Kim Ristolainen

Narrative Triggers of Information Sensitivity

Aboa Centre for Economics

Discussion paper No. 156

Turku

December 2022

The Aboa Centre for Economics is a joint initiative of the economics departments of the University of Turku and Åbo Akademi University.

ISSN 1796-3133

Printed in Uniprint
Turku
December 2022

Kim Ristolainen

Narrative Triggers of Information Sensitivity

Aboa Centre for Economics
Discussion paper No. 156
December 2022

ABSTRACT

Economic research has shown that debt markets have an information sensitivity property that allows these markets to work properly when price discovery is absent and opaqueness is maintained. Dang, Gorton and Holmström (2015) argue that sufficiently "bad news" can switch debt to become information sensitive and start a financial crisis. We identify narrative triggers in the news by utilizing machine learning methods and daily information about firm default probability, the public's information acquisition and newspaper articles. We find state-specific generalizable triggers whose effect is determined by the language used by journalists. This language is associated with different psychological thinking processes.

JEL Classification: G01, G14, G41

Keywords: information sensitivity, debt markets, financial crisis, machine learning, news data, primordial thinking process

Contact information

Kim Ristolainen (corresponding author)
Department of Economics, Turku School of Economics, University of Turku, Finland.
Email: kim.ristolainen@utu.fi

Acknowledgements

I gratefully acknowledge the research assistance provided by Eetu Laakso and Robert Pylkkänen, as well as the financial support provided by the Emil Aaltonen Foundation and Yrjö Jahnsson Foundation.

1 Introduction

Dang, Gorton and Holmström (2015) argue that debt markets are by design information insensitive. During normal periods, the cost of acquiring precise information about the collateral of the debt contract is higher than the value of that information. When information insensitivity is preserved, money markets function properly, as agents can trade without acquiring information because they do not have to fear that other agents acquire information on the value of the underlying collateral. A key element in preserving debt information insensitivity is opaqueness. Dang et al. (2017) show that banks keep their loans secret so that demand deposits remain money-like trading at par as agents do not want to produce costly private information about the loans. However, when sufficiently large negative news about the value of the debt collateral arrives, the debt turns information sensitive, as the value of the information about the collateral becomes larger than the cost of acquiring that information. When debt becomes information sensitive, money markets freeze, and a financial crisis occurs, as the quantities adjust to zero instead of the prices because no one wants to hold debt due to fears of adverse selection.

The theory-implied relationships between information sensitivity and information acquisition, non-price adjustments and opaqueness have all been confirmed by empirical research[1]. However, the sufficiently bad news that triggers switches from an information-insensitive state to information-sensitive state and vice versa has not been previously studied. In this paper, we analyze the precise narrative triggers of information sensitivity. We use a machine learning algorithm and daily credit default swap (CDS) spreads and Google search data as proxies for default probability and public information acquisition about a firm to find different states characterised by these two variables. We identify different persistent states, which can be characterized as an information (in)sensitive state with (low) high default probability and (low) high public information acquisition. We label each company-day observation as one of these states and we further identify the days when a switch

[1]Dang, Gorton and Holmström (2020) survey this literature in extent.

to either state has occurred.

To identify the general factors that trigger switches between these states, we merge the daily data on the information sensitivity states of 576 financial and non-financial companies with news article data from the *Wall Street Journal*. We use natural language processing techniques and machine learning methods to first identify 80 latent topics with their daily frequencies in 1890–2022 and to then extract the unexpected attention to each news topic on a given day with each article's contribution to this surprise. Unexpected attention is defined as attention that could not have been predicted with past news attention information.

By estimating local projection regressions (Jorda, 2005), we identify a number of topics that increase the share of companies that are in an information-sensitive after the news is published. These narrative triggers that work with varying lags include unexpected attention to topics about periodical financial figures and changes in company ownership. On the other hand, unexpected attention to news related to private equity consulting, bull market speculation, and data gathering/analysis are strong triggers for companies to return to an information-insensitive state. The strength of the triggers varies from 2 to 10 percentage point changes in the share of information-sensitive companies in the economy after a 1 percentage point surprise attention to a topic in the news.

To measure each article's and journalist's thinking process in the primary–conceptual thinking processes continuum first introduced by Freud (1938) in the field of psychology, we utilize Martindale's (1975) regressive imagery dictionary. The primary process has been theorized to emerge to satisfy primary urges from the first part of a person's personality that evolved during the early years of childhood, called the id. This part of personality is seen as a primitive part of the mind that contains buried memories. On the other hand, the conceptual thinking process is seen to relate to another part of a person's mind, called the ego, whose purpose is to adapt the primary and unreasonable needs of the id to the real world. The primary thinking process is seen as irrational, impulsive, common in fantasy, sensational, concrete and unconcerned with purpose, whereas the conceptual thinking process is asso-

ciated with rationality, reality, problem solving, logic, conceptual and narrow focus.

Although the majority of the journalists in our sample do not clearly lean to either thinking processes, there are large groups of hundreds of journalists who clearly use more language associated with one of the thinking processes throughout their careers. We exploit this difference and estimate local projection regressions for unexpected news attention to specific topics in articles written by specific thinking process groups. We find that most of the narrative triggers stop being triggers when the unexpected news attention is for articles written by authors in one of the extreme thinking process groups. In addition, some topics become triggers when the related articles are written by authors in these groups.

Our results add to the empirical research on information sensitivity. First, to our knowledge, we are the first to measure an individual firm's daily information sensitivity state. Second, we identify general triggers of information sensitivity and insensitivity that can be measured at a daily frequency. Third, we show that these triggers work in one direction, vary in strength and do not work immediately, but rather with a delay of weeks, implying initial underreaction by economic agents. Finally, we show that the thinking process, and thus the language used by the journalist, determines whether a topic serves as a trigger.

The remainder of the paper is organized as follows. Section 2 focuses in identifying the daily information sensitivity state for individual firms. Section 3 describes the formation of a measure for unexpected attention to news topics from text data. Section 4 describes the journalists' thinking processes. Section 5 presents the empirical results for information sensitivity triggers. Finally, Section 6 concludes.

2 Identifying Information Sensitivity

Information sensitivity has been investigated extensively from an empirical point of view. This literature focuses on studying the predictions that the theories of Dang, Gorton and Holmström (2015) make about informa-

tion sensitivity. These empirical papers confirm that more information is produced about debt collateral when a switch to an information-insensitive state occurs (Brancati and Macchiavelli, 2019; Gallagher et al., 2020), that the quantity rather than the price of debt adjusts when bad news arrives (Gorton, 1988; Perignon, Thesmar and Vuillemey, 2018) and that opaqueness maintains and transparency destroys information insensitivity (Baghai, Giannetti and Jäger, 2022; Cipriani and La Spada, 2021). This empirical literature is thoroughly surveyed by Dang, Gorton and Holmström (2020). Although the existence of an information sensitivity property and the characteristics of this property have been empirically confirmed in many studies, the actual "bad news" that triggers these state switches have not been previously examined. Dang, Gorton and Holmström (2015) argue that when debt suddenly becomes information sensitive, investors adjust their lending to zero due to fear of adverse selection, and a funding freeze (financial crisis) occurs. Thus, the identification of generalizable triggers is important to further understand the dynamics of these very harmful events.

To identify and examine possible information sensitivity triggers from an empirical point of view, we first have to measure the information sensitivity state of a firm and potential trigger candidates through time. Previous empirical studies have shown the presence of information sensitivity via significant effects in different regression frameworks that test a relationship between key variables predicted by the theories of Dang, Gorton and Holmström (2015). We also utilize the theoretically predicted and empirically shown relationship between information production and bad news about a company to identify information sensitivity, but we go further and label each company-day with a specific information sensitivity state characterized by its default probability measured with CDS spreads and the public's information acquisition about the company measured by Google searches.

From the characteristics of the information sensitivity property, one could argue that there are four possible states in the default probability (DPR)–public information acquisition (PIA) space of a firm's debt. First, there are days when information is not acquired, and the default probability of a company is low. During this type of day, the company can be argued

to be in an information-insensitive state. Second, there can be days when information about a company is acquired, but the default probability of the company still remains low, indicating that the company is trending due to other reasons, and the company's debt is still information insensitive. Third, when a company's debt becomes information sensitive, then information is acquired in vast amounts, and the default probability of the company also increases rapidly. Finally, when a company is no longer trending but has a very high and relatively stable CDS spread, the company is in a default state. Our aim is simply to categorize each company-day observation into one of these classes for further analyses.

2.1 Gaussian Mixture Model

To identify different information sensitivity states, we utilize the Gaussian mixture model (GMM), which is a popular choice among the various mixture models and has been used, for example, in modeling stock returns (Kon, 1984; Malevergne, Pisarenko and Sornette, 2005; Behr, 2007). The GMM assumes that in each state m the variables (the CDS spread and Google trends) are from a different multivariate normal distribution with their own means $(\mu_1, \mu_2, ..., \mu_M)$ and covariance matrices $(\Sigma_1, \Sigma_2, ..., \Sigma_M)$. We denote $x^1_{i,t}$ as the daily CDS spread and $x^2_{i,t}$ as the daily Google trend index value of company i in period t. Then the GMM can be written more formally as follows:

$$f(x) = \sum_{m=1}^{M} \theta_m g(x; \mu_m, \Sigma_m), \tag{1}$$

where θ represent the mixing proportions of each multivariate normal distribution g. The unknown mixing proportions θ_m, the means μ_m and the covariance matrices Σ_m for each state are estimated with the expectation-maximization (EM) algorithm. In the EM algorithm, we first set the initial guesses $\hat{\theta}_m, \hat{\mu}_m, \hat{\Sigma}_m$ for the unknown variables. Then in the *expectation* step, we form the so-called *responsibility* or the conditional expectation of an observation belonging to a specific state given the initial guesses of the unknowns

$(\hat{\theta}_m, \hat{\mu}_m, \hat{\Sigma}_m)$ and our data (x^1, x^2). In the *maximization* step, we use the *responsibility* of each observation to calculate updated values for the unknown parameters, and then we repeat these two steps until the process converges. More detailed information on the GMM and the EM algorithm can be found in Hastie, Tibshirani and Friedman (2009).

