with historical large price increases in many countries (Shiller, 2015). Nimark and Pitschner (2019) find that major events shift the general news focus and make coverage more homogeneous. Nyman et al. (2021) find increasing topic consensus around the strongly positive narrative prior to the global financial crisis, implying a growing and increasingly dominant new paradigm narrative. Similarly, Bertsch et al. (2021) find that "narratives tend to consolidate around a dominant explanation during expansions and fragment into competing explanations during contractions".

3.3 NLP Proxies

Most of the narratives and the important features mentioned above can be identified and operationalized with various data processing and NLP techniques. Firstly, narratives could be represented by a collection of textual contents, which could be identified using keywords/n-grams, manual annotations, and machine learning classifications or clustering techniques. Secondly, the narrative volumes can be represented by the amount of textual contents, which could be estimated using the count of documents (e.g Baker et al. (2016)), the length of documents or the aggregation of topic prevalences (if using topic models like LDA) or semantic similarities (if using semantic embedding methods). Thirdly, investor sentiment or emotions could be proxied by textual sentiment/tones of textual contents, estimated with lexicon-based methods or machine learning based methods (see Kearney and Liu (2014), Loughran and McDonald (2016) and Shapiro et al. (2020)). Fourthly, using narratives, the belief divergence can be operationalized by measures like tone dispersion (e.g. Xiong et al. (2020)). Lastly, topic consensus can be proxied by entropy measures (see Nimark and Pitschner (2019), Nyman et al. (2021), and Bertsch et al. (2021)).

We thus consider a small number of interpretable narrative features to represent the most useful information in market narratives. For the purpose of bubble detection, we firstly select country-specific market narrative intensity, textual sentiment level and dispersion, and narrative topic consensus, given their revealed importance in the literature. In addition to general market narratives, inspired by Shiller (2000), we consider the intensity measures of a small set of specific popular competing economic narratives, namely, "risk narratives", "bubble narratives" and "opportunity narratives", to proxy the two important investor attitudes in speculative markets - bubble expectations and confidence.

To represent the popular narratives, we use business news from major media resources. To constrain the content to be relevant to the narratives of interest, we rely on a combination of search query (on databases), keywords (n grams), named-entity recognition, LDA topic modelling and semantic similarity measures. For the narrative features, we partly follow Chen et al. (2022), with some modifications.

The first narrative variable is the market narrative intensity (NI). It is used to proxy the investors' attention/interest to a market. We simply operationalize it using the count of market narrative documents (sentences, as described in the data section):

$$NI_{c,t} = D_{c,t}. \tag{1}$$

Here, $NI_{c,t}$ represents the general market narrative intensity for country c at time t and $D_{c,t}$ represents the number of documents (sentences in this case) containing a geographical entity of country c at time (in this case, month) t.

The second feature included is the textual sentiment/tone (NT). It proxies investors' opinion on a optimism/pessimism basis. We construct the variable using the lexicon-based (the Loughran and McDonald (2011) word lists) approach. With the state-of-the-art finance-specific word lists, we estimate the sentiment measure using the following formula.

$$NT_{c,t} = \frac{\sum_{d=1}^{D_{c,t}} Count_{d,positive} - Count_{d,negative}}{\sum_{d=1}^{D_t} Count_{d,total}}. \tag{2}$$

In Equation [2], $NT_{c,t}$ represents the polarity score of narrative tone, $Count_{d,positive}$ and $Count_{d,negative}$ represents the count of words in document d that are listed in the word-lists of positive and negative tones, $Count_{d,total}$ represents the count of words in document d that are included in the Loughran and McDonald (2011) master dictionary, and $D_{c,t}$ refers to the total number of documents at time t bearing a country tag of c. We handle the negation terms by reversing their signs, that is, a term belonging to the positive dictionary is treated as negative if it is preceded by a negation.

Based on narrative tones, we include the narrative tone dispersion as the third feature, to proxy the investor disagreement (NTD). Consistent with Xiong et al. (2020), we estimate tone dispersion using standard deviation.[11]

[11] See-To et al. (2017) use standard deviation while Xiong et al. (2020) use variance.

$$NTD_{c,t} = \sqrt{\frac{\sum_{d=1}^{D_{c,t}} (NT_{d,t} - \overline{NT_{c,t}})^2}{D_{c,t}}}. \tag{3}$$

Here, $NTD_{c,t}$ represents the dispersion of tones for country c at time t. $\overline{NT_{c,t}}$ represents the average tone of country c at time t. $NT_{d,t}$ represents document d's tone score.

The fourth feature we consider is topic consensus (TC), a measure proxying the consensus level of investor attention, on different aspects of the market. Following Nyman et al. (2021), we measure topic consensus using the Shannon entropy. We take the negative value of the original Shannon entropy so that higher the value, higher the consensus:

$$TC_{c,t} = (-1) \cdot Entropy_{c,t} = \sum_{k=1}^{K} p_{k,c,t} \log p_{k,c,t}, \tag{4}$$

where $TC_{c,t}$ represents the topic consensus measure, and $p_{k,c,t}$ is the average probability of topic k over all documents published at time t with a country tag of c. The topic probabilities are the outputs of a tuned LDA model (Blei et al., 2003).[12] We train the LDA model using all sentences without country tags to identify economic topics that are not country and time specific to avoid endogeneity issues.

Lastly, we include the relative intensity measure for the competing narratives of opportunity, risk and bubble. It could be viewed as the opportunity narrative intensity, net of the risk and bubble narratives (NI^{NO}). We separately estimate the "opportunity narratives", "risk narratives", and "bubble narratives" intensities based on semantic similarity measures (SBERT). [13] Specifically, for each narratives, we calculate the average of the similarity score of a document to one representative seed narrative.[14] We aggregate by country and month

[12]We use coherence value to tune the number of topics, k.

[13]https://www.sbert.net

[14]The seed narratives are "What about the stock market?", "There is an asset bubble", "It is a good time to invest in the market", and "The risk level in the market is high". See the sample narratives in Appendix

to generate the narrative intensity measures for analysis.

$$Sim_{i,c,t} = \frac{\sum_{d=1}^{D_{c,t}} Sim_{i,d,t}}{D_{c,t}}.$$

Here, $Sim_{i,c,t}$ stands for the average similarity measure of narrative i of country c at time t, and $Sim_{i,d,t}$ represents document d's similarity score to narrative i. Each document has a single time stamp t. $D_{c,t}$ refers to the total number of documents with country tag of c and time stamp of t. $Sim_{i,c,t}$ have high correlations due to the common driver of similarity to market narratives, we construct a measure for the relative intensity of competing narratives between "opportunity" and "risk & bubble":

$$NI_{c,t}^{NO} = Sim_{Opportunity,c,t} - (Sim_{Risk,c,t} + Sim_{Bubble,c,t}). \tag{5}$$

We also calculate the specific narrative intensity measures by orthogonalizing the similarity measures to the similarity measure of "market narrative". First, we fit a linear regression:

$$Sim_{i,c,t} = \alpha + \beta \cdot Sim_{Market,c,t} + \epsilon_{c,t}.$$

Second, we record the residuals to proxy the narrative intensities:

$$NI_{i,c,t} = Sim_{i,c,t} - (\hat{\alpha} + \hat{\beta} \cdot Sim_{Market,c,t}). \tag{6}$$

In summary, the major set of narrative features contains NI, NT, NTD, TC, and NI^{NO}. We also include NI^{Risk}, NI^{Bubble}, and $NI^{Opportunity}$ to reveal the signs and pre-

B.

dictive power of the competing narratives. The succinct definitions of the features are presented in Table [2]. Our use of the simple and classic measurements has the advantage on replicability. Potential improvements on the measurements are expected to further enhance the identified relationships.

4 Augmented Bubble Forecasting

This section introduces the narrative-augmented strategy for bubble forecasting. We make use of narratives in bubble forecasting and empirically test the approaches on the representative stock market indices in over 40 countries. Firstly, we test the marginal predictive power of narratives by including market narrative features in predictive regressions of market bubbles. Secondly, we test the forecasting power for the scale of market drops. Thirdly, we compare the out-of-sample forecasting performances with the benchmark models. Lastly, we predict the probability of an ex-post bubble, conditional on the signal of bubbles.

4.1 Bubble Measures

Our starting point is to proxy the bubbliness with the price explosiveness measure obtained from the PSY algorithm. However, we also consider two other common measures, price elevation (Greenwood et al., 2019) and price deviation (Jordà et al., 2015) for robustness check. Greater the measures, higher the probability of a time being in a bubble regime. Price explosiveness is estimated by the PSY method (Phillips et al., 2015). Following Brunnermeier et al. (2020), we use price series as the input, but the results mostly hold when we use price-to-dividend ratios as the input. As in Greenwood et al. (2019), drastic long-run price elevation is used as the indicator of potential bubbles. Finally drawing on Jordà et al. (2015), large deviations of price from a long-run trend is used for bubble discovery. We use the continuous measures of "bubbliness" in regressions to investigate the relationship between narratives and bubbles. We use the binary classification measures based on the PSY method in a logistic regression to examine the predictability of ex-post bubble events.

With the PSY algorithm, we proxy the bubbliness by the backward sup ADF (BSADF, henceforth) statistics,

$$BSADF_{c,t}(w_0) = \sup_{w_1 \in [0, t-w_0]} ADF^t_{c,w_1}. \tag{7}$$

Here, $BSADF_{c,t}(w_0)$ represents the BSADF statistic of country c at time t, with a minimum estimation window of w_0 observations. As suggested by Phillips et al. (2015), we decide the minimum window using the ratio of $0.01 + 1.8\sqrt{obs}$, in which obs is the number of observations. ADF^t_{c,w_1} represent the Augmented Dickey Fuller test statistics of the price series of country c from time w_1 to time t. By comparing the BSADF statistics with the critical values, we have a binary classification variable for the bubble signal,

$$Signal_{PSY,c,t} = \mathbb{1}(BSADF_{c,t} > cv_{c,t}), \tag{8}$$

in which $cv_{c,t}$ is the critical value calculated using the PSY algorithm for the price series of country c and $\mathbb{1}(\cdot)$ is the indicator function. We use the 95 % critical value calculated using a wild bootstrap method (Phillips and Shi, 2018).

Using the price-elevation method, we proxy the bubbliness by a 24 month price run-up,

$$Elev_{c,w,t} = \frac{PI_{c,t} - PI_{c,(t-w)}}{PI_{c,(t-w)}}. \tag{9}$$

Following Jordà et al. (2015), we also proxy the bubbliness by the price deviation from the price trend.

$$Dev_{c,t} = \log PI_{c,t} - \log PI_{c,trend,t}. \tag{10}$$

Here, $Dev_{c,t}$ is the detrended log price index of country c. Following Jordà et al. (2015), we use a Hodrick–Prescott filter to estimate the trends. [15]

4.2 Predicting the Bubbliness and Future Drops

With the narrative features, we examine whether this information could help predict the common measures of "bubbliness". Given that the bubbliness measures, namely, $BSADF_{c,t}$, $Elev_{c,t}$ and $Dev_{c,t}$ are not stationary, we use the first difference in the regressions.

