BBVA Research

Creating Opportunities

22/05

Working Paper

Measuring news media sentiment using Big Data for Chinese stock markets

Shulin Shen, Le Xia, Yulin Shuai, Da Gao

July 2022

Measuring news media sentiment using Big Data for Chinese stock markets

Shulin Shen / Le Xia / Yulin Shuai / Da Gao[1]

April 2022

Abstract

We construct and assess new time series measures of news media sentiment based on Global Data on Events, Location, and Tone (GDELT) using Data Science techniques. Five sentiment measures representing the news media Tone, Optimism, Attention, Tone Dispersion, and Emotional Polarity of Chinese stock markets are constructed based on article tone scores and media coverages from GDELT. All these news media sentiment measures are shown to have significant predictive power for Chinese stock market returns and volatilities. We also document substantial asymmetric sentiment effects on the Chinese stock market returns and volatilities. Sentiment extended EGARCH models are shown to improve market return and volatility forecasting accuracy significantly.

Keywords: China; News sentiment; Big data; Stock market; GDELT.

JEL classification: G10, G40, C15, C32.

1: Please address all correspondence to Da Gao, School of Economics, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, Hubei, China 430074; Email: gaoda@hust.edu.cn. Shulin Shen, School of Economics, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, Hubei, China 430074; Email: shen_shulin@hust.edu.cn. Le Xia, BBVA, Unit 9507, BBVA, Level 95, ICC, 1 Austin Road West, Kowloon, Hong Kong; Email: le.xia@hhva.com. Yulin Shuai, School of Economics, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, Hubei, China 430074; Email: shuaiyulin1998@gmail.com. All remaining errors and omissions are our own.

1. Introduction

Given the important role of sentiment or narratives in influencing agent's decisions as so to affecting business cycle fluctuations (Pigou, 1927; Keynes, 1936; Shiller, 2017), extensive studies have extended the analysis to study the impact of investor sentiment and attention on financial markets (e.g., Baker and Wurgler, 2006; Andrei and Hasler, 2015; Huang et al., 2015; Chen et al., 2022; Li et al., 2019). Meanwhile, measuring sentiment and accessing their impacts on market activities are the key part of the related empirical research (Baker and Wurgler, 2007).

According to Li et al. (2019), compared with market-based measures and survey-based measures, sentiment measures based on textual data do not directly rely on equilibrium market conditions and hence represent a much more primitive sentiment in much higher frequencies. However, a main drawback of the textual approach is that the data sources are not available from standard databases. Many empirical studies use quite different data sources for textual sentiment analysis, such as search volumes (Da et al., 2011; Da et al., 2015; Fang et al., 2020; Gao et al., 2020), social media posts (Mao et al., 2011; Ackert et al., 2016; Renault, 2017; Li et al., 2019; López-Cabarcos et al., 2019; Shen et al., 2019; Guégan and Renault, 2020), and news articles (Tetlock, 2007; Baker et al., 2016; Soo, 2018; You et al., 2018; Shapiro et al., 2020).

As argued in Shapiro et al. (2020) and many other studies, news media reports may represent a much less biased estimate of sentiment as well-informed investors will not be searching for it in search engines or posting social media posts but instead focusing more on specialist coverages. However, previous studies on news sentiment pre-select a certain corpus of mainstream news articles to extract sentiment. For example, Shapiro et al. (2020) choose 16 major U.S. newspapers to construct economic sentiment to predict future economic activity. Tetlock (2007) uses the “Abreast of the Market” column in the *Wall Street Journal* to extract a textual pessimism measure to predict future stock returns of the U.S. market. Gracía (2013) uses the fraction of positive and negative words in two columns of financial news from the *New York Times* to conclude that the predictability of stock returns using news content is concentrated in recessions.

Given the increasing availability of digital news, there is no reason to restrict the textual sentiment analysis within the scope of mainstream newspapers. In this paper, we use a novel source of news reports - Global Data on Events, Location, and Tone (GDELT)[2] - to study whether news media sentiment from this massive news source impacts Chinese stock market returns and volatilities. Assisted by real-time translation of the world’s news in 100 languages, measurement of more than 2,300 emotions and themes from every article, and a massive inventory of the non-Western world’s media, GDELT provides a comprehensive digital dataset beyond mainstream newspapers for constructing news media sentiment of Chinse stock markets.

The majority of aforementioned sentiment research focuses on developed markets with a few exceptions such as Chen et al. (2014), Fang et al. (2020), and Li et al. (2019) on Chinese stock markets. Compared with developed markets, Chinese stock markets provide an excellent test site for investor or market sentiment for the following four reasons. Firstly, the Chinese stock markets are generally regarded as a highly speculative market dominated by individual investors who are more subject to irrational sentiment. Secondly, as launched in the 1990s, the Chinese stock markets are still symbolized by weak institutional organizations and stringent supervision such as short-sales constraints. More restrictive short-sales constraints result in high short-selling costs, hindering institutional investors from engaging in price stabilizing activities by trading against noise traders in China. Thirdly, a study on this largest emerging market can provide supplementary evidence about the impacts of market sentiment on stock market returns. Extant literature focuses on more developed markets. A study of the Chinese stock markets can extend the boundary of the existing research on market sentiment and provide more applicable insights for emerging markets. Lastly, the mere size of the Chinese stock markets merits such a study. With the second-largest market capital,

2: https://www.gdeltproject.org/

China has had the world's largest IPO market since 2009. Chinese stock markets have attracted increasing academic attention due to their increasing global influence and distinct institutional background.

Although there exist some studies on investor or market sentiment of Chinese stock markets, most of these studies either rely on market-based proxies (Chen et al, 2014; Zhu and Niu, 2016; Han and Li, 2017), search volumes (Fang et al., 2020), or social media posts (Li et al., 2019) to construct investor or market sentiment. It remains unclear how sentiment based on a massive database of news reports from a much broader coverage of digital sources other than mainstream newspapers affects the Chinese stock markets.

This paper tries to fill this gap by using the novel GDELT database to construct news sentiment measures for the Chinese stock markets. As introduced on the official website of GDELT and documented in Leetaru and Schrodt (2013), GDELT is an open-access database on international news that pins down and processes news in broadcast, print, and web media globally in over 100 languages on a daily basis. Thousands of emotions, organizations, locations, counts, news sources, events across the world, and average tones of analyzed news articles are identified in GDELT (see more details in Section 2).[3] It uses "directional" dictionaries measuring words associated with positive and negative connotations based on more than 40 refined sentiment dictionaries included in Wordnet.[4] The advantage of this approach is that it can directly estimate media sentiment without a pre-set emotional dictionary, and thus help reduce forecasting bias caused by subjective interferences (Schumaker et al., 2012; Rao et al., 2014).

GDELT has two types of databases, the Event database, and the Global Knowledge Graph (GKG) database. The Event database records georeferenced societal-scale behavior in more than 300 categories (such as protest, arrest, etc.) for all countries starting from 1979, while the GKG database (also called Global Beliefs database) records detailed emotional and thematic latent undercurrents of global activity starting from 2015. In this paper, we use the GKG database to extract narrative emotions from print news reports and open web articles with the theme of Chinese stock markets.

To construct news media sentiment for Chinese stock markets from the GKG database of GDELT, we first extract news articles originating from China containing the theme of "Stock Market", "IPO", and "Economic Bubble" using Google BigQuery. We then construct four variables from these news articles to measure the news media tone (daily average tone change), optimism (ratio of news reports with positive tones), attention (number of news reports), and tone dispersion (standard deviation of article tones) and estimate Exponential Generalized Autoregressive Conditional Heteroskedasticity (EGARCH) models of Nelson (1991) to explore the impacts of these sentiment variables on Chinese stock markets.

We show that these news media sentiment variables have significant impacts on Chinese stock market returns and volatilities. A larger news media tone and more news reports with positive tones indicate higher future market returns and a less volatile market condition. More media attention and a larger media tone dispersion indicate lower future market returns and a more volatile market condition. More importantly, we document the existence of asymmetric sentiment effects on aggregate Chinese stock market returns and conditional volatilities. Chinese stock market returns and volatilities tend to overreact to negative shocks to news media sentiment, and these asymmetric sentiment effects are more profound for the Shenzhen stock market. These results are robust to an alternative news media tone measure which excludes neutral words in the news article's total word count.

3: Research using GDELT includes work by Casanova et al. (2017) who use GDELT to construct a Chinese Vulnerability Index and document that this index constitutes a good indicator to assess the vulnerability sentiment of China.
4: http://wordnet.princeton.edu/. Refined dictionaries include Harvard-IV, the Fin-Neg list of Loughran and McDonald (2011), and the Federal Reserve Financial Stability list of Correa et al. (2017).

Noticing the recent work by Hasan et al. (2021) who stress the importance of integral emotions such as "Excitement" and "Anxiety" on portfolio decisions and asset prices, we also construct an emotional polarity index for the Chinese stock markets by using the percentage of emotionally charged words in each article. Even though this polarity index is by no means a substitute for the emotion index as in Hasan et al. (2021), our results show that a higher emotional polarity index is correlated with a more pessimistic outlook of news media reports and a broader tone dispersion among these reports. Moreover, this emotional polarity measure indicates higher future returns and can help improve the forecasting performance of market returns and volatilities.

