Social media effect, investor recognition and the cross-section of stock returns

Xiangtong Meng

Collage of Management and Economic,Tianjin University, Tianjin, 300072, China

Wei Zhang

Collage of Management and Economic,Tianjin University, Tianjin, 300072, China

China Center for Social Computing and Analytics, Tianjin 300072, China

Youwei Li

Hull University Business School, University of Hull, HU6 7RX, United Kingdom

Xing Cao

Collage of Management and Economic,Tianjin University, Tianjin, 300072, China

Xu Feng*

Collage of Management and Economic,Tianjin University, Tianjin, 300072, China

China Center for Social Computing and Analytics, Tianjin 300072, China

Abstract

Investor recognition affects cross-sectional stock returns. In informationally incomplete markets, investors have limited recognition of all securities, and their holding of stocks with low recognition requires compensation for being imperfectly diversified. Using the number of posts on the Chinese social media platform Guba to measure investor recognition of stocks, this paper provides a direct test of Merton's investor recognition hypothesis. We find a significant social media premium in the Chinese stock market. We further find that including a social media factor based on this premium significantly improves the explanatory power of Fama-French factor models of cross-sectional stock returns, and these results are robust when we control for the mass media effect and liquidity effect. Finally, we find that investment strategies based on the social media factor earn sizable risk-adjusted returns, which signifies the importance of the social media premium in portfolio management.

*Corresponding author: fengxu@tju.edu.cn

Keywords: social media; investor recognition; asset pricing

JEL Classification: G12; G14; G15

1. Introduction

Media coverage is an important determinant of cross-sectional stock returns. In the mass media era, information disseminated by media such as newspapers and television affects investors' trading activities and stock prices. Fang and Peress (2009) document the impact of mass media coverage on cross-sectional stock returns and propose what they call the "mass media effect": stocks with lower mass media coverage earn higher returns. This effect has been demonstrated in various countries (Ferguson et al., 2015; Aman et al., 2018; Griffin et al., 2011; Turner et al., 2018; Zou, Cao and Wang, 2018). Fang and Peress (2009) and others explain the mass media effect through the " investor recognition hypothesis" of Merton (1987). According to this argument, mass media coverage broadens investor recognition. Stocks with lower mass media coverage is required to offer higher returns to compensate investors for holding imperfectly diversified portfolios.

With the advent of the Internet, social media have become a mainstream platform for investors to post, gather, and exchange information (Allcott and Gentzkow, 2017; Lazer et al., 2018; Lee et al., 2016). It is thus important to study the effect of social media on cross-sectional stock returns.[1] Compared with mass media, social media may provide a better proxy for investor recognition of stocks. First, mass media disseminate information via a one-way route: it is difficult to know whether investors have received information, and it is not easy to determine how many investors recognize the securities in the market and to what extent. However, when investors post messages about a stock on social media platforms, they usually have a certain understanding of the stock, and it is easier to quantify their recognition of the stock. Second, social media facilitate interaction among investors, and this leads to lower information asymmetry (Miller and Skinner, 2015). Through investors' interactions and responses to online messages, those who initially had limited recognition of a stock will gain a certain level of knowledge about it. Finally, social media platforms can record investor interactions

[1] There are extensive studies of the relationship between social media and securities markets. Early studies discuss whether social media postings contain private information, and find that social media content is noise-based and has no statistically significant impact on stock returns (Tumarkin and Whitelaw, 2001; Das and Chen, 2007; Antweiler and Frank, 2004). In a recent study, Sprenger et al. (2014) use computational semantic analysis to classify S&P 500-related Twitter messages into good and bad news. They find that good news has the power to predict stock returns, which may indicate that social media information is not full of noise but contains private information.

and responses to online messages. These records serve as a direct measure of investor recognition, one that is not readily available in mass media.

This paper studies the social media effect on cross-sectional stock returns in China. We draw on online message posting data from Eastmoney Guba[2] (henceforth Guba), the largest and most active financial social media platform in China. The Guba platform is compatible with the Guba trading app, which is widely used by Chinese investors, especially individual investors. The influence of Guba is extensive: it occupies about 5% of Web traffic in all financial websites in China, and information from Guba influences 12 million to 22 million investors per day.[3] We collect nearly 100 million observations of investors posting data on the platform, which covers more than 3,000 stocks. For each stock, there is a firm-specific sub forum in Guba. So we are able to calculate monthly number of postings in each stock's sub-forum, which can be regarded as a direct measure of investor recognition of this stock. By sorting stocks into portfolios with different degrees of social media coverage, we can investigate the social media premium. If social media disseminate information to a wider audience and broaden investor recognition, we would expect to observe a significant social media premium. Next, we examine the impact of the social media premium on cross-sectional stock returns. By testing the informativeness of the social media premium, this paper identifies the risk associated with investor recognition in explaining cross-sectional stock returns. It also provides a direct test of the investor recognition hypothesis of Merton (1987).

The Chinese stock market is still developing, but China has witnessed considerable progress in applying social media to financial markets. According to a survey by McKinsey, 91% of Chinese respondents have visited a social media site recently.[4] 83.6% of individual investors use Internet-based social media on their computers and mobile phones as their preferred source of investment information.[5] The widespread use of social media platforms such as Guba in China ensures a huge data set of online postings, which is ideal for measuring investor recognition. Furthermore, the Chinese

[2] http://guba.eastmoney.com. In Chinese, "Gu" means stocks and "ba" means pub.
[3] Data Source: iResearch Consulting Group. http://www.iresearchchina.com
[4] Data Source: McKinsey & Company. 2014 https://www.mckinsey.com
[5] Data Source: Shenzhen Stock Exchange Report on Individual Investor Status 2016

stock market is dominated by individual investors (Han and Li, 2017), as individual trading accounts for more than 85% of the total volume.[6] Social media mainly influence the recognition of individual investors rather than institutional investors. This unique feature facilitates a direct test of Merton's investor recognition hypothesis.

In this paper, we report a significant social media premium in the Chinese stock market. Based on this premium, we further identify the risk associated with investor recognition by constructing a social media factor. We find that adding the social media factor to the well-known Fama-French factor model significantly improves its explanatory power for cross-sectional stock returns. The results are robust when we control for the mass media effect and liquidity effect. Moreover, using social media data to measure investor recognition, we provide direct evidence in support of Merton's investor recognition hypothesis. Lastly, we find that investment strategies based on the social media factor are able to generate significant profit on a risk-adjusted basis. This paper illustrates the importance of using social media data to understand cross-sectional stock returns. It points the way for future work along these lines, including research in other areas in finance.

While we view online posting as a measure of investor recognition, we note that it is also often associated with investor sentiment[7]. Bollen et al. (2011) interpret social media posting as a measure of investor sentiment. They extract daily public sentiment data from Twitter and find that public sentiment has a significant correlation with market return. Among studies linking the social media effect to cross-sectional stock returns, Kim and Kim (2014) test the predictability of extracted investor sentiment for stock returns based on 32 million messages on the Yahoo! message board from 2005 to 2010. Leung and Ton (2015) use a sample of 2.5 million observations from HotCopper, an Australian financial social media platform, from 2003 to 2008, and find that stock returns are only positively related to the sentiments expressed on social media in the case of poor performance and small-cap

[6] Data Source: Shanghai Stock Exchange Statistics Annual 2016

[7] Sentiment is not considered a rational pricing factor in many studies (Hirshleifer, 2001; Xu and Green, 2012) and has no predictive power for cross-sectional stock returns (Kim and Kim, 2014). Moreover, it is hard to accurately extract investor sentiment from online messages; for instance, an accuracy of 88% is obtained by Antweiler and Frank (2004) and about 60% by Das and Chen (2007).

stocks. We note that these studies focus mainly on the content of online postings rather than the number of online postings. However, to further control for possible sentiment effect, in this study we adopt common measures of sentiment such as the MAX effect (Bali et al., 2011; Fong and Toh, 2014), turnover ratio, and trading volume (Hong and Stein, 2007). We find that our main results remain unchanged after controlling for sentiment effect. This supports our interpretation of the number of online postings as a measure of investor recognition and provides further support for our arguments.

The remainder of this paper proceeds as follows. Section 2 describes in detail the data used. Section 3 provides the empirical results. Section 4 explains the social media effect based on the investor recognition hypothesis. Section 5 explores the profitability of investment strategies based on the social media premium. Section 6 concludes the paper.

2. Data

We collect social media data from Guba, the largest and most comprehensive financial social media platform in China, which features the highest network traffic of any comparable financial social media in China. Guba is established in 2006 to provide a financial information platform for investors to share information and opinions. At present, Guba has millions of registered users who make more than 100,000 posts per day, making it approximately 10 times larger than its closest competitors.[8] Guba includes all stocks that are currently trading or have been traded in the market, with independent sub-forums named according to the code or name of each stock. To post messages on Guba, investors need to know at least the name and code of the stock, which ensures that they have a certain degree of recognition of this stock before posting. We retrieve visible content from all Guba sub-communities, and use about 150 million postings. Noted that we regard all of postings in the sub-forum of one stock are associated with this stock because users should enter the sub-forum on their own initiative. Our key variable for a given stock, *Gubapost*, is the monthly number of all postings in its own Guba sub-forum. We use *Gubapost* as a proxy for investor recognition.

[8] Data Source: iResearch Consulting Group. http://www.iresearchchina.com.

Our sample consists of all listed companies in the Chinese A-share market from January 1, 2012, to June 30, 2017. During the sample period, social media exploded into the mainstream, which ensures enough data for our research. We exclude stocks that have been listed less than a year, those with negative book value, those subject to long-term suspension[9] or de-listing, and special-treatment stocks[10], to ensure that stock returns correctly reflect changes in the value of the company. Given that there are little stock in these abnormal situations, the average number of stocks in our sample is about 2,500, which covers most of the stocks listed in the market and over 90% of overall market capitalisation in each period. Therefore, the firms in our sample are representative. The accounting and trading data on listed companies are obtained from the Wind database.[11]

To compare the effects of social media with those of mass media, we construct a media coverage exposure variable following Fang and Peress (2009). We collect data from the "Daily News" section of the RESSET database. This section collects public news on specific stocks from well-known media sources, including Shanghai Securities News, Securities Times, China Economic Net, Sina Finance, and Tencent Finance.

Table 1 reports summary statistics on *Gubapost* and media coverage. Each stock sub-forum in Guba features an average of 700 postings per month. Social media and mobile networks have become increasingly popular in China since 2012, and *Gubapost* has grown during this period. *Gubapost* reached its peak in 2015, when the Chinese stock market boom attracted a huge number of investors, leading to a surge in social media postings. Compared with social media postings, the average number of news reports for each stock is small (about 7.8 per month). Similar to Fang and Peress (2009), about 20% of stocks are not covered by mass media over a given month. In contrast, every stock has social media coverage in every month in our sample.

[Table 1 about here.]

[9] Defined as over 10 days of suspension in a month.

[10] In the Chinese stock market, a special-treatment stock is an indication of an anomaly in its financial or other position that affects the company's going-concern assumption, or indicates that it will be forced to delist in the near future.

[11] We list all of the variables' definitions in Table 17 in the appendix.

In addition, we also test the correlation between Gubapost and firm characteristics. The result shows Gubapost is highly correlated to news coverage (the coefficient is 46%) and shareholder bases(the coefficient is 49%), which indicates Gubapost is related to investor recognition. It might be reasonable to assume that stocks with big size, high volatility and high return may attract more Gubapost, we find that correlation coefficient is 23%, 22% and 11% respectively.. Gubapost also associate with analyst coverage with the coefficient for 9%. We don't report the table for brevity.