To predict the change of the bubbliness measures, we use the models below.

$$\triangle B_{c,t+1} = \alpha_c + \boldsymbol{\theta} \cdot \boldsymbol{X_{c,t}} + \gamma \cdot N_{c,t} + \epsilon_{c,t+1} \tag{11}$$

As the baseline model, we include the controls ($X_{c,t}$) to predict the change of the bubbliness measure in the next month ($\triangle B_{c,t+1}$). De-trended market level ratios and macroeconomic variables, along with the lag of the dependent variable, are included in the controls. Specifically, we include dividend yield (DY), price to earning ratio (PE), price to book value (PB), interest rate (IR), inflation (Inflation), real GDP growth (RGDP), investment to GDP

[15] We set lambda as 14,400, given monthly time series.

growth, trading volume (VO), and the CLI indicator from OECD. Subscript "c" stands for country/region. In the narrative augmented model, we add the de-trended narrative series ($N_{c,t} \in \{NI_{c,t}, NT_{c,t}, NTD_{c,t}, TC_{c,t}, NI_{c,t}^{NO}, NI_{c,t}^{Opportunity}, NI_{c,t}^{Risk}, NI_{c,t}^{Bubble}, EPU_{c,t}\}$) to predict the same.[16] Although the correlations between the pairs of narrative features are not high, the series tend to shift together around significant events. For this reason, we include only one of the features at a time for better identification.

To determine the early warning credentials f our approach, we study whether our narrative measures could help predict the scale of near-future declines. We define the scale of maximum future drop as

$$MaxDrop_{c,t} = -\min(\frac{PI_{c,t+j}}{PI_{c,t+i}}, 0),\ \forall i \in [0, 11],\ j \in [1, 12]\ and\ i < j.$$

With similar specifications to the bubbliness model, we predict the change in the scale of market drop using the following regression,

$$\triangle MaxDrop_{c,t+12|t} = \alpha_c + \boldsymbol{\theta} \cdot \boldsymbol{X_{c,t}} + \gamma \cdot N_{c,t} + \epsilon_{c,t+1}. \tag{12}$$

Critically, we do not include the lag of the dependent variable in the controls. The reason is that the information contained in this variable is not available to investors until one year later.

Given our aim of bubble forecasting, predictive power is the main concern. However, to better understand the predictive relationships, we also check the signs the coefficients. We select the small set of narrative features based on theories/hypothesises from the literature, if the variables do carry predictive information on bubbles, we would expect the signs to

[16] We detrend the narrative series by subtracting their n-month weighted moving average. We let $n = 12$.

be consistent with those theories/hypothesises. Specifically, with more investor attention (Vozlyublennaia, 2014), we expect the market to be more efficient, which leads to a decrease of bubbliness. On the contrary, with higher level of investor optimism or opportunity narrative intensity (Shiller (2015); Aliber and Kindleberger (2015)), opinion disagreement (Harrison and Kreps (1978); Scheinkman and Xiong (2003)) or topic consensus (Nyman et al., 2021; Bertsch et al., 2021), we expect bubbliness to increase. That means we expect NI, NI^{Risk} and NI^{Bubble} to have negative signs and NT, NI^{NO} and $NI^{Opportunity}$, NTD and TC to have positive signs.

We begin our analysis examining the U.S.. The coefficients are tested out-of-sample using international panel data. For the U.S. only regression, we report the Newey-West standard errors, while for the panel regressions, following Brunnermeier et al. (2020), we include country/region fixed effects and report the coefficients with the standard errors clustered by both country/region and month. [17]

We examine the out-of-sample forecasting performance of the narrative-augmented models over the benchmark models. For each country, using the expanding window method, we fit the predictive regression models using the information available at one point of time and predict the change of bubbliness one month ahead. The forecasts are compared with actual values. We use the Mean Squared Errors (MSE) as the loss function and compare the forecasting performance with the benchmark model using the Diebold & Mariano test (Diebold and Mariano, 2002) or the Clark & West test (Clark and West, 2007). We consider forecasting models (1) with historical means, (2) with only the control variables, (3) with a single narrative variable, (4) with only narrative variables, (5) with both narrative and control variables, and (6) with ensemble method averaging all (1) - (5) predictions.

[17] For the convenience of coefficient interpretation and comparison, we normalize all variables by subtracting the average and dividing the standard deviation.

4.3 Predicting ex-post Bubbles

We use logistic regressions to estimate the probability of being in a bubble regime (after-the-fact) conditional on a signal of a bubble. With logistic regression, the equation is:

$$Prob(Bubble_{c,t}|Signal_{c,t}) = \frac{1}{e^{-(\alpha+\boldsymbol{\theta}\cdot\boldsymbol{X_{c,[t-w,t]}}+\gamma\cdot N_{c,[t-w,t]}+\epsilon_t)}}. \tag{13}$$

On the left hand side is the probability of a time with a bubble signal turned out to be followed by a crash in 24 months.[18] The predictive variables on the right hand side (within the parentheses) are similar with those in Equation (11) and Equation (12). The difference, is that we calculate the average of the predictive variables within the window w. For sequential bubble signals within a two year window, we only consider the first observation. As the ratio of real bubbles over signals is sometimes much smaller than 50 %, we use over-sampling strategy to balance the data before running the regression. [19] To maximize the number of observations, we consider a minimum number of controls in $X_{c,t}$ - dividend yield, price-to-earning ratio, price-to-book ratio, and trading volume.

5 Data

5.1 Narrative Data

Media news published on major newspapers and publications is the source of our narratives. There are mainly two reasons for choosing this data source. First, as argued by Nimark and Pitschner (2019), media news are filtered by editors to be highly correlated with popular

[18]We set 40% as the threshold of a market crash.

[19]For over-sampling, we use the Synthetic Minority Oversampling Technique (SMOTE).

narratives. Second, major news sources are less subject to noise information. We collect the media news satisfying the requirements from LexisNexis, a popular textual database platform that has been frequently used in the literature (e.g. Ardia et al. (2019) and Shapiro et al. (2020)).

Using the "advanced searching" function on LexisNexis's website, we design the searching query to satisfy the following conditions. First, the article should mention "market" at least 5 times. This is to limit the acquired news to be "market news". Second, the article should have both "Economy & Economic indicators" and "Financial Market Updates" subject tags. This condition ensures the articles to be more relevant to fundamentals instead of pure updates. Third, the news sources include 'Major U.S. Newspapers", "Major Non-U.S. Newspapers", "Major Newspapers", "Major Publications", "Major World Newspapers" and "Major World Publications". We incorporate all to include all major sources. We further set the "relevance" option to be "Major Terms Only". We retrieve 765,645 articles that are published between January 1st 1975 and December 31st 2021.

We pre-process the news articles to remove the articles without a date stamp precise to day. 748,970 articles are left in the dataset. We break down the articles to sentences for the purpose of narrative identification.[20] To avoid meaningless sentences, we keep those with character length between 10 and 550. Given that we will need the narratives for each country, we drop the sentences without a country tag. We use named-entity recognition to add country tags to each sentence. 4,668,295 country related sentences are identified for textual analysis. For our analysis, we select the 48 countries/regions with the largest number of sentences. We report the meta data in Appendix A.

The sentence level documents are processed by the SentenceTransformers framework to

[20] Using the sent_tokenize function from the "nltk" Python package, we obtain the sentences.

calculate semantic similarity scores with narrative queries - “The risk level in the market is high”, “There is an asset bubble” and “It is a good time to invest in the market”. The higher the similarity score, the more relevant the sentence is to the specific market narratives. We report some random samples that has high semantic similarity in Appendix B. With our data selection, a large portion of the sentences should be market narratives. The boxplot in Figure [2] proves a larger intensity of the “market narratives”. It also shows that “risk narratives” and “opportunity narratives” have similar scales, while “bubble narratives” are least popular among the four narratives. We calculate the narrative intensity measures by orthogonalizing the similarity measures to the similarity measure of “market narrative”.

[Insert Figure [2] about here]

We then extract four primary narrative features from the market narratives - market narrative intensity (NI), textual sentiment (NT), tone dispersion (NTD), topic consensus (TC), opportunity, risk and bubble narratives ($NI^{Opportunity}$,NI^{Risk},NI^{Bubble}), and the relative intensity of competing narratives (NI^{NO}). Additionally, we also collect the international EPU indices (Baker et al., 2016). [21] We include EPU in the narrative feature set given its textual-based nature for comparison. A visual description of the monthly narrative information for the United States is displayed in Figure [3].

[Insert Figure [3] about here]

The trends of the sentiment measure (on panel B) are also consistent with those in Ardia et al. (2019) and Barbaglia et al. (2022).[22] The sentiment dispersion measures are in general negatively correlated with sentiment. The topic consensus measure upwardly deviated from its trend in 2008, along with sudden increases of the “bubble narratives”

[21] https://www.policyuncertainty.com

[22] Ardia et al. (2019), Figure. 3, page 1378; sentiment measure for “economy” in Barbaglia et al. (2022), Figure. 1, page 7.

and “risk narratives” shown on panel D. We show the trends in Appendix B. During the dot-com bubble, the “bubble narratives” kept increasing since around 1996 while the “risk narratives” and “opportunity narratives” decrease. The same pattern was observed before the 2008 - 2009 Great Recession.

We report the correlations between the narrative features in Table [1]. The correlations are not large and the signs are all consistent with economic intuition. The signs of correlations are consistent with expectation. For instance, NI is negatively correlated with NT and NI^{NO}, implying “more news is bad news”. It is also negatively correlated with TC and positively correlated with NTD, implying that more news is usually accompanied with more dispersed topics and sentiment. Interestingly, NT and NTD are negatively correlated, implying that prevailing positive sentiment is accompanied with less disagreement.

[Insert Table [1] about here]

A further investigation of the data reveals that the between-country correlations of the narrative features are also modest or small. It is because we identify country-relevant sentences with clear named entities, leading to small data overlap. As shown in Appendix A, the sentence co-occurrence rate is low between the countries/regions. Narrative features extracted from the country-relevant sentences more precisely capture country-specific characteristics, but as a trade-off, it neglects other sentences in the articles, causing a potential downward pressure to the predictive power. We report one example of the correlation heat map in the Appendix in Figure [5]. Relatively higher correlations are observed between the countries with the largest media exposure or between countries geographically close.

5.2 Economic & Financial Data

We use country-level price index data for bubble determination. For the countries/regions with enough media exposure, we collect their representative price indices, dividend yield, price-to-earning ratio, price-to-book ratio, and trading volume from Datastream. For each country/region, we select either the Datastream Market index or the MSCI index (as seen in Appendix A), or both, to maximize the length of variables of interest. The length of available data varies for different countries. We also collect macroeconomic data from OECD, including inflation rate, interest rate (10-year government bond rate), real GDP, investment to GDP ratio and the Composite leading indicator (CLI). Our analysis uses monthly data, so we forward fill the quarterly macroeconomic data to represent the data available to investors in real time. We report the number of observations, the number of countries covered and the average data length per country for both the financial and narrative variables in Table [2]. We also visualize the variables for the U.S. in Appendix D.

[Insert Table [2] about here]

6 Empirical Results

6.1 The U.S. Evidence

6.1.1 Predicting Bubbliness

We start with the predictive regression in Equation [11] and report the results in Table [3]. All of the five main narrative features (NI, NT, NTD, TC, NI^{NO}) have statistically significant coefficients. On a predictive basis, the market narrative intensity (NI), the narrative tone dispersion (NTD) and the topic consensus (TC) are negatively associated with the change of bubbliness, while the narrative sentiment (NT) and the relative intensity for opportunity narratives (NI^{NO}) are positively associated with the change of bubbliness. In terms of the magnitude of effects, NT, NI and NI^{NO} has the greatest coefficient in an absolute sense. From the regressions with financial and economic controls, a one standard deviation increase of narrative tone (NT) is associated with 22% standard deviation increase of the change of bubbliness next month.