Our contributions are threefold. Firstly, to the best of our knowledge, our article is the first study to use the big database of GDELT to examine the relationship between news media sentiment and Chinese stock market returns and volatilities. Compared with work by Fang et al. (2020) and many others using (the Baidu) search index to construct investor sentiments for Chinese stock markets, media sentiments constructed in this paper explore emotional responses from a broad range of news reports. Besides a general sentiment measure, we also construct an alternative optimism sentiment measure, a media attention measure, a tone dispersion measure as well as an emotional polarity measure by the richness of the dataset.

Unlike Shaprio et al. (2020) who use 16 major mainstream U.S. newspapers to construct a news sentiment index for the U.S., we do not pre-select newspapers but include all the relevant digital news reports on Chinese stock markets from GDELT. This much broader source of news reports includes print news reports as well as open web articles, which supplement the traditional mainstream newspapers substantially. Compared with Li et al. (2019) based on social media posts, we examine news media sentiment's predictive ability on Chinese stock markets. Even though we both use textual analysis to extract sentiments from either social media posts (Li et al., 2019) or news media reports (this paper), the sentiment contents extracted are different. This paper focuses on news media reports on Chinese stock markets and examines if sentiments embedded in these reports can predict Chinese stock markets.

Secondly, we add to the literature of textual analysis by examining the role of news media sentiment in a less developed market—the Chinese stock markets. As stated in Li et al. (2019) and mentioned earlier, previous research on investor sentiment and more specifically on news media sentiment (including Shaprio et al., 2020) mainly focus on more developed countries. Given the size and the growing importance of China's economy, it is necessary to examine the role of media sentiment in the finance market of the world's largest developing country for external validity.

Thirdly, we add to the literature by examining one more aspect of the news media sentiment—the emotional polarity of news reports. Given that the literature on market emotions is relatively sparse except Hasan et al. (2021), Taffler et al. (2021), and Nyman et al. (2021), we make a preliminary examination of the impacts of emotional charge in news reports on Chinese stock markets. Even though our emotional polarity measure is not in the exact spirit of the market emotional index as in Hasan et al. (2021) or Taffler et al. (2021), our empirical results show that this emotional polarity measure of news reports reveals a more pessimistic media outlook and an enlarged disagreement among news reports. It may exert contemporaneous downside pressure on stock prices and hence imply higher future stock returns.

In short, we contribute to the literature by showing that news sentiment constructed from the big database GDELT can help explain and forecast Chinese stock market returns and volatilities. This massive dataset of news reports can provide alternative market sentiment measurements in addition to sentiments embedded in mainstream newspapers, social media posts, or search volumes. The richness of this dataset enables us to measure different aspects of news reports such as the general tone or optimism, the media coverage, the tone dispersion as well as the emotional polarity. Most of our empirical findings are consistent with findings in the literature though our

sentiment measures are from a unique massive dataset. We also show that the emotional polarity of news media reports can also help predict future market returns and volatilities.

The remainder of the paper is structured as follows. Section 2 describes the data collection and variable construction process. Section 3 discusses the methodology and forecasting procedures. Section 4 presents the empirical findings. Section 5 exhibits robustness checks of alternative news media measures. Section 6 concludes.

2. Data and variables

In this section, we describe the dataset used in our empirical studies. Section 2.1 discusses the stock market indices we use to represent the Chinese stock markets. Section 2.2 discusses the construction of news media sentiment measures from the GDELT database. Section 2.3 provides basic summary statistics on these stock market returns and sentiment proxies.

2.1 Stock market data

We use the daily Shanghai Stock Exchange Composite Index (SSEI) and Shenzhen Stock Exchange Composite Index (SZEI) from the China Stock Market and Accounting Research (CSMAR) database to represent these two Chinese stock markets. The daily returns of stock markets are calculated as in Equation (1), where $ClosePrice_{it}$ represents the closing price of the stock index i on day t:

$$Return_{it} = 100 * \ln\left(ClosePrice_{it}/ClosePrice_{i,t-1}\right). \qquad (1)$$

In the above equation and during the following empirical examinations, we scale up the daily log returns by 100. Hence, these stock market returns are in percentage points.

2.2 News media sentiment measures

The news media sentiment measures are constructed from the GKG database of GDELT. To construct the news media sentiment for Chinese stock markets, we first extract news articles originating from China containing the theme of “Stock Market”, “IPO”, and “Economic Bubble” using Google BigQuery for each day in our sample period. The sample spans from June 1, 2016, to December 31, 2021. The GKG database of GDELT was launched in 2015. However, the number of news articles it tracked was relatively small and unstable in the beginning year. So, we choose to start our sample in the middle of 2016 to allow for enough news reports for each transaction day in our sample period. A total number of 4,254,080 news articles are sourced from GDELT, with a daily average number of 3,128 news reports over the 1360 transaction days.

To be specific, we briefly summarize the tone calculation steps of GDELT as follows. Firstly, print and web news media in other languages are translated in real-time to English and parsed. The algorithm in GDELT will segment the words of each article into nouns, verbs, and adjectives, and clean texts such as “The”, “a”, and some URL links. Secondly, the GDELT provides a basic tone score based on a simple algorithm first presented by Shook et al

(2012). With the algorithm, the Positive Score of an article is calculated as the percentage of all words in the article that were found to have a positive emotional connotation. In the same way, the Negative Score of an article is calculated as the percentage of all words that were found to have a negative emotional connotation. Both Positive and Negative Scores can range from +100 to -100, but common values range from +10 to -10.

Lastly, the Tone score of article j on day t is calculated as Positive Score minus Negative Score:

$$Tone\_score_{t,j} = 100 * \frac{(\Sigma PositiveWords - \Sigma NegativeWords)}{\Sigma TotalWords}. \qquad (2)$$

The range of Tone scores is generally from -10 to +10. A tone score of zero can be the result of a neutral language or a balancing of some extremely positive expressions compensated by negative ones.

To remove the persistence in the tone scores, we define the news media sentiment regarding Chinese stock markets for day t (denoted by $Tone_t$) as the change in the daily average tone score:

$$Tone_t = Average\_tone_t - Average\_tone_{t-1}, \qquad (3)$$

where $Average\_tone_t = 1/M_t \sum_{j=1}^{M_t} Tone\_score_{t,j}$ with M_t the number of articles posted on day t.

The rationale for using innovations (changes) to raw tone scores is that while raw news media sentiment is proven to have downward pressure on market prices (Tetlock, 2007), it is innovations (i.e., first differences) to raw sentiments that exert significant impacts on market returns. Examples of using first differences of raw sentiment measures include Da et al. (2015), López-Cabarcos et al. (2019), and Soo (2018). Da et al. (2015) use the daily change in search volumes from Google Trends to construct a Financial and Economic Attitudes Revealed by Search (FEARS) index as a new measure of investor sentiment. By using messages posted on StockTwits.com, López-Cabarcos et al. (2019) extract an investor sentiment based on a natural language processing software called Stanford Core NLP and examine the impacts of the variations in this sentiment on Bitcoin volatilities. The study by Soo (2018) applies textual analysis to local housing news articles to construct local housing sentiment indices for 34 U.S. cities. In the empirical settings of Soo (2018), the housing price appreciations are regressed on the innovations of these local housing sentiment indices and results indicate that the media housing sentiment has significant predictive power for future housing prices.

Inspired by the media pessimism factor of Tetlock (2007) and the sentiment measure proxied by positive and negative word counts of Gracía (2013), we propose a media optimism factor by calculating the proportion of news articles with positive tones for each day (denoted by $Optimism_t$) as an alternative measure of the news media sentiment. In Tetlock (2007), words in newspaper articles of the *Wall Street Journal* "Abreast of the Market" column are categorized into 77 predetermined categories, and then these 77 categories are collapsed into a single media factor by the principal component analysis. The single media factor is found to strongly correlate with words associated with a negative outlook and is hence referred to as the pessimism factor by Tetlock (2007).

In contrast to the pessimism factor of Tetlock (2007), the media optimism factor in our paper reflects the confidence level of news media about the future stock market. And the media optimism factor ($Optimism_t$) in our paper is easy to calculate. We classify articles into two categories, i.e., one with a positive tone and the other with a non-positive tone, and then simply calculate the ratio of articles with a positive tone over the total article count for each transaction day. The abundance of articles in the GDELT database enables us to implement this simple definition of the media optimism factor.

Similar to prior work such as Li et al. (2019) and Da et al. (2011), we define the media attention variable (denoted as $Attention_t$) as the natural logarithm of the total number of news articles for day t. Unlike attention measures based on social media posts such as those in Li et al. (2019) or Google search volumes, our news media attention variable measures news media coverage on stock market-related topics. It is also a "revealed" attention measure but is a less noise estimator than those based on social media posts or search volumes in that news articles are more formal statements.

Started from the theoretical work on the implications of divergence of opinion on stock returns by Miller (1977), and as emphasized in Hong and Stein (2007), Harris and Raviv (1993), Kandel and Pearson (1995), and Li et al. (2019), investor disagreement plays an important role in stock market activities. In the same spirit as the disagreement measure of Li et al. (2019), we construct a media disagreement measure by the standard deviation of news article tones for each day:

$$Dispersion_t = \sqrt{1/M_t \sum_{j=1}^{M_t} (Tone\_score_{t,j} - Average\_tone_t)^2}\,. \quad (4)$$

The spread of article tones ($Dispersion_t$) for a given day varies across time and could convey information on the sentiment coherence across all news media publishers.