The M number of different states must be pre-specified before the unknown parameters of the model are estimated. There is no one correct way to define the optimal number of states or components when no prior knowledge about this hyperparameter exists. A common procedure is to select the model for which the Bayesian information criterion increases the most. We select the number of components to be four, as this number will most likely estimate the simplest model that can approximate the states that we hypothesize in the beginning of this section.

We collected all available spreads of 5-year CDS contracts on Refinitiv Datastream, including both non-financial and financial companies from the time interval 2006–2022. We use the CDS spread as a measure of the *default probability (DPR)* of a company. To measure the *public information acquisition (PIA)*, we collected daily Google trend data to approximate the public's information acquisition for some specific company[2]. After merging the CDS and Google trend data to form a panel of matched daily observations of both variables, we end up with 576 companies and slightly more than 1.9 million daily observations to use in the model estimation.

2.2 Characteristics of Information Sensitivity States

Panel A of Table 1 reports the estimated values for the unknown parameters given our extensive dataset. These results confirm our hypothesis that there are four very clearly separable and easily characterized states: an information sensitive state with high default probabilities and public information acqui-

[2]The daily trend data from 12/2007 to 2/2022 are collected by first downloading the monthly data for the entire search period for the search term and then gathering the daily data per month. The daily data are then made comparable between months by multiplying the daily data by the monthly search volume and dividing by 100. This is done because Google handles large trend requests by smoothing the data to measure only by monthly frequency.

Table 1: Information sensitivity states of companies.

Panel A	Variable	Mean	SD	N	Share %
Trending for other	CDS spread	72.9	34.9	941,001	48.9
	Google searches	36.1	22.9	941,001	48.9
Insensitive state	CDS spread	101.7	64.0	331,201	17.2
	Google searches	0	0.1	331,201	17.2
Default state	CDS spread	3,259.5	3,500.7	75,916	3.9
	Google searches	14.5	18.5	75,916	3.9
Sensitive state	CDS spread	307.5	178.4	575,884	29.9
	Google searches	28.9	22.9	575,884	29.9

Panel B	Trending for other$_{t-1}$	Insensitive state$_{t-1}$	Default state$_{t-1}$	Sensitive state$_{t-1}$	Total share %
Trending for other$_t$	90.986	8.525	0	0.463	100
Insensitive state$_t$	24.22	69.076	0	6.661	100
Default state$_t$	0	0	97.946	2.022	100
Sensitive state$_t$	0.747	3.836	0.267	95.123	100

sition, an information-insensitive state with low default probabilities and public information acquisition, a default state with very high CDS spreads and low information acquisition and a state in which the company is trending for other reasons unrelated to the default probability with high information acquisition and low default probabilities. The observations in the sample after they are classified into different states are plotted in Figure 1. This figure illustrates the described characteristics of each state with respect to default probability and public information acquisition clearly.

To investigate the persistence of each state, in Panel B of Table 1, we report the probabilities of a firm being in a specific state at period t conditional on the state of the previous period $t-1$. Although the model was not given any information about time, the states seem to be very persistent, as in the clear majority of the company-period observations the previous state was the same as the current one. The model also fits the assumed evolution of the information sensitivity that the default state is most often preceded by an information-sensitive state, and the latter is most often preceded by the insensitive state. It basically never happens that a firm goes from the default state to an information-insensitive state (either with high or low PIA). The same thing occurs in the other direction: The default state is never preceded by the information-insensitive state, implying that the firm first switches to

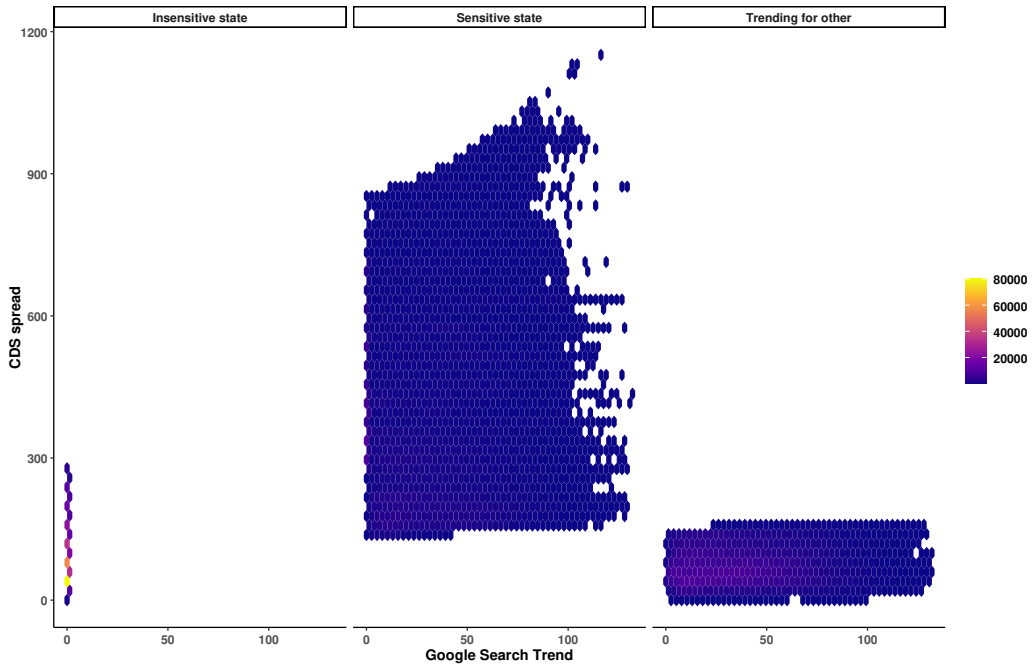


Figure 1: Density of different Information Sensitivity States in different values of CDS spreads and Google Search Trends.

the information-sensitive state before the default state. The most common states in the economy are the information-insensitive state and trending for other reasons, which accounts for a total share of almost 66.1% of the firm-days in the dataset. Default states are the rarest, with a 3.9% share, but the information-sensitive firm-days are not that uncommon, with a nearly 30% share of the total observations.

The evolution of the CDS spreads of six non-financial and six financial corporations from 2008 onward with the specific information sensitivity state the firm has been classified by the model in each day is plotted in Figures 2 and 3. The model captures the changes from calm to turbulence efficiently without giving "wrong" labels in the midst of a specific state. Next, we characterize the evolution of events for a few example cases when a switch in information sensitivity occurs.

Macy's had a troubling year in 2015, as its sales fell in the second half

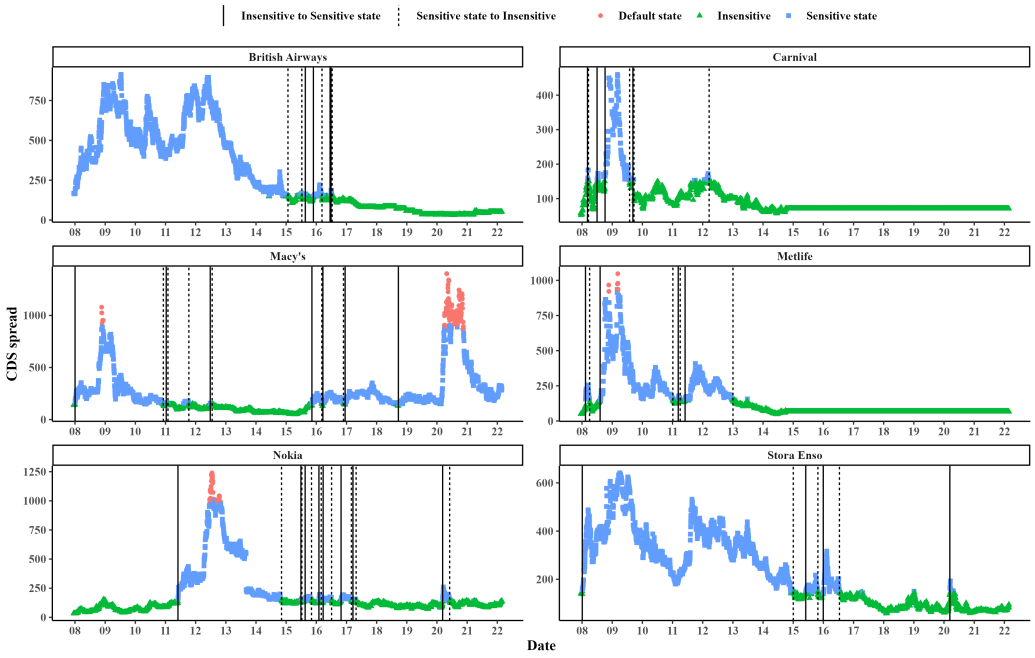


Figure 2: Information sensitivity states and CDS spreads for non-financial firms.

of the year. Our measure indicates that a clear switch to information sensitivity occurred on November 9, 2015–two days before the release of the company's disappointing third quarter earnings. Although the company reported a sales drop of 5.2% after several years of strong growth (FT.com, 2015), the news two days before about financial analysts revising their price targets for many retail companies due to excess inventory building and unusually warm weather (Kapner, 2015) is the likely cause of Macy's switch to an information-sensitive state.

Another very significant switch to information sensitivity occurred on Wednesday, June 1, 2011, for Nokia. On the previous day, the company had given a profit warning, which was mostly due to the increasing success of phones using Android as their operating system in the European market (Lawton and Efrati, 2011). On October 31, 2008, the cruise vacation company Carnival reported that it would not pay dividends to collect more cash

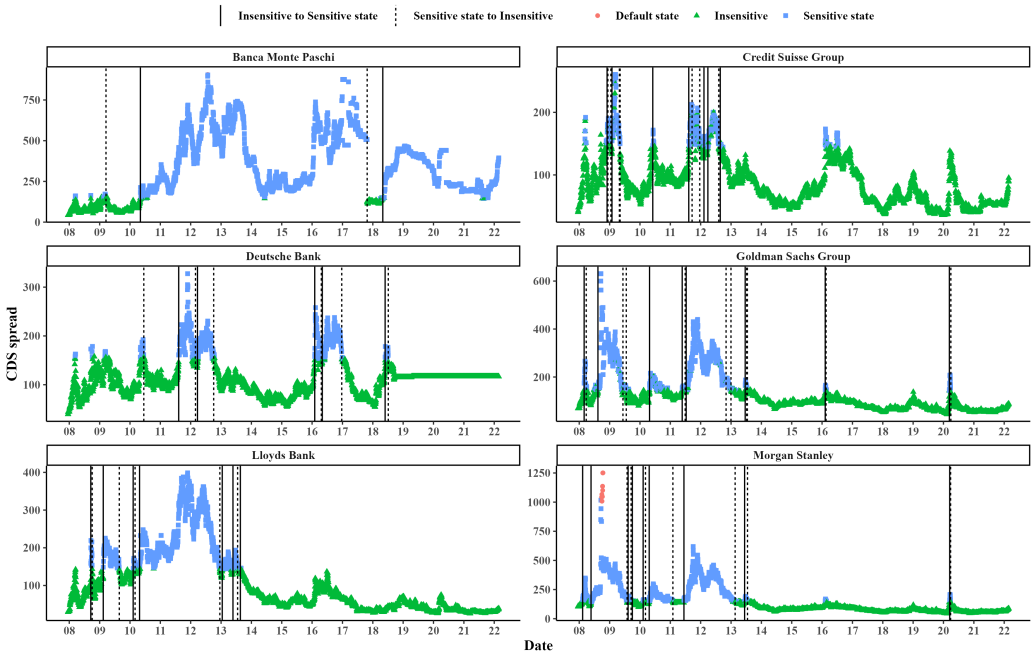


Figure 3: Information sensitivity states and CDS spreads for financial firms.

so that the company does not have to rely on the capital markets in the future (Curran, 2008). Stocks reacted negatively to this announcement, but our measure implies that the firm became information sensitive almost four weeks earlier on October 6.