3. Empirical results

In this section, we investigate social media effects on cross-sectional stock returns. We first look into the average excess returns of portfolios double-sorted by firm characteristics and social media coverage, and then study the risk-adjusted returns of these portfolios.

We use a nonparametric study to examine the effects of *Gubapost* across famous anomalies including $Size$, B/M, $Liquidity$, $Volatility$, $Price$, $Institution\ holding$ ratio, $Turnover$, $Dollar\ trading\ volume$, max daily return over the previous month R_{max}, monthly return of previous month R_{t-1}, $Gross\ profitability$ and $Analyst\ coverage$. Following Fang and Peress (2009), we double-sort stocks first by these firm characteristics and then in each group we sort stocks by *Gubapost*.[12] Table 2 reports average one-month-ahead returns in excess of the risk-free rate for 3×3 value-weighted (VW) portfolios double-sorted by firm characteristics and *Gubapost*. We find that as *Gubapost* increases, the average return decreases. The column "L-H" reports the average returns of a long-short portfolio (long the lowest postings portfolio and short the highest postings portfolio in each size tercile), with significantly positive returns of up to 2.07% per month in the small tercile.[13] The results show that the social media effect exists across different sizes of firm, except for the top quintile. Similar patterns of social media effects can be observed across other firm characteristics. The long-short portfolios with less significant

[12] When we single sort stocks by Gubapost, we find that the low-Gubapost portfolio has an average monthly excess return of 2.21% while high-Gubapost portfolio has only 0.67%. The difference is 1.54% and is significant at the 5% level.

[13] The results are robust across 4×4 and 5×5 value-weighted portfolios, and 3×3, 4×4 and 5×5 equal-weighted portfolios. These results are not shown in table for brevity and available upon request.

returns are the portfolios of stocks with big firm size, high B/M, high liquidity, low volatility, low turnover, and high sentiment. These stocks are likely to have more analyst coverage, more mass media coverage, and less information asymmetry, and thus better investor recognition. Accordingly, social media have limited effects on such stocks. In the same way, high returns will attract investors to focus the information about stocks, which also increase the investor recognition. Therefore, social media plays an indispensable role in providing information on less recognized stocks, and the premium of the social media effect on these stocks is significantly higher.

[Table 2 about here.]

Next, we identify the risk associated with investor recognition by constructing a Social Media Factor (SMF) and examining social media effects on the cross-sectional stock returns. Similar to Fang and Peress (2009), we divide all of the sample stocks into three groups according to *Gubapost* and calculate the one-month-ahead value-weighted return of the portfolio of the lower 1/3 *Gubapost* stocks minus those of the higher 1/3 as the value of SMF[14]. We calculate the returns of the long-short portfolio every month to obtain a time series, which represents the risk factor of the social media effect. For purpose of comparison, we also construct a Mass Media Factor (MMF) following Fang and Peress (2009). We also construct factors for the three-factor model of Fama and French (1993), four-factor model of Carhart (1997), five-factor model of Fama and French (2015), and liquidity factor (LIQ) of Pastor and Stambaugh (2003). All of the factors and their definitions are listed in Table 17. Table 3 reports the summary statistics of these factors. In our sample period, SMF accounts for a risk premium of 1.54% per month, which is higher than the premiums for size (1.45%), growth (0.21%), and other factors.[15]. Penal B of Table 3 reports the correlation matrix among SMF and other factors, which shows SMF is highly correlated with many known factors such as SMB and HML. So it's important

[14] We construct SMF exactly the same way as "mass media effect factor" in Fang and Peress (2009) so that the results are directly comparable. In addition, we construct another "SMF2" follow the approach in Fama and French (1993) to control for firm size. The correlation of two factors is 0.96, the empirical results on "SMF" and "SMF2" are similar. So, we only report the results for "SMF". We thank an anonymous referee for pointing this out.

[15] As a robustness check, we also use factors constructed from the CSMAR database and Guo et al. (2017), and obtain almost identical results.

to examine whether social media effect has a significant risk-adjusted returns .

[Table 3 about here.]

To investigate whether the new *SMF* indeed adds explanatory power to the previously established factors, we run a factor redundancy test by regressing *SMF* on other factors, following Fama and French (2015) and Fang and Peress (2009). If *SMF* provides new information, then the intercept of the regression will be significant. Table 4 reports the results of regression on Fama-French's three factors, Carhart's four factors and Fama-French's five factors, respectively, in the sub-columns of column (1); we add the liquidity factor in column (2) and the mass media factor in column (3). We see from the regression on Fama-French's three factors that they are correlated with *SMF*, and more importantly, that the intercept is significantly positive, with a coefficient of 0.01. We observe the same results for the four-factor and five-factor models with and without the liquidity factor. After controlling for market, size, book-to-market ratio, momentum, profitability, and investment factors, the significant positive intercept still exists and barely changes. These results show that *Gubapost* measures risk that cannot be explained by factors from the Fama-French and Carhart models. Moreover, the intercept remains significant after controlling for liquidity risk in column (2), which indicates that the social media factor cannot be explained by liquidity risk. When we add the mass media factor, we see that it has a significantly positive coefficient; the intercept drops from 0.010 to 0.009 but remains significant.[16] In summary, we find that *SMB*, *HML*, *LIQ*, and *MMF* are significantly correlated with *SMF*, and thus *SMF* provides new information beyond the existing well-known factors and measures a new type of risk of cross-sectional stock returns.

[Table 4 about here.]

Next, we examine the economic significance of the social media effect. We calculate the risk

[16] Hou et al. (2015) develop a q-factor model and show that of these factors, size, profitability, and investment cannot be explained by the Fama-French five factors in the Chinese market (Hou et al., 2019). We implement the factor redundancy test on the q-factor model, and obtain results similar to those for the Fama-French factors. We report the results in Table 15.

premium of *SMF* using a Fama-Macbeth two-step regression. We use the 25 *Size-Gubapost* portfolios as the left-hand-side (LHS) portfolios, and run time-series regressions of factor models to estimate betas for each factor. Then, in each period, we run a cross-sectional regression of portfolio returns on estimated betas to obtain the risk premium.

[Table 5 about here.]

Table 5 presents the Fama-Macbeth regression results. Column (1) reports the baseline results for the Fama-French three-factor, Carhart four-factor, and Fama-French five-factor models. We see that the R^2 of the three-factor models ranges from 49% to 66%. With *SMF* , R^2 increases by 10-20%, indicating that *SMF* effectively explains portfolio returns. We also note that the risk premium of *SMF* is 0.02 and significant at the 1% level with an annualized return of about 26.8%.[17] Note that the improvement on R^2 of the mass media effect of Fang and Peress (2009) is not as high as that of *SMF*. The results suggest that the social media effect is economically significant, as it generates a high premium per unit of risk.

Our study illustrates that the social media effect's impact on asset pricing is both economically and statistically significant. We further evaluate the impact by comparing the performance of factor models with and without *SMF*. We use the GRS statistic of Gibbons et al. (1989). The spirit of the GRS test is to determine whether the intercepts in regressions of portfolio returns on factor returns are jointly indistinguishable from zero, which is also the key point of the asset pricing model. The null hypothesis of the GRS test is that the intercepts of regressions are jointly indistinguishable from zero. To fully capture different characteristics of portfolios, we construct three kinds of LHS portfolios according to 5×5 *Size-B/M*, 5×5 *Size-Gubapost*, and 2×4×4 *Size-B/M-Gubapost*. We use four metrics to measure the performance of the model. The first metric $p(GRS)$ is the p-value of the GRS statistic. The second metric is $A|\alpha_i|$, the averaged absolute value of α, which is the excess returns not explained by risk factors (see Fama and French, 2015). The third measure is the indicator of total mean absolute

[17] Harvey et al. (2016) argue that a new factor needs to clear a much higher hurdle, with a t-statistic greater than 3.0. In our model, the t-statistic of β of Gubapost is about 4.0.

pricing error ($MAPE$) measured by $|\alpha| + \frac{1}{N}|\epsilon|$ in Adrian et al. (2014). The last measure, adj.R^2, is the average adjusted R^2 of the time-series regression, which measures pricing power simultaneously. Obviously, better model performance would be indicated by larger values of $p(GRS)$ and adj.R^2 and smaller values of $A|\alpha_i|$ and $MAPE$. Table 6 provides the statistics of the four measures.

[Table 6 about here.]

In Table 6, column (1) reports test statistics for the Fama-French three-factor, Carhart four-factor, and Fama-French five-factor models, and Column (2) reports the statistics for the same models augmented by *SMF*. The table shows that models with *SMF* perform better. For instance, in the *Size-Gubapost* portfolios, adding *SMF* to the Fama-French three-factor model changes the *p*-value of the GRS statistic from 0.0060 to 0.0272. The results suggest that *SMF* enhances the performance of asset-pricing models and has a pricing power for cross-sectional returns. Even in *Size-B/M* groups, which are independent of *Gubapost*, *SMF* improves model performance to certain extent.

In summary, we demonstrate that the social media effect exists across firm characteristics; the social media factor provides new information not conveyed by traditional risk factors and generates a significant positive risk premium. Asset pricing models that include the social media factor perform better in explaining cross-sectional stock returns. We now turn to exploring the economic interpretation of the social media factor.

4. Economic interpretation

The social media premium may provide compensation for imperfect diversification. The investor recognition hypothesis proposed by Merton (1987) holds that in an incomplete information environment, investors only know a subset of the available stocks, so stocks with lower recognition need to provide a risk premium to compensate their holders for being imperfectly diversified. Thus, stocks with lower recognition should have a higher rate of return. In empirical studies, because "investor recognition" is intangible, most literatures use shareholder base as the indicator to show the investor recognition of last period (Bodnaruk and Ostberg, 2009; Zhu and Jiang, 2018). However, shareholder base is only reported quarterly, so previous studies have had to select proxy variables from different perspectives. These have included media coverage (Fang and Peress, 2009; Barber and Odean, 2008), advertising investment (Grullon et al., 2004; Chemmanur and Yan, 2010), and company name (Green and Jame, 2013). [18]

In our study, the key variable *Gubapost* measures the number of postings about a stock in Guba. In order to post something, investors must know the name or code of the stock, which ensures they have a certain degree of recognition of the stock. It is from this perspective that we argue that the number of social media postings is a direct measure of investor recognition of a stock. In addition, social media provide a convenient recognition channel with low information cost to thousands of investors. The mean and median number of readers for each post in Guba is 2,068 and 1,240, respectively, indicating that information in a post will spread to about 1,000 to 2,000 investors. Thus, a higher amount of social media postings means the information will reach more investors, thereby increasing the level of investor recognition. Recognition of a stock leads investors to take the stock into consideration in their portfolio allocation. Thus, social media postings should be positively related to the shareholder bases and other investor recognition measures.