The above tone dispersion measure is also similar to the information uncertainty measure of Rahman, Oliver, and Faff (2020) who use firms' non-earnings news releases to study whether CEOs strategically increase information uncertainty surrounding their insider stock purchases. As stated in Rahman, Oliver, and Faff (2020), the second moment of news tone, i.e., tone dispersion, can reveal important information about the market which may be disguised by the first moment of the news tone. For a certain period, a neutral news tone with a high news tone dispersion, resulting from balanced releases of good news and bad news in the same period, may render more uncertain prospects of the underlying assets. Hence, when news dispersion increases, information uncertainty is high, and thus low stock prices may be expected. In contrast, in the case of zero tone dispersion, there is no induced uncertainty and hence impacts on stock returns are not expected.

To avoid the impacts of extreme values on estimation results, we winsorize all the news media sentiment measures at the upper 1% and lower 1% levels.

2.3 Summary statistics

Figure 1 and Figure 2 depict the sentiment measurements with stock market indices SSEI and SZEI, respectively. While the sentiment measure $Tone_t$ does not show close co-movements with these two market indices, the alternative sentiment measure $Optimism_t$ moves in a very similar pattern with these two indices. The media attention measure $Attention_t$ follows a similar trend with these two indices after mid-2018 but seems to experience a systematic shift after the middle of 2020. As the bottom right panels of these figures show, the tone dispersion measure also seems to experience a systematic shift at the year-end of 2017. It is also not obvious that this tone dispersion measure shows any forward-looking behavior for these two market indices.

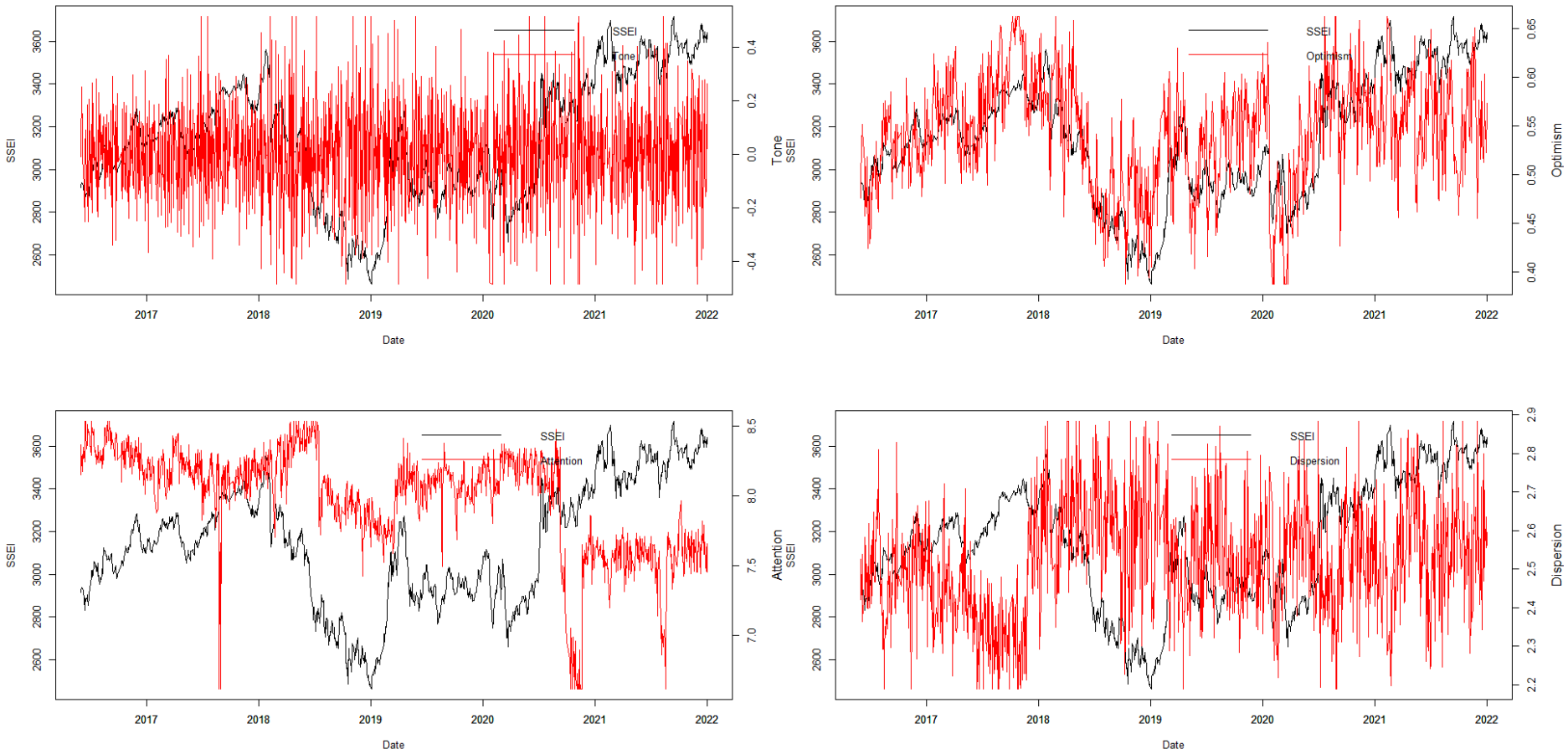
We report the descriptive characteristics of these sentiment variables and market returns in Table 1. $Return_{SSEI}$ and $Return_{SZEI}$ represent the daily log returns of these two stock markets (in percentage points), with means of 1.63% and 2.21% over the sample period, respectively. Both market returns are associated with a negative skewness and excess kurtosis. The sentiment variable $Tone_t$ has a mean of 0.000954 and a standard deviation of

0.189, while about 54% of news reports on Chinese stock markets are associated with a positive tone on average. The average daily tone spread of news reports is around 2.519 and the attention measure $Attention_t$ fluctuates around 7.99 over the sample period.

In Table 2, we tabulate the correlation coefficients among these news media sentiment measures and the two market returns. The two market returns tend to be positively correlated with the sentiment measures $Tone_t$ and $Optimism_t$, and tend to be negatively correlated with the attention measure $Attention_t$ and the tone dispersion measure $Dispersion_t$. The sentiment measure $Optimism_t$ shows significant positive correlations with $Tone_t$, while $Dispersion_t$ and $Attention_t$ exhibits negative correlations with other sentiment measures.

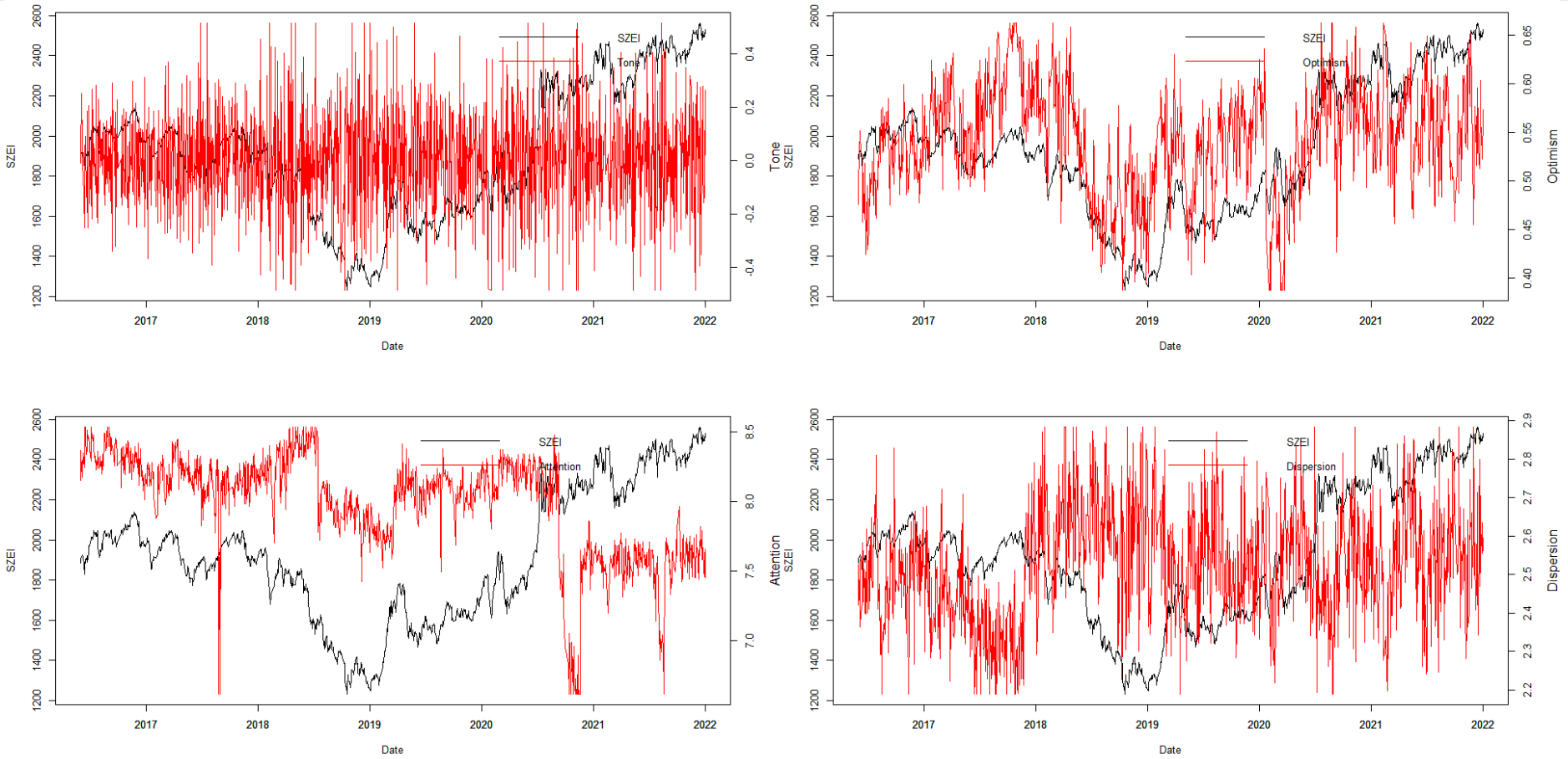
Unit root tests (ADF tests and Phillips-Perron tests) results on stock market returns and sentiment variables are depicted in Table 3. As shown in Table 3, all the time series in the sample period are stationary at the 1% significance level, which provides the validity for our EGARCH models with sentiment extended mean and conditional volatility equations.