3 Measuring News Surprises

In the previous section, we showed that the measure of information sensitivity captures the timing of switches between very persistent and identifiable states characterized by a company's default probability and public interest in acquiring information about the company. Although the measure co-occurs with many famous information sensitivity state switches, we want to identify information sensitivity-triggering news content at a general level that can be numerically measured. To achieve this, we have to first identify generalizable content in historical news articles and measure the prevalence of specific

content in a specific time period or an individual news title.

3.1 Attention to Economic News Topics in 1890–2022

To measure the attention that a news topic had on a specific day between 1890 and 2022, we estimate an extension of the most commonly used topic model, Blei, Ng and Jordan's (2003) latent Dirichlet allocation (LDA) model. Topic models are unsupervised learning models that try to uncover latent topics from a collection of text documents. These models assume that each text in the corpus is generated by a specific generative process. In the LDA, each text document d can consist of multiple topics k, and each topic k has a word distribution β stating how likely it is to observe a specific word from the fixed vocabulary V that holds all the unique words found in our text corpus. In addition, each document d has a topic distribution θ_d that represents the proportions of each topic that the document consists of.

The generative process works in the following way. First, a topic assignment $z_{n,d}$ is generated for each word position n for each document d from the topic distribution θ_d. Then, a word assignment $w_{n,d}$ is generated from the word distribution β_z given the topic assignment $z_{n,d}$. Both β and θ are assumed to be distributed according to a Dirichlet distribution with parameters α and η. These parameters influence how focused the Dirichlet distribution is either on the middle (documents with multiple topics) or on the corners of the distributions (documents with few topics). More formally, with a corpus of M documents with N words and K topics, the probability of observing a corpus can be written as follows:

$$P(\theta, \beta, Z, W) = \prod_{k=1}^{K} P(\beta_k|\eta) \prod_{d=1}^{M} P(\theta_d|\alpha) \prod_{i=1}^{N} P(z_{d,n}|\theta_d) P(w_{d,n}|\beta, z_{d,n}). \quad (2)$$

Given the word assignments $w_{d,j}$ and the number of topics K, the unknown parameters are estimated with Gibbs sampling.

The LDA has several important limitations. First, it assumes that the topics are uncorrelated. This is a relatively unrealistic assumption, as observ-

ing a specific topic in a document might give us information that it is likely to discuss topics that are related to the observed topic rather than completely unrelated topics. For example, if the corpus included lifestyle magazines, then if we observe a car topic without knowing that it is in a men's magazine, we would think that it is more likely to also have content about sports rather than women's fashion in the magazine. To account for this issue, Blei and Lafferty (2005) introduced the correlated topic model (CTM), which allows topics to be correlated. The CTM generative process differs from that of the LDA. The topic distributions θ_d are not from a Dirichlet distribution, but they are distributed according to a logistic normal distribution with a mean μ and a covariance matrix Σ with K dimensions. The covariance matrix enables the model to capture correlations between topics.

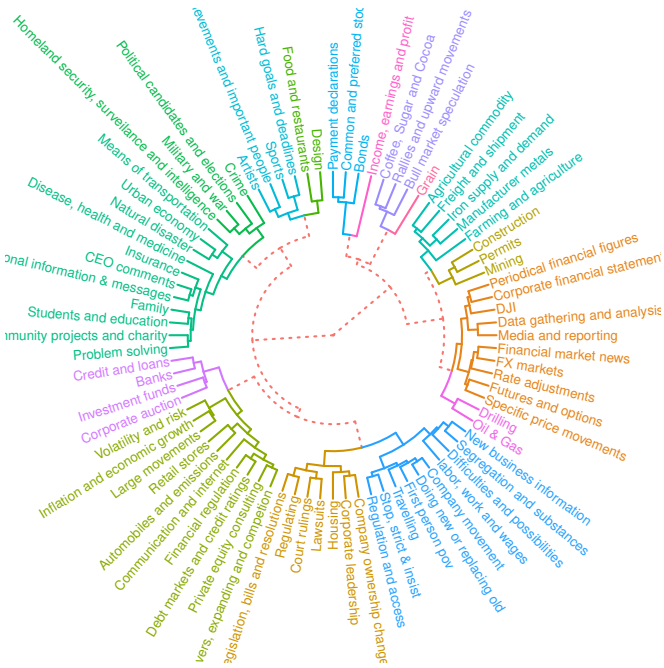


Figure 4: Hierarchical clustering of topics. The dendrogram plots the result of a hierarchical clustering model estimated with the topic word-distributions.

We estimate the CTM with a corpus that includes the titles of all news articles published in the *Wall Street Journal* in the period 1890–2022. The

Table 2: Topic labels and the predictability of the attention to topics. The table presents the most common words for each topic calculated by the term frequency measure that takes into account the exclusivity and the frequency of each word. MAE refers to the Mean Absolute Error from the rolling window out-of-sample forecast made by the machine learning algorithm with information from past topic frequencies. The last column presents the percentage share of positive surprises (prediction errors) from all surpises.

Topic	Common words	MAE	Positive surprises %
Rate adjustments	rate fix reduction percentage reduce adjust effect	0.186	60
Construction	line construction facility terminal bridge construct connect	0.155	65
Hard goals and deadlines	goal tough role tie deadline match hurdle	0.279	63
Freight and shipment	import loading shipment coin gold bulk freight	0.114	76
Bonds	certificate reorganization proceed obligation date bondholder coupon	0.186	58
Segregation and substances	give pacific reading tobacco observe telegraph carbide	0.05	66
FX markets	sterling mark check lira call guilde belgian	0.134	66
Regulation and access	free regulation commerce incentive barrier postal unlimited	0.065	71
Rallies and upward movements	trend rubber narrow curb rally dealing early	0.168	53
Urban economy	resident town city casino rural shelter gambling	0.28	65
Housing	estate building property owner rent office tenant	0.295	57
Legislation, bills and resolutions	tax adopt legislation bill resolution eliminate impose	0.312	63
Inflation and economic growth	economy economic spending inflation economist boost euro	0.897	58
Credit and loans	credit card loan lender borrower paper commercial	0.312	55
Difficulties and possibilities	country difficulty circle possibility doubt circumstance prosperity	0.142	65
Corporate financial statement	expect revenue cost gross result margin operating	0.301	51
Natural disaster	fire disaster waste escape blame earthquake explosion	0.319	62
Oil & Gas	natural gasoline gas energy pipeline gallon fuel	0.34	61
Financial market news	small telephone clearing demand premium flat keen	0.061	58
Family	family learn church mother son child lesson	0.357	71
Common and preferred stock	stockholder stock common purchase prefer capitalization share	0.24	53
Court rulings	decision ruling appeal answer legal question court	0.405	62
Permits	permit road application railway grant permission necessary	0.104	68
Financial regulation	financial regulator regulatory unit crisis collapse oversee	0.228	58
Stop, strict & insist	rule mail implement stop strict insist editorial	0.097	65
Students and education	student university op graduate education job academic	0.443	69
Doing new or replacing old	take new step advantage symbol change helm	0.135	57
Lawsuits	bankruptcy protection lawsuit filing file sue seek	0.291	59
Periodical financial figures	increase figure year less compare period decrease	0.2	57
Disease, health and medicine	drug study patient doctor disease cancer researcher	0.73	55
Sports	team game player league pitch baseball football	0.73	59
Community projects and charity	co complex community project found founder charity	0.147	66
Agricultural commodity	wool textile sight cotton worth fertilizer south	0.069	63
Retail stores	store retailer retail brand consumer shopper competition	0.353	59
Large movements	rise drop jump climb lift surge index	0.442	51
Iron supply and demand	iron pig inquiry concession basic foundry trade	0.07	60
Crime	criminal prosecutor probe crime arrest jury guilty	0.8	60
Achievements and important people	event honor revolution fortune moral self forget	0.334	66
Bull market speculation	advance buying bull tendency selling speculative considerable	0.144	57
labor, work and wages	worker labor wage employ strike employment salary	0.178	67
Media and reporting	press receive detail appear available comparison arrive	0.159	53
Specific price movements	price level low high ounce pound depressed	0.24	57
Volatility and risk	analyst emerge volatility investor risky cap volatile	0.418	51
Automobiles and emissions	auto vehicle emission battery truck solar carbon	0.45	62
Design	designer wall bedroom shirt clothe window style	1.339	54
Means of transportation	air train airline flight plane driver passenger	0.436	61
Personal information & messages	information personal employee message foundation define social	0.198	66
Artists	music artist theater theatre novel song dance	1.215	65
Farming and agriculture	farmer condition agricultural farm progress normal excellent	0.179	70
First person pov	go get way many want one come	0.344	57
Regulating	require law comply regulate state prohibit commissioner	0.207	66
Data gathering and analysis	accord number datum analysis collect person release	0.155	62
Income, earnings and profit	depreciation equal income equivalent net earn taxis	0.221	50
Communication and internet	network ad search web advertising phone mobile	0.517	56
Corporate leadership	president board vice member chairman secretary senior	0.343	61
Private equity consulting	partner strategy brokerage equity client top consulting	0.241	52
Debt markets and credit ratings	term raise debt rating finance swap downgrade	0.216	58
Investment funds	invest mutual investment fund pension manage asset	0.538	55
Coffee, Sugar and Cocoa	point close bag unchanged sugar steady closing	0.099	55
Corporate auction	bid municipal corporate auction offer bidder syndicate	0.219	58
Company ownership change	plan acquire merger announce company ownership shareholder	0.288	55
Military and war	military army troop regime rebel militant civilian	0.746	67
Travelling	visit cross trip vacation rich bind obstacle	0.159	62
Payment declarations	pay declare distribution payment declaration annual sum	0.146	60
Grain	bushel wheat northwest grain winnipeg argentine visible	0.05	62
New business information	business make understand important fact present enterprise	0.163	60
Political candidates and elections	party election campaign presidential candidate voter debate	0.95	63
Takeovers, expanding and competition	stake takeover venture expand customer competitor compete	0.311	51
Insurance	care insurance familiar people insurer coverage matter	0.341	69
Drilling	drill sand province drilling spill gravity deep	0.14	64
Banks	banking bank banker saving institution currency central	0.281	55
Food and restaurants	food restaurant ice eat drink cup dog	0.886	56
Mining	coal copper mine ore lake shipping vessel	0.134	66
Homeland security, surveillance and intelligence	agency official administration citizen protect intelligence ministry	0.351	67
Dow Jones Industrial Average	current value industrial average appreciation list secondary	0.103	69
Futures and options	contract near commodity deliver dealer derivative position	0.1	53
Problem solving	solution problem tool process experiment try rely	0.22	60
CEO comments	chief executive address interview conference say comment	0.303	61
Company movement	begin put set recent several bring similar	0.03	59
Manufacturer metals	material steel manufacturer capacity manufacture scrap tin	0.108	59

