Stock Pricing with Textual Investor Sentiment: Evidence from Chinese Stock Markets

Ziwei Li[1] and Yuan Wu[2,*]

[1]*Department of Economics, University of Ulsan. 93 Daehak-Ro, Nam-Gu, Ulsan, Korea.*

[2]*Department of Financial Management, Guangdong Baiyun University, Guangzhou, 510450, China.*

Abstract: With the application of big data, attention has been paid in recent years to the role of unstructured data such as investors' language for asset pricing. We use the most comprehensive online stock bar text data to measure investor sentiment and use it to extend and modify the Chinese three-factor model. The modified four-factor model is subjected to Fama-Macbeth regression tests and GRS regression tests, and the following results are obtained: (1) At least half of the text-based proxies for investor sentiment are significant in the regressions of 25 portfolios. (2) In the Fama-Macbeth regression test, E/P is found to be more significant than B/M. (3) In the GRS regression test, the modified four-factor model is found to pass the GRS test for Size and investor sentiment dimensions. Therefore, we conclude that (1) The textual investor sentiment can be a useful factor in China. (2) E/P is more suitable as a value factor than B/M. (3) The textual investor sentiment factor improves the multifactor model's explanatory strength for stock portfolios' excess returns and is only valid for extreme portfolios.

Keywords: Asset pricing; CAPM; Chinese three-factor model; investor sentiment; textual investor sentiment

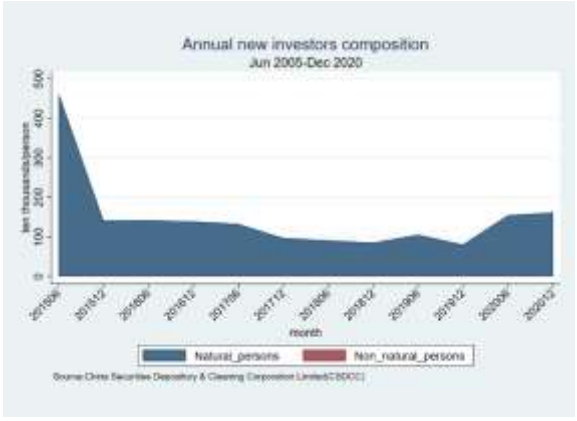
1. INTRODUCTION

The Capital Asset Pricing Model (CAPM), a theoretical foundation of contemporary financial economics, has been widely used in market investment decision-making, corporate investment, and financial management. Its primary purpose is to study the interrelationship between expected asset returns and risky assets in the stock market. Two essential assumptions underpin this theory: the efficient market hypothesis (EMH) and the investors' rationale and derive a linear relationship between the expected rate of return on securities and systematic risk. The theory provides investors with a way to analyze the risk of securities markets and use risk control to maximize investment returns. The core of the capital asset pricing model is to price only systematic risk in a portfolio on a risk-free return basis while not compensating for unsystematic risk market (Lintner, 1965a; Lintner, 1965b; Mossin, 1966; Sharpe, 1964).

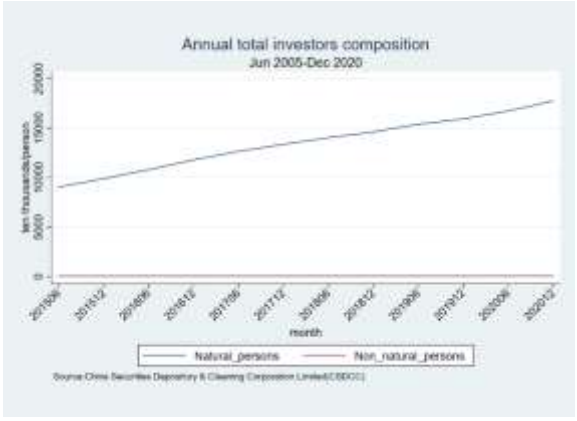
The empirical findings of the CAPM model are not satisfactory, and only a few empirical results support the theoretical model. For example, Black et al. (1972) confirmed the positive relationship between the beta coefficient and stock expected return using pre-1969 data. Fama and MacBeth (1973) demonstrated the positive relationship between the beta coefficient and stock anticipated return on the one hand. Further, the relationship between a beta squared term and unsystematic risk, and the results show a linear relationship between stock expected return and systematic risk. However, the mean value of the intercept term is more significant than the risk-free returns. Other than this, the results of different empirical tests fail to support the capital asset pricing model.

Most empirical research results cannot support the capital asset pricing model. The fundamental reason is that the two basic assumptions of the theoretical model, investors are entirely rational, and the market is completely efficient, which is inconsistent with reality. Investors are far less rational than we would like to believe. Research in psychology has shown that human decision-making is prone to bias and systematic error, leading to some bad choices without knowing them. Moreover, emotions play a much more critical role in human decision-making than we think, and it is easy to create the illusion of rational decision-making. In addition, neuroscience has confirmed that every thought that enters the human brain is assigned an emotional value, allowing humans to immediately give value to it based on the most critical emotional responses: friend or foe, war or flight, disgust, or anticipation. This branched mechanism in the brain is very effective but cannot be described as a rational thought process. Still, in theory, investors can never think entirely rationally (LeDoux, 2012). Similarly, it is impossible to find a completely efficient market, and the market is full fill with a large amount of asymmetric information. An efficient market with complete information does not exist, and it is possible to beat the market. Therefore, it is crucial to consider the irrationality of investors, and it is also essential to consider the combination of behavioral finance and asset pricing (Grossman & Stiglitz, 1980).

*Address correspondence to this author at the Department of Financial Management, Guangdong Baiyun University, Guangzhou, 510450, China; E-mail: ryan_ng@foxmail.com



(a)Annual new investors composition



(b)Annual total investors composition

Fig. (1). China Registered Investors Statistics from Jun 2005 to Dec 2020: (a) Annual new investors composition; (b) Annual total investors composition.

Behavioral finance theory came into being when the capital asset pricing model was questioned and challenged. Behavioral finance better explained the financial market anomalies such as the "equity premium puzzle" (Mehra & Prescott, 1985) and the "risk-free rate puzzle" (Weil, 1989), which the CAPM could not explain. Behavioral finance has two realistic theoretical bases: limited arbitrage and limited rationality of investors. Investors' bounded rationality is the expression of investors' psychological and behavioral factors, usually called investor sentiment. For a long time, investor sentiment has been the most important theoretical basis for explaining the financial market anomalies in behavioral finance. Investor sentiment is a good explanation for many anomalies that traditional finance cannot explain. Many behavioral finance studies have proved an inevitable link between investor sentiment and people's thinking judgment and investment behavior. Investor sentiment will have an impact on their specific investment decisions. More importantly, when these investment decisions involve risks and uncertainties, Investor sentiment is the most important determinant of stock market price and market trend changes. Investor sentiment impacts their irrational investment behavior (Brown & Cliff, 2004). This effect is more significant for individual investors with limited technology and information. Taking China as an example, the number of individual investors in the stock market has already exceeded 90%, and the proportion of circulating shares held by individual investors (excluding corporate claims) has long been in place (see Fig. **1**). Therefore, the importance of retail investors cannot be underestimated in the Chinese market, as they can influence the market's direction. This phenomenon differs significantly from developed markets such as Europe and the United States.