[Insert Table [3] about here]

The signs imply that positive sentiment and large volume of opportunity narratives are potential drivers of bubbles, consistent with the argument of Aliber and Kindleberger (2015) and Shiller (2015, 2017, 2019), that is, a sense of "investment opportunity" along with positive expectations of future cash flows trigger the investor enthusiasm to ride a speculative bubble. The negative sign of general market narratives (NI) is in line with the hypothesis that increased investor attention improves market efficiency (Vozlyublennaia, 2014). The presence of dominant topics (TC) and dispersed sentiment (NTD), however, are negatively associated with the change of bubbliness next month. The negative sign of TC is inconsistent with our expectation (as implied by Bertsch et al. (2021) and Nyman et al. (2021)) and is likely because of its correlation with negative events. As documented by Nimark and Pitschner (2019), major events shift the news focus and make coverage more homogeneous. Lastly, the negative sign of NTD is inconsistent with the predictions of disagreement-based models (Harrison and Kreps (1978); Scheinkman and Xiong (2003)), but consistent with Shiller (2015, 2017), that is, "significant market events generally occur when there is similar thinking among large groups of people" (Shiller (2015), p101)

Incorporated individually, our narrative measures provide excellent predictive power to future change of bubbliness, comparing with the financial and economic controls. In the univariate regressions, NI, NT and NI^{NO} have the highest adjusted R-squared, which are 3.43%, 3.68% and 2.37%, respectively. Transforming the narrative information set using their first three principal components, we have similarly high explanatory power and statistically significant coefficients, as reported in Appendix F. As a comparison, the highest adjusted R-squared obtained using the controls is 0.65% (by interest rate), and none of the controls have statistically significant coefficients. Interestingly, by comparing with the results of the contemporaneous regressions (as reported in Appendix E), we find that many financial and economic variables are statistically significantly associated with the bubbliness measure, but they do not have any predictive power. By contrast, narrative variables are stronger in predicting than explaining the bubbliness. The result indicates that narrative information carry additional predictive power on bubbliness. As suggested in our conceptual framework, narratives contain the perceived fundamental information and reveal investor beliefs, leading to additional predictive power to future economic events.

6.1.2 Granger Causality Tests

To show the direction of the causal relationships, we run Granger causality tests between the variables and the change of bubbliness. For each variable we considered in the regressions, we pair it with the change of bubbliness in a vector autoregression model. We set the maximum of lag as 15 and select the number of lags using the Akaike information criterion. After fitting the model, we report the p-value of the Wald test in Table [4]. Interestingly, all of our main narrative features (NI, NT, NTD, TC, NI^{NO}) reject the null hypothesis of not Granger causing the change of bubbliness, while there is no evidence of Granger causality from the other direction. Using the financial and economic variables, the pattern flipped. Only interest rate and the CLI indicator reject the null hypothesis of not Granger causing the change of bubbliness, while there are evidence of Granger causality from the change of bubbliness to the price-to-book ratio, the real GDP growth and the investment to GDP ratio. This result alludes that the financial and economic variables are likely influenced by the change of bubbliness, while the narrative variables are possibly impacting the change of bubbliness.

[Insert Table [4] about here]

6.1.3 Predicting Future Drops

Lastly, we report the estimated coefficients and adjusted R-squared of the maximum drop prediction regressions (Equation [12]) in Table [5]. Out of the five main narrative variables, NI and NT are statistically significant, and NTD and NI^{NO} are marginally significant. However, the signs are identical with those in the bubbliness prediction regressions. This result implies that the predictive power and the direction of the narrative information on bubbliness naturally extend to the consequences of bubbles. As in bubbliness prediction, we also transform the narrative information set using their first three principal components, we have similarly high explanatory power and statistically significant coefficients, as reported in Appendix F.

[Insert Table [5] about here]

Interestingly, the model with EPU index obtained the highest adjusted R-squared, with its coefficient being statistically significant. This implies that for the left-tail events, the narrative aspect on uncertainty provides additional predictive power.

6.2 International Evidences

The relationships we identified using the U.S. data mostly hold when we run the analysis with international data. We report the estimated coefficients and adjusted R-squared for the panel regressions in Table [6] and Table [7]. With the international panel data, market narrative intensity (NI), narrative tones (NT), narrative tone dispersion (NTD) and the relative intensity for opportunity narratives (NI^{NO}) are still statistically significant in predicting the change of bubbliness, with or without controls. When we exclude the U.S. from the panel, only NI becomes marginally statistically significant, with other variables keep being statistically significant. NT, and NI^{NO} are still statistically significant in predicting the change of maximum market drop, with or without controls, with or without the U.S. data. NI is always statistically significant but becomes marginally significant when we exclude the U.S. data and add controls. NTD is only statistically significant without controls. EPU is also statistically significant in all specifications.

[Insert Table [6] about here]

[Insert Table [7] about here]

This results suggest that the predictive relationships between narratives and bubble development are robust and universal.

6.3 Predictive Power Out-of-sample

We report the out-of-sample bubbliness forecasting performance in Table [8] and the market drop forecasting performance in Table [9]. For the bubbliness exercise, all of our narrative models outperform the model with only the financial and economic variables. When we compare with the "historical mean" and "zero" benchmarks, the model with only controls does not reject the null hypothesis of no outperformance. However, most of the narrative models or narrative augmented models outperform the benchmarks. The ensemble method always outperform the benchmarks with the lowest MSE. Our results indicate that the narrative features, taken individually or collectively, always outperform the financial and economic measures in terms of bubble prediction. By augmenting the predictive models with narrative information and implementing ensembles, we generate the best out-of-sample prediction performance.

For the maximum drop forecasting exercise, most of the models outperform the "zero" benchmark. However, only the narrative augmented model with both controls and narrative variables outperform the model with only controls. The ensemble approach also generates the lowest MSE. This result indicates that for the prediction of future market drops, financial and economic variables are important, and narratives are also important and can add additional predictive power.

[Insert Table [8] about here]

[Insert Table [9] about here]

6.4 Ex-post Bubbles Prediction

We report the results from the logistic regressions in Table [10]. We find four of the main narrative variables are statistically significant predictors in at least one specification, namely, narrative tones (NT), narrative tone dispersion (NTD) topic consensus (TC) and the relative intensity for opportunity narratives (NI^{NO}). When we observe a bubble signal, by evaluating the narrative attributes in the past 2 years, we expect that, the probability of the bubble candidate ends up with a crash (i.e. an after-the-fact bubble) is high if we have negative narrative sentiment, low sentiment dispersion, high topic consensus and low intensity of opportunity narratives. It is a strong evidence of the predictive power of narratives on ex-post bubble events. That means narratives contain information about whether a bubble signal indicate a true bubble, which is confirmed by a crash. Our finding speaks to Greenwood et al. (2019) by showing that narrative attributes before the bubble signal also help forecast an eventual crash. It also speaks to Goetzmann (2015) because the finding implies that the narrative attributes help differentiate "good booms" from bubbles (booms that went bad). Interestingly, comparing with the results in ex-ante bubbliness predition regressions, most of the signs flipped. It implies that the prediction of ex-ante and ex-post bubbles should be conducted differently.

[Insert Table [10] about here]

6.5 Robustness Tests with Alternative Measures

To test the robustness of the relationships between narratives and bubbles, we substitute the bubbliness proxy with the alternative measures and report the results of the predictive regressions and the out-of-sample performance. Firstly, we report the correlations between the pairs of the bubblliness measures in Table [11] and visualize the time series in Figure [4]. The correlations are large and statistically significant, and there are observable commen trends. As a comparison, we also plot the price-to-dividend explosiveness. Comparing with the other measures, this measure is flat for most of the time, and it has an incorrect jump after the 2008 financial crisis. We do not include it as a proxy because its first difference has low information-to-noise ratio.

[Insert Table [11] about here]

[Insert Figure[4] about here]

We report the predictive regression results in Table [12] and the out-of-sample predictive performance in Table [13]. The coefficients are very similar with those in the baseline analysis. When price elevation is used, all of the five main narrative features are statistically significant. When price deviation is used, market narrative intensity (NI), narrative tones (NT), narrative tone dispersion (NTD) and the relative intensity of opportunity narratives (NI^{NO}) are statistically significant. The signs are identical with those in the baseline analysis and the scales are also similar. In terms of the out-of-sample forecasting performance, the outperformance of narrative or narrative-augmented models over all benchmarks still hold.

[Insert Table [12] about here]

[Insert Table [13] about here]

In our unreported tests (available upon request), we replace the textual sentiment analysis method by the alternatives such as BERT; we change the NI proxy to be Sim_{Market}; we modify the training set selection and the number of topics for the LDA analysis; we also use alternative seed narratives to generate the opportunity, risk and bubble narratives. Our main results hold with those changes.

7 Conclusion

With the guide of bubble theories, we extract interpretable and important features from market narratives to predict country-level bubble measures. Our results confirm the robust forecast improvement of the narrative-augmented models. Market narratives, as a nexus linking investor beliefs and economic events, possess the capability to not only forecast ex-ante but also to retrospectively assess bubble measures. Given the centrality of investor beliefs in bubble theories, incorporating information about such beliefs is crucial, and narratives serve as a valuable repository of this information. Narratives have the potential to help evaluate and prepare for asset bubbles.

References

D. Abreu and M. K. Brunnermeier. Bubbles and crashes. *Econometrica*, 71(1):173–204, 2003.

G. A. Akerlof and R. J. Shiller. *Animal spirits: How human psychology drives the economy, and why it matters for global capitalism.* Princeton University Press, 2010.

R. Z. Aliber and C. P. Kindleberger. *Manias, panics, and crashes: A history of financial crises.* Springer, 2015.

A. K. Anundsen, K. Gerdrup, F. Hansen, and K. Kragh-Sørensen. Bubbles and crises: The role of house prices and credit. *Journal of Applied Econometrics*, 31(7):1291–1311, 2016.

D. Ardia, K. Bluteau, and K. Boudt. Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting*, 35(4):1370–1386, 2019.

C. Azariadis. Self-fulfilling prophecies. *Journal of Economic Theory*, 25(3):380–396, 1981.

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

S. R. Baker, N. Bloom, and S. J. Davis. Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636, 2016.

L. Barbaglia, S. Consoli, and S. Manzan. Forecasting with economic news. *Journal of Business & Economic Statistics*, pages 1–12, 2022.

B. M. Barber and T. Odean. Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1):261–292, 2001.

N. Barberis, A. Shleifer, and R. Vishny. A model of investor sentiment. *Journal of Financial Economics*, 49(3):307–343, 1998.

N. Barberis, R. Greenwood, L. Jin, and A. Shleifer. Extrapolation and bubbles. *Journal of Financial Economics*, 129(2):203–227, 2018.

T. Basse, T. Klein, S. A. Vigne, and C. Wegener. Us stock prices and the dot. com-bubble: Can dividend policy rescue the efficient market hypothesis? *Journal of Corporate Finance*, 67:101892, 2021.

D. Berger and H. J. Turtle. Sentiment bubbles. *Journal of Financial Markets*, 23:59–74, 2015.

C. Bertsch, I. Hull, and X. Zhang. Narrative fragmentation and the business cycle. *Economics Letters*, 201:109783, 2021.

O. J. Blanchard and M. W. Watson. Bubbles, rational expectations and financial markets. *NBER working paper*, (w0945), 1982.

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

M. Brunnermeier, S. Rother, and I. Schnabel. Asset price bubbles and systemic risk. *The Review of Financial Studies*, 33(9):4272–4317, 2020.

Y. Chen, D. Bredin, V. Potì, and R. Matkovskyy. Covid risk narratives: a computational linguistic approach to the econometric identification of narrative risk during a pandemic. *Digital Finance*, (4):17–61, 2022.

T. E. Clark and K. D. West. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311, 2007.

R. S. Dale, J. E. Johnson, and L. Tang. Financial markets can go mad: evidence of irrational behaviour during the south sea bubble. *The Economic History Review*, 58(2):233–271, 2005.

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

J. B. De Long, A. Shleifer, L. H. Summers, and R. J. Waldmann. Positive feedback investment strategies and destabilizing rational speculation. *The Journal of Finance*, 45(2): 379–395, 1990.

Y. Deng, E. Girardin, R. Joyeux, and S. Shi. Did bubbles migrate from the stock to the housing market in china between 2005 and 2010? *Pacific Economic Review*, 22(3): 276–292, 2017.