text data were gathered from Proquest Historical Newspapers using their text and data mining (TDM) tool. The news titles were cleaned[3] before they were transformed into a numerical format as data feature matrices (DFMs) that are used as inputs in a topic model. Each element of a DFM represents a word count, where the rows correspond to individual documents, and the columns represent unique words found in the corpus. We select the optimal number of topics with Mimno and Lee's (2014) algorithm. This algorithm utilizes the assumption that each topic has a specific anchor word that appears only in that specific topic. The authors show that using their algorithm to find the anchor words and then using these words in the estimation process of the topic model results in better topics quantified with many different measures.

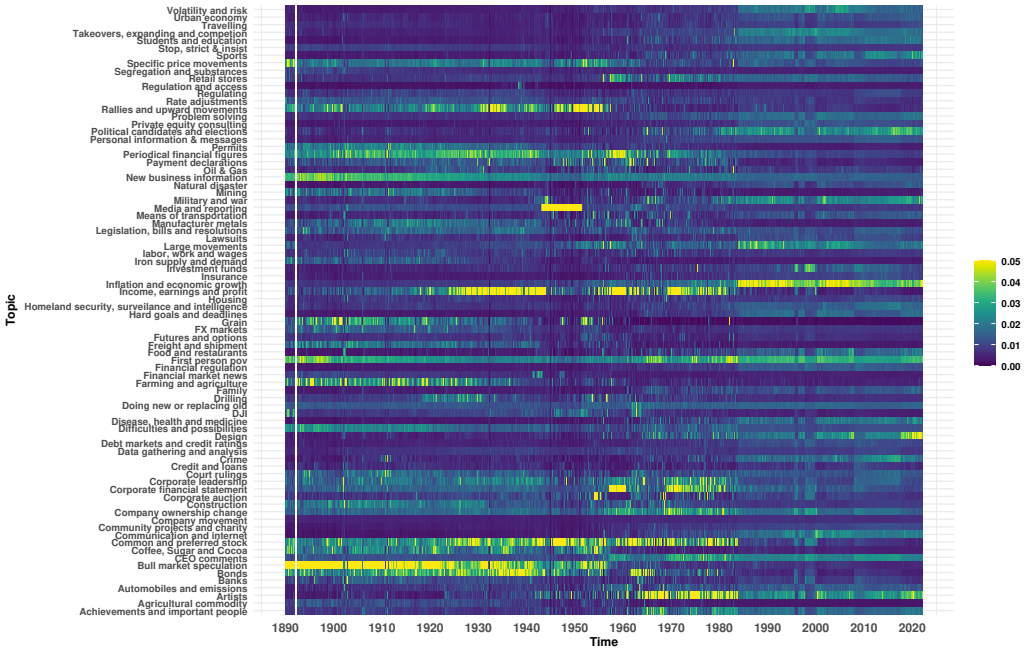


Figure 5: Prevalence of news topics in Time. The figure plots the topic distributions of each topic k aggregated to a monthly level across the period 1890–2022.

Table 2 presents the 80 topics of the estimated topic model with the labels

[3]This process is described in detail in the appendix.

and the most common words of each topic. The majority of the topics are highly identifiable from the most common words and are also quite separable from other topics. This can be seen in the fan dendrogram of Figure 4, which visualizes the topics with a hierarchical clustering algorithm that uses information from each topic's word distributions e.g. topics whose vocabulary is more similar are more likely to be grouped together. The model seems to capture a vast spectrum of different topics found in economic news in the past 130 years, ranging from insurance, debt markets, inflation, financial regulations and banking to natural disasters, crime, court rulings, political campaigns, military, wars and diseases. The model also identifies topics that are likely irrelevant for the economy, such as food, family, music, art, design and sports. Finally, the heatmap of topic provenances in Figure 5 visualizes economic news reporting over time.

3.2 Unexpected News Content

The output of the topic model that we want to utilize is the topic-word distributions β_k for each topic k and the document-topic distributions θ_d for each news article title d. The former can be used to label the topics and the latter to see which topics a specific new title consists of. We further aggregate the topic distribution information to a daily topic attention series by averaging the share of each topic for each day for the entire time period. As we are interested in the possible triggers of information sensitivity switches, we prefer to have a measure of news that enables us to state more about causal relationships, not just correlations. Therefore, we form measures of unexpected attention to different news topics. Unexpected attention means that this attention could not have been foreseen with prior information. This type of measure captures, for example, the start of the sudden increase in disease and medication news due to the COVID-19 pandemic, but then quite quickly normalizes after the beginning of the reporting, as then the attention to that topic is no longer a surprise.

We form this measure in the following way. First, we estimate the expected topic proportions for each topic for each day given the information on

past news. This is done so that a flexible elastic net model is estimated with cross-validation to predict tomorrow's topic distribution, given the information on past topic distributions of the last 5 years. Then, an out-of-sample prediction is made for the next day's topic distribution. The out-of-sample prediction error is used to measure the unexpected share of attention each topic has on a given day. Next, we discuss in detail how we extract unexpected news from the news topic data. Our approach is very similar to the procedure that Bianchi, Ludvigson and Ma (2022) used to extract biases in people's beliefs.

i. An elastic net model (Zou and Hastie, 2005) is estimated to predict the average share $Y_{k,t}$ of topic k in the news on day t with information X_{t-1} about all topic distributions[4] up to day $t-1$. The elastic net model can be formally presented as

$$\min_{\beta_0,\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \beta X_i)^2 + \lambda \sum_{j=1}^{P} (\frac{(1-\alpha)}{2} \beta_j^2 + \alpha |\beta_j|),$$

where λ is a regularization parameter that determines how much shrinkage and sparsity are introduced to the model via Ridge regression and least absolute shrinkage and selection operator (LASSO) penalties. The optimal value for *lambda* is estimated with 5-fold cross-validation, where each 20% proportion of data is reserved once as a validation set, and the model is estimated with the remaining 80% of the data. The prediction error for the validation set is collected, and the *lambda* that minimizes the average mean squared error (MSE) for these five validation errors is chosen as the optimal one. The model is estimated with the data from the previous 5 years.

ii. Step i is repeated for each day t and topic k in a rolling window fashion to get an out-of-sample prediction for the topic proportion in period t,

[4]The predictors X_{t-1} include the mean topic proportions of the previous 3 days ($t-3$ to $t-1$), and the mean and the standard deviation of the topic proportions of the previous week ($t-8$ to $t-1$), month ($t-30$ to $t-1$) and 6 months ($t-180$ to $t-1$) for all K topics, implying a total of 720 predictors with 80 different topics.

with an elastic net that was estimated with data available only before period t.

iii. Finally, to extract the unpredictable part of the attention to a topic, we collect the out-of-sample prediction error for each topic k for each day t.

To clarify, our purpose is not to measure whether a specific news title or event was completely unexpected but whether the daily attention to a specific general topic was unexpected.

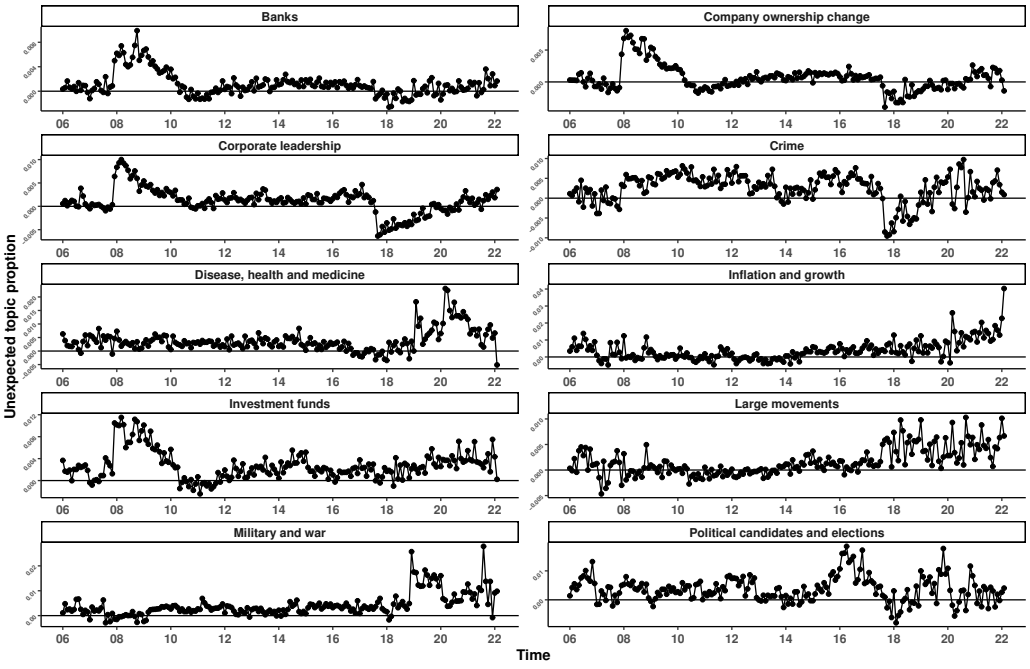


Figure 6: Evolution of Unexpected News in Selected Topics. The figure plots the unpredictable part of topic's daily prevalence for each topic aggregated to a monthly level across the period 2006–2022.