The existing papers on asset pricing theory and empirical research based on investor sentiment are mainly from two aspects. The first is to construct an asset pricing model including investor sentiment factors using various methods to build sentiment indexes, including questionnaire method, single market proxy index, Baker and Wurgler index (Baker & Wurgler, 2006), and composite sentiment index improved based on BW index. Then empirically test the impact of adding the investor sentiment factor to the asset pricing model. The second is to explore the effects of different sentiment measurement methods on asset pricing models and prove the ability of each other sentiment measurement methods to capture financial market anomalies such as scale effect, value effect, liquidity effect, and momentum effect. The core of these two aspects of research is to improve the pricing efficiency of asset pricing models based on investor sentiment. We will adopt the first method, which takes the capital asset pricing model as the basic theory, calculates the textual investor sentiment index, and adds it as a factor to the Chinese three-factor model (Liu et al., 2019). We empirically discuss the efficiency of asset pricing with the textual investor sentiment factor and analyze and test the significant role of textual investor sentiment in improving the efficiency of capital asset pricing.

How to construct an accurate investor sentiment index is the key to the current research on investor sentiment. Although data at the broad level provide suggestive evidence, it is challenging to measure investor sentiment directly, and all existing methods to gauge investor sentiment are proxy methods (Zhou, 2018). For example, the BW index, a composite investor sentiment index, based on multiple market variables to represent investor sentiment. Although this index is the most representative indicator in the method based on market variables, all market variables are only proxy variables. With the emergence of big data and data mining technology, data mining of investor sentiment based on Internet platforms (such as online message boards, Twitter, and other social media) has become a hot topic in recent years. The use of unstructured text data to construct investor sentiment is the most accurate measurement method because these texts are written by investors and represent their thoughts so that they can measure investor sentiment more directly and accurately. And a specific online message board has an evident tendency for a single emotion. It isn't easy to reflect the broad overview of investment sentiment objectively and comprehensively, so we choose all the existing online message boards of stocks in China as the data source of the textual investor sentiment index. At the same time, considering that many individual investors have a neutral attitude in their online comments, although unbiased comments cannot reflect the differences of opinion of investors, they still represent the attention to the stock. Therefore, we integrate Ant-

weiler and Frank's bullish index (Antweiler & Frank, 2004) and neutral comment to calculate the BU's textual investor sentiment index (BU et al., 2018). To the best of our knowledge, we are the first to use such comprehensive Chinese stock message board data to measure a textual investor sentiment index.

After calculating the textual investor sentiment index, we substitute it as a factor into the three-factor pricing model (Liu et al., 2019), commonly known as the Chinese three-factor model, and empirically test the revised capital asset pricing model, including textual investor sentiment factors. We select individual stocks in non-financial industries in China's main-board market (Shanghai and Shenzhen Stock Market, A-share) and second-board market (Growth Enterprise Market, GEM) as objects. To investigate the significantly different role of textual investor sentiment in various portfolios. In the revised capital asset pricing model, we do not exclude the stocks ranked in the bottom 30% of market capitalization, which is also the difference between the treatment of the sample pool in this paper and the Chinese three-factor pricing model.

Due to the lengthy application period for listing, Chinese enterprises often adopt reverse mergers of listed companies with low market value to achieve the purpose of listing, so the code of listed companies becomes a scarce "shell resource." To avoid the impact of "shell resources," the Chinese three-factor model deletes the 30% of stocks with the smallest market value in the Chinese three-factor model. However, we believe this data processing method may magnify the significant effect of the value factor variable, earning-to-price (E/P), in the Chinese three-factor model. And weaken the power of the value factor variable, book-to-market (B/M), in the Fama and French three-factor pricing model. The reason is that B/M benefits small companies more than large companies (Asness et al., 2015). In addition, China introduced a company registration system in 2015 and added the Science and Technology Innovation Board for SEMs to the list in 2018, which significantly shortened the application period for Chinese enterprises. Therefore, we use all non-financial stocks in the main-board and second-board markets to construct different portfolios and use Fama-Macbeth cross-sectional regression to compare the power of E/P and B/M as two value factor variables. The result supports the conclusion that E/P is more suitable as a value factor variable than B/M.

We record some critical conclusions. First, after constructing a capital asset pricing model including textual investor sentiment, it is found that this model can explain the expected excess returns under different portfolios to a large extent. Second, the model's market, size, and value factors significantly positively affect the expected excess returns of other portfolios. However, textual investor sentiment is negatively related to the expected excess return of the portfolio. This result means that stocks listed for more than six months that individual investors are bullish on are giving them unsatisfactory returns. Finally, textual investor sentiment seems to have a significant effect only on the highest or lowest value portfolios but not on the portfolios in non-extreme positions. We believe that these findings help explain the frequent yield booms and busts among retail investors in China.

The rest of the paper is as follows. Section 2 literature review. Section 3 introduces materials and methods. Section 4 provides the empirical analyses of the modified four-factor model. Section 5 discussion. Section 6 concludes.

2. LITERATURE REVIEW

2.1. Asset Pricing Models Under Neoclassical Economics

Based on the efficient market hypothesis (Fama, 1965, 1970), the portfolio theory and established a complete mean-variance analysis framework and investor behavior paradigm (Markowitz, 1952). The approach clearly defined the two concepts of return and risk. The stock price in the portfolio is a random variable; the mean represents the return, and the variance is the risk. Investors seeking utility maximization decide the proportion of investment in various securities to maximize investment returns and minimize risk. The securities portfolios are less risky than individual securities. Markowitz's pioneering work made it possible to delve into the theory of stock pricing. Since then, the focus of financial economists has evolved into how to match risk and return.

Under the assumption that all investors have the same utility function, the CAPM (Fama, 1965; Lintner, 1965a; Lintner, 1965b) proves that the return rate of the security portfolio is mean-variance efficient, the firm-specific risk of the security is fully diversified, and only the systematic risk cannot be diversified. Therefore, the CAPM believes that portfolio risk (beta coefficient, that is, the regression coefficient of portfolio returns rate and market return rate) can explain the expected return rate of a portfolio, and there is a positive correlation between the two,

$$E(R_i) = R_f + \beta[E(R_m - R_f)] \qquad (1)$$

Where $\beta = \frac{cov(R_i, R_m)}{Var(R_m)}$; $E(R_i)$ represents the expected return rate of individual stocks; R_f represents the risk-free interest rate; $E(R_m)$ represents the expected rate of return of the whole market.

However, the assumptions of CAPM are rigorous: first, investors are entirely rational and strictly follow Markowitz's portfolio model for diversification and select market portfolios at a certain point on the efficient market boundary. Second, capital markets are efficient, with no frictions-transaction costs, and taxes that discourage investment. The two assumptions are too strict about realizing.

The intertemporal Capital Asset Pricing Model (ICAPM) is based on the CAPM (Merton, 1973). Using the utility maximization to obtain accurate multi-factor forecasts of expected securities returns. While CAPM assumes that investors choose portfolios that will generate returns in the future, in the ICAPM, investors must consider not only their end-of-period returns but also their opportunities to consume or invest in the returns. Thus, when choosing a portfolio at time t-1, the investors consider how their wealth at time t varies with future variables, including labor income, consumer goods prices, and the nature of portfolio opportunities at time t.