B. T. Diba and H. I. Grossman. Explosive rational bubbles in stock prices? *The American Economic Review*, 78(3):520–530, 1988.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1):134–144, 2002.

K. Eliaz and R. Spiegler. A model of competing narratives. *American Economic Review*, 110(12):3786–3816, 2020.

T. Engsted, S. J. Hviid, and T. Q. Pedersen. Explosive bubbles in house prices? Evidence from the OECD countries. *Journal of International Financial Markets, Institutions and Money*, 40:14–25, 2016.

G. W. Evans. Pitfalls in testing for explosive bubbles in asset prices. *The American Economic Review*, 81(4):922–930, 1991.

R. P. Flood and R. J. Hodrick. On testing for speculative bubbles. *Journal of Economic Perspectives*, 4(2):85–101, 1990.

K. A. Froot and M. Obstfeld. Intrinsic bubbles: The case of stock prices. *American Economic Review*, 81(5):1189–1214, 1991.

S. Giglio, M. Maggiori, and J. Stroebel. No-bubble condition: Model-free tests in housing markets. *Econometrica*, 84(3):1047–1091, 2016.

W. N. Goetzmann. Bubble investing: Learning from history. Technical report, National Bureau of Economic Research, 2015.

R. Greenwood, A. Shleifer, and Y. You. Bubbles for Fama. *Journal of Financial Economics*, 131(1):20–43, 2019.

J. M. Harrison and D. M. Kreps. Speculative investor behavior in a stock market with heterogeneous expectations. *The Quarterly Journal of Economics*, 92(2):323–336, 1978.

H. Hong and J. C. Stein. Disagreement and the stock market. *Journal of Economic Perspectives*, 21(2):109–128, 2007.

H. Hong, J. Scheinkman, and W. Xiong. Asset float and speculative bubbles. *The Journal of Finance*, 61(3):1073–1117, 2006.

Y. Hu and L. Oxley. Do 18th century 'bubbles' survive the scrutiny of 21st century time series econometrics? *Economics Letters*, 162:131–134, 2018a.

Y. Hu and L. Oxley. Bubble contagion: Evidence from japan's asset price bubble of the 1980-90s. *Journal of the Japanese and International Economies*, 50:89–95, 2018b.

Ò. Jordà, M. Schularick, and A. M. Taylor. Leveraged bubbles. *Journal of Monetary Economics*, 76:S1–S20, 2015.

C. Kearney and S. Liu. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185, 2014.

J. S. Kim, D. Ryu, and S. W. Seo. Investor sentiment and return predictability of disagreement. *Journal of Banking & Finance*, 42:166–178, 2014.

V. H. Larsen, L. A. Thorsrud, and J. Zhulanova. News-driven inflation expectations and information rigidities. *Journal of Monetary Economics*, 117:507–520, 2021.

V. Leone and O. R. de Medeiros. Signalling the dotcom bubble: a multiple changes in persistence approach. *The Quarterly Review of Economics and Finance*, 55:77–86, 2015.

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

T. Loughran and B. McDonald. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230, 2016.

J. Ma, X. Li, L. Lu, W. Wu, and X. Xiong. Individual investors' dispersion in beliefs and stock returns. *Financial Management*, 51(3):929–953, 2022.

E. M. Miller. Risk, uncertainty, and divergence of opinion. *The Journal of Finance*, 32(4): 1151–1168, 1977.

K. P. Nimark and S. Pitschner. News media and delegated information choice. *Journal of Economic Theory*, 181:160–196, 2019.

R. Nyman, S. Kapadia, and D. Tuckett. News and narratives in financial systems: exploiting big data for systemic risk assessment. *Journal of Economic Dynamics and Control*, 127: 104119, 2021.

E. Pavlidis, A. Yusupova, I. Paya, D. Peel, E. Martínez-García, A. Mack, and V. Grossman. Episodes of exuberance in housing markets: in search of the smoking gun. *The Journal of Real Estate Finance and Economics*, 53(4):419–449, 2016.

E. G. Pavlidis, I. Paya, and D. A. Peel. Using market expectations to test for speculative bubbles in the crude oil market. *Journal of Money, Credit and Banking*, 50(5):833–856, 2018.

P. C. Phillips and S.-P. Shi. Financial bubble implosion and reverse regression. *Econometric Theory*, 34(4):705–753, 2018.

P. C. Phillips, S. Shi, and J. Yu. Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500. *International Economic Review*, 56(4):1043–1078, 2015.

J. A. Scheinkman and W. Xiong. Overconfidence and speculative bubbles. *Journal of Political Economy*, 111(6):1183–1220, 2003.

E. See-To, Y. Yang, et al. Market sentiment dispersion and its effects on stock return and volatility. *Electronic Markets*, 27(3):283–296, 2017.

A. H. Shapiro, M. Sudhof, and D. J. Wilson. Measuring news sentiment. *Journal of Econometrics*, 2020.

R. J. Shiller. Measuring bubble expectations and investor confidence. *The Journal of Psychology and Financial Markets*, 1(1):49–60, 2000.

R. J. Shiller. *Irrational Exuberance*. Princeton University Press, 2015.

R. J. Shiller. Narrative economics. *American Economic Review*, 107(4):967–1004, 2017.

R. J. Shiller. *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*. Princeton University Press, 2019.

S. Ter Ellen, V. H. Larsen, and L. A. Thorsrud. Narrative monetary policy surprises and the media. *Journal of Money, Credit and Banking*, 2021.

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

N. Vozlyublennaia. Investor attention, index performance, and return predictability. *Journal of Banking & Finance*, 41:17–35, 2014.

X. Xiong, J. Han, X. Feng, and Y. An. Sentiment dispersion and asset pricing error: Evidence from the chinese stock market. *Emerging Markets Finance and Trade*, 56(4):820–839, 2020.

J. Yu. Disagreement and return predictability of stock portfolios. *Journal of Financial Economics*, 99(1):162–183, 2011.

Table 1: Correlations between narrative variables

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
NI	1.0***	-0.07***	0.1***	-0.21***	-0.11***	0.05***	0.04***	0.04***	0.2***
NT	-0.07***	1.0***	-0.24***	0.0	0.12***	0.11***	-0.09***	-0.1***	-0.15***
NTD	0.1***	-0.24***	1.0***	-0.32***	-0.2***	0.05***	0.07***	0.14***	0.09***
TC	-0.21***	0.0	-0.32***	1.0***	0.17***	-0.19***	-0.09***	-0.08***	-0.09***
NI^{NO}	-0.11***	0.12***	-0.2***	0.17***	1.0***	0.19***	-0.48***	-0.54***	0.01
$NI^{Opportunity}$	0.05***	0.11***	0.05***	-0.19***	0.19***	1.0***	0.25***	0.2***	0.03**
NI^{Bubble}	0.04***	-0.09***	0.07***	-0.09***	-0.48***	0.25***	1.0***	0.23***	0.01
NI^{Risk}	0.04***	-0.1***	0.14***	-0.08***	-0.54***	0.2***	0.23***	1.0***	0.16***
EPU	0.2***	-0.15***	0.09***	-0.09***	0.01	0.03**	0.01	0.16***	1.0***

Notes: This table presents the Pearson correlations between the narrative variables. ***, ** and * indicate significance level of 1%, 5% and 10%, respectively.

Table 2: Data coverage

	Obs.	Number of countries	Avg. Obs.	Category	Description
NI	26,189	48	545	Narrative	Narrative intensity
NT	26,189	48	545	Narrative	Narrative tone
NTD	25,341	48	527	Narrative	Narrative tone dispersion
TC	26,189	48	545	Narrative	Topic consensus
NI^{NO}	26,189	48	545	Narrative	Net opportunity narrative intensity
$NI^{Opportunity}$	26,189	48	545	Narrative	Bubble narrative intensity
NI^{Bubble}	26,189	48	545	Narrative	Opportunity narrative intensity
NI^{Risk}	26,189	48	545	Narrative	Risk narrative intensity
EPU	7,317	21	348	Narrative	Economic Policy Uncertainty Index
DY	19,918	46	433	Fin./Econ.	Dividend yield
PE	19,643	46	427	Fin./Econ.	Price-to-earning ratio
PB	17,833	45	396	Fin./Econ.	Price-to-book ratio
IR	12,676	31	408	Fin./Econ.	Interest rate
Inflation	17,112	31	552	Fin./Econ.	log difference of CPI
RGDP	12,135	23	527	Fin./Econ.	Real GDP growth
Investment	15,861	32	495	Fin./Econ.	Investment
GDP	16,035	34	471	Fin./Econ.	GDP
VO	17,169	46	373	Fin./Econ.	Trading Volume
CLI	15,161	31	489	Fin./Econ.	OECD composite leading indicator

Notes: This table presents the total number of observations, the number of countries with coverage, and the observations per country for both narrative and financial or economic measures.

Table 3: Bubbliness prediction with U.S. data

Panel A: Univariate regressions

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.19***	0.197***	-0.105*	-0.091**	0.16***	0.094*	-0.071	-0.048	-0.107
	(0.069)	(0.057)	(0.058)	(0.043)	(0.059)	(0.057)	-0.045	(0.039)	(0.07)
Controls	No	No	No	No	No	No	No	No	No
No. of obs.	503	503	503	503	503	503	503	503	443
Adj. R-squared	3.43%	3.68%	0.90%	0.63%	2.37%	0.69%	0.30%	0.03%	0.93%

Panel B: Regressions with controls

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.193***	0.22***	-0.132**	-0.102**	0.156***	0.086	-0.054	-0.044	-0.138
	(0.07)	(0.064)	(0.06)	(0.044)	(0.056)	(0.058)	(0.048)	(0.039)	(0.092)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of obs.	501	501	501	501	501	501	501	501	442
Adj. R-squared	5.22%	5.87%	3.35%	2.65%	3.99%	2.35%	1.93%	1.84%	3.18%

Panel C: Univariate regressions using the control variables

	DY	PE	PB	IR	Inflation	RGDP	InvGDP	VO	CLI
Coefficient	0.019	-0.039	-0.067	-0.094*	-0.0	0.045	0.029	-0.002	0.049
	(0.038)	(0.043)	(0.053)	(0.053)	(0.035)	(0.052)	(0.048)	(0.047)	(0.044)
No. of obs.	503	503	503	503	502	503	503	503	503
Adj. R-squared	-0.16%	-0.05%	0.25%	0.65%	-0.20%	0.00%	-0.12%	-0.20%	0.04%

Notes: This table presents the estimated OLS coefficients for

$$\triangle B_{c,t+1} = \alpha_c + \boldsymbol{\theta} \cdot \boldsymbol{X_{c,t}} + \gamma \cdot N_{c,t} + \epsilon_{c,t+1},\ c = U.S.,$$

with Newey–West standard errors reported in parentheses. The dependent variable ($\triangle B_{c,t+1}$) is the change of bubbliness, which was proxied by the $BSADF$ statistics of the price index. $\boldsymbol{X_{c,t}}$ is a set of financial/economic variables, and $N_{c,t}$ is one of the narrative features - market narrative intensity (NI), textual sentiment (NT), tone dispersion (NTD), topic consensus (TC), opportunity, risk and bubble narratives ($NI^{Opportunity}$,NI^{Risk},NI^{Bubble}), and the relative intensity of competing narratives (NI^{NO}). All variables by subtracting the average and dividing the standard deviation. All independent variables are lagged. ***, ** and * indicate significance level of 1%, 5% and 10%, respectively.