The daily unexpected news attention series for a group of selected topics aggregated to a monthly frequency is plotted in Figure 6. The measure seems to work well as it captures some highly significant and unexpected shifts in news reporting. The start of the global financial crises of 2008 can be clearly

seen in the figures, as the banks, corporate leadership, investment funds and company ownership topics receive more unexpected attention in the news during those periods. This occur seems to occur during the 2016 and 2020 U.S. elections, when the political candidates and elections topic receive unexpectedly large attention relative to previous elections. The disease, health and medicine topic seems to peak in early 2020 when the COVID-19 pandemic began. In addition, the inflation and growth topic surprisingly receives much attention in 2022, when inflation started to rise astonishingly fast.

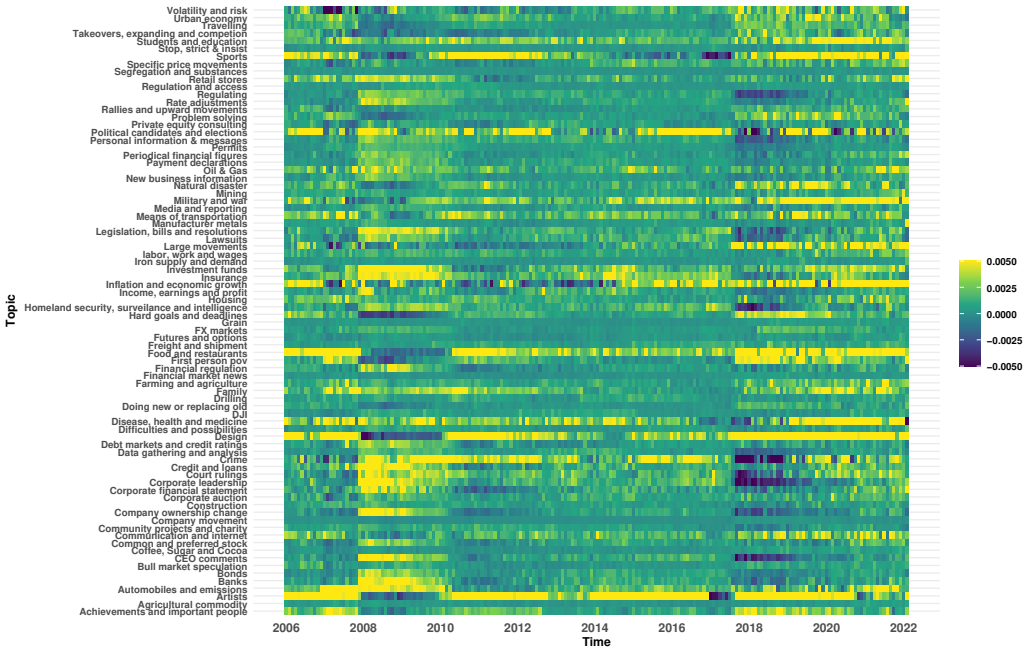


Figure 7: Evolution of Unexpected News in Time. The figure plots the unpredictable part of topic's daily prevalence for each topic k aggregated to a monthly level across the period 2006–2022.

Columns 3 and 4 of Table 2 report the out-of-sample mean absolute errors (MAEs) for the prediction of each topic by the elastic net model and the share of positive surprises in attention to each topic for the entire time span. It seems that it is more common that an increase rather than a decrease in attention to a specific topic is unexpected. The results imply that some topics

are, in general, clearly more unpredictable than others. For example, the attention to commodity, agricultural, exchange rate, manufacturing material and work, labor and wages is much more predictable than the attention to large movements, research and education, disease, health and medicine, military and war, and inflation and growth topics. This makes sense, as specific topics often relate to periodical and seasonal reporting and events, and others are more unpredictable by nature. With these observations, we infer that our measure captures the unexpected attention to news topics sufficiently.

4 Journalists' Thinking Processes

News is supposed to be an objective source of information about events of different levels of importance. However, the writing style, creativity and language used can vary across journalists, and even among articles written by the same journalist. In addition to the news content, these aspects of the text can affect the signals that the economic agents receive from news articles. This variation in writing style can be a result of external (mood and other personal events) and news content–related (journalist subjective opinion/view about the news and its possible effects on the world) factors specific to the journalist. The meaning of news content varies across reporters. Psychological literature explains why a journalist's personal relationship with the news content can materialize in the way the news article is written.

Freud (1938) argued that a person's personality consists of the id, the ego and the superego. The id is seen to be the most primitive part of the personality, and it is the first part of the personality that evolves when a human is born. According to Freud, the so-called primary thinking process is a way for the id to handle the primitive urges that the pleasure principle creates. When a person grows older, the ego and the superego play a larger role in a person's personality, and the secondary or conceptual thinking process emerges to tackle the urges to satisfy primary needs that are not suitable in the real world. These two thinking processes where introduced in the psychological literature by Freud (1938) and further discussed in Goldstein

(1939) and Werner (1948).

The primordial or primary thinking process has been seen to relate to thinking that is irrational, free-associative, sensational, impulsive, concrete and unconcerned with a purpose. Primordial thinking is thought to be free of time, space, real world and social institutions; thus, it is more common during dreams, fantasy and the use of drugs. On the other hand, conceptual or secondary thinking is rational, reality-oriented, problem solving, logical, conceptual and narrowly focused (Svensson, Archer and Norlander, 2006; Granger, 2011; Kopcsó and Láng, 2019). Primary thinking has been associated with creativity (Martindale, 1998). Katz (1997) argued that the primary process is used during the inspirational, incubation and illumination phases of the creative process, whereas the conceptual thinking process is used later during a verification phase. Journalists' primary feelings related to a news event might trigger the primary process during the writing process and emerge as a specific type of language used in the text. For example, a journalist might have strong feelings or opinions about specific politics, laws, or natural disasters that span from her id that developed early in her childhood. There might be a primary need to react to the news content, and the journalist's primary process facilitates this urge during the writing process.

To measure a journalist's mental thinking process, we utilize the regressive imagery dictionary developed by Martindale (1975). The dictionary is a collection of words that are seen to relate to either primordial or conceptual thinking. Many papers have validated this dictionary by showing that primary process words are more common in written text during coprolalic verbal ticks symptoms of people with Gilles de la Tourette's syndrome (Martindale, 1977), during the use of marijuana (West et al., 1983), in stories that are more creative (Martindale and Dailey, 1996) and among people who are writing in the dark and suffer from the fear of dark relative to texts written in well-lit areas (Kopcsó and Láng, 2019). The words of the thinking processes can be further divided into different subcategories. Examples of the subcategories of primary thinking words are vision, concreteness, unknown, brink passage, general sensation, hard, soft, consciousness alteration, diffusion, narcissism, concreteness, passivity, voyage, random movement, chaos,

timelessness, diffusion, touch, taste, odor, sound, cold and conscious. Secondary process words are about abstraction, social behavior, instrumental behavior, restraint, order, temporal references and moral imperatives (Martindale, 1977).

As primary process thinking is related to specific aspects, such as creativity, impulsiveness, irrationality etc., the share of primary and secondary thinking processes words among the texts in news articles discussing the economy and companies whose debt the agents hold (or whose debt is the collateral for the debt they own) can give signals that distort, emphasize, diminish, magnify, raise doubt, confuse or elucidate the message about the fundamental content of the news. In addition, the primary thinking process can emerge from agents who are the subjects of the news. For example, there where a lot of different ways that Mario Draghi could have given the message in his famous speech on July 26, 2012. If he had left out the phrases the *ECB will do whatever it takes* and *you better believe it is enough* from the speech, then the message may not have been as persuasive, and the European debt markets might have remained in turmoil.

We measure the thinking process TP_d behind document d as the difference between the shares of primordial thinking process words and conceptual thinking process words. More formally,

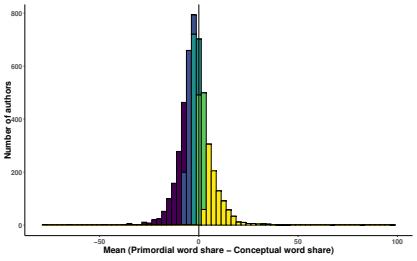
$$TP_d = Conceptual\ words\ share\ \%_d - Primary\ words\ share\ \%_d. \quad (3)$$

We aggregate this measure to daily TP_t and author-level TP_a measure the following:

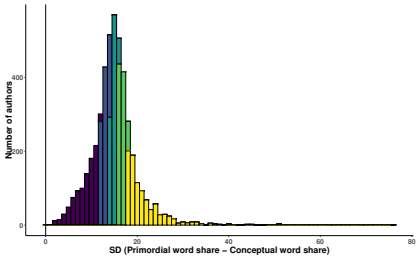
$$TP_t = \sum_{d \in t} TP_d, \quad (4)$$

$$TP_a = \sum_{d \in a} TP_d. \quad (5)$$

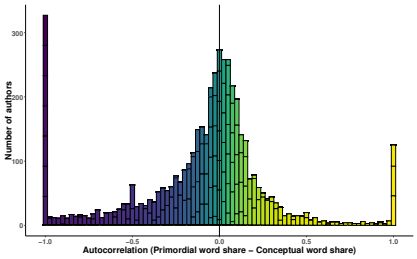
This measure captures in which direction on the primordial–conceptual thinking process continuum the news article texts lean. Different statistics char-



(a) Mean TP_d.



(b) SD of TP_d.