We therefore test whether *Gubapost* is positively related to other proxies of investor recognition used in pervious studies, including the number and the growth of shareholders, mass media coverage and advertising cost. In China, listed companies are required to disclose the number of shareholders and other financial indicator in their quarterly reports. For every quarter, we run a cross-sectional regression of the proxies on the logarithm of *Gubapost* and control for firm size, firm age, share price (the inverse of median price of the company stock over the previous year), return, and volatility (the

[18] There are studies provide evidence not in support of Merton's investor recognition hypothesis (see Shapiro, 2002; Sun et al., 2010). We note that empirical results in Shapiro (2002) that investor recognition effect is weak not in the whole market but in big-size (visible) stocks, is consistence with our results. In addition, Shapiro (2002) uses pension fund investment as proxy of investor recognition. Sun et al. (2010) studies EREITs market which has different investor structure to that of Chinese stock market. We view these studies and our paper are complementary to gain a better understanding of Merton's hypothesis.

average standard deviation of stock returns in the past three months), as in Bodnaruk and Ostberg (2009). Table 7 reports the average estimates over the sample quarters.

[Table 7 about here.]

Column (1) in Table 7 shows that there is a significant positive relationship between the number of social media postings and the number of shareholders. An increase of *Gubapost* by one logarithmic unit corresponds to a 0.454 logarithmic unit increase in the shareholder base. Because shareholder base is a highly persistent variable, we also test whether *Gubapost* has positive relationship with the change of shareholder bases. Column (2) shows *Gubapost* has a significant positive relationship with the change of shareholder bases. An increase of *Gubapost* by one logarithmic unit will bring 3.1% increase in total shareholder base. In addition, we use media coverage of Fang and Peress (2009) and Barber and Odean (2008), advertising investment of Grullon et al. (2004) and Chemmanur and Yan (2010) as proxies of investor recognition. Again, we find significant positive relationships in Column (3) and (4). These results show that *Gubapost* is positively related to conventional proxies of investor recognition.

In addition, there are some advantages of *Gubapost* measure over other measurements of investor recognition. First, media coverage and advertising investment are based on information dissemination and it can't be sure that investors receive the information and recognize the stock. While *Gubapost* is a direct measure that investors must have a certain recognition of a stock before they post on Guba-forum. Second, shareholder base of a stock reflects the number of investors who have taken the stock in their portfolios. However, it ignores potential investors who have recognition of the stock and may add it to their portfolios in the future. *Gubapost* reflects stock recognition of both current and future investors. Third, it is easy to get both historical and live data of *Gubapost* and it is not subject to restrictions of quarterly disclosure suffered by other measures.

We argue that social media postings broaden investor recognition and increase the shareholder base. However, stocks with a large shareholder base may garner more postings on social media, which may introduce an endogenous issue. Therefore, we double-sort stocks by shareholder base and *Gubapost* into 2×2 sub-samples, and run the preceding regression for each sub-sample. Table 8 reports the results[19].

[Table 8 about here.]

In each sub-sample in Table 8, there is still a significantly positive relationship between the number of social media postings and the number of shareholders. It is worth noting that in columns (2) and (3), for the sub-sample of high shareholder base and low *Gubapost* and the sub-sample of low shareholder base and high *Gubapost*), the relationship is still significant, which suggests that the positive relationship is robust across sub-samples double-sorted by shareholder base and *Gubapost*. This lends further support for the argument that *Gubapost* is a good proxy for investor recognition.

In the literature, there are other interpretations of the social media effect. Kim and Kim (2014) interpret the social media effect on cross-sectional returns in terms of sentiment. To evaluate the possible sentiment effect in *Gubapost*, we classify *Gubapost* into three groups (positive, neutral, and negative) using a Naive Bayesian Classification method of textual analysis (see Kim and Kim, 2014; Leung and Ton, 2015).[20] We classify 149,569,708 postings; the highest proportion are neutral in sentiment (about 74.8%), while negative postings account for 16.6% and positive for 8.6%. If the social media effect is related more closely to recognition than sentiment, we should observe the social media premium across positive, negative, and neutral sentiment. Similar to Table 2, we report average one-month-ahead returns in excess of the risk-free rate for 25 value-weighted portfolios double-sorted by *Size* and *Gubapost* under different sentiments in Table 9.

[19] We also report regression results in Table 7 using other proxies of investor recognition as dependent variables, we find consistent results. These results are available upon request.

[20] Details of the classification method are provided in the appendix. Feng et al. (2015) check the accuracy of the classification and find that the in-sample accuracy reaches 85.4%. They randomly choose and manually classify another 1,000 messages for an out-of-sample test. The accuracy rate of the out-of-sample test is also high, at 77.9%. This result is better than that in prior studies of English classification (e.g. Das and Chen, 2007; Kim and Kim, 2014). More importantly, the number of messages classified as the opposite sentiment (positive messages classified as negative and vice versa) is low (0.4% and 0.2%), which suggests a low level of systematic error.

[Table 9 about here.]

The results in Table 9 indicate that a social media premium exists across different sentiment groups. Panel B of Table 9 shows that the premium of sentiment-neutral social media postings has a similar pattern to that of the whole sample of social media in Table 2. In each panel, there is a similar pattern: as *Gubapost* increases, the average return of the stock portfolio has a significant decrease. That is, as long as the postings increase, no matter how the sentiment changes, the corresponding investor recognition will increase and the premium will decrease. The results suggest that rather than sentiment, the social media effect is induced by investor recognition.

We also investigate other sentiment proxies, for example $Turnover$(Baker et al., 2012), *Dollar trading volume*(Scheinkman and Xiong, 2003; Baker et al., 2012), and max daily return in the past months (the MAX effect) (Bali et al., 2011; Fong and Toh, 2014). They find that past high sentiment (high turnover, volume and R_{max}) results in an overestimation of stock price, which leads to a low return in the following month. We further test this to see if the social media effect can be subsumed by the sentiment effect. We construct *TurnoverF* from the return of a portfolio of 1/3 low turnover stocks minus the return of a portfolio of 1/3 high turnover stocks. We construct *VolumeF* and *MAXF* similarly. We regress *SMF* on these sentiment factors to see if it produces a significant positive α, and control for commonly used factors.

[Table 10 about here.]

Table 10 reports the regression results. In each column, we control for *MMF* and *LIQ* and other commonly used factors, as in Column (3) of Table 4. The results indicate that the sentiment effect is positively associated with the social media effect. We observe a positive correlation between *VolumeF* and *SMF*.[21] Overall, we see that the intercepts of the regressions are all significantly positive, which

[21] Due to multicollinearity with *VolumeF*, *TurnoverF* and *MAXF* show little or no correlation with *SMF*. Further tests find that each single sentiment proxy has a significant positive correlation with *SMF*. We note that the multicollinearity has little impact on the significance of the intercept; the results are available upon request.

indicates that the social media effect cannot be subsumed by the sentiment effect.[22]

In summary, we find that *Gubapost* is positively related to the number of shareholders, and the relationship is robust across sub-samples double-sorted by number of shareholders and *Gubapost*. We also find that the social media effect cannot be subsumed by the sentiment effect, which supports the view that *Gubapost* is a measure of investor recognition rather than a measure of sentiment. We therefore conclude that *Gubapost* is a better measure of investor recognition than others proposed in the literature; thus, this paper provides the first direct support for Merton's investor recognition hypothesis.

5. Profitability

We find that stocks with low social media postings earn a return premium. Here, we examine the out-of-sample profitability of portfolios of long stocks with low social media postings and short stocks with high social media postings. We develop decile portfolios on social media postings and obtain the one-month-ahead value-weighted returns of these portfolios. The monthly performance of the portfolios is reported in Table 11.

[Table 11 about here.]

Column (1) of Table 11 reports one-month-ahead value-weighted excess returns (to the risk-free rate) of each portfolio. The portfolio of the lowest 10% of *Gubapost* has an average monthly excess return of 2.6%. As the postings increase, the excess returns of the portfolios decrease.[23] The difference of returns between the lowest 1/10 and highest 1/10 *Gubapost* is 2.4% per month and it is significant. In addition, we compare the results with the returns of portfolios grouped by tercile, and the long-short

[22] Some researches document that *VolumeF* and *TurnoverF* are also the measures of investor disagreement (see Miller, 1977; Hong and Stein, 2007). So our results also indicate that social media effect cannot be subsumed by the disagreement effect.

[23] Note that there is no monotonic decrease around D7; however, the p-value of difference D7-D6 is 0.33, far from significant, while those of D1-D7 and D7-D10 are significant, which indicates that the nonlinear trend around D7 is caused by outliers.

portfolios still have a significant return of about 1.5%. Controlling for three, four, five and liquidity factors, we report the risk-adjusted return and intercept (alpha) in Columns (2)-(5). The alphas are consistently positive, and are both economically and statistically different from zero for most portfolios. The alphas range between 1.3% (the lowest postings portfolio in Column (5)) and -0.36% (the highest postings portfolio in Column (5)) with a premium of 1.66%. Column (6) of Table 11 reports the winning ratio of the portfolios' outperformance of the VW market portfolio, indicating that in 70% of the months in our sample, the long position of the portfolio with the lowest level of social media postings can outperform the market portfolio. The winning ratio for the long-short portfolio is 63%, which indicates that in most cases, the long-short portfolio outperforms the market portfolio. The last two columns reflect the risk of the portfolio, i.e., the standard deviation and Sharpe ratio of the portfolios.

We further examine whether this profitability can be sustained across different sub-samples of firm characteristics. We build the portfolio by taking a long position on the lowest 10% and a short position on the highest 10% of *Gubapost*.[24] We use firm $Size$, $Book-to-Market\ ratio$, and $Momentum$ as the characteristics of the firms, and sort them into three sub-samples. In each sub-sample, we report the portfolio's abnormal returns (alpha) as estimated by different factor models.

[Table 12 about here.]

Table 12 reports the significant risk-adjusted returns of the long-short strategy across sub-samples of different firm characteristics. We note that Fang and Peress (2009) tests the illiquidity hypothesis for the media coverage effect. The illiquidity hypothesis holds that the effect reflects mispricing and can be eliminated by arbitrage trading. Similarly, we examine whether the social media effect supports the illiquidity hypothesis. If the abnormal returns are concentrated among the most illiquid stocks, the social media effect can be eliminated by arbitrage, which supports the illiquidity hypothesis. We

[24] In terms of investability, the portfolio has an average stock number of 400, 200 for the long position and 200 for the short position. The small number of stocks is easy to manage and rebalance. The long-short portfolio also ensures initial zero investment.

use long stocks with the lowest 10% of *Gubapost* and short stocks with the highest 10%. We use $Amihud's\ Illiquidity\ Ratio$ and $Dollar\ Trading\ Volume$ as proxies for illiquidity. We examine the risk-adjusted returns of the portfolios in samples sorted by these measures. Panels A and B of Table 13 report the risk-adjusted returns of portfolios under different factor models. The reported alphas are significantly positive among all illiquidity sub-samples. For example, the alphas range from 1.2% to 1.7% in the first column of Panel A, which indicates that there is not enough evidence to support the illiquidity hypothesis.

We also examine the impacts of the sentiment effects on the profitability of the social media effect trading strategy. As before, we sort stocks according to sentiment proxies such as $Turnover$, R_{max}, and $Dollar\ Trading\ Volume$. (see Scheinkman and Xiong, 2003; Baker et al., 2012; Bali et al., 2011; Fong and Toh, 2014). Panels B and C of Table 13 report the risk-adjusted returns of portfolios under different factor models. The significantly positive alphas indicate that the excess returns of the long-short social media effect strategy cannot be explained by the sentiment effect.

[Table 13 about here.]