Table 4: Granger causality tests

Panel A: Narrative Variables

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
To bubbliness	***0.0%***	***0.0%***	**2.6%**	**2.4%**	***0.1%***	**2.0%**	25.3%	15.5%	***0.0%***
From bubbliness	38.9%	57.7%	63.1%	61.7%	23.7%	**2.7%**	75.1%	38.0%	69.9%

Panel B: Financial/Economic Variables

	DY	PE	PB	IR	Inflation	RGDP	InvGDP	VO	CLI
To bubbliness	59.8%	95.8%	17.1%	**2.1%**	59.5%	77.1%	90.9%	92.4%	***0.6%***
From bubbliness	18.3%	*6.0%*	**1.3%**	39.4%	45.2%	***1.0%***	**1.6%**	52.0%	80.0%

Notes: This table presents the Wald p-values for the Granger causality tests between the considered variables and the bubbliness measure, using U.S. data. Bold and italic values indicate significance at the 1% level, Bold values indicate significance at the 5% level, and italic values indicate significance at the 10% level.

Table 5: Max drop prediction with U.S. data

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.178**	0.14**	-0.109*	-0.039	0.114*	0.11*	-0.031	-0.015	-0.185**
	(0.09)	(0.07)	(0.057)	(0.055)	(0.066)	(0.057)	(0.056)	(0.062)	(0.091)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of obs.	490	490	490	490	490	490	490	490	431
Adj. R-squared	7.16%	5.91%	5.25%	4.25%	5.35%	5.25%	4.19%	4.12%	8.36%

Notes: This table presents the estimated OLS coefficients for

$$\triangle MaxDrop_{c,t+12|t} = \alpha_c + \boldsymbol{\theta} \cdot \boldsymbol{X_{c,t}} + \gamma \cdot N_{c,t} + \epsilon_{c,t+1},\ c = U.S.,$$

with Newey–West standard errors reported in parentheses. The dependent variable ($\triangle MaxDrop_{c,t+12|t}$) is the change of the magnitude of future max drop. $\boldsymbol{X_{c,t}}$ is a set of financial/economic variables, and $N_{c,t}$ is one of the narrative features - market narrative intensity (NI), textual sentiment (NT), tone dispersion (NTD), topic consensus (TC), opportunity, risk and bubble narratives ($NI^{Opportunity}$,NI^{Risk},NI^{Bubble}), and the relative intensity of competing narratives (NI^{NO}). All variables by subtracting the average and dividing the standard deviation. All independent variables are lagged. ***, ** and * indicate significance level of 1%, 5% and 10%, respectively.

Table 6: Bubbliness prediction with international data

U.S. data excluded

Panel A: Univariate regressions

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.053**	0.163***	-0.061***	-0.001	0.087***	0.081***	0.009	-0.051***	-0.063***
	(0.025)	(0.034)	(0.013)	(0.017)	(0.028)	(0.022)	(0.016)	(0.017)	(0.022)
Controls	No	No	No	No	No	No	No	No	No
No. of obs.	5,140	5,140	5,140	5,140	5,140	5,140	5,140	5,140	3,488
Adj. R-squared	0.31%	2.62%	0.41%	0.04%	0.77%	0.68%	0.04%	0.29%	0.43%

Panel B: Regressions with controls

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.028*	0.174***	-0.085***	-0.022	0.107***	0.108***	0.022	-0.064**	-0.058**
	(0.016)	(0.056)	(0.02)	(0.025)	(0.041)	(0.032)	(0.021)	(0.028)	(0.023)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of obs.	2,371	2,371	2,371	2,371	2,371	2,371	2,371	2,371	2,135
Adj. R-squared	1.96%	4.73%	2.59%	1.93%	2.97%	2.99%	1.93%	2.27%	2.29%

U.S. data included

Panel C: Univariate regressions

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.08***	0.164***	-0.06***	-0.004	0.091***	0.081***	0.004	-0.05***	-0.066***
	(0.026)	(0.033)	(0.013)	(0.016)	(0.028)	(0.022)	(0.016)	(0.016)	(0.025)
Controls	No	No	No	No	No	No	No	No	No
No. of obs.	5,642	5,642	5,642	5,642	5,642	5,642	5,642	5,642	3,931
Adj. R-squared	0.66%	2.61%	0.38%	0.02%	0.81%	0.66%	0.02%	0.26%	0.45%

Panel D: Regressions with controls

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.085**	0.171***	-0.081***	-0.029	0.108***	0.103***	0.011	-0.057**	-0.068**
	(0.035)	(0.051)	(0.02)	(0.025)	(0.038)	(0.029)	(0.021)	(0.023)	(0.028)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of obs.	2,873	2,873	2,873	2,873	2,873	2,873	2,873	2,873	2,578
Adj. R-squared	2.58%	4.61%	2.53%	1.96%	2.99%	2.88%	1.89%	2.19%	2.23%

Notes: This table presents the estimated OLS coefficients for

$$\triangle B_{c,t+1} = \alpha_c + \boldsymbol{\theta} \cdot \boldsymbol{X_{c,t}} + \gamma \cdot N_{c,t} + \epsilon_{c,t+1},$$

with clustered (by country/region and month) standard errors reported in parentheses. The dependent variable ($\triangle B_{c,t+1}$) is the change of the bubbliness measure, which was proxied by the $BSADF$ statistics of the price index. $\boldsymbol{X_{c,t}}$ is a set of financial/economic variables, and $N_{c,t}$ is one of the narrative features - market narrative intensity (NI), textual sentiment (NT), tone dispersion (NTD), topic consensus (TC), opportunity, risk and bubble narratives ($NI^{Opportunity}$,NI^{Risk},NI^{Bubble}), and the relative intensity of competing narratives (NI^{NO}). All variables by subtracting the average and dividing the standard deviation. All independent variables are lagged. ***, ** and * indicate significance level of 1%, 5% and 10%, respectively.

Table 7: Max drop prediction with international data

U.S. data excluded

Panel A: Univariate regressions

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.11**	0.2***	-0.067***	-0.012	0.131***	0.125***	-0.025	-0.044**	-0.109***
	(0.046)	(0.036)	(0.02)	(0.021)	(0.034)	(0.031)	(0.017)	(0.018)	(0.042)
Controls	No	No	No	No	No	No	No	No	No
No. of obs.	5,248	5,248	5,248	5,248	5,248	5,248	5,248	5,248	3,547
Adj. R-squared	1.22%	3.92%	0.45%	0.01%	1.67%	1.55%	0.06%	0.19%	1.21%

Panel B: Regressions with controls

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.099*	0.164***	-0.024	-0.013	0.146***	0.107***	-0.049*	-0.056*	-0.077*
	(0.056)	(0.035)	(0.028)	(0.031)	(0.043)	(0.035)	(0.028)	(0.033)	(0.045)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of obs.	2,302	2,302	2,302	2,302	2,302	2,302	2,302	2,302	2,066
Adj. R-squared	3.60%	5.24%	2.69%	2.65%	4.73%	3.76%	2.86%	2.93%	3.14%

U.S. data included

Panel A: Univariate regressions

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.1**	0.196***	-0.066***	-0.012	0.13***	0.124***	-0.026	-0.043**	-0.115***
	(0.039)	(0.036)	(0.019)	(0.021)	(0.034)	(0.03)	(0.017)	(0.018)	(0.043)
Controls	No	No	No	No	No	No	No	No	No
No. of obs.	5,739	5,739	5,739	5,739	5,739	5,739	5,739	5,739	3,978
Adj. R-squared	1.01%	3.79%	0.44%	0.01%	1.64%	1.53%	0.07%	0.18%	1.34%

Panel B: Regressions with controls

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.099***	0.157***	-0.027	-0.013	0.143***	0.108***	-0.047*	-0.051*	-0.086*
	(0.036)	(0.034)	(0.027)	(0.03)	(0.041)	(0.034)	(0.027)	(0.031)	(0.047)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of obs.	2,792	2,792	2,792	2,792	2,792	2,792	2,792	2,792	2,497
Adj. R-squared	3.72%	5.12%	2.81%	2.75%	4.75%	3.87%	2.95%	2.99%	3.40%

Notes: This table presents the estimated OLS coefficients for

$$\triangle MaxDrop_{c,t+12|t} = \alpha_c + \boldsymbol{\theta} \cdot \boldsymbol{X_{c,t}} + \gamma \cdot N_{c,t} + \epsilon_{c,t+1},$$

with clustered (by country/region and month) standard errors reported in parentheses. The dependent variable ($\triangle MaxDrop_{c,t+12|t}$) is the change of the magnitude of future max drop. $\boldsymbol{X_{c,t}}$ is a set of financial/economic variables, and $N_{c,t}$ is one of the narrative features - market narrative intensity (NI), textual sentiment (NT), tone dispersion (NTD), topic consensus (TC), opportunity, risk and bubble narratives ($NI^{Opportunity}$,NI^{Risk},NI^{Bubble}), and the relative intensity of competing narratives (NI^{NO}). All variables by subtracting the average and dividing the standard deviation. All independent variables are lagged. ***, ** and * indicate significance level of 1%, 5% and 10%, respectively.

Table 8: OOS performance for bubblliness prediction

	Benchmark: Controls			Benchmark: Historical mean			Benchmark: Zero		
Model	MSE ratio	DM/CW statistic	p-value	MSE ratio	DM/CW statistic	p-value	MSE ratio	DM/CW statistic	p-value
Controls	1.000			1.218	0.47	31.92%	1.221	0.42	33.88%
NI	0.823	2.163	**1.53%**	1.002	2.59	***0.48%***	1.004	2.54	***0.55%***
NT	0.857	1.776	**3.79%**	1.044	4.35	***0.00%***	1.046	4.28	***0.00%***
NTD	0.831	2.175	**1.48%**	1.013	1.27	10.17%	1.015	1.13	12.94%
TC	0.821	2.342	***0.96%***	1.000	2.24	**1.25%**	1.002	2.12	**1.68%**
NI^{NO}	0.826	2.280	**1.13%**	1.006	3.24	***0.06%***	1.008	3.19	***0.07%***
$NI^{Opportunity}$	0.807	2.599	***0.47%***	0.983	3.23	***0.06%***	0.985	3.15	***0.08%***
NI^{Bubble}	0.818	2.369	***0.89%***	0.996	1.98	**2.38%**	0.998	1.80	**3.60%**
NI^{Risk}	0.836	2.157	**1.55%**	1.018	0.81	20.87%	1.021	0.69	24.56%
EPU	0.833	2.197	**1.40%**	1.015	1.32	*9.37%*	1.017	1.29	*9.87%*
Narrative	0.882	1.423	*7.73%*	1.074	5.06	***0.00%***	1.076	5.02	***0.00%***
Controls + Narra	1.042	3.690	***0.01%***	1.269	3.63	***0.01%***	1.272	3.60	***0.02%***
Ensemble	0.780	3.151	***0.08%***	0.950	2.22	**1.33%**	0.952	2.16	**1.55%**

Notes: This table presents the out-of-sample bubbliness forecasting performance. "Narra" stands for narrative variables. A model is tested using the Clark and West (2007) test when the benchmark model is a "nested model" of the tested model. The ensembles model was tested using the Diebold and Mariano (2002) tests. Bold and italic values indicate significance at the 1% level, Bold values indicate significance at the 5% level, and italic values indicate significance at the 10% level.