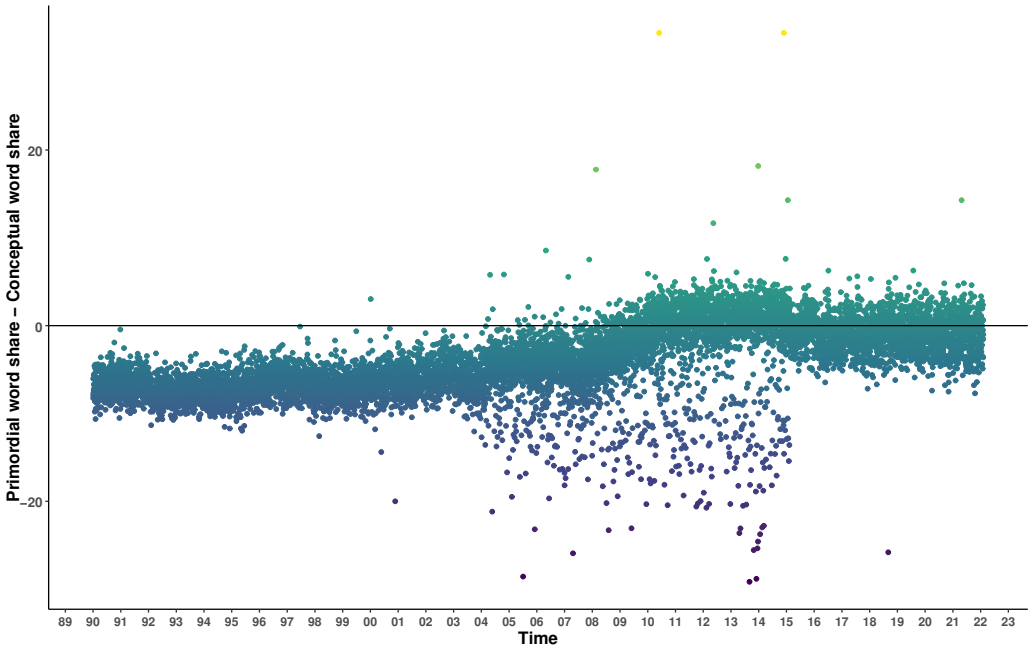
(c) Autocorrelation of TP_d.

Figure 8: Distribution of the primordial - conceptual word share difference across authors. A total amount of 4,654 authors and 91 articles per author on average.

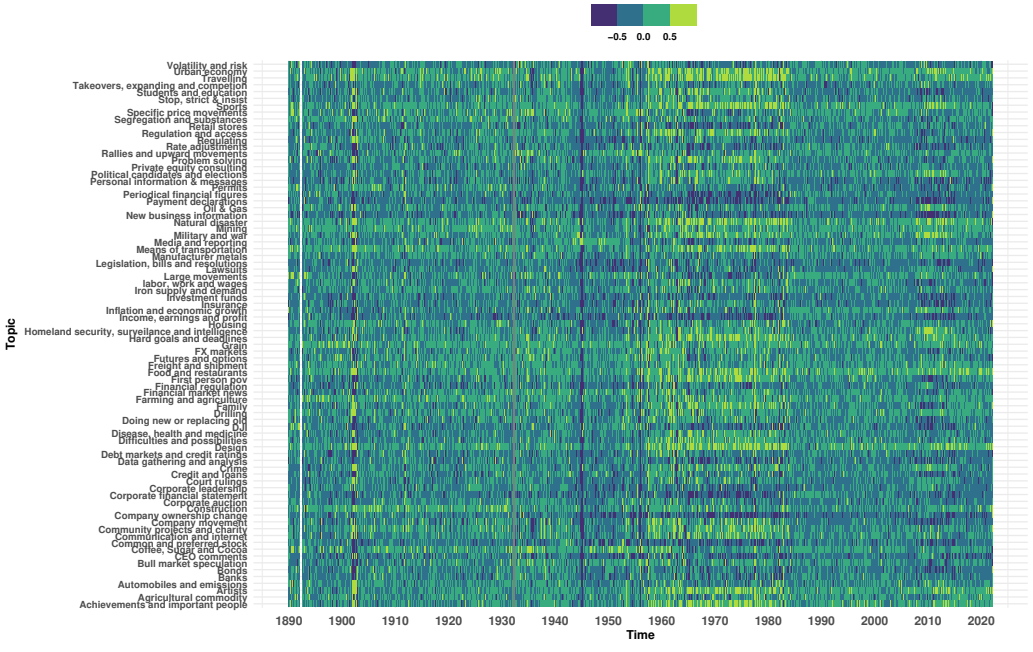
acterizing TP_a across authors are plotted in Figure 8. Figure 8a–8c reveal that the largest share of authors do not lean to either thinking processes on average, but thousands of journalists use either the primary or conceptual thinking process more on average. The large dispersion in author-specific standard deviations implies that the thinking process is in no way constant and varies substantially for each author and more, among others. Interestingly, for a large share of authors, the thinking process is very persistent (a positive auto-correlation) and for the large majority, it is not that persistent. There are also authors whose thinking process across time has a negative autocorrelation, implying that they switch persistently to the other process after each news article. This descriptive evidence point to the fact that these two thinking processes are present in news articles.

The average monthly TP_t across time from 1990 onward is plotted in Figure 9a[5]. It seems that there was a large shift from more conceptual thinking process language from 1990 to 2009, and then there was a permanent shift to language that does not lean toward either processes. However, the share of primordial thinking process language was clearly elevated and more present compared to conceptual thinking process language during 2010–2014. The figure also shows that there are large deviations across days, but this dispersion was extremely high from around 2005 to 2015. It is also plausible that these different thinking processes are more common in some topics than in others. Figure 9b displays the monthly correlation of a topic's prevalence and TP_d across topics and time. What is striking is that although there is variation in the correlation across topics, it seems to be high during specific longer time periods. For example, the language in news articles was clearly leaning toward the conceptual thing process in the 10 years following the Second World War. In addition, the primordial thinking process was relatively more present from 1955 to 1985.

[5]The author data on articles are not that detailed before this time period.



(a) Daily thinking process leaning across time. The figure plots TP_t for each day in the period 1990–2022.



(b) Monthly correlation of topics and primordial - conceptual word share difference across time and topics.

Figure 9: Primordial-conceptual word share difference across time.

5 Triggers of Information Sensitivity

What did eventually calm the European money markets? Governor Draghi's statement "we will do whatever it takes – and you better believe it is enough". This is as opaque a statement as one can have. There were no specifics on how calm would be reestablished, but the lack of specific information is, in the logic presented here, a key element in the effectiveness of the message. So was the knowledge that Germany stood behind the message – an implicit guarantee that told the markets that there would be enough collateral. A detailed, transparent plan to get out of the crisis, including rescue funds, which were already there, might have invited differences in opinion instead of convergence in views.

–Holmstrom (2015)

To examine how unexpected attention to different news topics affects companies' information sensitivity, we utilize Jorda's (2005) local projection method and estimate the following specification:

$$\Delta_h Y_t = \alpha^h + \beta_k^h \sum_{k=1}^{80} A_{k,t} + \epsilon_t \quad \text{for} \quad h = 1, ..., 90. \tag{6}$$

The dependent variable $\Delta_h Y_t$ is the change in the percentage share of information-sensitive companies from period t to $t + h$. The main explanatory variable $A_{k,t}$ is the daily unexpected attention to different topics k on day t. The unexpected part of the attention to a specific topic is defined as the deviation of the realized topic attention T_t^k from the predicted daily aggregate share $E(T_t^k|\xi_{t-1})$ of that topic among all articles published in period t given past news ξ_{t-1}:

$$A_t^k = T_t^k - E(T_t^k|\xi_{t-1}). \tag{7}$$

The coefficients β_k^h capture the effect that unexpected attention to a news topic k has on the aggregate share of information-sensitive companies in the economy for different horizons, keeping other news surprises constant. We use heteroskedasticity and autocorrelation robust standard errors in the inference. The data used in the estimation include 3,487 daily observations

aggregated from a dataset of 1,923,992 day-company observations for 576 companies from the period 17.12.2006–25.2.2022. There are a total of 1596 separate switches to an information-sensitive state and 1,442 switches to an information-insensitive state.

A trigger might not always simply be a specific topic, for example, large movements in sales or profitability, but instead, a topic combined with how it is discussed and then perceived by economic agents. As an example, investors are reading about a company CEO's statement about the firm's plan about the future in a declined economic state specific to that firm, for example, Nokia's plans when Android and IPhone were taking over the market. It might be crucial whether the CEO uses very concrete language in his statement rather than relatively opaque and visionary language about the company's future plans. As information insensitivity, and thus the debt markets, relies on opaqueness to function, the type of language might be very important in shaping agents' beliefs and the actual underlying fundamentals. A 2% percent decline in sales might be perceived differently if it is described as *a rather modest decrease* or *a never-before-seen drop*. News about a supply shortage in materials used to make phones might trigger an agent's information acquisition when it is described as a severe shortage without very precise information, but debt related to mobile phone manufacturers might stay information sensitive if the shortage is described with precise and relatively neutral terms.

As it is likely that not only what is talked about but also how it is talked about matters for economic agents, we estimate the local projection model in Equation 6 for different variations of the topic data. The descriptive information in Section 4 reveals that the thinking process, and thus the language used in the news articles, varies significantly across the journalists in our dataset. Although the majority of authors are close to the middle in the primary–conceptual thinking process language continuum, there are very large groups (hundreds of authors) that clearly use more conceptual or primary thinking process language during their journalists' careers. We incorporate this aspect into our analysis to reveal whether the thinking process and the language used by journalists affect whether a specific topic serves as

a trigger of information sensitivity.

The analysis is conducted for different subsets of the data in the following way. First, the 4,654 authors who wrote 91 articles, on average, are divided into three thinking process groups according to their average TP_a score across the news articles that they wrote. The two highest terciles are labeled primary thinking process authors, the lowest two terciles are labeled conceptual thinking process authors and the remaining 60% of the authors between these two groups are labeled normal authors. Next, the unexpected daily topic attention series for each topic is reconstructed so that they use only topic frequency information from the articles written by an author in that specific group. Finally, the local projection model of Equation 6 is estimated for 1- to 90-day horizons separately with each thinking process group's topic series data to see whether the triggers depend on the language used, keeping the content (topic) constant, and whether some triggers are more immediate, slower, transitory or persistent than others.