We also develop two trading strategies to test whether the profitability driven by the social media premium is persistent. The first is a buy-and-hold strategy to long stocks with the lowest 10% of social media postings. The second strategy holds a zero-investment portfolio and uses a long-short strategy, that is, to long stocks in the lowest decile of *Gubapost* and to short the highest decile. We rebalance the two portfolios monthly according to updated *Gubapost*. Figure 1 shows the cumulative returns of the two strategies compared with the VW market portfolio.

[Figure 1 about here.]

We see from Figure 1 that the two strategies achieve returns of about 300% for the whole sample period. Despite drawdowns, they appear to outperform the market portfolio. Next, we expand the portfolio holding period to 12 months. Table 14 reports the annual returns of the two strategies. Table

14 shows that the buy-and-hold strategy can achieve nearly 40% annualized return and 22.6% factor-free excess return with a winning ratio of 90% against the VW market portfolio. In addition, we use three market indices as benchmark to evaluate portfolio performance, that is CSI 300 (300 companies with biggest size), CSI 500 (500 companies with medium size) and CSI 1000 (1000 companies with small size). The long only strategy have excess returns of 27.3%, 19.5%, 12.9% respectively. For the long-short strategy, the annualized return is about 32%, and the Sharpe ratio is 1.168.

[Table 14 about here.]

Overall, the performance of the two strategies shows that portfolios constructed by the social media premium are able to generate economically significant profits. Portfolio managers and market practitioners should therefore take the social media premium into account in their portfolio and risk management.

6. Conclusion

Previous studies of the social media effect on asset pricing have mainly been undertaken in relation to the time series effect and sentiment effect. This paper focuses on the social media effect on cross-sectional stock returns. Using *Gubapost* as a measure of social media coverage of the Chinese stock market, we find a significant social media premium. We further identify the risk associated with investor recognition by constructing a social media factor based on the social media premium, and find that the social media factor provides new information not covered by the well-known Fama-French factors and the mass media factor of Fang and Peress (2009), and this factor cannot be subsumed by sentiment effects. Fama-French factor models augmented by the social media factor are able to generate significant risk-adjusted returns. The risk pricing test finds that per unit of risk, the social media effect can bring about a 2% change in returns of market portfolios. The GRS statistics also show that the social media factor enhances the performance of existing factor models.

Why the social media factor can improve the explanation power of cross-sectional asset returns of the well-known Fama-French factor models? We show that the mechanism is the risk associated with investor recognition. In informationally incomplete markets, investors have limited recognition of all stocks, their portfolios are imperfectly diversified, and they need to be compensated for holding such portfolios. We find that the number of social media postings is a better measure of investor recognition than existing ones in the literature. Therefore, this paper provides direct evidence in support of the investor recognition hypothesis of Merton (1987). Furthermore, we find that trading strategies based on social media premium can produce a sizable profit, which suggests that market practitioners should take the social media effect into account in their portfolio and risk management.

Appendices

A. The textual analysis process

To obtain data on the sentiment of posts collected from Guba, we classify them into three types according to their sentiment: positive, neutral, and negative. Naive Bayesian Classification is used for textual analysis. Naive Bayesian Classification is a simple and efficient method widely used for sentiment classification (Antweiler and Frank, 2004; Feng et al., 2015). We follow the method from the aforementioned papers, using the Chinese sentiment dictionary of Feng et al. (2015) based on FudanNLP, the HowNet Chinese sentiment dictionary, and 1,043 special online terms that are considered Key Sentiment Words on Chinese stock message boards. Five thousand training sentences are chosen and manually classified into three groups as the training sample.

Naive Bayesian Classification assumes that the occurrences of words are independent of each other. The conditional probability that one message contains the keyword W_I, which is included in Key Sentiment Words belongs to the sentiment group T_C, $C \in \{Positive, Neutral, Negative\}$, is as follows:

$$P(T_C|W_I) = \frac{P(W_I|T_C)P(T_C)}{P(W_I)} = \frac{P(T_C)\sum_{k=1}^{I} P(w_k|T_C)}{\sum_{k=1}^{I} P(w_k)} \tag{1}$$

where w_k is a word from the sequence W_I, and I is the total of W_I. We can calculate the sentiment probability of each message and choose the one with the maximum probability as its sentiment type and set its value to {-1,0,1}.

$$Type(W_I) = Max\{P(T_c|W_I)\}, \quad C \in \{Positive, Neutral, Negative\} \tag{2}$$

where:

$$Sentiment_I = \begin{cases} 1 & W_I = Positive \\ 0 & W_I = Neutral \\ -1 & W_I = Negative \end{cases} \tag{3}$$

B. The redundancy test for q-factors

Retail investors are highly responsive to earnings and investment decisions. Firm profitability and investment are thus powerful ex post explanatory variables in cross-sectional returns. However, the profitability and investment factors in Fama and French (2015) seem to have limited explanatory power for the Chinese stock market (Guo et al., 2017). Hou et al. (2015) develop a q-factor model, and find that the size, profitability, and investment factors among the q-factors cannot be explained by the Fama-French five factors (Hou et al., 2019). Their model thus provides a better explanation of the cross-sectional returns of the Chinese stock market.

We implement the redundancy test on the q-factors constructed from Hou et al. (2015), and report the results in Table 15. We see that the q-factors have little correlation with *SMF* except the q-size factor. In Columns (1) and (2) the q-size factor has a positive relationship with *SMF*, while other q-factors contribute little to explaining *SMF*. However, the relationship disappears after controlling for *MMF*. It is worth noting that the intercepts of the regressions remain significantly positive, which indicates that the q-factors cannot perfectly explain the social media effect. Overall, the results for the q-factors are similar to those for the Fama-French factors, showing that the investor recognition effects reflected by the social media factor cannot be subsumed by the usual factors in the literature on cross-sectional stock returns.

[Table 15 about here.]

C. Parsimonious model and short-term reversal factor.

Short-term reversal effect is an important anomaly in financial market, which has been reported in Chinese market (Nusret et al., 2017; Han et al., 2019). Pervious literature shows that the short-term reversal effect is related to overreaction, fads, or simply cognitive errors as well as price pressure when the short-term demand and supply curve changing(Da et al., 2011; Jegadeesh and Titman, 1993). Stock with high return may attract more discussion in social media, so it's reasonable doubt that low *Gubapost* comes from the low returns of the stock in the prior month, which leads to high returns in the subsequent month. In order to examine whether social media effect can be explained by short-term reversal effect, we construct short-term reversal factor($STREV$) following Jegadeesh and Titman (1993) and Han et al. (2019), and we implement a redundancy regression on *SMF*.

[Table 16 about here.]

Table 16 shows the regression result, and Column (1) shows a negative relationship between *SMF* and $STREV$. In other words, *Gubapost* and return is not absolute correlation, higher return may not be able to attract more investor posts. The result also suggests that investor recognition is a long-term process that is not dominated by the influence of short-term returns. Further, intercepts of the regressions remain significantly positive, which indicates that the short-term reversal effect cannot explain the social media effect.

In Column (2) of Table 16, we employ three kinds of parsimonious factor model to check the robustness of regression. The inspirit of parsimonious factor model is to explain data with a minimum number of variables. Hence, we construct the model in three ways: models (1) include the significant factors in pervious regression; (2) exclude the factors seem not being priced in Chinese market[25]; (3) factors in (2) and add mass media factor and liquidity factor. The result shows in each model of redundant test, intercepts remain significant, which indicates the result is robust.

D. Variable Definitions

[Table 17 about here.]

[25] CMA and UMD have little effect in China following Guo et al. (2017)

Adrian, T., Etula, E. and Muir, T. (2014), 'Financial Intermediaries and the Cross-Section of Asset Returns', *The Journal of Finance* **69**(6), 2557–2596.

Allcott, H. and Gentzkow, M. (2017), 'Social media and fake news in the 2016 election', *Journal of Economic Perspectives* **31**, 211–236.

Aman, H., Kasuga, N. and Moriyasu, H. (2018), 'Mass media effects on trading activities: television broadcasting evidence from Japan', *Applied Economics* **0**(0), 1–18.

Antweiler, W. and Frank, M. Z. (2004), 'Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards', *The Journal of Finance* **59**(3), 1259–1294.

Baker, M. and Wurgler, J. and Yuan, Y. (2012), 'Global, local, and contagious investor sentiment', *Journal of Financial Economics* **104**(2), 272–287.

Bali, T. G. and Cakici, N. and Whitelaw, R. F (2011), 'Maxing out: Stocks as lotteries and the cross-section of expected returns', *Journal of Financial Economics* **99**(2), 427–446.

Barber, B. M. and Odean, T. (2008), 'All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors', *Review of Financial Studies* **21**(2), 785–818.

Bodnaruk, A. and Ostberg, P. (2009), 'Does investor recognition predict returns?', *Journal of Financial Economics* **91**(2), 208-226.

Bollen, J., Mao, H. and Zeng, X. (2011), 'Twitter mood predicts the stock market', *Journal of Computational Science* **2**(1), 1–8.

Carhart, M. (1997), 'On persistence in mutual fund performance', *The Journal of Finance* **52**(1), 57–82.

Chemmanur, T. J. and Yan, A. (2010), 'Advertising, Investor Recognition, and Stock Returns', *SSRN Electronic Journal* .

Da, Z., Liu, Q. and Schaumburg, E. (2011), 'Decomposing Short-Term Return Reversal', *SSRN Electronic Journal*.

Das, S. R. and Chen, M. Y. (2007), 'Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web', *Management Science* **53**(9), 1375–1388.

Fama, E. F. and French, K. R. (1993), 'Common risk factors in the returns on stocks and bonds', *Journal of Financial Economics* **33**(1), 3 – 56.

Fama, E. F. and French, K. R. (2015), 'A five-factor asset pricing model', *Journal of Financial Economics* **116**(1), 1–22.

Fang, L. and Peress, J. (2009), 'Media Coverage and the Cross-section of Stock Returns', *The Journal of Finance* **64**(5), 2023–2052.

Feng, X., He, X.-Z., Lin, S. and Wang, J. (2015), 'Online Sentiment Contagion in China ', *working paper* pp. 1–47.

Ferguson, N. J., Philip, D., Lam, H. Y. and Guo, J. M. (2015), 'Media Content and Stock Returns: The Predictive Power of Press', *Multinational Finace Journal* **19**(1), 1–31.

Fong, W. M. and Toh, B. (2014), 'MAX, Lottery-like stocks, Investor sentiment, Investor optimism, Market efficiency', *Journal of Banking & Finance* **46**(C) 190–201.

Gibbons, M. R., Ross, S. A. and Shanken, J. (1989), 'A Test of the Efficiency of a Given Portfolio', *Econometrica* **57**(5), 1121.

Green, T. C. and Jame, R. (2013), ‘Company name fluency, investor recognition, and firm value’, *Journal of Financial Economics* **109**(3), 813–834.

Griffin, J. M., Hirschey, N. H. and Kelly, P. J. (2011), ‘How important is the financial media in global markets?’, *The Review of Financial Studies* **24**(12), 3941–3992.

Grullon, G., Kanatas, G. and Weston, J. P. (2004), ‘Advertising, Breadth of Ownership, and Liquidity’, *Review of Financial Studies* **17**(2), 439–461.

Guo, B., Zhang, W., Zhang, Y. and Zhang, H. (2017), ‘The five-factor asset pricing model tests for the Chinese stock market’, *Pacific-Basin Finance Journal* **43**(C), 84–106.