Table 9: OOS performance in max drop prediction

	Benchmark: Controls			Benchmark: Zero		
Model	MSE ratio	DM/CW statistic	p-value	MSE ratio	DM/CW statistic	p-value
Controls	1.000			1.088	4.81	***0.00%***
NI	0.902	0.79	21.43%	0.981	2.61	***0.46%***
NT	0.936	0.52	30.11%	1.018	1.58	*5.72%*
NTD	0.922	0.64	26.19%	1.002	0.62	26.82%
TC	0.932	0.55	29.09%	1.013	1.47	*7.06%*
NI^{NO}	0.913	0.71	23.86%	0.993	2.83	***0.23%***
$NI^{Opportunity}$	0.903	0.80	21.24%	0.982	3.07	***0.11%***
NI^{Bubble}	0.925	0.61	27.22%	1.006	-1.17	87.96%
NI^{Risk}	0.934	0.54	29.60%	1.016	0.29	38.48%
EPU	0.896	0.84	20.07%	0.975	3.04	***0.12%***
Narra	0.923	0.63	26.53%	1.003	3.34	***0.04%***
Controls + Narra	1.008	1.86	**3.14%**	1.097	4.90	***0.00%***
Ensemble	0.873	1.07	14.22%	0.950	3.42	***0.03%***

Notes: This table presents the out-of-sample market drop forecasting performance. "Narra" stands for narrative variables. A model is tested using the Clark and West (2007) test when the benchmark model is a "nested model" of the tested model. The ensembles model was tested using the Diebold and Mariano (2002) tests. Bold and italic values indicate significance at the 1% level, Bold values indicate significance at the 5% level, and italic values indicate significance at the 10% level.

Table 10: After-the-fact bubble prediction

	1 month	6 months	12 months	24 months	1 month	6 months	12 months	24 months
NI	-0.029	0.022	0.041	-0.031	-0.256	-0.218	0.04	-0.032
	(0.099)	(0.112)	(0.122)	(0.174)	(0.176)	(0.171)	(0.114)	(0.149)
NT	-0.747**	-0.306	-0.309	-0.116	-1.003***	-0.406	-0.353	-0.13
	(0.336)	(0.257)	(0.252)	(0.234)	(0.363)	(0.381)	(0.392)	(0.351)
NTD	-0.376	0.139	-0.158	0.018	-0.648**	0.166	0.259	0.13
	(0.234)	(0.239)	(0.249)	(0.276)	(0.287)	(0.326)	(0.263)	(0.348)
TC	0.021	0.645*	0.631**	0.612*	0.875	1.487***	1.594***	1.271**
	(0.336)	(0.352)	(0.306)	(0.333)	(0.605)	(0.537)	(0.479)	(0.509)
NI^{NO}	-0.622***	-0.546**	-0.641***	-0.421**	-0.608*	-0.384	-0.497	-0.322
	(0.222)	(0.24)	(0.221)	(0.209)	(0.367)	(0.312)	(0.306)	(0.29)
Obs. of signals	125	125	123	121	102	102	101	98
Obs. of crashes	30	30	29	28	24	24	24	24
Pseudo R-squared	6.19%	5.93%	6.19%	3.84%	16.21%	15.12%	11.83%	9.12%
Controls	No	No	No	No	Yes	Yes	Yes	Yes

Notes: This table presents the coefficients of the narrative variables in predicting ex-post bubbles conditional on bubble signals:

$$Prob(Bubble_{c,t}|Signal_{c,t}) = \frac{1}{e^{-(\alpha+\boldsymbol{\theta}\cdot\boldsymbol{X_{c,[t-w,t]}}+\gamma\cdot N_{c,[t-w,t]}+\epsilon_t)}}.$$

The headers refer to the length of windows for the estimation of narrative attributes. ***, ** and * indicate significance level of 1%, 5% and 10%, respectively.

Table 11: Correlations between the bubbliness measures

	$BSADF$	$Elev$	Dev
$BSADF$	1.0***	0.52***	0.32***
$Elev$	0.52***	1.0***	0.25***
Dev	0.32***	0.25***	1.0***

Notes: This table presents the correlations between the bubbliness measures, namely, price explosiveness ($BSADF$), price elevation ($Elev$) and price deviation (Dev). ***, ** and * indicate significance level of 1%, 5% and 10%, respectively.

Table 12: Alternative bubbliness prediction with panel data

Panel A: Price Elevation

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.089***	0.249***	-0.11***	-0.05**	0.181***	0.155***	-0.028	-0.096***	-0.082**
	(0.026)	(0.041)	(0.026)	(0.024)	(0.039)	(0.03)	(0.025)	(0.027)	(0.035)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of obs.	2,793	2,793	2,793	2,793	2,793	2,793	2,793	2,793	2,576
Adj. R-squared	4.45%	9.46%	4.87%	3.92%	6.81%	5.98%	3.76%	4.55%	4.51%

Panel B: Price Deviation

	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.127***	0.35***	-0.089***	-0.025	0.211***	0.193***	-0.032	-0.094***	-0.138***
	(0.036)	(0.044)	(0.024)	(0.036)	(0.034)	(0.033)	(0.031)	(0.029)	(0.048)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of obs.	2,873	2,873	2,873	2,873	2,873	2,873	2,873	2,873	2,578
Adj. R-squared	6.70%	16.51%	5.89%	5.18%	9.39%	8.67%	5.21%	5.96%	6.72%

Notes: This table presents the estimated OLS coefficients for

$$\triangle B_{c,t+1} = \alpha_c + \boldsymbol{\theta} \cdot \boldsymbol{X_{c,t}} + \gamma \cdot N_{c,t} + \epsilon_{c,t+1},$$

with clustered (by country/region and month) standard errors reported in parentheses. The dependent variable ($\triangle B_{c,t+1}$) is the change of the bubbliness measure, which was proxied by the Elv or Dev statistics of the price index. $\boldsymbol{X_{c,t}}$ is a set of financial/economic variables, and $N_{c,t}$ is one of the narrative features - market narrative intensity (NI), textual sentiment (NT), tone dispersion (NTD), topic consensus (TC), opportunity, risk and bubble narratives ($NI^{Opportunity}$,NI^{Risk},NI^{Bubble}), and the relative intensity of competing narratives (NI^{NO}). All variables by subtracting the average and dividing the standard deviation. All independent variables are lagged. ***, ** and * indicate significance level of 1%, 5% and 10%, respectively.

Table 13: OOS forecasting performance for alternative bubblliness measures

	Price Elevation						Price Deviation					
	Benchmark: Controls			Benchmark: Historical mean			Benchmark: Controls			Benchmark: Historical mean		
Model	MSE ratio	DM/CW statistic	p-value	MSE ratio	DM/CW statistic	p-value	MSE ratio	DM/CW statistic	p-value	MSE ratio	DM/CW statistic	p-value
Controls	1.000			1.124	3.07	***0.11%***	1.000			1.288	0.46	32.43%
NI	0.858	2.27	**1.16%**	0.964	4.19	***0.00%***	0.737	1.98	**2.37%**	0.949	4.36	***0.00%***
NT	0.838	2.54	***0.56%***	0.942	6.72	***0.00%***	0.719	2.14	**1.61%**	0.926	7.18	***0.00%***
NTD	0.874	2.06	**1.99%**	0.983	3.64	***0.01%***	0.773	1.73	**4.20%**	0.995	2.26	**1.19%**
TC	0.884	1.90	**2.90%**	0.993	3.11	***0.09%***	0.765	1.78	**3.78%**	0.986	3.11	***0.10%***
NI^{NO}	0.866	2.15	**1.58%**	0.973	4.85	***0.00%***	0.748	1.93	**2.69%**	0.963	4.69	***0.00%***
$NI^{Opportunity}$	0.866	2.22	**1.31%**	0.973	4.46	***0.00%***	0.731	2.08	**1.86%**	0.942	4.94	***0.00%***
NI^{Bubble}	0.893	1.74	**4.06%**	1.004	0.37	35.50%	0.770	1.74	**4.10%**	0.992	2.21	**1.35%**
NI^{Risk}	0.891	1.77	**3.82%**	1.001	2.32	**1.02%**	0.784	1.64	*5.05%*	1.009	1.10	13.57%
EPU	0.883	1.93	**2.71%**	0.993	3.20	***0.07%***	0.764	1.80	**3.56%**	0.984	2.66	***0.39%***
Narra	0.842	2.50	***0.63%***	0.946	7.49	***0.00%***	0.707	2.27	**1.16%**	0.911	7.20	***0.00%***
Controls + Narra	0.911	5.70	***0.00%***	1.024	7.67	***0.00%***	0.864	4.80	***0.00%***	1.113	5.14	***0.00%***
Ensembel	0.815	3.25	***0.06%***	0.916	5.98	***0.00%***	0.707	2.39	***0.83%***	0.910	4.12	***0.00%***

Notes: This table presents the out-of-sample forecasting performance for the alternative bubbliness measures. "Narra" stands for narrative variables. A model is tested using the Clark and West (2007) test when the benchmark model is a "nested model" of the tested model. The ensembles model was tested using the Diebold and Mariano (2002) tests. Bold and italic values indicate significance at the 1% level, Bold values indicate significance at the 5% level, and italic values indicate significance at the 10% level.

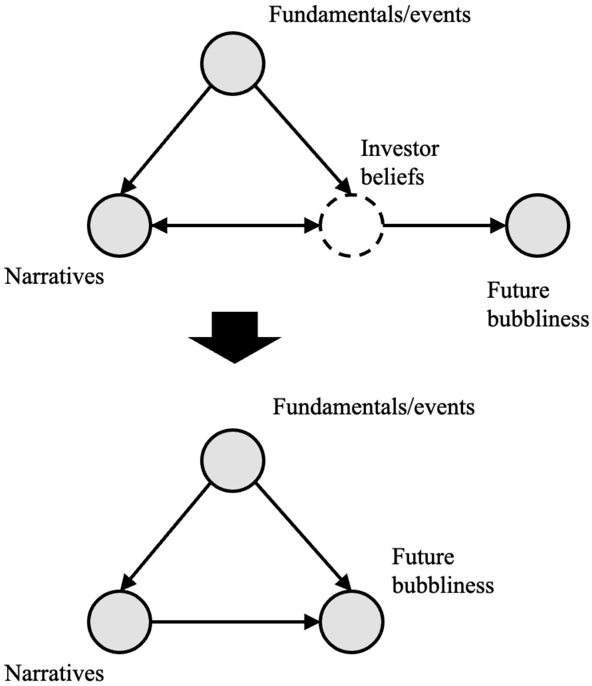

Figure 1: Directed acyclic graph

Notes: The figure displays the directed acyclic graph for the variables.

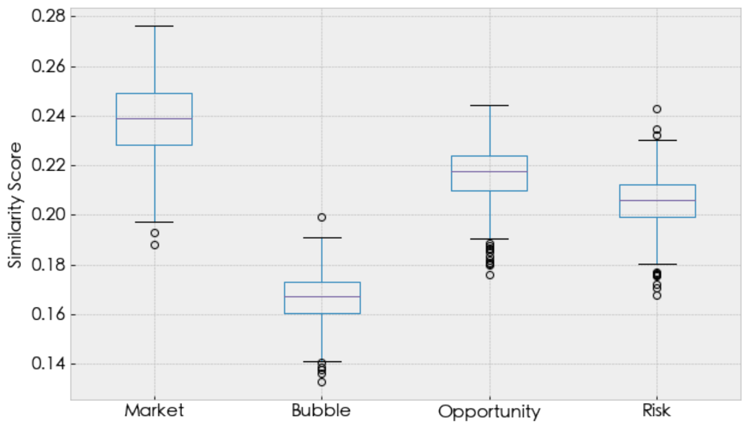

Figure 2: Boxplot for the Narrative Intensity

Notes: The figure displays the Boxplot for the monthly narrative intensity of four narratives. The intensity measures are calculated based on the monthly average of semantic similarity to four narrative queries - "How about the stock market?", "The risk level in the market is high", "It is a good time to invest in the market", and "There is an asset bubble" .