Figure 10 plots the dynamic reaction of the share of information-sensitive companies in the economy to a shock in the attention to a specific topic in the news for those topics that had at least one coefficient that was statistically significant at the 1% level across the 90-day horizon. The larger the absolute value of a significant coefficient, the darker its color in the plot. The figure displays three clear implications. First, there exist topics that serve as triggers of information sensitivity (*periodical financial figures, company ownership change, agricultural commodity, difficulties and possibilities and means of transportation*) or information insensitivity (*doing new or replacing old, and bull market speculation, data gathering and analysis, and private equity consulting*) in the economy. These results seem to fit our perception of what these information events could be, for example, bad periodical financial figures awaken the interest of investors, or bull market speculation during bad/uncertain times makes economic agents wonder whether the outlooks are starting to look better for specific companies.

Second, these triggers work with lags of varying lengths, with some more immediate (e.g., *difficulties and possibilities and doing new or replacing old*) and some more slower (e.g., *periodical financial figures and data gathering*

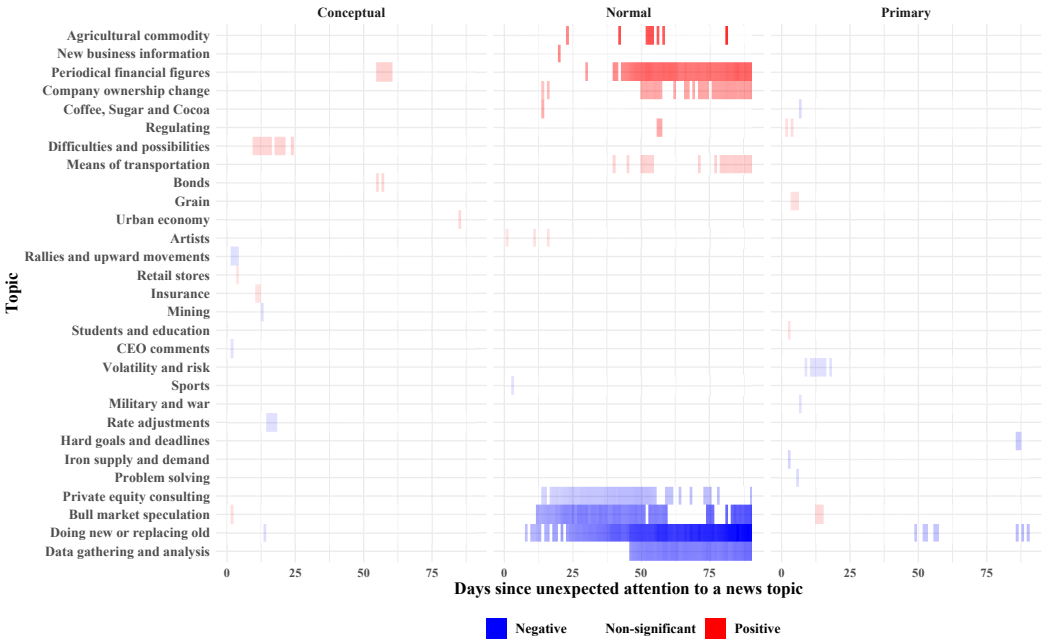


Figure 10: Journalists' thinking process and narrative triggers of information sensitivity. The figure plots the β_k^h coefficients of Equation 6 for different topics with at least one coefficient that is statistically significant at a 1% level between 1–90 day horizons. Statistically significant coefficients are displayed with either a blue colour (negative coefficient) or a red colour (positive coefficient). The Larger the absolute value of a coefficient is, the darker colour it has in the figure. Statistical significance is calculated with autocorrelation and heteroskedasticity robust standard errors.

and analysis). The lagging response indicates that people initially underreact to a news shock, and this phenomenon is more severe for other triggers. As the switch to information sensitivity is caused by economic agents when they decide to start acquiring information about a corporation given their information set, this delayed switch implies that economic agents underreact to news. Coibion and Gorodnichenko (2015) use survey data to show that professional forecasters' consensus underreact to aggregate news. Our measure of unexpected attention to a news topic can be seen to measure aggregate news, but we do not have a direct measure of consensus beliefs. However, our measure of information sensitivity is a measure of economic agents' actions driven by motives and information regarding a company; it measures

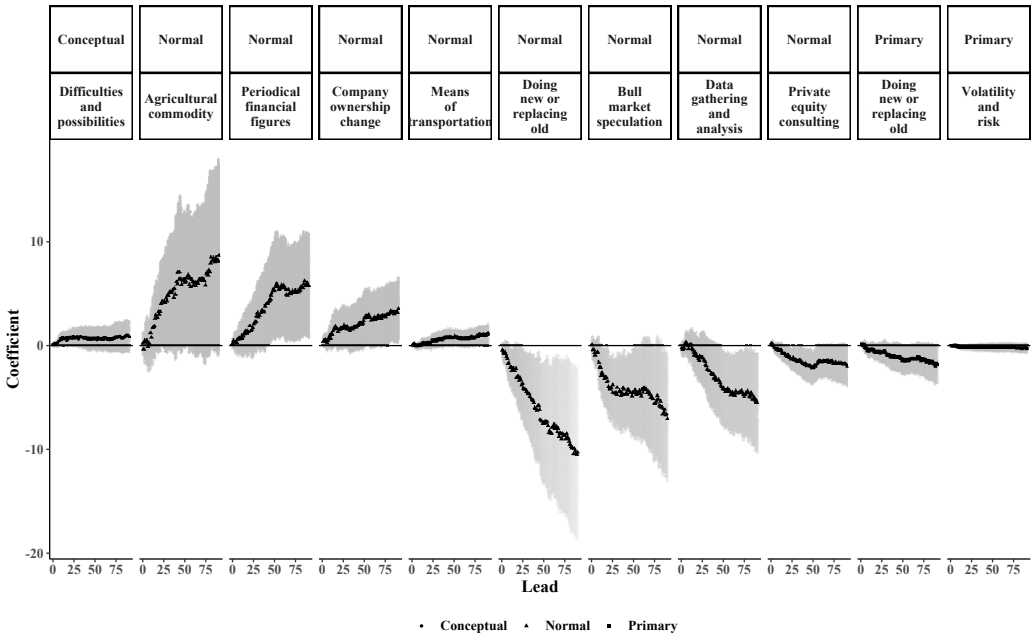


Figure 11: Journalists' thinking process and the strongest narrative triggers of information sensitivity. The figure plots the β_k^h coefficients of Equation 6 with 99% confidence intervals for topics with at least seven coefficients that are statistically significant at a 1% level between 1–90 day horizons. Statistical significance is calculated with autocorrelation and heteroskedasticity robust standard errors.

the change in aggregate beliefs. Our empirical evidence reveals that general triggers of information sensitivity do not work immediately (the same or the next day), but they sow the seeds of doubt in the minds of economic agents, and this doubt is realized with a long lag after growing for weeks.

Finally, it is clear that the thinking process evident from the language of the news text matters significantly, whether a topic serves as a trigger or not. The two clearest triggers of information sensitivity—unexpected attention to news about *periodical financial figures* or *company ownership change*—are completely absent when these topics are written by authors who, on average, use much more primary or conceptual language. In addition, the most immediate trigger of information sensitivity—*difficulties and possibil-*

ities topic—is present only when it is written by authors who are labeled the conceptual thinking process group. Interestingly, the four clearest triggers of information insensitivity—unexpected attention to news about *doing new or replacing old*, *bull market speculation*, *data gathering and analysis*, and *private equity consulting*—are not triggers when the news is written by an author of either extreme thinking process group. Bull market speculation switches from being a trigger of information insensitivity to a trigger of information sensitivity when the author of the news is from the primary thinking process group rather than the normal group. In addition, the *rallies and upward movements* and *rate adjustment* topics become triggers of information insensitivity only when the news is written by an author from the conceptual thinking process group.

If we assume that the news is randomly distributed for the journalists, then combined with the fact that the regressors are the unexpected (unpredictable) part of the attention to a news topic on a given day, these results imply a causal relationship between the news topic attention and the language used in the news and the prevalence of information sensitivity in the economy. In Figure 11, the largest significant coefficients are larger than 10, implying that a one-percentage-point surprise in the attention to a news topic increases (decreases) the share of firms in an information-sensitive state by 10 percentage points in the upcoming weeks. In the company sample, this is equal to around 57 companies. As the regressive imagery dictionary that we use to measure a writer's thinking process has been validated in a large number of studies in different contexts and time periods, it is very troublesome that non-fundamental factors related to the messenger of the news can have such a large effect on the economy. According to Freud (1938), a person uses the primary thinking process because of the urge to satisfy primary motives from the id part of a person's personality that emerged during the first years of a person's life. Highly individual-specific factors, experiences, and traumas from that time can affect how a person feels at an unconscious level about a news topic that he is going to write about. If the news content is related to primary urges from the id part of the personality, then the journalist might satisfy these urges by using more primary process language in the news.

These results imply that it matters whether the news about a company is written by an author who generally writes articles with language that is irrational or rational, non-reality or reality-oriented, illogical or logical, impulsive or sensible, sensational or ordinary, with or without a purpose. As information insensitivity is an essential characteristic of the debt markets that enables them to work properly, the indirect implications of these results that the switch back to information insensitivity does not happen when the underlying event is written with deviating language due to writer-specific factors from psychology is quite troublesome for financial stability and the economy. On the other hand, when these journalists write about news topics that often serve as triggers of information sensitivity, this troubling information event is absent. These results are highly relevant for mitigating bad news expanding to switches in companies' information sensitivity state and then to financial crises.

6 Conclusion

We measure the information sensitivity states of 576 financial and non-financial companies at the daily frequency level by utilizing machine learning methods with daily CDS spreads and Google search trends. We identify 3,038 days when a company switched either to an information-insensitive state or to an information-sensitive state. These states are highly persistent and capture many known information events.

To identify what triggers information sensitivity state switches in the economy, we estimate a CTM with all news article titles published in the *Wall Street Journal* from 1890 to 2022. We use the daily prevalence of an estimated 80 topics to form measures of unexpected attention to specific news topics. Unexpected news are defined as the part of the daily frequency of each topic that could not be predicted by a machine learning model with information about past frequencies of all topics. These measures capture the global financial crisis, the reporting on the Trump presidency, the COVID-19 outbreak, the start of the war in Ukraine and the recent surprising burst in inflation.