The CAPM model with only one source of risk for the return on assets -market risk -is oversimplified. Different from it, the arbitrage pricing theory (APT) (Ross, 1976) is an im-

provement of CAPM, and its assumptions are not as strict as it is. It no longer assumes that all investors carry out the same optimization process and there is a unique risk portfolio. That is, the market portfolio. In APT, the return rate of assets has multiple risk sources, which are represented by factors. Therefore, APT has a stronger ability to explain reality than CAPM.

On the basis of ICAPM and APT, the consumption-based capital asset pricing model (CCAPM) (Breeden, 1979) considers the investment behavior in the case of intertemporal consumption. By finding an equilibrium under a two-period consumption decision, the expected price of the asset can be determined given the investor's consumption preference in each period and the asset income in the next period.

Based on the production-based Capital Asset Pricing model (PCAPM), a time-continuous generalized equilibrium model for a simple and complete economy emerged and was used to test the behavior of asset prices (Cox et al., 1985). At the same time, this paper is one of the milestones in the generalized equilibrium theoretical approach to finance.

A monetary-based CAPM (MCAPM), finds that the amount of currency issued affects only the nominal price of an asset but not the real price of the asset, while the speed of currency issuance affects the nominal pricing of the asset (Lucas & Stokey, 1987).

But if we analyze asset prices from the perspective of corporate finance with the liquidity-based CAPM (LCAPM), which means not only do consumers make choices about assets, but also firms make choices about assets, and firms make choices about assets for their own liquidity. In this way we transform the question of how money affects asset pricing into a question of how liquidity affects asset prices (Holmström & Tirole, 2001).

No matter whether CAPM with too strict assumptions, APT with perfect form but complicated calculation, or other traditional asset pricing models developed based on the two, they have not achieved good results in empirical research. There are many "anomalies" that cannot be explained by the capital asset pricing model (Fama & French, 1996), among which the more famous ones are the "equity premium puzzle" (Mehra & Prescott, 1985) and the "risk-free return puzzle" (Weil, 1989). The so-called "equity premium puzzle" refers to the fact that used the data from 1889 to 1978 in 1979 to prove that the US stock market had a phenomenon of very high stock market return and low Treasury bonds at the same time. The equity premium is always maintained at about 6%, which is too high. Since then, many scholars have explained the "equity premium puzzle" from various perspectives and have made significant achievements (Epstein & Zin, 1989, 1991; John Y. Campbell & John H. Cochrane, 1999). The most influential is Fama and French (1992). They conducted cross-sectional regression on the stock samples of NYSE, AMEX, and NASDAQ markets from 1963 to 1990, and found that no significant relationship was found between the single variable test and the multivariate joint test, while the size factor (ME) and the book-to-market ratio factor (BE/ME) had strong explanatory power. Subsequently, Fama and French (1993) added these two factors to the CAPM model, which greatly increased the explanatory power of the CAPM model, and this model was also called the FF three-factor model,

$$E(R_{it}) - R_{ft} = \beta_i[E(R_{mt}) - R_{ft}] + s_i E(SMB_t) + h_i E(HML_t) \quad (2)$$

Fama and French (1993) cleverly separated company size from B/M by double independent ranking: first, the company's market value at the end of the first year of the selected period was ranked by quintile from small to large. Then the stocks in the same group are also sorted by quintile from small to large according to the year-end company B/M, and divided into five groups, so that a total of 5*5=25 groups of portfolios are obtained. Then, the average monthly return rate of each group is calculated as the return rate of the group, and the risk-free return rate is subtracted to obtain the excess return of the portfolio, that is, the explained variable. Repeat the above process in the second and third years..... In the last year, all explained variables $E(R_{it}) - R_{ft}$ are calculated by repeated grouping. At the meanwhile, the explanatory variables in the model, the three factors, are constructed as follows: the calculation of SMB and HML is like the grouping of explained variables, which is divided into six groups: two groups, small and big, according to market value; and three groups, high, middle, and low, according to book-to-market ratio, as shown in Table **1**.

Table 1. Independent Double Sorting in FF Three-Factor Model.

	High (30%)	**Middle (40%)**	**Low (30%)**
Small (50%)	SH	SM	SL
Big (50%)	BH	BM	BL

Then we can calculate the size factor and value factor, according to the grouping results above,

$$SMB = \frac{(SH+SM+SL)}{3} - \frac{(BH+BM+BL)}{3} \quad (3)$$

$$HML = \frac{(SH+BH)}{2} - \frac{(SL+BL)}{2} \quad (4)$$

The market factor, $E(R_{mt}) - R_{ft}$, is the excess return of the market portfolio as in the CAPM model.

After the three factors are calculated according to the above method, the regression on the explanatory variables is performed to obtain a time series regression model divided into 25 groups. In addition, the factor loadings β_i, s_i, and h_i are the slope coefficients formed in the time series regression of the three-factor model. The core idea of the FF three-factor model is that the excess expected return of the portfolio relative to the risk-free return can be explained by the three factors.

Fama and French verified the data of the United States, Europe, Australia, and other countries, proving that the two factors of company size and B/M have strong explanatory power for stock returns, and the three-factor model can explain the cross-sectional anomalies of stocks represented by the B/M effect. The FF three-factor model extends the description of systemic risk of CAPM model to the industry level, which partially overcomes the limitation of CAPM,

but the residual value of non-systematic risk of portfolio is still too large.

The FF three-factor model became one of the most important models in empirical asset pricing, pioneering the logic of finding anomalies - adding factors - explaining anomalies, and was widely accepted by subsequent researchers. It was widely accepted by subsequent researchers. Since then, a steady stream of anomalies has been discovered and explained, which greatly enriched the models of asset pricing, including the momentum factor proposed by Jegadeesh and Titman (1993), and Carhart (1997) four-factor model inspired by it, the Novy-Marx (2013) four-factor model with the profitability factor; the Fama and French (2013) four-factor model with the profitability factor and the investment factor. factor and investment factor; the Fama and French (2015) five-factor model, which adds the earnings factor and investment factor; unlike the Fama-French five-factor model using the discount cash flow extension, the q-factor asset pricing model proposed by Hou et al. (2014) from the historical investment perspective of the firm model, also known as Investment-based CAPM; and the recently popular behavioral finance-based asset pricing (Daniel et al., 2019; Liu et al., 2019; Stambaugh & Yuan, 2016).

Because of the unique IPO process in the Chinese stock market, the FF three-factor model prevalent in the U.S. market has been poorly applied in China, and Liu et al. (2019) argue that a Chinese version of the three-factor model is thus proposed. The Chinese three-factor model can better explain most of the cross-sectional anomalies of Chinese market returns found in academia, and its explanatory strength is much stronger than that of the FF three-factor model. The Chinese three-factor model as follows.

$$E(R_{it}) - R_{ft} = \beta_i[E(R_{mt}) - R_{ft}] + s_i E(SMB_t) + h_i E(VMG_t) \quad (5)$$

Where the value factor, VMG_t, is calculated after using E/P grouping in the same way as the value factor in the FF three-factor model.