Han, X. Li, K. and Li, Y. (2019), ‘Investor Overconfidence and the Security Market Line: New Evidence from China’, *SSRN Electronic Journal*

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

Harvey, C. R., Liu, Y. and Zhu, H. (2016), ‘. . . and the cross-section of expected returns’, *The Review of Financial Studies* **29**(1), 5–68.

Hirshleifer, D. (2001), ‘Investor Psychology and Asset Pricing’, *The Journal of Finance* **56**(4), 1533–1597.

Hong, H. and Stein, J. D. (2007), ‘Disagreement and the Stock Market’, *Journal of Economic Perspectives* **21**(2), 109–128.

Hou, K., Mo, H., Xue, C. and Zhang, L. (2019), ‘Which factors?’, *Review of Finance* **23**(1), 1–35.

Hou, K., Xue, C. and Zhang, L. (2015), ‘Digesting anomalies: An investment approach’, *The Review of Financial Studies* **28**(3), 650–705.

Jegadeesh, N. and Titman, S. (1993), ‘Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency’, *The Journal of Finance* **48**(1), 65-91.

Kim, S.-H. and Kim, D. (2014), ‘Investor sentiment from internet message postings and the predictability of stock returns’, *Journal of Economic Behavior & Organization* **107**, 708–729.

Lazer, D., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G. and Rothschild, D. (2018), ‘The science of fake news.’, *Science* **359**(6380), 1094–1096.

Lee, J. M., Hwang, B. H. and Chen, H. (2016), ‘Are founder ceos more overconfident than professional ceos? evidence from s&p 1500 companies’, *Strategic Management Journal* **38**(3), 751–769.

Leung, H. and Ton, T. (2015), ‘The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks’, *Journal of Banking & Finance* **55**(C) 37–55.

Merton, R. C. (1987), ‘A Simple Model of Capital Market Equilibrium with Incomplete Information’, *The Journal of Finance* **42**(3), 483–510.

Miller, E. M. (1977), ‘Risk, Uncertainty, and Divergence of Opinion’, *The Journal of Finance* **32**(4), 1151–1168.

Miller, G. S. and Skinner, D. J. (2015), ‘The Evolving Disclosure Landscape: How Changes in Technology, the Media, and Capital Markets Are Affecting Disclosure’, *Journal of Accounting Research* **53**(2), 221–239.

Nusret, C., Kalok, C. and Kudret, T. (2017), ‘Cross-sectional stock return predictability in China’, *The European Journal of Finance* **23**, 581-605.

Pastor, L. and Stambaugh, R. (2003), ‘Liquidity risk and expected stock returns’, *Journal of Political Economy* **111**(3), 642–685.

Scheinkman, J. A. and Xiong, W. (2003), ‘Overconfidence and Speculative Bubbles’, *Journal of Political Economy* **111**(6), 1183–1220.

Shapiro, A.(2002), ‘The Investor Recognition Hypothesis in a Dynamic General Equilibrium: Theory and Evidence’, *Review of Financial Studies* **15**(1), 97–141.

Sprenger, T. O., Sandner, P. G., Tumasjan, A. and Welpe, I. M. (2014), ‘News or Noise? Using Twitter to Identify and Understand Company-specific News Flow’, *Journal of Business Finance & Accounting* **41**(7-8), 791–830.

Sun, Q., Yung, K. and Rahman, H.(2010), ‘Investor recognition and expected returns of EREITs’, *Journal of Real Estate Portfolio Management* **16**(2), 153–169.

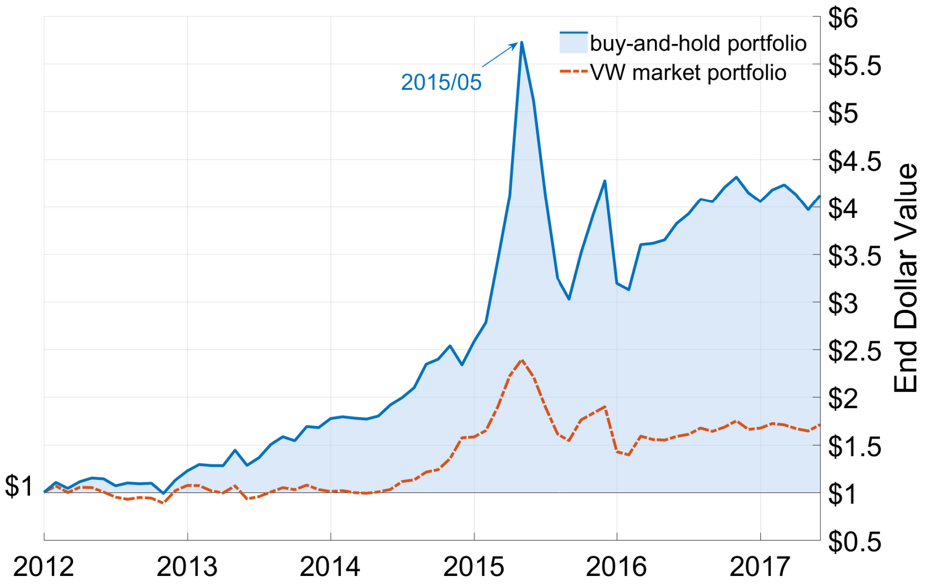
Tumarkin, R. and Whitelaw, R. F. (2001), ‘News or Noise? Internet Postings and Stock Prices’, *Financial Analysts Journal* **57**(3), 41–51.

Turner, J., Ye, Q. and Walker, C. (2018), ‘Media Coverage and Stock Returns on the London Stock Exchange, 1825–70’, *Review of Finance* **22**(4), 1605–1629.

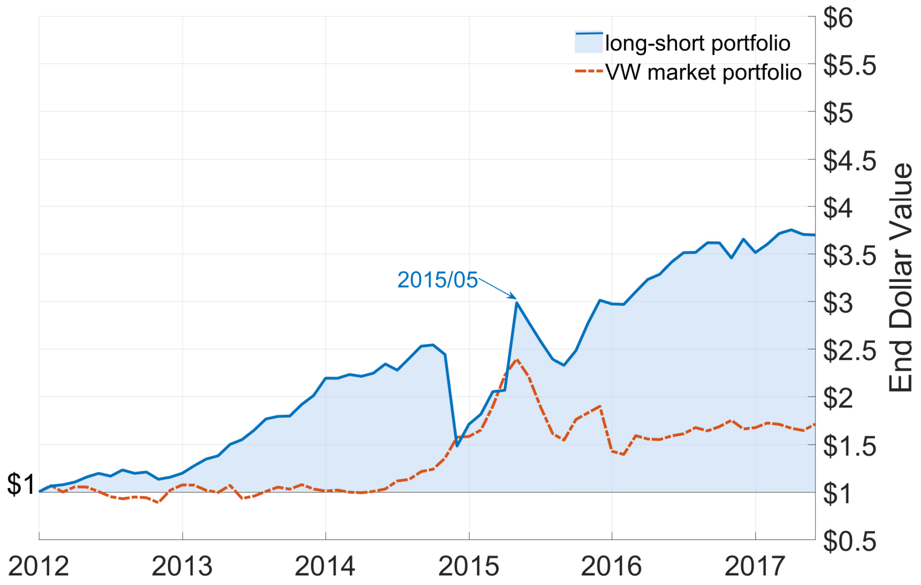
Xu, Y. and Green, C. J. (2012), ‘Asset pricing with investor sentiment: evidence from Chinese stock markets’, *The Manchester School* **81**(1), 1–32.

Zhu, H. and Jiang, L. (2018), ‘Investor recognition and stock returns: evidence from China’, *China Finance Review International*, **8**(2) 199-215.

Zou, L., Cao, K. D. and Wang, Y. (2018), ‘Media coverage and the cross-section of stock returns: The Chinese evidence’, *International Review of Finance*, **0**(0) 1–23.



(a) Buy-and-Hold Strategy



(b) Long-Short Strategy

Figure 1: Cumulative returns of strategies

This figure shows the cumulative returns of strategies with an initial investment of $1. There are two strategies shown: (a) the buy-and-hold strategy, which holds a long position on the lowest 10% of *Gubapost* for one month; (b) the long-short strategy, which holds a long position on the lowest 10% and a short position on the highest 10% of *Gubapost*. The blue shaded area represents the returns of the strategy, while the orange dashed line is the base returns of the value-weighted market portfolio.

Table 1: Summary statistics of social media and mass media coverage during 2012.1-2017.6

This table reports summary statistics for *Gubapost* and mass media coverage. The summary statistics include the average, standard deviation, minimum, percentile of 5% to 95%, maximum and minimum of *Gubapost*, and mass media coverage. The row "*overall*" reports the corresponding statistics for the full sample period.

Period	Mean	Std	Min	P5	P25	Median	P75	P95	Max
Panel A: Gubapost									
2012.1∼2012.6	492.8	843.8	11	61	154	293	551	1459	22867
2012.7∼2012.12	370.5	592.2	10	49	115	216	410	1138	17655
2013.1∼2013.6	439.9	677.5	7	60	144	260	487	1276	14440
2013.7∼2013.12	598.7	909.1	13	87	203	363	660	1825	34041
2014.1∼2014.6	426.2	547.8	12	64	149	270	490	1296	10653
2014.7∼2014.12	753.4	1018.2	1	96	249	464	876	2277	30154
2015.1∼2015.6	1289.6	2403.2	1	129	343	694	1410	4156	135392
2015.7∼2015.12	1002.0	1385.9	1	136	327	618	1150	3052	37582
2016.1∼2016.6	738.9	859.8	1	124	289	494	875	2087	18907
2016.7∼2016.12	737.9	1278.1	1	110	271	480	842	2094	89106
2017.1∼2017.6	623.3	801.6	1	98	229	400	705	1844	14676
overall	700.7	1186.7	1	82	214	406	784	2175	135392
Panel B: News									
2012.1∼2012.6	3.6	6.8	0	0	0	2	4	14	110
2012.7∼2012.12	3.6	7.4	0	0	0	1	4	14	195
2013.1∼2013.6	4.1	8.3	0	0	0	2	5	15	190
2013.7∼2013.12	6.6	12.4	0	0	1	3	8	23	517
2014.1∼2014.6	5.9	11.2	0	0	1	3	7	20	215
2014.7∼2014.12	6.8	12.0	0	0	1	3	8	23	200
2015.1∼2015.6	6.7	12.3	0	0	1	3	8	23	359
2015.7∼2015.12	10.9	21.5	0	0	2	5	12	40	603
2016.1∼2016.6	14.2	18.0	0	1	4	9	18	46	598
2016.7∼2016.12	12.4	25.0	0	0	3	7	14	41	1395
2017.1∼2017.6	4.1	8.4	0	0	1	2	5	15	193
overall	7.8	15.8	0	0	1	3	9	29	1395

Table 2: Monthly excess returns of (3 × 3) VW portfolios.