Figure 1. **SENTIMENT MEASUREMENTS AND SHANGHAI STOCK EXCHANGE COMPOSITE INDEX** (SSEI)



Note: SSEI (black) with Tone (upper left), Optimism (upper right), Attention (bottom left), and Dispersion (bottom right).
Source: BBVA Research

Figure 2. **SENTIMENT MEASUREMENTS AND SHENZHEN STOCK EXCHANGE COMPOSITE INDEX** (SZEI)



Note: SZEI (black) with Tone (upper left), Optimism (upper right), Attention (bottom left), and Dispersion (bottom right).
Source: BBVA Research

Table 1. **DESCRIPTIVE STATISTICS**

Variables	N	Mean	Standard deviation	Min	Max	Skewness	Kurtosis
$Return_{SSEI}$	1,360	0.0163	1.037	-8.039	5.554	-0.634	9.137
$Return_{SZEI}$	1,360	0.0221	1.333	-8.789	5.275	-0.763	7.060
Tone	1,360	0.000954	0.189	-0.484	0.516	0.0546	3.152
Optimism	1,360	0.540	0.0551	0.387	0.662	-0.284	2.898
Attention	1,360	7.990	0.366	6.620	8.535	-1.280	4.981
Dispersion	1,360	2.519	0.145	2.190	2.883	0.115	2.731
*Tone**	1,360	9.43e-05	0.0288	-0.0763	0.0755	0.0559	3.104
Polarity	1,360	6.168	0.269	5.614	6.934	0.416	2.971

Note: The sample period is from June 1, 2016, to December 31, 2021. Market returns are in percentage points. All sentiment variables are winsorized at the upper 1% and lower 1% levels.
Source: BBVA Research

Table 2. **CORRELATION MATRIX OF STOCK MARKET RETURNS AND SENTIMENT VARIABLES**

Variables	$Return_{SSEI}$	$Return_{SZEI}$	*Tone*	*Optimism*	*Attention*	*Dispersion*	*Tone**	*Polarity*
$Return_{SSEI}$	1.000							
$Return_{SZEI}$	0.908***	1.000						
Tone	0.310***	0.307***	1.000					
Optimism	0.236***	0.224***	0.268***	1.000				
Attention	-0.033	-0.026	-0.009	-0.128***	1.000			
Dispersion	-0.141***	-0.134***	-0.092***	-0.330***	-0.091***	1.000		
*Tone**	0.332***	0.332***	0.947***	0.282***	-0.010	-0.083***	1.000	
Polarity	-0.089***	-0.091***	-0.019	-0.296***	0.001	0.441***	-0.033	1.000

Note: Market returns are in percentage points. All sentiment variables are winsorized at the upper 1% and lower 1% levels. ***, ** and * denote statistical significance at 1%, 5% and 10% levels, respectively.
Source: BBVA Research

Table 3. **UNIT ROOT TESTS ON STOCK RETURNS AND SENTIMENT VARIABLES**

Variables	ADF	PP
$Return_{SSEI}$	-38.213***	-38.008***
$Return_{SZEI}$	-38.927***	-39.126***
Tone	-42.067***	-44.402***
Optimism	-11.114***	-10.538***
Attention	-7.063***	-6.408***
Dispersion	-15.625***	-15.496***
*Tone**	-42.526***	-45.113***
Polarity	-11.669***	-11.195***

Note: Market returns are in percentage points. All sentiment variables are winsorized at the upper 1% and lower 1% levels. ***, ** and * denote statistical significance at 1%, 5% and 10% levels, respectively.
Source: BBVA Research

3. Methodology

To examine the impacts of news media sentiment, optimism, attention, and tone dispersion on Chinese stock markets, we extend an EGARCH of order (2,1) as follows[5]:

Mean Equation:

$$R_t = \mu + \varphi_1 Sent_{t-1} + \varepsilon_t \quad \text{with } \varepsilon_t = \sigma_t z_t, \qquad (5)$$

Conditional Variance Equation:

$$\ln(\sigma_t^2) = \omega + \alpha_1 z_{t-1} + \gamma_1(|z_{t-1}| - E|z_{t-1}|) + \sum_{j=1}^{2} \beta_j \ln(\sigma_{t-j}^2) + \theta_1 Sent_{t-1}, \quad (6)$$

where $Sent_{t-1}$ represents lagged values of sentiment variables constructed in Section 2. To serve as a benchmark, we also estimate an EGARCH (2,1) model without any sentiment variables and denote the benchmark specification as the baseline model.

To explore the asymmetric effect of positive and negative news media sentiment shocks, we further extend the benchmark EGARCH model by including variable $Tone_{t-1}$ and its absolute value $|Tone_{t-1}|$ in the mean and variance equations as follows:

Mean Equation:

$$R_t = \mu + \varphi_1 Tone_{t-1} + \vartheta_1 |Tone_{t-1}| + \varepsilon_t \quad \text{with } \varepsilon_t = \sigma_t z_t, \quad (7)$$

Conditional Variance Equation:

$$\ln(\sigma_t^2) = \omega + \alpha_1 z_{t-1} + \gamma_1(|z_{t-1}| - E|z_{t-1}|)$$

$$+ \sum_{j=1}^{2} \beta_j \ln(\sigma_{t-j}^2) + \theta_1 Tone_{t-1} + \delta_1 |Tone_{t-1}|, \qquad (8)$$

where parameter ϑ_1 captures the asymmetric effect of $Tone_{t-1}$ on market returns while parameter δ_1 captures the asymmetric effect of $Tone_{t-1}$ on conditional volatilities.

Based on the above models, one-step-ahead predictions of Chinese stock market returns and volatilities are conducted to evaluate the forecasting performances of these sentiment variables. We first estimate the models with data from June 1, 2016, through December 11, 2019, and then predict the return and volatility for December 12, 2019. By rolling the sample ahead, we repeat the forecasting exercise for the rest of the sample period. In total, 500 out-of-sample one-step-ahead forecasts have been implemented. We report the mean squared error (MSE), mean absolute error (MAE), and directional accuracy (DAC) of the forecast versus realized returns and calculate the volatility MSE as the average squared differences between the predicted conditional volatilities and the absolute values of realized returns.

5: The EGARCH order of (2,1) is selected by the BIC criteria.

4. Empirical Results

4.1. In-sample estimation results

Table 4 comprises the estimation results of the EGARCH(2,1) model with six different specifications for the Shanghai stock market (Panel A) and the Shenzhen stock market (Panel B), respectively. The positive coefficient on $Tone_{t-1}$ in the mean equation of Model 1 of Panel A implies that a higher value of sentiment measure $Tone_{t-1}$ is associated with a higher future market return, while the negative coefficient in the conditional volatility equation means a higher sentiment measure is correlated with a less volatile market. The results for the Shenzhen stock market are similar as shown in Model 1 of Panel B.

Regarding Model 2 of Panel A, the media optimism measure, $Optimism_{t-1}$, is significant in the mean equation but not in the volatility equation for Shanghai stock market returns. For Shenzhen stock market returns, the impact of $Optimism_{t-1}$ is significant in the volatility equation at the 10% significance level but not in the mean equation as shown in Model 2 of Panel B. The signs of coefficients on $Optimism_{t-1}$ are the same as those of sentiment measure $Tone_{t-1}$.