In summary, the "anomalies" of the CAPM stem largely from the fact that the core assumptions of the CAPM - market efficiency and investor rationality - do not correspond to reality. Subsequent scholars have made many efforts to relax the assumptions, but none of them has yielded satisfactory empirical results. This situation was broken until Fama-French's three-factor model, which made FF three-factor model the most important empirical asset pricing model. Although FF three-factor model is widely applied in Europe and the United States, it does not work well in the Chinese market. In contrast, the Chinese three-factor model based on behavioral finance is applied much better than the FF three-factor model in China. Therefore, our follow-up study is based on the Chinese three-factor model.

2.2. Investor Sentiment

Although the research on investor sentiment has gone through three decades, there are various definitions of investor sentiment. There are two main definitions of Investors Sentiment in academic research :(1) the degree to which noise traders' expectations of future stock prices deviate from the beliefs of rational arbitrageurs (Long et al., 1990). (2) A belief formed by investors based on the expectation of future cash flow and investment risk of assets (Baker & Wurgler, 2006). In general, the specific expression of investors' mentality and behavior is investor sentiment.

The measurement methods of investor sentiment indicators mainly include the following three categories: (1) Indirect measures based on market indicators are used to represent investor sentiment. For example, using trading volume (Hou et al., 2009; Kumar & Lee, 2006), stock rise and fall (Arms Jr, 1989), closed-end fund discount (Lee et al., 1991) and other market-specific indicators to approximately measure investor sentiment. In addition, Baker and Wurgler (2006) used a combination of six indicators to capture investor sentiment: closed-end fund discount, trading volume, number of IPOs, first-day IPO returns, dividend yield, and security issuance. The BW index is currently the most widely used indirect measure of investor sentiment. (2) The direct method of the survey on investor cognition. For example, the US market typically uses two standard investor sentiment survey indices, the AAII survey constructed by the American Association of Individual Investors and the II (Investor Intelligence) Survey. These two indexes are generally considered as indicators of institutional investor sentiment (Lee et al., 2002). Two widely accepted investment sentiment indexes in the Chinese market are CCTV Observation and JUCHAO Investor Confidence Index. However, the above indicators all have certain defects. Indicators based on trading behavior are the equilibrium outcome of multiple market forces and reflect more than investor sentiment. Although subjective indicators can reflect the emotions of respondents when filling out questionnaires, they cannot fully reflect the emotions of investors in the investment process. Moreover, most institutions produce indices that are short lived or even no longer updated. With the development of the Internet and deep learning, investor sentiment based on text big data has received increasing attention. (3) Measuring investor sentiment based on news websites, social media, Internet stock message boards and other online platforms. For example, Dougal et al. (2012), Ahern and Sosyura (2014), Kelley and Tetlock (2017), measured investor sentiment through the lexical information of news websites; Antweiler and Frank (2004), Das (2007) and Chen et al. (2007) both mined information texts from Yahoo and other stock message boards to construct investor sentiment indicators. Meanwhile, Chen et al. (2014), Sprenger et al. (2014), and Bartov et al. (2018) constructed Twitter investor indicators by collecting posts published on Twitter. Zhou (2018) Compared with the calculation methods based on market and survey, it is surprising that the investor sentiment calculated based on text analysis performs best so far. This may indicate that the stock market is likely to ignore the latter information. Of course, more research is needed to prove this point and reveal more reasons why market data-based measures underperform textual data.

Bullish index proposed by Antweiler and Frank (2004) is currently the most widely used text sentiment indicator. They constructed a bullish sentiment indicator using information from Yahoo! Finance and Raging Bull stock message boards to construct bullish sentiment index, a total of three measures are proposed, as follows.

$$B_t^* \equiv \ln\left[\frac{1+M_t^{BUY}}{1+M_t^{SELL}}\right] = \ln\left[\frac{2+M_t(1+B_t)}{2+M_t(1-B_t)}\right] \approx B_t \ln(1+M_t) \quad (6)$$

Where M_t^{BUY}, M_t^{SELL} are the number of texts supported for purchase or sale, respectively; $M_t = M_t^{BUY} + M_t^{SELL}$; and $B_t \equiv \frac{M_t^{BUY} - M_t^{SELL}}{M_t^{BUY} + M_t^{SELL}}$, the value between -1 and +1.

In addition, Antweiler and Frank (2004) also found that there is a significant but negative contemporaneous correlation between the amount of information and stock returns. Although the impact of stock information on stock returns is statistically significant, the effect is negligible in the economic sense. This conclusion is consistent with Harris and Raviv (2015), that disagreement in published information is associated with an increase in trading volume.

BU et al. (2018) argue that the Bullish index does not take neutral message into account and that the neutral text is also an indication of the level of investor concern, which is also an indication of investor sentiment, so even if investors express neutral expectations, this information is still valuable. In view of this, BU et al. (2018) revised the Bullish index ,

$$B_t^{ATT} \approx B_t \ln(1 + M_t^{BUY} + M_t^{SELL} + M_t^{NEUTRAL}) \quad (7)$$

The empirical results of BU et al. (2018) reveal that: while investor sentiment has no predictive power on stock market returns, trading volume, and volatility, investor sentiment has a current impact on stock returns and trading volume; stock commentary sentiment in the pre-opening non-trading session has predictive power on the opening price, and stock commentary sentiment in the post-opening trading session has a more significant impact.

2.3. Investor Sentiment and Asset Pricing

Whether investor sentiment affects asset pricing is a long-standing debate between behavioral finance and neoclassical finance. The noise trader model (DSSW), proposed by Long et al. (1990), is the basis of behavioral finance. The DSSW model assumes that there are both rational investors and noise traders in the market, and the trading behavior of noise traders is mainly influenced by market sentiment, which can not only deviate asset prices, but also be systematic and unpredictable. On this basis, Shefrin (2001) links investor sentiment with behavioral pricing and derives a behavioral asset pricing model(BAPM) that includes investor sentiment.

Like the BAPM based on DSSW noise traders, there is also an asset pricing model based on investors' cognitive biases. However, the two theoretical models differ in that investors in the BAPM are similarly divided into two categories, information traders and noise traders, which interact with each other to determine asset prices. Information traders are rational traders who act strictly according to the CAPM and do not suffer from systematic biases; it is worth noting that noise traders, as defined by investor cognitive biases, are different from DSSW, where noise traders refer to investors who do not act according to the CAPM and make various cognitive bias errors, so this type of model is also called investor cognitive bias-based asset pricing model.

In addition, most of the remaining asset pricing models based on investor sentiment are empirical models that expand on FF three-factor model or other multi-factor models.

3. MATERIALS AND METHODS

3.1. Data Collecting and Processing

To consider the impact of textual investor sentiment on individual stocks, we select all individual stocks in non-financial sectors in China's main-board market (Shanghai and Shenzhen Stock market, A-share) and second-board market (growth enterprise market, GEM) as portfolio representatives to investigate the differential impact of textual investor sentiment on different portfolios of individual stocks.

We obtain financial and market structural data from the China Stock Market and Accounting Research Database (CSMAR) from 2008 to 2020, specifically monthly individual stock returns, individual stock market capitalization (ME), individual stock Earning-to-Price (E/P), individual stock Book-to-Market (B/M), and stock market yields, and risk-free rates.