This table reports the average monthly returns in excess of the risk-free rate for 3 × 3 value-weighted portfolios double-sorted by firm characteristics and *Gubapost*. Taking the Size × *Gubapost* portfolio for an example, at the beginning of each month stocks are allocated to three Size groups (Small to Big) according to their market caps, then in each size group the stocks are allocated equally to *Gubapost* groups (Low to High) based on the previous month's *Gubapost*. The intersections of the two sorts produce twelve value-weight Size-*Gubapost* portfolios. The row "Small" reports monthly excess returns of small firms across low, medium, and high *Gubapost* groups, and the column "L-H" reports the difference between high and low *Gubapost* portfolios of small firms. The table reports Size, B/M, Liquidity, Volatility, Price, Institutional holding ratio, Turnover, Dollar trading volume, max daily return over the past month R_{max}, monthly return of pervious month R_{t-1}, Gross profitability and Analyst coverage. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	Gubapost					Gubapost			
Factors	Low	Medium	High	L-H	Factors	Low	Medium	High	L-H
Size × Gubapost portfolio					*B/M × Gubapost portfolio*				
Small	0.0361	0.0277	0.0154	0.0207***	Low	0.0214	0.0130	0.0015	0.0199***
Medium	0.0260	0.0178	0.0085	0.0175***	Medium	0.0251	0.0142	0.0070	0.0182***
Big	0.0147	0.0141	0.0059	0.0087	High	0.0211	0.0169	0.0085	0.0126*
Liquidity × Gubapost portfolio					*Volatility × Gubapost portfolio*				
Low	0.0289	0.0278	0.0196	0.0093***	Low	0.0218	0.0183	0.0096	0.0121*
Medium	0.0197	0.0170	0.0135	0.0063**	Medium	0.0236	0.0137	0.0072	0.0164**
High	0.0146	0.0112	0.0046	0.0100*	High	0.0241	0.0091	-0.0034	0.0274***
Price × Gubapost portfolio					*Institution holding × Gubapost portfolio*				
Low	0.0230	0.0192	0.0077	0.0152**	Less	0.0267	0.0177	0.0094	0.0173**
Medium	0.0232	0.0136	0.0082	0.0150**	Medium	0.0207	0.0163	0.0106	0.0101
High	0.0219	0.0126	0.0036	0.0183**	More	0.0214	0.0138	0.0067	0.0147**
Turnover × Gubapost portfolio					*Dollar trading volume × Gubapost portfolio*				
Low	0.0204	0.0157	0.0099	0.0105	Low	0.0268	0.0238	0.0196	0.0071**
Medium	0.0244	0.0165	0.0085	0.0159*	Medium	0.0179	0.0170	0.0092	0.0087***
High	0.0247	0.0114	0.0014	0.0233***	High	0.0160	0.0102	0.0040	0.0120**
R_{max} × Gubapost portfolio					*R_{t-1}× Gubapost portfolio*				
Low	0.0236	0.0165	0.0085	0.0151**	Low	0.0249	0.0236	0.0100	0.0148**
Medium	0.0244	0.0148	0.0099	0.0144*	Medium	0.0246	0.0146	0.0082	0.0165**
High	0.0193	0.0132	0.0072	0.0122	High	0.0166	0.0114	0.0041	0.0125
Gross profitability × Gubapost portfolio					*Analyst coverage × Gubapost portfolio*				
Low	0.0212	0.0121	0.0072	0.0139**	Low	0.0265	0.0162	0.0063	0.0202***
Medium	0.0221	0.0127	0.0085	0.0136*	Medium	0.0221	0.0146	0.0073	0.0148**
High	0.0225	0.0157	0.0058	0.0167**	High	0.0197	0.0146	0.0075	0.0122

Table 3: Summary statistics for risk factors

This table shows the summary statistics and correlation matrix for the average monthly returns of the risk factors used in this study. The construction of the factors *Mkt*, *SMB*, *HML*, *UMD*, *RMW*, *CMA*, *LIQ* is consistent with the studies of Fama and French (1993), Carhart (1997), Fama and French (2015), and Pastor and Stambaugh (2003). *SMF* is the social media factor and is constructed using the difference between the VW returns of the lowest and highest 1/3 of the portfolios, sorted by the last month *Gubapost*. *MMF* is the mass media factor and is constructed using the difference between the VW returns of the portfolios of stocks with no media coverage and the portfolios of stocks whose media coverage exceeds the median level.

Panel A: Summary statistics

	SMF	MMF	Mkt	SMB	HML	UMD	RMW	CMA	LIQ
Mean	0.0154	0.0136	0.0077	0.0145	0.0021	0.0016	-0.0058	0.0045	0.0060
Std dev.	0.0636	0.0579	0.0738	0.0544	0.0454	0.0505	0.0312	0.0195	0.0258
Median	0.0135	0.0132	0.0097	0.0138	0.0025	0.0081	-0.0073	0.0036	0.0026
Min	-0.3039	-0.2513	-0.2511	-0.1999	-0.1648	-0.1974	-0.0868	-0.0378	-0.0393
Max	0.2905	0.2274	0.1670	0.2015	0.1915	0.1463	0.0914	0.0636	0.0880
Skewness	-0.6760	-0.7222	-0.4337	-0.2451	0.3486	-0.5066	0.2782	0.5100	1.1003
Kurtosis	15.9914	10.4936	4.7582	6.9479	8.7261	6.0287	4.1680	3.9467	4.9241

Panel B: Correlation matrix

	SMF	MMF	Mkt	SMB	HML	UMD	RMW	CMA	LIQ
SMF	1.0000								
MMF	0.8759	1.0000							
Mkt	0.1403	0.1627	1.0000						
SMB	0.8652	0.8924	0.3113	1.0000					
HML	-0.8925	-0.7842	-0.1851	-0.8144	1.0000				
UMD	0.3665	0.1717	-0.0373	0.3209	-0.3120	1.0000			
RMW	-0.6657	-0.7686	-0.4047	-0.8659	0.5886	-0.3214	1.0000		
CMA	-0.0562	0.1768	0.1721	0.1991	0.2180	0.0693	-0.4606	1.0000	
LIQ	-0.3946	-0.2223	0.1570	-0.1423	0.3470	-0.2721	-0.0233	0.3077	1.0000

Table 4: Factor redundancy test for SMF

This table shows the results of regressing the returns of the long-short portfolio of *Gubapost* on commonly used risk factors. A significant intercept of a regression indicates that the corresponding risk factors cannot fully explain the excess return of the portfolio. *SMF* is the dependent variable, and columns (2) and (3) add *LIQ* and *MMF* to control the liquidity risk and mass media effect, respectively. t-values are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

		(1)			(2)			(3)	
Mkt	-0.088**	-0.081*	-0.085*	-0.064	-0.061	-0.064	-0.035	-0.020	-0.031
	(-2.032)	(-1.860)	(-1.861)	(-1.564)	(-1.492)	(-1.501)	(-0.907)	(-0.521)	(-0.758)
SMB	0.537***	0.520***	0.588***	0.600***	0.590***	0.599***	0.305**	0.197	0.360**
	(5.395)	(5.167)	(3.407)	(6.330)	(6.067)	(3.712)	(2.363)	(1.427)	(2.158)
HML	-0.753***	-0.744***	-0.715***	-0.608***	-0.608***	-0.600***	-0.558***	-0.548***	-0.482***
	(-6.521)	(-6.430)	(-4.559)	(-5.208)	(-5.181)	(-3.998)	(-5.072)	(-5.088)	(-3.348)
UMD		0.069			0.031			0.115*	
		(1.077)			(0.507)			(1.915)	
RMW			0.045			-0.016			0.032
			(0.179)			(-0.056)			(0.156)
CMA			-0.059			-0.023			-0.145
			(-0.243)			(-0.079)			(-0.654)
LIQ				-0.392***	-0.379***	-0.392***	-0.356***	-0.299**	-0.342***
				(-3.220)	(-3.029)	(-3.136)	(-3.119)	(-2.593)	(-2.925)
MMF							0.336***	0.418***	0.359***
							(3.151)	(3.706)	(3.221)
Intercept	0.010***	0.010***	0.010***	0.011***	0.011***	0.011***	0.010***	0.010***	0.009***
	(2.919)	(2.943)	(2.689)	(3.415)	(3.394)	(3.240)	(3.370)	(3.375)	(3.019)
R^2	0.863	0.865	0.863	0.883	0.883	0.883	0.900	0.906	0.901

Table 5: Factor risk premium on Fama-Macbeth regression

This table shows the Fama-Macbeth two-step regression results. We use the time-series average returns of 25VW *Size-Gubapost* LHS portfolios as the dependent variable; the independent variable is the beta coefficients estimated from the first step time-series: $r_{p,t} = \alpha_p + \beta_{i,p} Factor_{i,p,t} + \epsilon_{p,t}$. The table presents the estimation results of $\bar{r_p} = \lambda_0 + \lambda_i \beta_{i,p} + \eta_p$. Column (1) reports the baseline model of the risk premium of factors from Fama and French (1993), Carhart (1997), and Fama and French (2015). Column (2) adds *SMF* to the regression, while column (3) adds *MMF*. We calculate the Newey-West standard errors. t-values are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	(1)			(2)			(3)		
	FF 3	Carhart 4	FF 5	FF 3	Carhart 4	FF 5	FF 3	Carhart 4Variable	FF 5
SMF				0.020***	0.020***	0.019***			
				(4.049)	(4.379)	(4.173)			
MMF							0.030*	0.030*	0.022
							(1.778)	(1.773)	(1.528)
Mkt	-0.028	-0.032	0.001	0.025	0.021	0.033	-0.023	-0.026	0.004
	(-1.475)	(-1.556)	(0.054)	(1.157)	(1.054)	(1.673)	(-1.140)	(-1.180)	(0.219)
SMB	0.013***	0.013***	0.012***	0.018***	0.019***	0.016***	0.013***	0.013***	0.012***
	(3.898)	(3.871)	(4.369)	(5.975)	(6.601)	(5.806)	(3.922)	(3.859)	(4.359)
HML	-0.007	-0.008	-0.011	0.014	0.017*	0.007	-0.002	-0.003	-0.007
	(-0.962)	(-0.984)	(-1.607)	(1.664)	(2.028)	(0.830)	(-0.194)	(-0.256)	(-0.765)
UMD		-0.005			-0.024			-0.012	
		(-0.229)			(-1.469)			(-0.531)	
RMW			0.01			-0.001			0.008
			(1.663)			(-0.212)			(1.106)
CMA			0.002			0.011			0.006
			(0.326)			(1.615)			(0.693)
Intercept	0.035*	0.039*	0.006	-0.018	-0.016	-0.026	0.029	0.032	0.002
	(1.750)	(1.810)	(0.331)	(-0.835)	(-0.782)	(-1.290)	(1.352)	(1.376)	(0.121)
R^2	0.490	0.498	0.662	0.688	0.737	0.765	0.509	0.513	0.673

Table 6: The GRS test of the factor models

This table reports the test results of model performance of asset pricing models for the 25 (5 × 5) and 32 (2 × 4 × 4) VW portfolios. The models are the Fama and French (1993) three-factor model with *Mkt*, *SMB* and *HML*, Carhart (1997) four-factor model with *Mkt*, *SMB*, *HML* and *UMD*, and Fama and French (2015) five-factor model with *Mkt*, *SMB*, *HML*, *RMW* and *CMA*. The test statistics for models with and without *SMF* are reported in columns (1) and (2), respectively. $p(GRS)$ is the p-value of the GRS statistic, which tests whether the expected values of all intercepts are zero. $A|\alpha_i|$ is the average absolute value of intercepts. $totalMAPE$ is the total mean absolute pricing error measured by $|\alpha| + \frac{1}{N}|\epsilon|$. $adj.R^2$ is average adjusted R-square of the time-series regression.