The above empirical evidence on the news media tone and the optimism measure is consistent with findings in the literature on textual tones and stock returns. For example, Loughran and McDonald (2011) show that a more negative tone in a corporate 10-K report implies lower stock returns, while Tetlock (2007) shows that the textual pessimism extracted from the “Abreast of the Market” column in the *Wall Street Journal* predicts negative returns the next day. By using information extracted from two columns in the New York Times, Gracía (2013) report similar findings. Li et al. (2019) develop textual sentiment measures for the Chinese stock market by extracting the textual tone of a large amount of online investor forum posts. Their empirical results also indicate a positive correlation between social media sentiment and future aggregate stock market returns. Even though the sentiment tone variable (and the optimism measure) in this paper is constructed based on textual analysis of print and open web news, our results are consistent with empirical findings using firm-level documents as in Loughran and McDonald (2011), mainstream news reports as in Tetlock (2007) and Gracía (2013), and online investor forum posts as in Li et al. (2019).

The news media attention measure, $Attention_{t-1}$, shows significant predicting power for stock returns for both markets as shown in Model 3. A larger value of the attention variable indicates a lower future return. However, this news media attention measure seems to have no significant impact on the market volatilities of both stock markets. The sign of investor attention’s impact on future stock returns is not conclusive in the literature. Our estimated negative impact of news media coverage ($Attention_{t-1}$) on future stock returns is in line with results from Peng and Xiong (2006), Da et al. (2011), and Chen et al. (2022) which support the argument by Barber and Odean (2008) that high attention leads to contemporaneous positive price pressure and thus lower future returns.

Likewise, the news media tone disagreement measure, $Dispersion_{t-1}$, is only significant in the mean equations of both two market returns with a negative sign and it shows positive impacts (but not significant) on the conditional volatilities for both markets as depicted in Model 4. In the conclusive Model 6 with all these four sentiment measures included, the signs on $Dispersion_{t-1}$ in the mean equations for both markets become positive. However, we tend to believe more in the results from Model 4 instead of Model 6 since these four sentiment variables are correlated and thus estimation results of the conclusive Model 6 may give misleading conclusions. Hence, we conclude that a larger news media tone dispersion will indicate lower future market returns.

The negative impacts of tone dispersion on future stock market returns are consistent with the result in Rahman, Oliver, and Faff (2020) which show that increased information uncertainty is associated with lower stock prices. This result is also consistent with findings from a broader literature on investor disagreement and stock returns. For example, Yu (2011) finds that portfolio disagreement measured from individual-stock analyst forecast dispersions is negatively related to the ex-post market return. Using trading volume as a proxy for differences in investor opinions, Lee and Swaminathan (2000) find that an enlarged opinion divergence predicts lower future returns, while Chen et al. (2002) reach the same conclusion using the breadth of mutual fund ownership as a measure of disagreement among investors. Diether et al. (2002) also come to the same conclusion by using the dispersion in analysts' earnings forecasts as a proxy of investor opinion divergence.

In the conclusive Model 6 with all these sentiment variables included, we see all these four sentiment variables have significant predicting power on the returns of these two markets (except that $Attention_{t-1}$ not significant for the Shenzhen stock returns). For both stock markets, a higher new media tone ($Tone_{t-1}$), more news reports with positive tones ($Optimism_{t-1}$), less media attention ($Attention_{t-1}$), imply higher future stock returns. As to the impacts of the tone dispersion ($Dispersion_{t-1}$), we rely on the results of Model 4 and conclude that a broader news media tone dispersion implies lower future stock returns.

For the volatility equation in Model 6, all the four sentiment variables are not significant for both stock market markets except $Tone_{t-1}$. Even though the impacts of the other three sentiment measures on each market's volatility are not significant, we can conclude that a higher new media tone ($Tone_{t-1}$), more news reports with positive tones ($Optimism_{t-1}$), less media attention ($Attention_{t-1}$), and a smaller media tone spread ($Dispersion_{t-1}$), indicate less volatile markets.[6]

Regarding Model 5, we find news media sentiment $Tone_{t-1}$ has asymmetric impacts on stock returns and volatilities for the Shenzhen stock market. The coefficients on the absolute value of $Tone_{t-1}$ are not significant for the Shanghai stock market returns. For the Shenzhen stock market, when $Tone_{t-1}$ decreases by one unit, we expect the Shenzhen stock market return to decrease by about 2.27 percentage points, while for $Tone_{t-1}$ to increase by one unit, the Shenzhen stock market return is predicted to increase only by 1.59 percentage points. The asymmetric effect of $Tone_{t-1}$ on conditional volatilities is also only significant for the Shenzhen stock market. For the Shenzhen stock market, a unit increase in $Tone_{t-1}$ will lead the log conditional variance to decrease by about 1.18 units, while a unit decrease in $Tone_{t-1}$ will lead the conditional variance to increase by about 1.60 units. These findings are consistent with the fact that investors tend to overreact when market sentiments go down.

This asymmetric impact of news media tone on conditional volatility echoes previous findings of "asymmetric" or "leverage" volatility effects in Black (1976), Nelson (1991), Engle and Ng (1993), Glosten et al. (1993), and more recently Li et al. (2019). Also, this result verifies the excess-volatility hypothesis of Li et al. (2019), meaning unusually high or low news media sentiment predicts higher volatility. This finding adds to the literature on excess volatility by extending the hypothesis to news media sentiment.

To summarize, our empirical findings indicate that a higher news media tone ($Tone_{t-1}$), and more news reports with positive tones ($Optimism_{t-1}$), indicate higher future market returns and a less volatile market condition (smaller volatilities). More intensive media attention ($Attention_{t-1}$), and a larger media tone spread ($Dispersion_{t-1}$), indicate lower future market returns and a more volatile market condition (larger volatilities). More importantly, market returns and volatilities tend to overreact to negative shocks to news media sentiment, and these asymmetric sentiment effects are more profound for the Shenzhen stock market.

6: Again, we rely on estimates from Model 2 and Model 3 to draw the conclusion on the impacts of news media optimism and news media attention on the conditional volatilities, respectively, when the signs in the conclusive Model 6 are not consistent with estimates from individual models.

Table 4. **ESTIMATION OF EXTENDED EGARCH MODELS**

Panel A- for Shanghai stock market returns

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Mean equation (R_t)						
$Tone_{t-1}$	1.4334***				1.4565***	1.2948***
$Optimism_{t-1}$		2.9574***				2.0041***
$Attention_{t-1}$			-0.0448***			-0.0343***
$Dispersion_{t-1}$				-0.0287***		0.0538***
$\|Tone_{t-1}\|$					-0.2707	
Variance equation						
$Tone_{t-1}$	-1.1881***				-1.1520	-1.1147***
$Optimism_{t-1}$		-0.6126				0.0641
$Attention_{t-1}$			0.0049			-0.0021
$Dispersion_{t-1}$				0.1385		0.0687
$\|Tone_{t-1}\|$					0.1274	

Panel B- for Shenzhen stock market returns

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Mean equation (R_t)						
$Tone_{t-1}$	1.9631***				1.9332***	1.7287***
$Optimism_{t-1}$		4.0717				2.7939***
$Attention_{t-1}$			-0.0526***			-0.0003
$Dispersion_{t-1}$				-0.1195***		0.1442***
$\|Tone_{t-1}\|$					-0.3415***	
Variance equation						
$Tone_{t-1}$	-1.4104***				-1.3888***	-1.3343***
$Optimism_{t-1}$		-1.3269*				-0.0197
$Attention_{t-1}$			0.0139			-0.0027
$Dispersion_{t-1}$				0.1732		0.0169
$\|Tone_{t-1}\|$					0.2132***	

Note: Estimates of α_1, β_1, β_2, γ_1, μ, ω are not shown to save space. Market returns are in percentage points. $|Tone_{t-1}|$ denotes the absolute value of $Tone_{t-1}$. All sentiment variables are winsorized at the upper 1% and lower 1% levels. ***, ** and * denote statistical significance at 1%, 5% and 10% levels, respectively.
Source: BBVA Research

4.3. Out-of-sample forecasting performance

Table 5 displays forecasting performances for stock returns of these two stock markets. For both market returns, all the sentiment extended models improve the forecasting accuracy significantly compared to the baseline model. Among these four sentiment-extended models, the model with $Tone_{t-1}$ shows the largest improvement of MSE and MAE for both market returns (and DAC for the Shanghai stock market), while the model with news media optimism shows the largest improvement of DAC for the Shenzhen Stock market. The conclusive model with all the four sentiment variables improves over all the single sentiment extended models on MSE for both market returns (and MAE for the Shenzhen stock market and DAC for Shanghai stock market returns).

Table 6 shows the forecasting performances for the conditional volatilities of these two stock markets. For the Shanghai stock market volatilities, these sentiments extended models all improve over the baseline model, while the model with $Tone_{t-1}$ shows the largest improvement of MSE for the Shanghai stock market conditional volatilities. Meanwhile, results for Shenzhen stock market volatilities indicate that all the extended models also outperform the baseline model. The asymmetric sentiment model (Model 5) achieves the best forecasting results for Shenzhen stock market volatilities.

Overall, our one-step-ahead forecasting results show that sentiment variables constructed from the GDELT database can improve return and volatility forecasting for the Chinese stock markets.