(1) Monthly individual stock returns. There are usually two ways to calculate monthly stock returns: one is to consider cash dividends reinvested, and the other is to disregard cash dividends. Here we choose the former method, which is calculated as follows.

$$r_{n,t} = \frac{P_{n,t}}{P_{n,t-1}} - 1 \quad (8)$$

Where $P_{n,t}$ is the comparable price of the daily closing price of stock n on the last trading day of the month t considering the reinvestment of cash dividends; $P_{n,t-1}$ is the comparable price of the daily closing price of stock on the last trading day of the month $t-1$ considering the reinvestment of cash dividends.

(2) Stock Market Capitalization (ME). The product of the number of outstanding shares of stock and its monthly closing price.

(3) Earning-to-Price ratio (E/P). Use the inverse of the Price-to-Earning (P/E), where Price is the daily closing price of stock on the last trading day of the month t; Earning, we use the ending value of the previous year' s net income.

(4) Book-to-Market ratio (B/M). B/M is the inverse of the Price-to-Book ratio and the ending value of the prior year's net assets is used in the calculation.

(5) Market return. The market return is the weighted average return of all stocks. Its weighting is calculated by the equal-weighted average method, the market capitalization-weighted average method, and the total market-weighted average method. We choose the market capitalization-weighted average method. The following equation calculates the market return.

$$R_{n,t} = \frac{\sum_n w_{n,t} r_{n,t}}{\sum_n w_{n,t}} \quad (9)$$

When $w_{n,t}$ denotes the market capitalization of stock outstanding in month $t-1$, then $R_{n,t}$ is the market capitalization-weighted average return on outstanding. Where $w_{n,t} = V_{n,t-1} * P_{n,t-1}$; $V_{n,t-1}$ is number of outstanding shares of stock n in month $t-1$; $P_{n,t-1}$ is closing price of stock n at the end of month $t-1$. $r_{n,t}$ denotes the return on individual stocks considering cash dividend reinvestment, while $R_{n,t}$ is the weighted average market capitalization return on market capitalization outstanding considering cash dividend reinvestment.

(6) Risk-free rate. We obtain this by dividing the one-year regular whole deposit rate by 12 months.

In addition, we collected text data from the Chinese listed company stock commentary database (GUBA) from 9 a.m. to 3 p.m. on trading days for calculating individual investor sentiment. The database makes judgments based on machine learning methods and counts the total number of corresponding sentiment posts. And the database includes statistics on public company postings on all Internet stock message boards since 2008.

We need to filter and screen the selected stocks further to construct the factors. The first filter excludes financial stocks, which we do not consider due to their significant differences from ordinary companies in terms of business models and accounting accounts; the second filter excludes stocks that have been listed for less than six months to avoid the impact of IPOs, and the third filter excludes data for stocks with less than 15 trading days in the previous month or less than 120 trading days in the past 12 months (excluding that month)—avoiding the special treatment of stocks.

3.2. The Modified Four-factor Model

The modified four-factor model incorporates the previously measured textual investor sentiment as a factor in the Chinese three-factor model. We place emphasis on textual investor sentiment for two reasons: first, retail investors dominate in absolute numbers in the Chinese capital market, but the information available to retail investors is often lagging or scarce. When information access is limited, retail investors may make wrong decisions for a long period of time (Hirshleifer & Teoh, 2003), which explains to some extent why Chinese stock markets often experience large fluctuations. Second, the Internet has become the main channel of information access for retail investors in China. According to a survey report released by the China Internet Information Center (CNNIC) on December 31, 2020, China ranked first in the world with 0.989 billion Internet users and a 70.4% Internet penetration rate. Through the Internet, Chinese retail investors can access information at low cost, in a timely and efficient manner, and the lag of information is greatly reduced. The emergence of social media, such as Internet stock postings, has brought together investors with the same information needs to exchange information and share opinions, and the interactive behavior among investors has led to an explosion of information. In addition, many listed companies are using the Internet to push information about company characteristics to investors, which reduces the time cost of information collection for investors and improves the pricing efficiency of the capital market. In the Internet era, the increase of investors' interactive behaviors and statements brought by the widespread use of social media will affect the stock market, including its pricing issues. Therefore, we believe that constructing an asset pricing model that incorporates textual investor sentiment can help to better understand this emerging capital market in China. The modified four-factor model is specified as follows.

$$E(R_{it}) - R_{ft} = \beta_i[E(R_{mt}) - R_{ft}] + s_i E(SMB_t) + h_i E(VMG_t) + p_i E(PMN_t) \quad (10)$$

Where PMN_t represents the monthly returns with positive textual tendencies minus monthly returns with negative textual tendencies (Positive minus Negative).

We independently double-sorting the stocks, such as market capitalization and value factor, market capitalization, and individual investor sentiment factor. We follow the sorting method from Fama and French. The Market factor, MKT, is consistent with the construction method of models such as the CAPM model, which is the market portfolio return minus the risk-free return. The size factor, SMB, uses the median market capitalization outstanding as the breakpoint at the end of t − 1 and divides all stocks into two groups, big and small, and use small minus big to construct SMB. The value factor, VMG, uses the 30% and 70% quartiles of a stock's E/P as breakpoints, and all stocks are divided into three groups: the first 30% (Value group), the middle 40% (Middle group), and the last 30% (Growth group). And value minus growth can get the VMG. The textual investor factor, PMN, uses the 30% and 70% quartiles of the BU index (B_t^{ATT}) as breakpoints, and all stocks are divided into three groups: the first 30% (Positive group), the middle 40% (Middle group), and the last 30% (Negative group). And positive minus negative achieves the PMN.

Eventually, we obtain twelve asset portfolios through the 2*3 independent double sorting method, according to two market cap groupings, three E/P groupings, and three sentiment groupings, which are equally weighted using market cap outstanding.

After constructing the factors, we averaged the groupings according to their quintiles after sorting them from smallest to largest according to the size and E/P dimensions to obtain 25 portfolios. We take the expected returns of the 25 portfolios as explanatory variables and obtain their factor coefficients by regressing the factors on the 25 portfolios. The significance of the factor coefficients allows us to determine whether the factors are helpful or not. Likewise, we ran the same regression on the 25 portfolios of the size AND BU index dimensions.

3.3. Testing Methodology

3.3.1. Gibbons-Ross-Shanken Regression Testing

The Gibbons-Ross-Shanken regression testing (Gibbons et al., 1989) is time-serial regression testing. It states that if the asset pricing model is valid, i.e., captures all the stock changes and changes in average returns. The intercept term of the regression model must be all equal to zero, and its

proposed GRS test statistic satisfying $T-N-K$ freedom and obeying the F distribution,

$$\text{GRS statistic} = \frac{T}{N} \times \frac{T-N-L}{T-L-1} \times \frac{\alpha' \epsilon^{-1} \alpha}{1+\mu' \Omega^{-1} \mu} \quad (11)$$

The statistic obeys a non-central F distribution with degrees of freedom N and $T-N-L$, where T is the number of units of time, i.e., 155 months; N is the number of portfolios on the left side of the equation of the regression equation, i.e., the 25 portfolios described previously; L is the number of factors; α is a matrix consisting of intercept terms with dimension (N, 1); $\in$ is a matrix consisting of the variance and covariance of the residual terms with dimension (N, N); μ is the average return matrix of the factors with dimension $(L, 1)$; Ω is a matrix consisting of the variance and covariance of μ with dimension (L, L).