	(1)				(2)			
	$p(GRS)$	$A\|\alpha_i\|$	$MAPE$	$adj.R^2$	$p(GRS)$	$A\|\alpha_i\|$	$MAPE$	$adj.R^2$
25 Size-B/M portfolios								
FF 3	0.0127	0.0027	0.0173	0.9597	0.0351	0.0029	0.0171	0.9610
Carhart 4	0.0099	0.0027	0.0169	0.9614	0.0293	0.0029	0.0168	0.9628
FF 5	0.0125	0.0028	0.0172	0.9598	0.0353	0.0027	0.0168	0.9612
25 Size-Gubapost portfolios								
FF 3	0.0060	0.0061	0.0222	0.9427	0.0272	0.0043	0.0191	0.9525
Carhart 4	0.0032	0.0061	0.0221	0.9436	0.0180	0.0042	0.0189	0.9537
FF 5	0.0150	0.0057	0.0217	0.9431	0.0485	0.0040	0.0186	0.9532
32 Size-B/M-Gubapost portfolios								
FF 3	0.0714	0.0058	0.0237	0.9649	0.2069	0.0045	0.0213	0.9684
Carhart 4	0.0818	0.0058	0.0233	0.9651	0.2234	0.0046	0.0208	0.9682
FF 5	0.1429	0.0056	0.0233	0.9641	0.3227	0.0044	0.0209	0.9678

Table 7: Investor recognition and Gubapost

This table shows Fama-Macbeth regressions of investor recognition proxies on *Gubapost* and mass media coverage. The dependent variables are logarithm of shareholder bases, the change of shareholder bases, mass media coverage and advertisement. *Gubapost* and *Mass media* are the respective amounts of social media posts and mass media coverage. Size is the market cap at the beginning of the month. Age is the number of days from the company's founding. 1/(share price) is the inverse of the median price in a given month. Return and Volatility are respectively the average and standard deviation of stock returns over the past three months. The regressions are estimated quarterly. We use the Newey-West method to adjust standard errors. *t*-values are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	(1)	(2)	(3)	(4)
	ln(Shareholder)	ΔShareholder	ln(Mass media)	ln(Advertisement)
ln(Gubapost)	0.454***	0.031***	0.212***	0.079**
	(25.018)	(5.604)	(23.111)	(2.781)
ln(Mass media)	-0.022**	0.013***	-	0.149***
	(-2.120)	(5.075)	-	(3.494)
ln(Size)	0.397***	0.003	0.521***	-0.096***
	(28.015)	(0.637)	(29.646)	(-3.064)
ln(Age)	0.230***	-0.019***	-0.124***	0.170***
	(13.985)	(-3.791)	(-6.148)	(3.603)
1/(share price)	5.699***	-0.072	-1.554***	-4.827***
	(22.331)	(-0.790)	(-5.392)	(-3.568)
Return	-0.367***	-0.194***	0.198**	0.210
	(-8.756)	(-8.454)	(2.568)	(0.938)
Volatility	-10.978***	7.769***	32.054***	-27.309***
	(-6.813)	(10.063)	(12.759)	(-3.286)
Intercept	-3.876***	-0.387***	-10.522***	2.907***
	(-14.301)	(-3.926)	(-20.903)	(3.336)
R^2	0.816	0.155	0.485	0.045

Table 8: Shareholder base regressions grouped by shareholder base and Gubapost

We double-sort stocks into four groups according to the shareholder base and *Gubapost*. This table shows Fama-Macbeth shareholder base regression results for each group. The variables are the same as in Table 7. The regressions are estimated quarterly: $\ln(numbers\ of\ shareholders) = \alpha + \beta \ln(Gubapost) + \gamma_i control_i + \epsilon$. We calculate the Newey-West standard errors. t-values are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	(1)	(2)	(3)	(4)
	Low Gubapost	Low Gubapost	High Gubapost	High Gubapost
	Low Shareholders	High Shareholders	Low Shareholders	High Shareholders
ln(Gubapost)	0.408***	0.161***	0.189***	0.320***
	(25.749)	(10.599)	(13.654)	(11.610)
ln(Mass media)	-0.021**	-0.021***	-0.009	-0.007
	(-2.259)	(-3.958)	(-1.251)	(-0.856)
ln(Size)	0.239***	0.295***	0.182***	0.373***
	(24.342)	(19.638)	(13.647)	(30.586)
ln(Age)	0.174***	0.100***	0.122***	0.141***
	(11.655)	(9.112)	(7.736)	(9.751)
1/(share price)	6.040***	3.287***	4.279***	3.863***
	(16.203)	(22.915)	(17.603)	(17.401)
Return	-0.387***	-0.334***	-0.221***	-0.154***
	(-9.981)	(-3.941)	(-6.563)	(-4.530)
Volatility	-9.155***	-4.209**	-1.067	-9.581***
	(-6.356)	(-2.671)	(-1.043)	(-6.469)
Intercept	0.174	2.065***	3.303***	-1.242***
	(0.633)	(5.527)	(7.217)	(-6.460)
R^2	0.650	0.484	0.418	0.723

Table 9: Monthly excess returns of 25 VW portfolios in different sentiment groups

This table reports average monthly excess returns for 25 (5 × 5) VW portfolios formed on Size and *Gubapost* over 65 months. At the beginning of each month, stocks are allocated to five Size groups (Small to Big) by market cap, and allocated equally to five *Gubapost* groups (Low to High) by the average number of Guba posts in the last month. The intersections of these two sorts produce three 25 value-weight Size-*Gubapost* portfolios. Columns L-H show the difference between the returns of the low and high quintile of *Gubapost* portfolios. Correspondingly, rows S-B show the average returns of Small size portfolios minus those of Big size portfolios. Panels A-C separately report the results of different sentiment groups (negative, neutral, positive) . *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A Size × Negative Gubapost portfolio

		Gubapost					
		low	2	3	4	High	L-H
Size	small	0.037	0.034	0.028	0.021	0.010	0.027***
	2	0.032	0.026	0.021	0.016	0.002	0.030***
	3	0.027	0.022	0.017	0.011	0.007	0.020***
	4	0.021	0.017	0.013	0.014	0.007	0.013***
	big	0.017	0.015	0.016	0.013	0.004	0.013**
	S-B	0.020**	0.019**	0.012	0.008	0.006	

Panel B Size × Neutral Gubapost portfolio

		Gubapost					
		low	2	3	4	High	L-H
Size	small	0.039	0.033	0.026	0.021	0.014	0.025***
	2	0.034	0.024	0.023	0.016	0.000	0.034***
	3	0.027	0.021	0.017	0.013	0.006	0.021***
	4	0.021	0.016	0.015	0.011	0.009	0.012**
	big	0.013	0.014	0.014	0.015	0.006	0.008
	S-B	0.026***	0.019**	0.012*	0.006	0.008	

Panel C Size × Positive Gubapost portfolio

		Gubapost					
		low	2	3	4	High	L-H
Size	small	0.038	0.035	0.026	0.020	0.017	0.021***
	2	0.033	0.027	0.021	0.016	0.000	0.034***
	3	0.025	0.026	0.016	0.010	0.008	0.017***
	4	0.021	0.015	0.016	0.012	0.008	0.014***
	big	0.013	0.014	0.013	0.013	0.006	0.006
	S-B	0.025***	0.021***	0.013*	0.007	0.010	

Table 10: Factor redundancy test for SMF on sentiment factors.

This table shows the results of regressing the returns of the long-short portfolio of *Gubapost* on sentiment factors: $SMF_t = \alpha + \beta Sentiment\ Factors_t + \lambda Control\ Factors + \epsilon_t$. A significant intercept in the regression indicates that the sentiment factors cannot fully explain the excess return of the portfolio. *SMF* is the dependent variable and *TurnoverF*, *VolumeF*, and *MAXF* are the independent variables. We also control for commonly used factors, the mass media factor, and the liquidity factor. t-values are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	(1)	(2)	(3)
TurnoverF	0.201	0.210*	0.228*
	(1.558)	(1.784)	(1.765)
VolumeF	0.436***	0.451***	0.417***
	(3.135)	(3.270)	(3.032)
MAXF	-0.010	-0.027	-0.031
	(-0.084)	(-0.261)	(-0.271)
Mkt	0.024	0.035	0.023
	(0.400)	(0.640)	(0.393)
SMB	0.215	0.098	0.245
	(1.278)	(0.580)	(1.277)
HML	-0.535***	-0.523***	-0.494***
	(-3.574)	(-3.524)	(-2.868)
UMD		0.117***	
		(2.808)	
RMW			-0.055
			(-0.303)
CMA			-0.149
			(-0.689)
MMF	0.103	0.183*	0.125
	(0.956)	(1.713)	(1.376)
LIQ	-0.354***	-0.297***	-0.347***
	(-4.720)	(-4.804)	(-4.546)
Intercept	0.006**	0.006**	0.006**
	(2.451)	(2.344)	(2.267)
R^2	0.928	0.934	0.929

Table 11: Monthly one-month-ahead return of Gubapost portfolio decile

This table reports the monthly returns of 10 portfolios sorted by *Gubapost* of the previous month and returns of two long-short portfolios of the low 1/10 - high 1/10 and low 1/3 - high 1/3. We make use of profitability measurements of Return (the portfolio excess over the risk-free rate), Alpha (the intercept of three, four, five, and PS liquidity factor regression: $r_p - r_f = \alpha + \beta Factors + \epsilon$), and Odds (the ratio of the portfolio's outperformance of the value-weighted market portfolio). The table also reports the risk measurements of Std (the standard deviation of return) and the Sharpe Ratio. t-values are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Decile	Return	Alpha(3)	Alpha(4)	Alpha(5)	Alpha (5+LIQ)	Odds	Std	Sharpe Ratio
D1(low)	0.0267	0.0117***	0.0118***	0.0120***	0.0131***	0.7077	0.0984	0.2713
D2	0.0214	0.0098***	0.0098***	0.0094***	0.0104***	0.7538	0.0895	0.2391
D3	0.0202	0.0097***	0.0097***	0.0098***	0.0103***	0.6615	0.0896	0.2248
D4	0.0176	0.0078***	0.0077***	0.0075***	0.0082***	0.5077	0.0903	0.1947
D5	0.0140	0.0060**	0.0060**	0.0059**	0.0067**	0.5846	0.0875	0.1597
D6	0.0149	0.0064*	0.0063*	0.0061*	0.0069**	0.5077	0.0849	0.1756
D7	0.0183	0.0114***	0.0112***	0.0112***	0.0107***	0.5538	0.0827	0.2211
D8	0.0096	0.0023	0.0023	0.0027	0.0030	0.5231	0.0802	0.1200
D9	0.0093	0.0030*	0.0029*	0.0028	0.0024	0.4154	0.0747	0.1244
D10(high)	0.0025	-0.0034	-0.0034	-0.0030	-0.0036	0.2923	0.0768	0.0325
D1-D10	0.0242**	0.0150***	0.0152***	0.0150**	0.0166***	0.6308	0.0872	0.2777
	(2.2387)	(2.804)	(2.880)	(2.653)	(3.065)			
Tercile1-Tercile3	0.0154**	0.0099***	0.0100***	0.0096***	0.0108***	0.5846	0.0636	0.2428
	(1.9573)	(2.919)	(2.943)	(2.689)	(3.240)			

Table 12: Trading profits grouped by firm characteristics

This table examines the profitability of a social media effect trading strategy in sub-samples of stocks sorted by various firm characteristics. Each month, stocks are sorted into 10 *Gubapost* portfolios, and the portfolio goes the long position on the lowest 10% and a short position on the highest 10% of *Gubapost*. The long and short legs of the portfolio invest an equal amount in each underlying stock, and portfolio weights are rebalanced monthly. We report alphas from regressing the resulting time series of long-short portfolio returns on the market factor, the Fama-French three-factor, Carhart four-factor, Fama-French five-factor, and Pastor-Stambaugh liquidity factor models. *t*-values are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