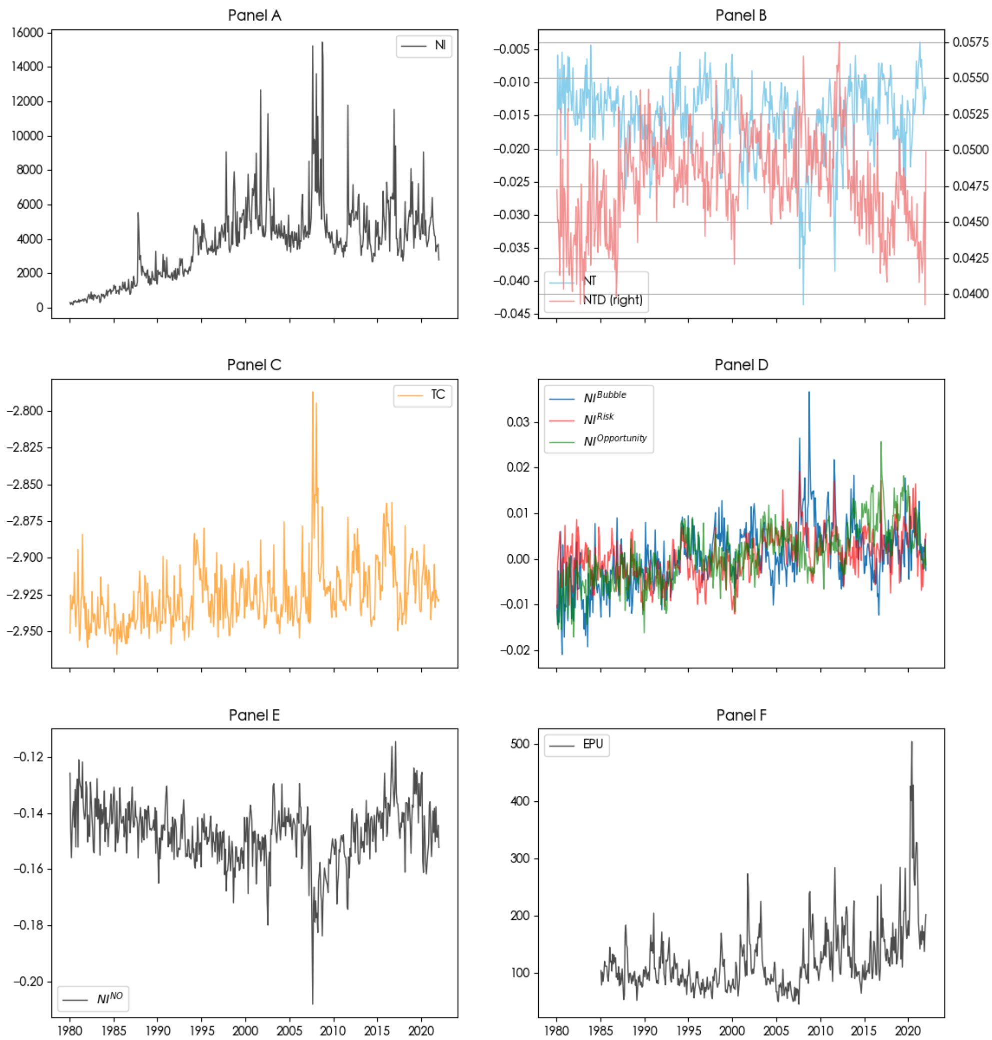

Figure 3: Narrative information sets (for the U.S.)

Notes: The figure displays the time series of the narrative features for the United States. Panel A shows the total number of sentences. Panel B shows the monthly average and the monthly standard deviation L&M sentiment measures of the sentences. Panel C shows the topic consensus. Panel D shows the intensity measures of three narratives. The intensity measures are calculated based on the monthly average of semantic similarity to three narrative queries - "The risk level in the market is high", "It is a good time to invest in the market", and "There is an asset bubble". Panel E shows the relative narrative intensity of the competing narratives between "opportunity" and "risk/bubble". Panel F shows the EPU index.

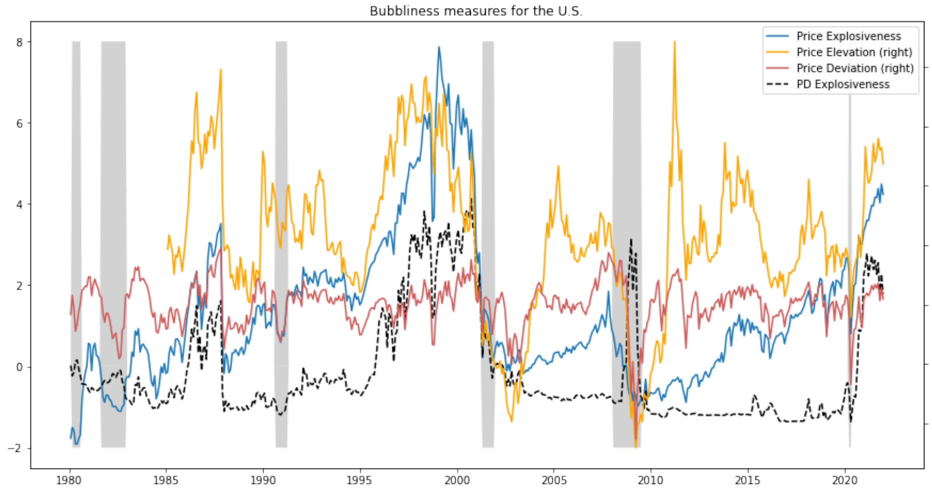

Figure 4: Alternative bubbliness measures for the U.S.

Notes: The figure displays the four bubbliness measures for the U.S. The shaded areas indicate the NBER recession periods.

Appendix A: International Data

Table 14: Countries/Regions Included

Country/Region	Obs. of sentence	Ratio of sentence	Obs.of article	SCO avg	Codes
US	1,914,881	7.99%	502,105	12%	TOTMKUS
UK	676,450	2.82%	233,996	7%	TOTMKUK
China	459,447	1.92%	142,763	7%	TOTMKCH
Japan	327,077	1.36%	147,638	5%	TOTMKJP
Australia	226,136	0.94%	108,596	3%	TOTMKAU
Canada	212,087	0.88%	79,003	4%	TOTMKCN
HK	185,272	0.77%	87,044	4%	TOTMKHK
Germany	158,073	0.66%	87,373	5%	TOTMKBD
Singapore	116,485	0.49%	55,850	4%	TOTMKSG
France	108,445	0.45%	65,320	6%	TOTMKFR
Russia	97,212	0.41%	40,548	4%	MSRUSSL
India	87,569	0.37%	41,818	3%	TOTMKIN
South Korea	78,394	0.33%	44,329	4%	TOTMKKO
Greece	74,743	0.31%	23,481	4%	MSGREEL,TOTMKGR
Spain	70,057	0.29%	33,478	6%	MSSPANL,TOTMKES
Italy	69,396	0.29%	35,920	5%	TOTMKIT
Brazil	69,238	0.29%	36,674	4%	MSBRAZL,TOTMKBR
Malaysia	64,201	0.27%	32,591	4%	TOTMKMY
New Zealand	54,827	0.23%	28,962	3%	MSNZEAL,TOTMKNZ
Indonesia	50,957	0.21%	29,183	5%	MSINDFL,TOTMKID
Thailand	48,625	0.20%	26,856	4%	TOTMKTH
Taiwan	42,904	0.18%	26,737	3%	TOTMKTA
Saudi Arabia	40,341	0.17%	17,931	2%	TOTMKSI
Mexico	37,543	0.16%	20,431	4%	MSMEXFL,TOTMKMX
Nigeria	33,425	0.14%	13,272	1%	TOTMKNG
Ireland	32,997	0.14%	17,541	2%	TOTMKIR
Turkey	29,625	0.12%	14,944	3%	MSTURKL,TOTMKTK
Philippines	28,446	0.12%	18,438	3%	TOTMKPH
South Africa	28,113	0.12%	15,334	1%	TOTMKSA
Belgium	26,870	0.11%	18,105	2%	TOTMKBG
Switzerland	26,810	0.11%	19,222	3%	TOTMKSW
Argentina	23,706	0.10%	11,401	1%	TOTMKAR
Poland	23,284	0.10%	10,020	3%	MSPLNDL,TOTMKPO
Netherlands	22,576	0.09%	17,695	3%	TOTMKNL
Portugal	19,294	0.08%	11,686	3%	MSPORDL,TOTMKPT
United Arab Emirates	17,514	0.07%	8,678	1%	TOTMKAE
Vietnam	15,561	0.06%	7,584	1%	TOTMKVI
Sweden	15,159	0.06%	10,678	3%	MSSWDNL,TOTMKSD
Israel	15,153	0.06%	7,386	2%	TOTMKIS
Kuwait	14,792	0.06%	8,423	2%	TOTMKKW
Venezuela	14,314	0.06%	8,248	2%	TOTMKVE
Hungary	14,187	0.06%	6,823	2%	TOTMKHN
Austria	13,436	0.06%	9,140	2%	TOTMKOE
Egypt	12,815	0.05%	6,840	2%	MSEGYTL,TOTMKEY
Kenya	12,535	0.05%	5,515	2%	MSKNYAL
Qatar	11,396	0.05%	5,690	1%	TOTMKQA
Chile	10,266	0.04%	5,809	1%	MSCHILL,TOTMKCL
Norway	10,105	0.04%	7,101	1%	MSNWAYL,TOTMKNW

Notes: This table presents the meta data for the countries/regions included in the analysis. "SCO" stands for "sentence co-occurrence ratio". It represents the average ratio of other countries' documents that mention this country. The last column lists the codes of country indices on Datastream.

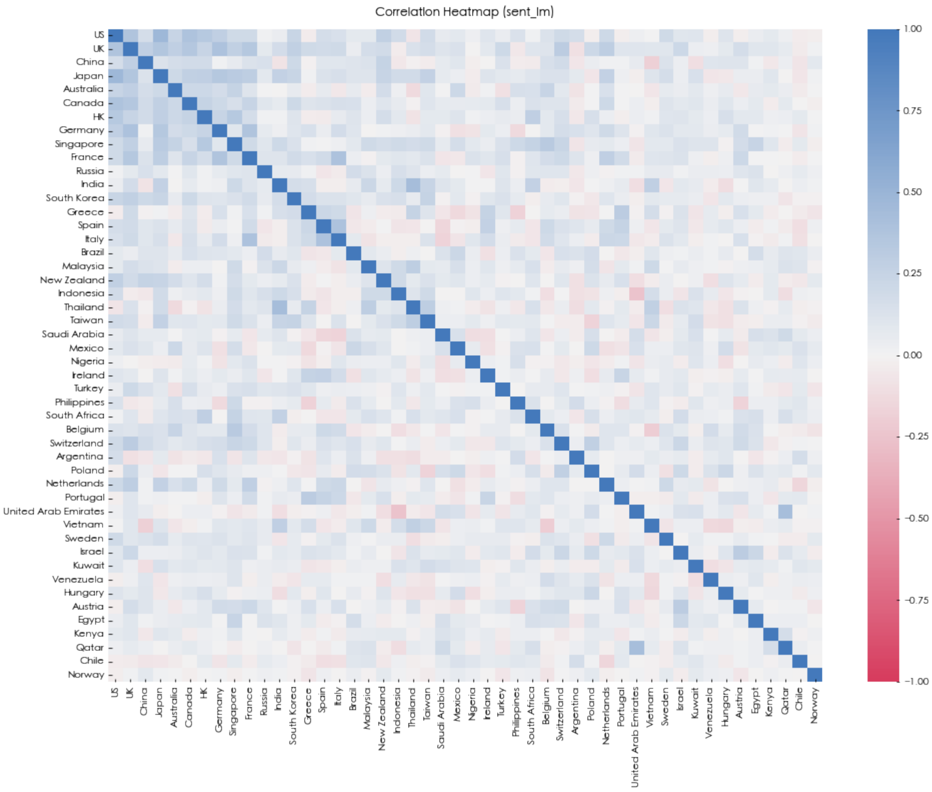

Figure 5: Correlation of the L&M sentiment between countries

Notes: The figure displays the correlation heat map for the L&M sentiment between countries