The null hypothesis tested by the GRS statistic is $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_N = 0$. The smaller the GRS statistic, the lower the probability that the null hypothesis is rejected, which implies the stronger the explanatory power of the tested asset pricing model.

3.3.2. Fama-Macbeth Regression Testing

Fama-Macbeth regression testing (Fama & MacBeth, 1973) is widely used because it cleverly excludes the influence of residual correlation on the cross-sectional error. The following is a brief explanation of the Fama-MacBeth regression using a multifactor model as an example.

Under the framework of the multifactor model, the expected excess return rate of any risky asset i is entirely determined by K factors included in the scope of the investigation,

$$E(R_i) = \sum_K (\lambda_j) \beta_{i,j}, \forall i \quad (12)$$

Where R_i is the excess return on asset i. λ_j is the price of risk for factor K (state variable K). $\beta_{i,j}$ is estimated by the following regression, and this also called the first stage time-series regression,

$$R_{i,t} = a_i + \sum_K \left(\beta_{i,f^K}\right) f_t^K + \varepsilon_{i,t}, \forall i \quad (13)$$

Where f_t^K is the observed value of factor K at the end of period t.

In the second stage of cross-sectional regression, the model is set up as follows:

$$E[R_{i,t}] = \sum_K (\lambda_j) \hat{\beta}_{i,j}, j = 1, \ldots, K \quad (14)$$

Where $\hat{\beta}_{i,j}$ is the systematic risk corresponding to factor j obtained from the first-stage regression.

The estimated risk premium and standard deviation of factor F_j are:

$$\hat{\lambda}_j = \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}_t \quad (15)$$

$$\sigma^2(\hat{\lambda}_j) = \frac{1}{T^2} \sum_{t=1}^{T} (\hat{\lambda}_{j,t} - \hat{\lambda}_j)^2 \quad (16)$$

The pricing error and standard deviation of asset i are:

$$\hat{\varepsilon}_i = \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_{i,t}$$

$$\sigma^2(\hat{\varepsilon}_i) = \frac{1}{T^2} \sum_{t=1}^{T} (\hat{\varepsilon}_{i,t} - \hat{\varepsilon}_i)^2 \quad (18)$$

The greatest advantage of the Fama-MacBeth regression is that it excludes the effect of residual cross-sectional correlation on the standard errors. However, it does not do anything about the correlation of residuals in the time series. Common treatments are Shanken correction (Hansen, 1982; Shanken, 2015) and Newey-West adjustment (Newey & West, 1987) to remove the effect of heteroskedasticity and autocorrelation to resolve correlation on time series. We have chosen to use the Newey-West adjustment in the latter part of the paper.

4. EMPIRICAL ANALYSES OF THE MODIFIED FOUR-FACTOR MODEL

4.1. Summarize and Correlation Analysis

We have statistically described four factors in the modified four-factor model: market, size, value, and textual investor sentiment, as shown in Table **2**. We can see that the mean values of the four factors are 0.347, 0.894, -0.380, and 0.409, respectively. The monthly standard deviations of CHSMB and CHVMG are 3.031 and 2.541, respectively, both of which are about half and one-third of the MKT standard deviation of 7.844. And the three factors, MKT, CHVMG, PMN, have left long-tailed distributions. Moreover, all the factors have flatness distributions. And the last three columns count the minimum, 50th percentile, and maximum values.

Table 2. Summary Statistics for the Modified Four-factor.

Factor	count	mean	sd	skewness	kurtosis	min	p50	max
MKT	155	0.347	7.844	-0.589	4.660	-26.835	0.966	18.485
CHSMB	155	0.894	3.031	0.178	6.001	-11.596	0.921	12.244
CHVMG	155	-0.380	2.541	-0.226	4.757	-9.908	-0.321	8.546
PMN	155	0.409	1.633	-0.553	5.566	-6.684	0.515	4.213

Before performing correlation analysis among the four factors, the factors are first tested for stationarity on the time series, and the empirical results are shown in Table **3**. From the results, the factor's returns significantly reject the original hypothesis of the existence of a unit root in the time series, which indicates that the factor's returns are stationary in the time series, i.e., they pass the stationarity test. In response, a series of studies including correlation tests can be conducted using the factor data.

Table 3. Dickey-Fuller Test for Unit Root.

Factor	Test statistic	1% critical value	5% critical value	10% critical value	Stationary
MKT	-11.109	-4.022	-3.443	-3.143	YES
CHSMB	-11.966	-4.022	-3.443	-3.143	YES

CHVMG	-12.304	-4.022	-3.443	-3.143	YES
PMN	-12.405	-4.022	-3.443	-3.143	YES

On the basis of the four factors passing the stationarity test, Table **4** shows the matrix of correlation coefficients calculated between the four factors. Among them, the Pearson product moment correlation coefficients are shown at the bottom left of the main diagonal.

Table 4. Pearson Correlation.

Factor	MKT	CHSMB	CHVMG	PMN
MKT	1.000			
CHSMB	0.054	1.000		
CHVMG	0.185**	0.432***	1.000	
PMN	-0.168**	-0.193**	-0.163**	1.000

* p<0.1, ** p<0.05, *** p<0.01.

From the correlation coefficient matrix, it can be tentatively confirmed that there is a positive linear relationship between CHSMB and CHVMG and MKT. Higher market factor MKT means higher market excess return, higher size factor CHSMB means larger market capitalization, higher value factor CHVMG means more mature company. There is a significant negative linear relationship between PMN and the other three factors. It means that the higher the market factor, size factor and value factor, the lower the sentiment factor PMN means the more negative investor sentiment.

4.2. Empirical Regression and Analysis

We first report the average monthly excess returns of the 25 portfolio groups to visualize the differences in the explanatory power of the different factors. The trend of the average monthly excess returns of each portfolio can directly reflect the difference in the explanatory power of different factors. The results are shown in Table **5**.

Table 5. Average Monthly Excess Return of the Portfolios.

	Growth	2	3	4	Value	Value - Growth
Small	2.034	1.457	1.754	1.549	1.608	-0.426
						(-1.579)
2	0.871	1.036	1.180	1.216	1.298	0.427*
						(-1.933)
3	0.071	0.442	0.740	0.925	1.039	0.968***
						(-4.019)
4	0.050	0.274	0.433	0.726	0.641	0.591***
						(-2.779)
Big	0.021	0.175	0.242	0.360	0.556	0.535
						(-1.623)
Small - Big	2.012***	1.282**	1.512***	1.189**	1.051**	
	(-3.575)	(-2.273)	(-2.797)	(-2.291)	(-2.027)	

* p<0.1, ** p<0.05, *** p<0.01.

The change in portfolio excess returns makes it easy to see the different effects. First, the size effect is significant. The small-cap outperformers return more than portfolios of Big-cap outperformers. It is evident in all groups. Second, the value effect is also generally apparent, with portfolios of high-value companies generally returning more than portfolios of growth companies. Finally, the sentiment effect is significant in the smaller and larger portfolios.

Next, we regress the excess returns of 25 portfolios consisting of the dimension of size and E/P on the four factors according to the four-factor model. The regression results are shown in Table **6**.

Table 6. The Modified Four-factor Model Regression Results Based on Size and E/P.