FF 3	Carhart 4	FF 5	LIQ
Panel A: By Firm Size			
Small			
0.025***	0.025***	0.022***	0.023***
(4.265)	(4.267)	(3.683)	(3.828)
Medium			
0.023***	0.023***	0.020***	0.021***
(3.953)	(3.948)	(3.428)	(3.437)
Big			
0.012***	0.012***	0.011**	0.012***
(2.689)	(2.672)	(2.463)	(2.770)
Panel B: By Book-to-Market			
Low			
0.019***	0.019***	0.018***	0.019***
(3.486)	(3.499)	(3.151)	(3.540)
Medium			
0.017***	0.017***	0.014**	0.016***
(2.965)	(3.009)	(2.485)	(2.855)
High			
0.010*	0.010*	0.009*	0.010*
(1.933)	(1.934)	(1.755)	(1.851)
Panel C: By Past 12-Month Momentum			
Low			
0.011**	0.011**	0.011**	0.011**
(2.304)	(2.310)	(2.111)	(2.205)
Medium			
0.012*	0.012*	0.011*	0.012*
(1.952)	(1.938)	(1.784)	(1.858)
High			
0.021***	0.021***	0.020**	0.022***
(2.897)	(2.869)	(2.653)	(2.837)

Table 13: Trading profits grouped by liquidity and sentiment characteristics

This table examines the profitability of a social media effect trading strategy in sub-samples of firms sorted by various liquidity and sentiment proxies. Each month, stocks are sorted into 10 *Gubapost* portfolios, and the portfolio goes long on the lowest 10% and short on the highest 10% of *Gubapost*. The long and short legs of the portfolio invest an equal amount in each underlying stock, and portfolio weights are rebalanced monthly. The reported numbers are alphas obtained by regressing the resulting time series of zero-investment portfolio returns on the market factor, Fama-French three-factor, Carhart four-factor, Fama-French five-factor, and Pastor-Stambaugh liquidity factor models. Note that in Panel B, $Dollar\ Trading\ Volume$ is a proxy of both liquidity and sentiment. t-values are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

FF 3	Carhart 4	FF 5	LIQ
Panel A: By Amihud's Illiquidity Ratio			
Low			
0.012**	0.012**	0.010*	0.012**
(2.203)	(2.184)	(1.822)	(2.061)
Medium			
0.015***	0.016***	0.012**	0.014***
(2.995)	(3.000)	(2.407)	(2.739)
High			
0.017***	0.017***	0.016***	0.016**
(3.028)	(3.020)	(2.720)	(2.654)
Panel B: By Dollar Trading Volume			
Low			
0.015***	0.015***	0.014***	0.016***
(3.108)	(3.238)	(2.845)	(3.285)
Medium			
0.013**	0.013**	0.011*	0.012**
(2.306)	(2.331)	(1.892)	(2.147)
High			
0.014**	0.014**	0.013**	0.014**
(2.518)	(2.493)	(2.215)	(2.426)
Panel C: By Turnover			
Low			
0.011**	0.011**	0.012**	0.013***
(2.265)	(2.247)	(2.439)	(2.679)
Medium			
0.014**	0.014**	0.011*	0.012**
(2.437)	(2.652)	(1.911)	(2.022)
High			
0.025***	0.025***	0.022***	0.024***
(3.179)	(3.280)	(2.798)	(2.955)
Panel D: By R_{max}			
Low			
0.012***	0.012***	0.013***	0.013***
(2.731)	(2.709)	(2.886)	(2.890)
Medium			
0.011*	0.011*	0.011	0.012*
(1.761)	(1.756)	(1.645)	(1.887)
High			
0.017**	0.018**	0.014*	0.014**
(2.467)	(2.450)	(1.982)	(2.031)

Table 14: Annual performance of social media effect strategy

This table reports the annual performance of the social media effect strategy. “Mean num of Stock” is the rounded average number of stock trades in the strategy. Five annualized returns in excess of the risk-free rate, VW market portfolio return, CSI 300, 500, 1000 (big, medium and small size) index returns are reported in the followed rows. “Odds” is the winning ratio by which the portfolio outperforms the VW market portfolio. “Sharpe Ratio” is the Sharpe ratio of the strategy. *t*-values are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	buy-and-hold portfolio	long-short portfolio
Mean num of Stocks	200	200 long & 200 short
Annualizd Return	0.3966*** (6.8972)	0.3219*** (8.5848)
Excess Return relevant to VW-Market Portfolio	0.2265*** (9.0259)	0.1518** (2.0469)
Excess Return relevant to CSI 300 Index Return	0.2734*** (8.7269)	0.1987*** (2.7099)
Excess Return relevant to CSI 500 Index Return	0.1950*** (7.0832)	0.1203* (1.7039)
Excess Return relevant to CSI 1000 Index Return	0.1290*** (4.3559)	0.0544 (0.8066)
Odds	0.9074	0.7593
Sharpe Ratio	0.9386	1.1682

Table 15: Factor redundant test for SMF on q-factors.

This table shows the results of regression on the time-series returns on the long-short portfolio of *Gubapost* and profitability and investment risk q-factors. The significant coefficient of the intercept indicates that the profitability and investment risk factors cannot explain the excess return of the portfolio. The regression uses *SMF* as the dependent variable and q-factors as the independent variables, constructed from Hou et al. (2015). The regression is as follows: $SMF_t = \alpha + \beta Factors_t + \epsilon_t$. t-values are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	(1)	(2)	(3)
Mkt	-0.040	-0.036	-0.008
	(-0.682)	(-0.653)	(-0.163)
q_size	0.416***	0.429***	0.140
	(3.290)	(3.592)	(1.095)
q_inv	0.074	-0.003	-0.118
	(0.434)	(-0.021)	(-0.803)
q_prof	-0.168	-0.314	-0.003
	(-0.875)	(-1.665)	(-0.014)
LIQ		-0.396***	-0.328**
		(-2.872)	(-2.645)
MMF			0.473***
			(4.043)
Intercept	0.013***	0.014***	0.009**
	(3.210)	(3.780)	(2.589)
R^2	0.846	0.865	0.895

Table 16: Factor redundant test on STREV and parsimonious model test.

Column (1) of this table shows the results of regressing the returns of the long-short portfolio of *Gubapost* on sentiment factors: $SMF_t = \alpha+\beta STREV\ Factors_t+\lambda Control\ Factors+\epsilon_t$. A significant intercept in the regression indicates that the short-term reversal effect cannot fully explain the excess return of the portfolio. *SMF* is the dependent variable and *STREV* is the independent variables constructed by one-month stock returns . We also control for commonly used factors, the mass media factor, and the liquidity factor. Column (2) employ three kinds of parsimonious model combined significant factors as robustness check. t-values are in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

		(1)			(2)	
STERV	-0.030	-0.038	-0.029		0.011	-0.030
	(-0.540)	(-0.712)	(-0.521)		(0.188)	(-0.545)
Mkt	-0.042	-0.028	-0.038	-0.021	-0.081*	-0.037
	(-1.020)	(-0.693)	(-0.868)	(-0.521)	(-1.676)	(-0.851)
SMB	0.281**	0.163	0.338*	0.197	0.584***	0.333*
	(2.055)	(1.116)	(1.951)	(1.427)	(3.408)	(1.937)
HML	-0.575***	-0.569***	-0.501***	-0.548***	-0.730***	-0.554***
	(-5.001)	(-5.073)	(-3.359)	(-5.088)	(-5.544)	(-4.526)
UMD		0.118*		0.115*		
		(1.959)		(1.915)		
RMW			0.037		0.077	0.100
			(0.170)		(0.346)	(0.509)
CMA			-0.137			
			(-0.633)			
LIQ	-0.343***	-0.280**	-0.329***	-0.301**		-0.335***
	(-2.888)	(-2.327)	(-2.707)	(-2.593)		(-2.772)
MMF	0.362***	0.454***	0.384***	0.418***		0.372***
	(3.075)	(3.655)	(3.141)	(3.706)		(3.097)
Intercept	0.010***	0.010***	0.010***	0.010***	0.009**	0.010***
	(3.390)	(3.422)	(3.039)	(3.375)	(2.620)	(3.095)
R^2	0.900	0.907	0.901	0.906	0.863	0.901

Table 17: Variable Definition

Name	Definition
Key Variables	
Gubapost	The number of postings in each stock's sub-forum on Guba.eastmoney.com, over the previous month.
SMF	The social media factor: return of a portfolio of 1/3 low Gubapost stocks minus the return of a portfolio of 1/3 high Gubapost stocks.
MMF	The mass media factor: return of a portfolio of no-news-coverage stocks minus the return of a portfolio of 1/2 high-news-coverage stocks.
Control Variables	
Size	Natural log of the average market capitalization of equity over the previous calendar year.
Book-to-Market ratio (B/M)	Natural log of the book value of equity divided by the market value of equity, as of the previous year end.
Liquidity	The inverse of Amihud's (2002) illiquidity ratio.
Volatility	The standard deviation of returns over the past three months.
Price	Average closing price during the previous month.
Institutional holding	Percentage of the stock's outstanding shares owned by institutions.
Momentum	Past 12-month return.
Amihud's (2002) illiquidity ratio	Stock's absolute return divided by its daily dollar trading volume.
Dollar trading volume	Daily value of trades in a stock, averaged over all days in a month.
Turnover	Daily turnover of a stock, measured by the ratio of trading volume on total liquidity capital and averaged over all days in a month.
R_{max}	Max daily return of a stock in the past month.
R_{t-1}	Monthly return of a stock in previous month.
Gross profitability	Gross profits-to-assets ratio over the previous calendar year.
Analyst coverage	The number of analyst reports of a stock in the past month.

Mkt	Value-weighted return of our sample minus return on 1/12 annualized 3-month SHIBOR.
SMB	Return of a portfolio of 30% small stocks minus the return of a portfolio of 30% large stocks.
HML	Return on a portfolio of stocks with 1/2 high book-to-market ratio, minus return on a portfolio of stocks with 1/2 low book-to-market ratio.
UMD	Return on a portfolio of stocks with a high past 12-month return, minus the return on a portfolio of stocks with a low past 12-month return.
RMW	Profitability factor followed the construction of Fama and French (2015).
CMA	Investment factor followed the construction of Fama and French (2015).
LIQ	Traded liquidity factor followed the construction of Pastor and Stambaugh (2003).
TurnoverF	A sentiment factor: return of a portfolio of 1/3 low turnover stocks minus the return of a portfolio of 1/3 high turnover stocks.
VolumeF	A sentiment factor: return of a portfolio of 1/3 low volume stocks minus the return of a portfolio of 1/3 high volume stocks.
MAXF	A sentiment factor: return of a portfolio of 1/3 low R_{max} stocks minus the return of a portfolio of 1/3 high R_{max} stocks.
q_size	Size factor of the q-factor model followed the construction of Hou et al. (2015).
q_inv	Investment factor of the q-factor model followed the construction of Hou et al. (2015).
q_prof	Profitability factor of the q-factor model followed the construction of Hou et al. (2015).
STREV	Short-term reversal factor constructed by one-month stock returns followed Han et al. (2019).