We then utilized the regressive imagery dictionary to measure the amount of primary and secondary thinking process used by the authors during the writing process. After categorizing authors into different thinking process groups, we match the daily information sensitivity state data to the data on unexpected news attention to different topics to estimate a local projection regression to examine how different news topics affect the share of companies in an information sensitive state in the economy. The analysis reveals that surprise attention to specific topics acts as triggers of information sensitivity or information insensitivity in the economy. These triggers work with a lag of weeks, and the language used in the news article determines whether a topic acts as a trigger of information sensitivity or not. This implies that factors related to individual authors and their personalities can have a large effect on financial markets and financial stability.

A Appendix

A.1 Text Collection

- The data were collected from Proquest Historical Newspapers by using their TDM tool.
- The collection process was performed between 2022-03-28 and 2022-03-04.
- We collected all available titles, abstracts, page numbers, author names and dates of texts in the *Wall Street Journal* that we categorized as articles, features or news.
- The publication dates range from 1889-07-08 to 2022-02-05.
- After duplicated title-publication id pairs were removed, the corpus included 4,323,637 individual texts.

A.2 Text Preprocessing

1. We identified empty titles, titles that were duplicates but still individual news (different publication id), and titles of irrelevant news types. Due to this, we removed the titles that included the following patters: – No Title *OR* Dividends Rep *OR* ted *OR* Stocks Ex-Dividend—Stockholder Meeting Brief: *OR* Corrections Amplifications: *OR* Corrections & Amplifications: *OR TITLE BEGINS* REVIEW *OR TITLE BEGINS* REVIEW amp; OUTLOOK (Editorial): *OR TITLE BEGINS* Business Brief: *OR* Theater: *OR* Dividend News: *OR* Sports *OR* Film *OR* Letters to the Editor: *OR* Co. *TITLE ENDS OR* Corp. *TITLE ENDS OR* Inc. *TITLE ENDS OR* Bookshelf: *OR* Opera: *OR* Gardening: *OR* Seeing Stars: *OR* Thinking Things Over: *OR* WORD BEGINS Art *WORD ENDS OR TITLE BEGINS* Books: *OR* Television: *OR* financial briefing book: *OR* reporter's notebook.
2. Extra whitespace was removed from texts, and all letters were changed to lowercase.

3. If the abstract was not missing, then it was chosen as the text representing the article; otherwise, the title was chosen.

4. Non-duplicate texts with at least 20 words were included.

5. Python's Spacy library was utilized to parse the individual texts into individual parts of a sentence and identify the final list of words that we wanted to include.

6. All words with the following entity categorization were removed: CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE, LAW, LOC, MONEY, NORP, ORDINAL, ORG, PERCENT, PERSON, PRODUCT, QUANTITY, TIME and WORK_OF_ART.

7. All words that had the following universal tag for parts-of-speech were chosen for inclusion: Adjectives (ADJ), Nouns (NOUN) and Verbs (VERB).

8. The remaining words in each text were transformed into their lemma form.

9. Lemma forms that were stopwords, included only one character, or included numbers or punctuation were removed.

10. The lemmas that were among the lemmas with the highest 10,000 term frequency–inverse document frequency (tf-idf) values were included in the final corpus.

11. Texts that had fewer than 10 words/lemmas after the cleaning process were removed from the final corpus.

References

Baghai, Ramin P., Mariassunta Giannetti, and Ivika Jäger. 2022. "Liability Structure and Risk Taking: Evidence from the Money Market Fund Industry." *Journal of Financial and Quantitative Analysis*, 57(5): 1771–1804.

Behr, Andreas. 2007. "Assessing the stability of Gaussian mixture models for monthly returns of the S&P 500 index." *Applied Financial Economics Letters*, 3(4): 215–220.

Bianchi, Francesco, Sydney C. Ludvigson, and Sai Ma. 2022. "Belief Distortions and Macroeconomic Fluctuations." *American Economic Review*, 112(7): 2269–2315.

Blei, David M., and John D. Lafferty. 2005. "Correlated Topic Models." *NIPS'05*, 147–154. Cambridge, MA, USA:MIT Press.

Blei, David M., Andrew Y. Ng, and Michael J. Jordan. 2003. "Latent dirichlet allocation." *The Journal of Machine Learning Research*, 3: 993–1022.

Brancati, Emanuele, and Marco Macchiavelli. 2019. "The Information Sensitivity of Debt in Good and Bad Times." *Journal of Financial Economics*, 133(1): 99–112.

Cipriani, Marco, and Gabriele La Spada. 2021. "Investors' appetite for money-like assets: The MMF industry after the 2014 regulatory reform." *Journal of Financial Economics*, 140(1): 250–269.

Coibion, Olivier, and Yuriy Gorodnichenko. 2015. "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts." *American Economic Review*, 105(8): 2644–2678.

Curran, Rob. 2008. "Moving the Market Banks Stocks Rally; Carnival Slips 12% Dow Industrials Rise 144.32, in First Two-Session Gain Since September." Copyright - (c) 2008 Dow Jones Company, Inc. Reproduced with permission of copyright owner. Further reproduction or distribution is prohibited without permission; Last updated - 2020-11-20.

Dang, Tri Vi, Gary Gorton, and Bengt Holmström. 2015. "Ignorance, Debt and Financial Crises." Yale University, Massachusetts Institute of Technology, and Columbia University Unpublished Working Paper.

Dang, Tri Vi, Gary Gorton, and Bengt Holmström. 2020. "The Information View of Financial Crises." *Annual Review of Financial Economics*, 12(1): 39–65.

Dang, Tri Vi, Gary Gorton, Bengt Holmström, and Guillermo Ordoñez. 2017. "Banks as Secret Keepers." *American Economic Review*, 107(4): 1005–1029.

Freud, S. 1938. "The Interpretation of Dreams." New York: Random House.

FT.com. 2015. "Retailers punished as Wall Street sours on sector." *FT.com.* Copyright - Copyright The Financial Times Limited Nov 9, 2015; Last updated - 2020-11-19.

Gallagher, Emily, Lawrence Schmidt, Allan G Timmermann, and Russ Wermers. 2020. " Investor Information Acquisition and Money Market Fund Risk Rebalancing during the 2011–2012 Eurozone Crisi." *Review of Financial Studies*, 33(4).

Goldstein, K. 1939. *The organism.* Boston: Beacon.

Gorton, Gary. 1988. "Banking Panics and Business Cycles." *Oxford Economic Papers*, 40(4): 751–781.

Granger, C. W. J. 2011. "The Regressive Imagery Dictionary: A test of its concurrent validity in English, German, Latin, and Portuguese." *Literary and Linguistic Computing*, 26(1): 125–135.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Holmstrom, Bengt. 2015. "Understanding the role of debt in the financial system." Bank for International Settlements BIS Working Papers 479.

Jorda, Oscar. 2005. "Estimation and Inference of Impulse Responses by Local Projections." *American Economic Review*, 95(1): 161–182.

Kapner, Suzanne. 2015. "Macy's Seeks Answers As Sales Slide — After resisting, CEO tries discount stores; stock tumbles on earnings." *Wall Street Journal*. Name - Starboard Value LP; Macys Inc; Copyright - (c) 2015 Dow Jones Company, Inc. Reproduced with permission of copyright owner. Further reproduction or distribution is prohibited without permission; People - Lundgren, Terry J; Last updated - 2021-09-13.

Katz, A. N. 1997. "Creativity and the cerebral hemispheres." *The creativity research handbook*, , ed. M. A. Runco, 203–226. Cresskill, NJ: Hamption.

Kon, Stanley J. 1984. "Models of Stock Returns-A Comparison." *Journal of Finance*, 39(1): 147–65.

Kopcsó, Krisztina, and András Láng. 2019. "Uncontrolled Thoughts in the Dark? Effects of Lighting Conditions and Fear of the Dark on Thinking Processes." *Imagination, Cognition and Personality*, 39(1): 97–108.

Lawton, Christopher, and Amir Efrati. 2011. "Nokia's Latest Headache: Android." Copyright - (c) 2011 Dow Jones Company, Inc. Reproduced with permission of copyright owner. Further reproduction or distribution is prohibited without permission; Last updated - 2021-09-22.

Malevergne, Y., V. Pisarenko, and D. Sornette. 2005. "Empirical distributions of stock returns: between the stretched exponential and the power law?" *Quantitative Finance*, 5(4): 379–401.

Martindale, Colin. 1975. *Romantic progression: The psychology of literary history. Washington, D.C.: Hemisphere.* Washington, D.C.: Hemispher.

Martindale, Colin. 1977. "Syntactic and semantic correlates of verbal tics in Gilles de la Tourette's syndrome: A quantitative case study." *Brain and Language*, 4: 231–247.

Martindale, Colin. 1998. "Biological Bases of Creativity." *Handbook of Creativity*, , ed. Robert J.Editor Sternberg, 137–152. Cambridge University Press.

Martindale, Colin, and Audrey Dailey. 1996. "Creativity, primary process cognition and personality." *Personality and Individual Differences*, 20(4): 409–414.

Mimno, David, and Moontae Lee. 2014. "Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference." 1319–1328. Doha, Qatar:Association for Computational Linguistics.

Perignon, Christophe, David Thesmar, and Guillaume Vuillemey. 2018. "Wholesale Funding Dry-Ups." *The Journal of Finance*, 73(2): 575–617.

Svensson, Nina, Trevor Archer, and Torsten Norlander. 2006. "A Swedish Version of the Regressive Imagery Dictionary: Effects of Alcohol and Emotional Enhancement on Primary–Secondary Process Relations." *Creativity Research Journal*, 18(4): 459–470.

Werner, H. 1948. *The Comparative psychology of mental development.* New York: International University Press.

West, Alan, Colin Martindale, Dwight Hines, and Walton T. Roth. 1983. "Marijuana-Induced Primary Process Thought in the TAT." *Journal of Personality Assessment*, 47(5): 466–467.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2): 301–320.

The **Aboa Centre for Economics (ACE)** is a joint initiative of the economics departments of the Turku School of Economics at the University of Turku and the School of Business and Economics at Åbo Akademi University. ACE was founded in 1998. The aim of the Centre is to coordinate research and education related to economics.

Contact information: Aboa Centre for Economics, Department of Economics, Rehtorinpellonkatu 3, FI-20500 Turku, Finland.

www.ace-economics.fi

ISSN 1796-3133