	Growth	2	3	4	Value	Growth	2	3	4	Value
	a (intercept)					t(a)				
Small	0.786***	-0.177	-0.129	-0.273	-0.144	4.595	-0.712	-0.496	-1.120	-0.634
2	0.151	-0.015	0.041	-0.007	0.189	0.612	-0.063	0.193	-0.032	0.787
3	-0.395*	-0.194	-0.028	0.114	0.266	-1.696	-0.747	-0.104	0.501	1.160
4	-0.025	-0.321	-0.123	0.090	0.078	-0.100	-1.113	-0.431	0.364	0.339
Big	0.328	0.117	-0.021	0.054	0.254	1.310	0.526	-0.098	0.276	1.223
	b (MKT coefficient)					t(b)				
Small	1.066***	1.090***	0.997***	1.028***	1.078***	51.215	31.490	39.942	25.845	31.703
2	1.061***	1.059***	1.029***	1.003***	1.087***	38.006	30.866	31.180	22.146	39.533
3	1.077***	1.093***	1.043***	1.023***	1.084***	38.446	27.135	26.157	30.777	35.932
4	1.10***	1.089***	1.037***	1.037***	1.129***	27.779	28.789	26.497	32.388	33.980

Big	1.062***	0.949***	0.964***	1.083***	1.124***	35.276	32.608	25.949	40.912	41.138
	s (SMB coefficient)					t(s)				
Small	1.568***	1.698***	1.907***	1.758***	1.681***	24.547	12.436	17.357	16.556	19.626
2	0.986***	1.218***	1.231***	1.204***	1.087***	12.242	13.363	13.410	13.688	13.515
3	0.784***	0.793***	0.854***	0.880***	0.738***	10.636	6.997	8.619	9.284	9.872
4	0.325***	0.588***	0.585***	0.543***	0.385***	3.479	5.296	5.373	6.487	4.218
Big	-0.214**	-0.030	0.117	-0.029	-0.175**	-2.465	-0.326	1.504	-0.405	-2.197
	v (VMG coefficient)					t(v)				
Small	0.792***	0.651***	0.526***	0.289**	-0.006	10.973	9.542	6.293	2.228	-0.045
2	0.761***	0.751***	0.675***	0.307**	0.102	7.229	10.973	8.831	2.288	1.116
3	0.960***	0.813***	0.655***	0.437***	-0.042	9.524	5.862	5.063	4.286	-0.429
4	1.038***	0.787***	0.625***	0.209**	-0.121	9.029	6.239	5.247	1.998	-1.146
Big	0.968***	0.727***	0.285***	-0.174**	-0.632***	10.646	6.099	2.611	-2.208	-6.393
	p (PMN coefficient)					t(p)				
Small	-0.546***	-0.035	0.077	0.010	-0.312*	-4.735	-0.212	0.659	0.046	-1.758
2	-0.589***	-0.295**	-0.152	-0.253	-0.492***	-4.406	-2.232	-0.993	-1.394	-3.326
3	-0.598***	-0.351*	-0.263	-0.404**	-0.683***	-3.620	-1.741	-1.587	-2.039	-5.458
4	-0.494***	-0.023	-0.216	-0.320**	-0.536***	-4.344	-0.109	-1.161	-2.254	-3.662
Big	-0.284*	0.079	-0.165	-0.269**	-0.420***	-1.857	0.538	-0.916	-2.103	-4.390
	Adj-R^2					s(e)				
Small	0.965	0.930	0.926	0.920	0.941	2.105	3.003	2.980	3.002	2.559
2	0.942	0.949	0.932	0.924	0.938	2.480	2.364	2.662	2.764	2.465
3	0.943	0.915	0.914	0.919	0.927	2.487	3.031	2.897	2.726	2.579
4	0.934	0.896	0.891	0.910	0.934	2.628	3.242	2.897	2.717	2.432
Big	0.934	0.889	0.908	0.951	0.946	2.416	2.802	2.897	1.927	2.103

Table **6** includes the regression coefficients (β, s, v, p), intercept terms (α) and t values and adjusted goodness-of-fit ($Adj - R^2$) for the MKT, SMB, VMG and PMN.

The coefficient β of the MKT factor is all positive and strongly significant in 25 portfolios, and the value fluctuates in the range of 0.94-1.13. This result shows that the excess return of the stock portfolio changes in the same direction as the MKT factor, and proves that the MKT factor is highly significant.

The coefficient s of the SMB factor is positive in all but the largest size group. Its value is between -0.22 to 1.91. As the size increases, the coefficient s gets smaller and smaller. This result shows that the SMB factor has a more significant impact on the smaller portfolios, which can bring more excess returns. Except for three portfolios that are not significant, the remaining 22 portfolios are at least at the 5% significance level, indicating that the SMB factor is also an essential factor.

The coefficient v of the VMG factor is significant in 21 portfolios and the value is between -0.64 to 1.04. The results show that the VMG factor also has a more significant effect on lower E/P groups. And the VMG factor on the growth company is even better.

The coefficient p of the PMN factor is negative in most of the portfolios. 15 portfolios are significant, with values ranging from -0.69 to 0.08. The results suggest that the PMN factor logically moves in the opposite direction to the excess return of the stock portfolios. It is not difficult to find that the PMN factor does not have as much effect on the middle sentiment groups.

Finally, looking at $Adj - R^2$, the fit of the portfolio is about 0.92. The results indicate that the four-factor model has strong explanatory power for the excess return of the different stock portfolios.

Moreover, we also regress the excess returns of another 25 portfolios consisting of size and textual investor sentiment(BU index) on the four factors according to the four-factor model. The regression results are shown in Table **7**.

Table 7. The Modified Four-factor Model Regression Results Based on Size and BU Index.