Table 5. **FORECASTING PERFORMANCES FOR STOCK RETURNS**

	Shanghai stock market returns			Shenzhen stock market returns		
	MSE	MAE	DAC	MSE	MAE	DAC
Baseline model	1.2357	0.7845	0.528	1.9745	1.0315	0.562
Extended model with news media sentiment						
$Tone_{t-1}$	1.1472	0.7713	0.586	1.8507	1.0145	0.566
$Optimism_{t-1}$	1.2007	0.7758	0.554	1.9253	1.0206	0.576
$Attention_{t-1}$	1.2385	0.7881	0.506	1.9839	1.0334	0.490
$Dispersion_{t-1}$	1.2358	0.7849	0.510	1.9770	1.0329	0.520
$\lvert Tone_{t-1}\rvert$	1.1415	0.7730	0.580	1.8524	1.0228	0.568
Model 6	1.1319	0.7738	0.582	1.8307	1.0143	0.584
Robustness checks with alternative tone and emotional polarity measures						
$Tone^{*}_{t-1}$	1.1294	0.7684	0.606	1.8069	1.0087	0.582
$\lvert Tone^{*}_{t-1}\rvert$	1.1211	0.7676	0.600	1.7964	1.0128	0.590
Model 9	1.1199	0.7712	0.592	1.8020	1.0103	0.584
$Polarity_{t-1}$	1.1266	0.7697	0.608	1.8132	1.0174	0.578
Model 11	1.1199	0.7728	0.574	1.8365	1.0184	0.574

Note: Row "$\lvert Tone_{t-1}\rvert$" stands for forecasts based on Model 5 of Table 4. Row "Model 6" stands for forecasts based on Model 6 of Table 4. Row "$\lvert Tone^{*}_{t-1}\rvert$" stands for forecasts based on Model 8 of Table 7. Model 9 and Model 11 are from Table 7. MSE stands for mean squared error, MAE stands for mean absolute error, and DAC stands for directional accuracy of the forecast versus realized returns.
Source: BBVA Research

Table 6. **FORECASTING PERFORMANCES FOR CONDITIONAL VOLATILITIES**

	MSE of Shanghai stock market volatilities	MSE of Shenzhen stock market volatilities
Baseline model	0.7148	0.9802
Extended model with news media sentiment		
$Tone_{t-1}$	0.6331	0.9071
$Optimism_{t-1}$	0.6949	0.9504
$Attention_{t-1}$	0.7398	1.0276
$Dispersion_{t-1}$	0.6981	0.9685
$\|Tone_{t-1}\|$	0.6368	0.9036
Model 6	0.6509	0.9552
Robustness checks with alternative tone and emotional polarity measures		
$Tone^{*}_{t-1}$	0.6415	0.9253
$\|Tone^{*}_{t-1}\|$	0.6437	0.9099
Model 9	0.6555	0.9621
$Polarity_{t-1}$	0.6878	0.9568
Model 11	0.6705	0.9487

Note: Row "$|Tone_{t-1}|$" stands for forecasts based on Model 5 of Table 4. Row "Model 6" stands for forecasts based on Model 6 of Table 4. Row "$|Tone^{*}_{t-1}|$" stands for forecasts based on Model 8 of Table 7. Model 9 and Model 11 are from Table 7.
Source: BBVA Research

5. Robustness checks

5.1. Alternative tone measure

In Section 2, we construct the news media tone measure based on a basic tone score of each article defined as the difference between the Positive Score and the Negative Score of the article. Recall that the Positive Score of an article is calculated as the percentage of all words in the article that were found to have a positive emotional connotation, while the Negative Score of an article is calculated as the percentage of all words that were found to have a negative emotional connotation. From this calculation method, we can see that the neutral words enter this basic tone score calculation as a part of the total word count as can be seen from Equation (2).

Someone may worry that the presence of neutral words may contaminate the construction of the article tone score. Hence, we construct an alternative tone score for each article as the following:

$$Tone\_score^{*}_{t,j} = 100 * \frac{(\Sigma PositiveWords - \Sigma NegativeWords)}{(\Sigma PositiveWords + \Sigma NegativeWords)}. \qquad (9)$$

This alternative basic tone score measure is now calculated as the ratio of the difference between the positive word count and the negative word count over the sum of these two counts of an article. The neutral word count is now excluded in this alternative tone score calculation.

Based on this alternative basic tone score of each article, we define the following alternative news media tone regarding Chinese stock markets for day t (denoted by $Tone_t^*$) as the change of the alternative daily average tone score:

$$Tone_t^* = Average\_tone_t^* - Average\_tone_{t-1}^*, \qquad (10)$$

where $Average\_tone_t^* = 1/M_t \sum_{j=1}^{M_t} Tone\_score_{t,j}^*$ with M_t the number of articles posted on day t. We also winsorize this alternative news media tone measure at the upper 1% and lower 1% levels to remove potential outliers.

We report some basic summary statistics of this alternative news tone measure in Table 1. As Table 1 shows, this alternative news tone measure has a mean near zero, a much smaller standard deviation (with a value of 0.0288) than the original tone measure in our sample period. Its skewness and kurtosis are almost the same as those of the original tone measure.

Table 2 shows that this alternative tone measure is highly correlated with the original tone measure with a correlation coefficient of 0.947. Unit root tests results in Table 3 show that this variable is also stationary. We re-estimate Model 1, Model 5, and Model 6 of Table 4 replacing the original news media tone measure with this alternative tone measure, with results summarized in Model 7, Model 8, and Model 9 of Table 7, respectively.

As we can see from the first three columns of Table 7, the estimation results are qualitatively similar to those using the original news media tone measure. This alternative news media tone measure exerts positive impacts on two market stock returns and negative impacts on these two stock market volatilities. A slight difference occurs with regards to the asymmetric effects. Estimates of Model 8 in Table 7 show that both the Shanghai stock market returns and the Shenzhen stock market returns respond more to a negative news media tone shock, in contrast to an insignificant asymmetric effect of the original news media tone measure on the Shanghai stock market returns as shown in Model 5 of Table 4. However, the asymmetric tone effect on conditional volatilities becomes insignificant when we use this alternative tone measure. The estimation results of the conclusive Model 9 in Table 7 are also qualitatively similar to those in Model 6 of Table 4 in terms of the sign and significance of coefficients.

One-step-ahead forecasting results using this alternative tone measure are summarized in the first three rows of the “Robustness checks” panels of Table 5 and Table 6. These forecasting performances are quite close to those using the original tone measure, with forecasts using this alternative tone measure slightly outperforming those using the original tone measure for the stock returns.

Table 7. **ROBUSTNESS CHECKS OF ESTIMATION OF EXTENDED EGARCH MODELS**

Panel A- for Shanghai stock market returns

Variables	Model 7	Model 8	Model 9	Model 10	Model 11
Mean equation (R_t)					
$Tone^*_{t-1}$	9.5813***	9.8810***	8.7728***		8.7344***
$Optimism_{t-1}$			1.7199***		1.6501***
$Attention_{t-1}$			-0.0511***		-0.0531***
$Dispersion_{t-1}$			0.0163		0.0537***
$\lvert Tone^*_{t-1}\rvert$		-2.6021***			
$Polarity_{t-1}$				10.0440***	-0.1302***
Variance equation					
$Tone^*_{t-1}$	-8.3155***	-8.1633***	-7.8821***		-7.7698***
$Optimism_{t-1}$			0.0443		0.0499
$Attention_{t-1}$			-0.0021		-0.0017
$Dispersion_{t-1}$			0.0726*		0.07097
$\lvert Tone^*_{t-1}\rvert$		0.3586			
$Polarity_{t-1}$				0.0094	-0.0014

Panel B- for Shenzhen stock market returns

Variables	Model 7	Model 8	Model 9	Model 10	Model 11
Mean equation (R_t)					
$Tone^*_{t-1}$	13.076***	13.8310***	11.671***		1.6426***
$Optimism_{t-1}$			2.5012***		2.8018***
$Attention_{t-1}$			-0.0228***		0.0095***
$Dispersion_{t-1}$			0.1253***		0.2591***
$\lvert Tone^*_{t-1}\rvert$		-4.3168***			
$Polarity_{t-1}$				13.9460***	0.2815***
Variance equation					
$Tone^*_{t-1}$	-9.3813***	-9.3448***	-8.7395***		-1.2116***
$Optimism_{t-1}$			-0.0801		-0.2082
$Attention_{t-1}$			0.0043		-0.0128
$Dispersion_{t-1}$			0.0216		0.0266
$\lvert Tone^*_{t-1}\rvert$		0.1275			
$Polarity_{t-1}$				0.0335	0.0068

Note: Estimates of α_1, β_1, β_2, γ_1, μ, ω are not shown to save space. ***, ** and * denote statistical significance at 1%, 5% and 10% levels, respectively. Market returns are in percentage points. $\lvert Tone_{t-1}\rvert$ denotes the absolute value of $Tone_{t-1}$.
Source: BBVA Research

5.2. Emotional polarity of news reports

Recent work by Hasan et al. (2021) highlights the importance of integral emotions on portfolio decisions and asset prices. Using a dictionary of anxiety and excitement-related keywords, Hasan et al. (2021) construct a market emotion index defined as the ratio of the difference between excitement and anxiety word counts to the sum of these two word counts derived from news articles.