Appendix B: Sample Narratives

Table 15: Sample narrative snippets

Sentence	Sim Bubble	Sim Risk	Sim Opportunity	Sim Market	Date
There are asset bubbles all over the place, some old (London) and some new (I'm coming to this).	0.82	0.23	0.18	0.26	26/10/2013
Some market specialists worry that asset bubbles akin to the one that inflated and burst in the American housing market might be growing in places like China and Hong Kong.	0.78	0.38	0.31	0.37	30/12/2009
But talk of a US asset market bubble is seriously overdone.	0.77	0.44	0.47	0.50	23/10/1999
Market risks here are perceived to be less than those in the United States, Japan or even Europe.	0.31	0.70	0.40	0.44	13/02/2003
The risk is being priced back into the market, says Chong Yoon Chou, a Singapore-based analyst with Aberdeen Fund Managers.	0.31	0.68	0.49	0.45	02/01/2001
More analysts agree that the bigger danger comes from the US market, which they see as overvalued compared with ours, making it more vulnerable to a fall.	0.22	0.68	0.37	0.40	14/07/1999
Renowned U.S. value investor Mario Gabelli said in an interview with The Nikkei Veritas that a good time to invest is when the market is turbulent.	0.26	0.37	0.78	0.49	23/10/2014
"Go ahead and invest in the stock market," says Bob Hewitt, financial planner from Monterey, Calif. "If you're in for the long term, there are no bad times to invest.	0.22	0.40	0.78	0.54	24/05/1993
But given the intense spate of bad news, it may be a good time to invest if it's for the long term, said Chang who sees the U.S. market as having the biggest upside.	0.26	0.41	0.77	0.52	25/10/2008
The past 10 years have hardly been kind to the US stock market.	0.35	0.45	0.57	0.72	31/08/2012
But we do note with some wonder the resilience of the United States stock market.	0.27	0.50	0.51	0.72	24/11/1997
WHAT KEEPS the U.S. stock market so strong in the face of mediocre corporate earnings and a frightening trade deficit?	0.29	0.49	0.47	0.72	12/01/1987

Notes: This table presents the sample narrative snippets.

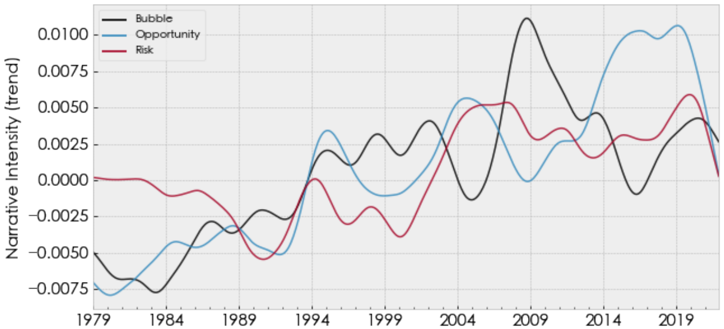

Figure 6: Time Series of the Narrative Intensity (U.S.)

Notes: The figure displays the smoothed time series of the monthly narrative intensity of three narratives for the United States. The intensity measures are calculated based on the monthly average of semantic similarity to three narrative queries - "The risk level in the market is high", "It is a good time to invest in the market", and "There is an asset bubble". The monthly series are smoothed using 30-month rolling average.

Appendix C: Example of application of the PSY Method

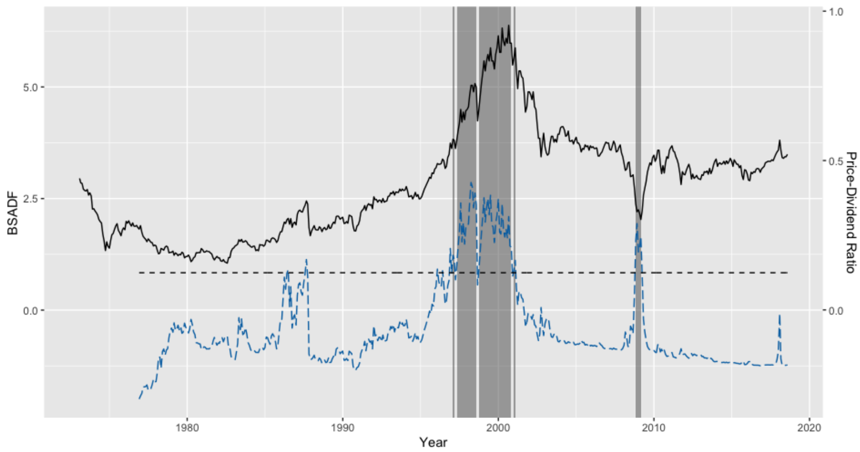

Figure 7: PSY results of the S&P 500 price to dividend ratios

Notes: The figure illustrates the explosive periods detected in the S&P 500 price to dividend ratio using the Phillips et al. (2015) method. The dark line represents the S&P 500 price to dividend ratio, and the grey shaded regions indicate the detected explosive periods, in which the BSADF statistics (blue dashed line) exceed the bootstrapped critical values (dark dashed line).

Appendix D: Economic & financial variables (for the U.S.)

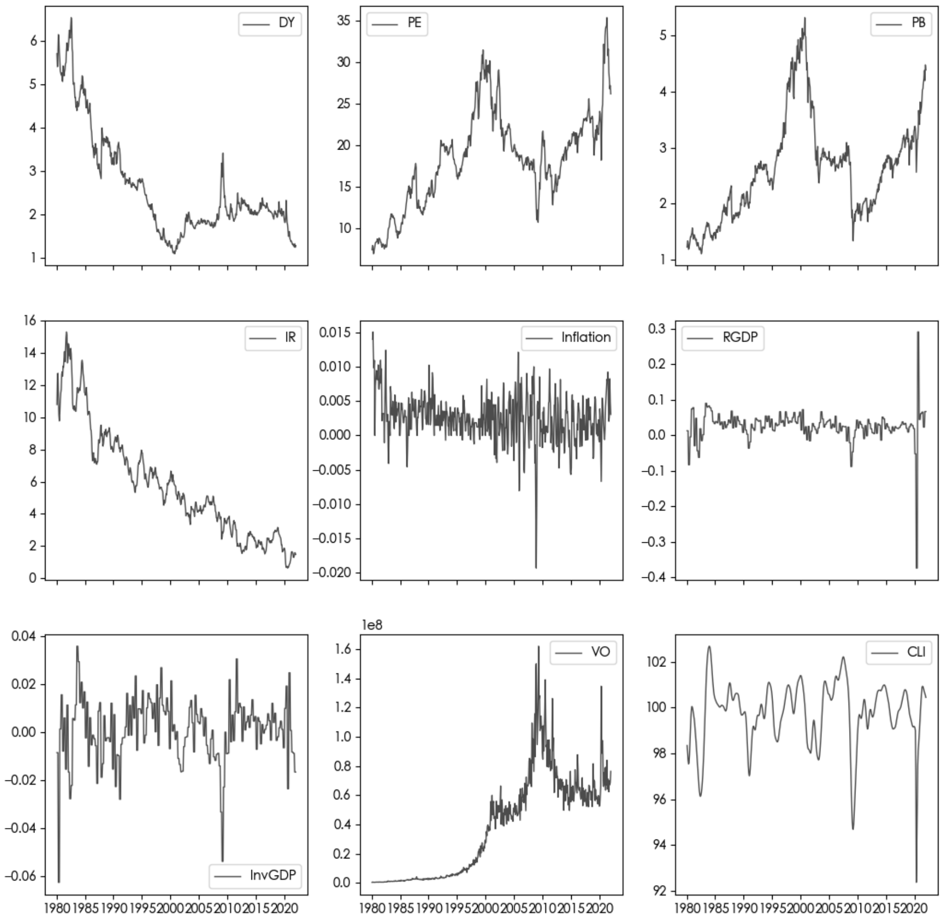

Figure 8: Economic & financial variables (for the U.S.)

Notes: The figure displays the time series of the economic & financial variables for the U.S.

Appendix E: Contemporaneous Regressions of Bubbliness

Table 16: Contemporaneous regressions of bubbliness with U.S. data

Panel A: Univariate regressions									
	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.186***	0.231***	-0.099*	-0.128***	0.106**	0.018	-0.048	-0.054	-0.22***
	(0.053)	(0.058)	(0.057)	(0.04)	(0.044)	(0.039)	(0.033)	(0.036)	(0.062)
Controls	No	No	No	No	No	No	No	No	No
No. of obs.	503	503	503	503	503	503	503	503	444
Adj. R-squared	3.26%	5.12%	0.78%	1.43%	0.93%	-0.17%	0.03%	0.09%	4.64%

Panel B: Regressions with controls									
	NI	NT	NTD	TC	NI^{NO}	$NI^{Opportunity}$	NI^{Bubble}	NI^{Risk}	EPU
Coefficient	-0.175***	0.194***	-0.064	-0.092**	0.11**	0.047	-0.044	-0.052	-0.171**
	(0.051)	(0.055)	(0.05)	(0.038)	(0.045)	(0.037)	(0.033)	(0.036)	(0.067)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of obs.	503	503	503	503	503	503	503	503	444
Adj. R-squared	14.11%	14.46%	11.46%	11.89%	12.23%	11.27%	11.25%	11.33%	12.46%

Panel C: Univariate regressions using the control variables									
	DY	PE	PB	IR	Inflation	RGDP	InvGDP	VO	CLI
Coefficient	-0.279***	0.269***	0.312***	-0.058	0.055	0.102**	0.039	-0.153***	0.129***
	(0.052)	(0.059)	(0.064)	(0.046)	(0.042)	(0.05)	(0.046)	(0.044)	(0.045)
No. of obs.	503	503	503	503	503	503	503	503	503
Adj. R-squared	7.62%	7.04%	9.55%	0.14%	0.10%	0.84%	-0.05%	2.14%	1.47%

Notes: This table presents the estimated OLS coefficients for

$$\triangle B_{c,t} = \alpha_c + \boldsymbol{\theta} \cdot \boldsymbol{X_{c,t}} + \gamma \cdot N_{c,t} + \epsilon_{c,t},\ c = U.S.,$$

with Newey–West standard errors reported in parentheses. The dependent variable ($\triangle B_{c,t+1}$) is the change of bubbliness, which was proxied by the $BSADF$ statistics of the price index. $\boldsymbol{X_{c,t}}$ is a set of financial/economic variables, and $N_{c,t}$ is one of the narrative features - market narrative intensity (NI), textual sentiment (NT), tone dispersion (NTD), topic consensus (TC), opportunity, risk and bubble narratives ($NI^{Opportunity}$,NI^{Risk},NI^{Bubble}), and the relative intensity of competing narratives (NI^{NO}). All variables by subtracting the average and dividing the standard deviation. ***, ** and * indicate significance level of 1%, 5% and 10%, respectively.

Appendix F: Prediction with PCA

Table 17: Prediction with U.S. data and PCA

	Bubbliness		Max drop	
	(1)	(2)	(1)	(2)
pc1	-0.186***	-0.271***	-0.183*	-0.185*
	(0.057)	(0.077)	(0.100)	(0.106)
pc2	0.040	0.054	-0.046	-0.009
	(0.045)	(0.049)	(0.058)	(0.056)
pc3	0.024	-0.043	-0.152**	-0.162**
	(0.054)	(0.080)	(0.064)	(0.078)
Controls	No	Yes	No	Yes
No. of obs.	443	443	431	431
Adj. R-squared	3.0%	7.5%	5.2%	9.4%

Notes: This table presents the estimated OLS coefficients, with Newey–West standard errors reported in parentheses. We only include NI, NT, NTD, TC, NI^{NO} and EPU in the information set to compute the principal components. All independent variables are lagged. ***, ** and * indicate significance level of 1%, 5% and 10%, respectively.