This innovative emotion index differs from previous sentiment proxies in that instead of using the positive/negative word dictionaries of Loughran and McDonald (2011), Hasan et al. (2021) use the context-specific keyword lexicons of excitement and anxiety of Taffler et al. (2021) to measure and quantify the market emotion. Emotions as defined in Taffler et al. (2021) include "Excitement", "Anxiety", "Mania", "Panic", "Blame", "Denial", and "Guilt" and cover over 835 keywords related to all these seven kinds of emotions in total. Hasan et al. (2021) employ the word counts of "Excitement" and "Anxiety" keywords from news articles on S&P 500 of 21 national and local level newspapers to derive an aggregate market-level emotion index. The most interesting finding of Hasan et al. (2021) is that it is the emotional intensity of investor engagement with a stock that is priced rather than simply its positive/negative valence.

By the same token, we could be able to construct an emotion index for the Chinese stock markets using the GDELT database. However, there is one insurmountable obstacle for us to construct an emotional index for the Chinese stock market as in Hasan et al. (2021) by using the GDELT GKG database. Even though the emotion dictionary keyword lists of Taffler et al. (2021) outline the exact list of words belonging to these two polar emotions "Excitement" and "Anxiety", it is hard to get the exact "Excitement" and "Anxiety" word counts for each article in the GDELT GKG database.

Instead of storing the textual content of each article, the GKG database of GDELT processes the original textual content and records the related emotional "scores" (word counts or percentages as of the total word counts) of each article through the GDELT Global Content Analysis Measures (GCAM) module. As claimed in the online codebook of the GKG database, the GCAM module brings together 24 emotional measurement packages that assess more than 2,300 emotions and themes from every article. This module does record the word counts of each article belonging to specific dimensions of a dictionary. However, the exact content (or word list) of a specific dimension of each dictionary is unclear. Thus, we could not match the keywords of the two polar emotions of Taffler et al. (2021) to these specified dimensions of the GCAM module. Hence, to our best knowledge, using the GCAM module to construct an emotion index as in Hasan et al. (2021) is not easy work without knowing the exact content of its dictionaries' specific dimensions.

However, the GKG database does report an alternative "emotional" polarity score for each article. The emotional polarity score is calculated as the percentage of words that had matches in the tonal dictionaries as an indicator of how emotionally polarized or charged the text is. If an article is associated with a high emotional polarity measure but with a neutral tone score, it would suggest the text was highly emotionally charged but had roughly the same numbers of positively and negatively charged emotional words. Mathematically, this emotional polarity score of each article is defined as the following:

$$Polarity\_score_{t,j} = 100 * \frac{(\Sigma PositiveWords + \Sigma NegativeWords)}{\Sigma TotalWords}. \qquad (11)$$

As can be seen from the above expression, this polarity score is calculated as the sum of the Positive Score and the Negative Score of an article. Based on the basic emotional polarity measure of each article, we calculate the

daily average of these polarity scores as an aggregate emotional polarity measure for Chinese stock markets.[7] Even though we call this measure "emotional polarity measure", we should keep in mind that still it relies on the word counts of the positive/negative word counts instead of the "Excitement" and "Anxiety" word counts as in Hasan et al. (2021).

Basic summary statistics and correlations with other variables of this emotional polarity measure are depicted in Table 1 and Table 2, respectively. As we can see, about 6.2% of article words are associated with emotional charges on average during our sample period, and this polarity measure exhibits significant negative correlations with these two stock market returns. More importantly, this polarity measure is significantly negatively correlated with the news media optimism measure (with a correlation coefficient of -0.296) and positively correlated with the news media tone dispersion measure (with a correlation coefficient of 0.441). When the news reports are more emotionally charged, the news media is generally more pessimistic and disagrees more on the market outlook.

Unit root test results show that this polarity measure is stationary. We estimate a polarity-extended EGARCH model as results summarized in Model 10 of Table 7. The estimated coefficients of this polarity indicate that an increase in the emotional polarity is associated with higher future stock returns for both markets and no significant changes in market volatilities. We re-estimate the conclusive Model 9 of Table 7 by adding in this polarity measure and report the result in Model 11. As we can see, the coefficient on polarity on Shanghai stock market return becomes negative. However, estimation results of the conclusive Model 11 should be interpreted with caution as sentiment variables are correlated.

At first glance, the positive impacts of emotional polarity on market returns seem inconsistent with the positive effect of the news media optimism measure (which is negatively correlated with the polarity measure) and the negative impact of the tone dispersion measure (which is positively correlated with the polarity measure) on stock market returns. However, given that a highly charged news media environment may lead to contemporaneous negative price pressure and thus higher future returns, the positive correlation between the emotional polarity index and future markets returns seems reasonable and non-counterintuitive.

One-step-ahead forecasting results as summarized in Table 5 and Table 6 indicate that this emotional polarity measure from the GDELT database can improve return and volatility forecasting for the Chinese stock markets over the baseline model.

In brief, the news media emotional polarity index, based on each article's percentage of emotionally charged words (positive and negative), represents the general emotional incongruity of the news media environment. When this emotional polarity index is high, the news media is generally more pessimistic and disagrees more on the market outlook. More importantly, the emotional polarity index is correlated with higher future returns for the Chinese stock markets and can help improve the forecasting performance of market returns and volatilities.

7: This emotional polarity measure is also winsorized at the upper 1% and the lower 1% levels.

6. Conclusion

This study provides new evidence regarding the impacts of news media sentiment on Chinese stock market returns and volatilities. Using the big database of news reports from GDELT, we construct four sentiment measures, namely, the general Tone (daily average tone change), Optimism (proportion of news reports with positive tones), Attention (number of news reports), and Tone dispersion (standard deviation of article tones) for the Chinese stock markets. We then extend the EGARCH model by including sentiment variables in the mean and the conditional volatility specifications of the stock market returns.

The results obtained suggest that news media sentiment plays a significant role in predicting future returns and volatilities in China. Higher news media sentiment (Tone and Optimism), fewer news media coverage (Attention), and a smaller news media tone dispersion imply higher next-day market returns. For market volatilities, lower news media sentiment (measured by Tone and Optimism), more news media attention, and a larger news media tone dispersion, may indicate much larger market volatilities. We also document the existence of asymmetric sentiment effects on stock market returns and volatilities in China. Market returns and conditional volatilities overreact to negative shocks to the news media sentiment, and these asymmetric sentiment effects are more profound for the Shenzhen stock market. These results are robust to an alternative news media tone measure excluding neutral words.

We also construct an emotional polarity index for the Chinese stock markets by using the percentage of emotionally charged words in each article. Even though this polarity index is by no means a substitute for the emotion index as in Hasan et al. (2021), a higher emotional polarity index reveals a more pessimistic outlook of news media reports and a broader tone dispersion among these reports. This emotional polarity measure may indicate a contemporaneous negative price pressure on the market and thus imply higher future returns. This emotional polarity measure can also help improve the forecasting performance of market returns and volatilities.

The richness of the GDELT dataset allows us to explore many different aspects of the effects of news media on the Chinese stock markets and document the importance of news media's role in forming the public's belief under different market conditions. More importantly, this dataset does not only include news reports about the Chinese financial markets only. Actually, this dataset covers over 1000 themes in international news from many different languages. So, introducing GDELT to the financial research group is one of our motivations for this paper, and exploring other international financial market-related topics by use of GDELT is on our next research agenda.

Appendix. Description of News media sentiment

Variables	Definitions
Tone	Innovation to the daily average of article tone scores.
Optimism	The proportion of news articles with positive tone scores for each transaction day.
Attention	Natural logarithm of the total number of news articles for each transaction day.
Dispersion	The standard deviation of tone scores for each transaction day.
*Tone**	Innovation to the daily average of alternative tone scores excluding neutral words.
Polarity	Daily average of news articles' percentages of positive and negative words.

Declarations of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (Grant No. 71903060).

References

Ackert, L. F., Jiang, L., Lee, H. S., & Liu, J. (2016). Influential investors in online stock forums. International Review of Financial Analysis, 45, 39-46.

Andrei, D., & Hasler, M. (2015). Investor attention and stock market volatility. The Review of Financial Studies, 28(1), 33-72.

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

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

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. Journal of economic perspectives, 21(2), 129-152.

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

Black, F. (1976). Studies of stock market volatility changes. 1976 Proceedings of the American Statistical Association Business and Economic Statistics Section, 177-181.

Casanova, C., Ortiz, A., Rodrigo, T., Xia, L., & Iglesias, J. (2017). Tracking Chinese vulnerability in real time using Big Data (No. 17/13).

Chen, H., Chong, T. T. L., & She, Y. (2014). A principal component approach to measuring investor sentiment in China. Quantitative Finance, 14(4), 573-579.

Chen, J., Hong, H., & Stein, J. C. (2002). Breadth of ownership and stock returns. Journal of Financial Economics, 66(2-3), 171-205.

Chen, J., Tang, G., Yao, J., & Zhou, G. (2022). Investor attention and stock returns. Journal of Financial and Quantitative Analysis, 57(2), 455-484.

Correa, R., Garud, K., Londono, J. M., & Mislang, N. (2017). Sentiment in Central Banks' Financial Stability Reports. Available at SSRN 3091943.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. The Journal of Finance, 66(5), 1461-1499.

Da, Z., Engelberg, J., & Gao, P. (2015). The sum of all FEARS investor sentiment and asset prices. The Review of Financial Studies, 28(1), 1-32.

Diether, K. B., Malloy, C. J., & Scherbina, A. (2002). Differences of opinion and the cross-section of stock returns. The Journal of Finance, 57(5), 2113-2141.

Engle, R. F., & Ng, V. K. (1993). Measuring and testing the impact of news on volatility. The journal of finance, 48(5), 1749-1778.

Fang, J., Gozgor, G., Lau, C. K. M., & Lu, Z. (2020). The impact of Baidu Index sentiment on the volatility of China's stock markets. Finance Research Letters, 32, 101099.

Gao, Z., Ren, H., & Zhang, B. (2020). Googling Investor Sentiment around the World. Journal of Financial and Quantitative Analysis, 55(2), 549-580. doi:10.1017/S0022109019000061

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. The Journal of Finance, 48(5), 1779-1801.

García, D. (2013). Sentiment during recessions. The Journal of Finance, 68(3), 1267-1300.

Guégan, D., & Renault, T. (2020). Does investor sentiment on social media provide robust information for Bitcoin returns predictability?. Finance Research Letters, 101494.

Han, X., & Li, Y. (2017). Can investor sentiment be a momentum time-series predictor? Evidence from China. Journal of Empirical Finance, 42, 212-239.

Harris, M., & Raviv, A. (1993). Differences of opinion make a horse race. The Review of Financial Studies, 6(3), 473-506.

Bin Hasan, S., Kumar, A., & Taffler, R. (2021). Anxiety, Excitement, and Asset Prices. Available at SSRN.

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

Huang, D., Jiang, F., Tu, J., & Zhou, G. (2015). Investor sentiment aligned: A powerful predictor of stock returns. The Review of Financial Studies, 28(3), 791-837.

Kandel, E., & Pearson, N. D. (1995). Differential interpretation of public signals and trade in speculative markets. Journal of Political Economy, 103(4), 831-872.

Keynes, J. M. (1936). General Theory of Employment, Interest, and Money. London: Palgrave Macmillan.

Lee, C. M., & Swaminathan, B. (2000). Price momentum and trading volume. the Journal of Finance, 55(5), 2017-2069.

Leetaru, K., & Schrodt, P. A. (2013, April). Gdelt: Global data on events, location, and tone, 1979–2012. In ISA annual convention (Vol. 2, No. 4, pp. 1-49). Citeseer.

Li, Jia, Chen, Yun, Shen, Yan, Wang, Jingyi, and Huang, Zhuo. (2019) Measuring China's Stock Market Sentiment. Available at SSRN: https://ssrn.com/abstract=3377684

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35-65.

López-Cabarcos, M. Á., Pérez-Pico, A. M., Piñeiro-Chousa, J., & Šević, A. (2019). Bitcoin volatility, stock market and investor sentiment. Are they connected?. Finance Research Letters, 101399.

Mao, H., Counts, S., & Bollen, J. (2011). Predicting financial markets: Comparing survey, news, Twitter, and search engine data. arXiv preprint arXiv:1112.1051.

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

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. Econometrica: Journal of the Econometric Society, 347-370.

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

Peng, L., & Xiong, W. (2006). Investor attention, overconfidence, and category learning. Journal of Financial Economics, 80(3), 563-602.

Pigou, A. C. (1927). Industrial Fluctuations, London: Palgrave MacMillan.

Rahman, D., Oliver, B., & Faff, R. (2020). Evidence of strategic information uncertainty around opportunistic insider purchases. Journal of Banking & Finance, 117, 105821.

Rao, Y., Lei, J., Wenyin, L., Li, Q., & Chen, M. (2014). Building emotional dictionary for sentiment analysis of online news. World Wide Web, 17(4), 723-742.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the US stock market. Journal of Banking & Finance, 84, 25-40.

Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. Decision Support Systems, 53(3), 458-464.

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). Measuring news sentiment. Journal of Econometrics.

Shen, D., Urquhart, A., & Wang, P. (2019). Does Twitter predict Bitcoin?. Economics Letters, 174, 118-122.

Shiller, R. (2017), "Narrative Economics". Cowles Foundation Discussion Paper N 2069.

Shook, E., Leetaru, K., Cao, G., Padmanabhan, A., & Wang, S. (2012). Happy or not: Generating topic-based emotional heatmaps for Culturomics using CyberGIS. In 2012 IEEE 8th International Conference on E-Science (pp. 1-6). IEEE.

Soo, C. K. (2018). Quantifying sentiment with news media across local housing markets. The Review of Financial Studies, 31(10), 3689-3719.

Taffler, R. J., Agarwal, V., & Obring, M. (2021). Narrative Economics and Market Bubbles. Working paper.

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

You, J., Coakley, J., Firth, M., Fuertes, A. M., & Shen, Z. (2018, July). Driving the Presence of Investor Sentiment: The Role of Media Tone in IPOs. At 31st Australasian Finance and Banking Conference.

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

Zhu, B., & Niu, F. (2016). Investor sentiment, accounting information and stock price: Evidence from China. Pacific-Basin Finance Journal, 38, 125-134.

Working paper

2022

22/05 **Shulin Shen, Le Xia, Yulin Shuai, Da Gao:** Measuring news media sentiment using Big Data for Chinese stock markets.

22/04 **22/04 Ángel de la Fuente:** Las finanzas autonómicas en 2021 y entre 2003 y 2021.

22/03 **José E. Boscá, José Cano, Javier Ferri:** Covid-19 in Spain during 2021: what have vaccines achieved and what is the health cost of vaccine hesitancy?.

22/02 **Adriana Haring and Mario Iparraguirre:** Argentina. Futuro de los sectores post pandemia.

22/01 **Ángel de la Fuente:** Series largas de algunos agregados económicos y demográficos regionales: actualización de RegData hasta 2020 (RegData y RegData Dem versión 6.1-2020).

2021

21/11 **Ángel de la Fuente and Rafael Doménech:** Cross-country data on skills and the quality of schooling: a selective survey.

21/10 **Ángel de la Fuente:** La evolución de la financiación de las comunidades autónomas de régimen común, 2002-2019.

21/09 **Ángel de la Fuente:** La liquidación de 2019 del sistema de financiación de las comunidades autónomas de régimen común.

21/08 **Rodolfo Méndez-Marcano:** A global vector autoregressive model for banking stress testing.

21/07 **Ali B. Barlas, Seda Guler Mert, Berk Orkun Isa, Alvaro Ortiz, Tomasa Rodrigo, Baris Soybilgen and Ege Yazgan:** Big Data Information and Nowcasting: Consumption and Investment from Bank Transactions in Turkey.

21/06 **Ángel de la Fuente y Rafael Doménech:** El nivel educativo de la población en España y sus regiones: actualización hasta 2019.

21/05 **Saidé Salazar, Jaime Oliver, Álvaro Ortiz, Tomasa Rodrigo and Ignacio Tamarit:**
ESP/ Patrones de Consumo de Efectivo vs Tarjeta en México: una aproximación Big Data.
ING / Cash Vs Card Consumption Patterns in Mexico: A Machine Learning Approach.

21/04 **Ángel de la Fuente:** La financiación autonómica en 2020: una primera aproximación y una propuesta de cara a 2021.

21/03 **Ángel de la Fuente:** Las finanzas autonómicas en 2020 y entre 2003 y 2020.

21/02 **Joxe Mari Barrutiabengoa, J. Julián Cubero and Rodolfo Méndez-Marcano:** Output-side GHG Emissions Intensity: A consistent international indicator.

21/01 **Ángel de la Fuente y Pep Ruiz:** Series largas de VAB y empleo regional por sectores, 1955-2019 Actualización de *RegData-Sect* hasta 2019.

CLICK HERE TO ACCESS THE WORKING DOCUMENTS PUBLISHED IN

Spanish and English

DISCLAIMER

This document has been prepared by BBVA Research Department. It is provided for information purposes only and expresses data, opinions or estimations regarding the date of issue of the report, prepared by BBVA or obtained from or based on sources we consider to be reliable, and have not been independently verified by BBVA. Therefore, BBVA offers no warranty, either express or implicit, regarding its accuracy, integrity or correctness.

Any estimations this document may contain have been undertaken according to generally accepted methodologies and should be considered as forecasts or projections. Results obtained in the past, either positive or negative, are no guarantee of future performance.

This document and its contents are subject to changes without prior notice depending on variables such as the economic context or market fluctuations. BBVA is not responsible for updating these contents or for giving notice of such changes.

BBVA accepts no liability for any loss, direct or indirect, that may result from the use of this document or its contents.

This document and its contents do not constitute an offer, invitation or solicitation to purchase, divest or enter into any interest in financial assets or instruments. Neither shall this document nor its contents form the basis of any contract, commitment or decision of any kind.

With regard to investment in financial assets related to economic variables this document may cover, readers should be aware that under no circumstances should they base their investment decisions on the information contained in this document. Those persons or entities offering investment products to these potential investors are legally required to provide the information needed for them to take an appropriate investment decision.

ENQUIRIES TO:
BBVA Research: Level 95, International Commerce Centre, Austin Road West, Kowloon, Hong Kong.
Tel. + 2582 3111 / Fax. +852-2587-9717
bbvaresearch@bbva.com www.bbvaresearch.com