	Negative	**2**	**3**	**4**	**Positive**	**Negative**	**2**	**3**	**4**	**Positive**
	a (intercept)					**t(a)**				
Small	0.003	-0.073	-0.102	0.118	0.152	0.013	-0.383	-0.443	0.567	0.663
2	0.046	-0.162	0.334	0.082	-0.037	0.207	-0.671	1.643	0.333	-0.177
3	-0.127	-0.008	0.054	-0.094	-0.030	-0.541	-0.030	0.193	-0.360	-0.122
4	0.015	-0.175	-0.001	-0.074	-0.052	0.074	-0.732	-0.004	-0.347	-0.197
Big	0.331	0.196	-0.096	-0.014	0.30	1.372	1.160	-0.515	-0.095	1.583
	b (MKT coefficient)					t(b)				
Small	1.043***	1.082***	1.050***	1.037***	1.043***	38.948	32.639	35.662	34.048	35.594
2	1.055***	1.071***	1.057***	1.003***	1.060***	39.957	27.172	36.855	33.231	28.056
3	1.059***	1.076***	1.082***	1.048***	1.041***	31.331	37.357	27.846	27.279	30.725
4	1.071***	1.087***	1.083***	1.103***	1.050***	31.763	30.229	29.883	28.924	27.679
Big	1.006***	1.042***	1.032***	1.072***	1.033***	40.369	46.555	42.728	54.200	55.779
	s (SMB coefficient)					t(s)				
Small	1.718***	1.626***	1.710***	1.795***	1.742***	21.085	16.032	18.734	12.721	19.936
2	1.049***	1.133***	1.140***	1.140***	1.311***	11.974	16.339	13.607	14.499	13.107
3	0.864***	0.762***	0.814***	0.816***	0.754***	9.968	7.616	6.684	10.636	8.317
4	0.464***	0.455***	0.474***	0.505***	0.521***	6.166	5.981	7.210	5.838	4.826
Big	0.048	-0.153**	-0.055	-0.118**	-0.058	0.708	-2.282	-0.735	-2.420	-0.831
	v (VMG coefficient)					t(v)				
Small	0.394***	0.475***	0.460***	0.534***	0.432***	4.857	5.469	4.590	6.030	4.419
2	0.557***	0.477***	0.499***	0.602***	0.460***	6.387	4.857	5.003	6.752	5.087
3	0.539***	0.522***	0.595***	0.533***	0.690***	5.536	6.195	4.523	4.266	7.157
4	0.432***	0.507***	0.516***	0.502***	0.589***	4.638	5.309	4.460	4.251	5.385
Big	0.30***	0.280***	0.175**	0.181***	0.220**	4.741	3.648	2.308	2.717	2.477
	Negative	2	3	4	Positive	Negative	2	3	4	Positive
	p (PMN coefficient)					t(p)				
Small	-0.591***	-0.346**	-0.192	0.055	0.284***	-3.907	-2.443	-1.187	0.307	2.729
2	-0.881***	-0.592***	-0.365**	-0.074	0.126	-5.047	-4.057	-2.420	-0.572	0.777
3	-0.876***	-0.699***	-0.462**	-0.192	-0.116	-5.525	-5.091	-2.485	-1.411	-0.705
4	-0.832***	-0.657***	-0.224	-0.048	0.152	-6.422	-4.382	-1.430	-0.280	0.816
Big	-0.992***	-0.511***	-0.206	0.156*	0.50***	-10.072	-5.689	-1.605	1.865	4.101
	Adj-R^2					s(e)				
Small	0.948	0.947	0.950	0.937	0.946	2.481	2.481	2.425	2.687	2.458
2	0.950	0.945	0.939	0.945	0.932	2.301	2.407	2.459	2.393	2.655
3	0.934	0.935	0.923	0.928	0.921	2.585	2.544	2.803	2.537	2.729

	Negative	2	3	4	Positive	Negative	2	3	4	Positive
4	0.924	0.932	0.930	0.917	0.907	2.672	2.541	2.803	2.788	2.855
Big	0.944	0.948	0.944	0.960	0.943	2.080	1.982	2.803	1.722	1.996

Table 8. The Fama-Macbeth regression testing.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variable	**R_Rf**	**R_Rf**	**R_Rf**	**R_Rf**	**R_Rf**	**R_Rf**	**R_Rf**	**R_Rf**
β	-0.684***		-0.672***	-0.659***	-0.665***	-0.660***	-0.668***	-0.655***
	(-4.998)		(-4.932)	(-4.902)	(-4.986)	(-4.904)	(-4.994)	(-4.967)
logME		-0.465***	-0.473***	-0.507***	-0.501***	-0.529***	-0.522***	-0.546***
		(-2.754)	(-2.863)	(-3.148)	(-3.049)	(-3.327)	(-3.207)	(-3.431)
logB/M				0.220*		0.224*		0.199
				(1.719)		(1.768)		(1.546)
L.EP					2.331**		2.237**	1.751*
					(2.272)		(2.175)	(1.703)
L.BU						0.022***	0.023***	0.023***
						(3.296)	(2.903)	(3.045)
Constant	1.521*	7.781***	8.650***	9.479***	8.999***	9.806***	9.322***	9.972***
	(1.939)	(2.646)	(2.986)	(3.320)	(3.139)	(3.481)	(3.282)	(3.559)
R^2	0.030	0.030	0.056	0.069	0.060	0.072	0.064	0.076

t statistics in parentheses
* p<0.1, ** p<0.05, *** p<0.01.

The results in table **7** are basically similar to Table **6**, with the slight difference that the coefficient p of PMN is more significant for the group with greater negative sentiments.

4.3. Empirical Testing of the Modified Four-factor Model

4.3.1. The Fama-Macbeth Regression Testing

We use the Fama-MacBeth regression to test the factor expected returns, i.e., whether the regression coefficients of the factors are significant, to verify the validity of the factors. We use the proxy variables to substitute the factors and run the regression with variables including pre-ranking β, logarithm of stock market value (logME), logarithm of stock book-to-market ratio (logB/M), logarithm of stock earning-to-price ratio(E/P), and the textual investor sentiment(BU index). The candidates for the value factor include B/M and E/P to verify E/P is a more suitable value factor than B/M. The results are shown in Table **8**.

First, the β coefficients are all significantly negative at the 1% significant level, indicating that the Chinese market has priced in market factor, and there may be low β anomalies. Second, logME is also significant in all eight regressions, indicating the importance of the market factor. Third, the E/P ratio can obtain a significant factor premium when the logB/M cannot be significant in column (8), which means E/P rather than B/M as the value factor. Fourth, the BU index is equally significant, indicating that the textual sentiment factor is effective for asset pricing. Finally, we can find the variables are all significant except logB/M in column (8), and this proves that the factors selected in our modified four-factor model are all valid.

To verify that the modified four-factor model is more effective, we further compare the modified four-factor model with the CAPM, the Chinese three-factor model using the GRS regression test.

4.3.2. The GRS Regression Testing

We compare the modified four-factor model with the CAPM and the Chinese three-factor models to see if the modified four-factor model outperforms the other two models. Results are shown in Table **9**.

If the pricing model can fully explain the excess returns of all stock portfolios in the cross-section, then the joint test of all portfolio regression intercept terms should not reject the original hypothesis of being simultaneously zero. In addition to the GRS statistic, the first indicator, alpha, is the average of the 25 regression intercept terms; the second indicator, abs_alpha, is the average of the absolute values of the 25 regression intercept terms. The smaller these two indicators indicate that the intercept term is closer to 0.

Table **9** compares the performance of the modified four-factor model with the CAPM and the Chinese three-factor model under different groupings and finds that the modified four-factor model has lower F-GRS values than the CAPM

Table 9. The GRS Regression Testing.

SIZE-E/P: 25 Portfolios	F-GRS	alpha	abs_alpha	P-value	Adj R2
MKT	2.964***	0.442	0.560	0.000	0.784
MKT CHSMB CHVMG	1.676**	-0.119	0.194	0.034	0.924
MKT CHSMB CHVMG PMN	1.970***	0.025	0.173	0.008	0.927
SIZE-BU: 25 portfolios	F-GRS	alpha	abs_alpha	P-value	Adj R2
MKT	2.631***	0.433	0.529	0.000	0.798
MKT CHSMB CHVMG	1.898**	-0.122	0.225	0.011	0.931
MKT CHSMB CHVMG PMN	1.386	0.023	0.107	0.124	0.937

* p<0.1, ** p<0.05, *** p<0.01.

in all cases, although the corresponding test values are slightly higher than those of the Chinese three-factor model in the size and E/P groupings; however, the corresponding test values are the lowest in the size and BU index groupings.

Combining the two intercept term indicators, alpha and abs_alpha, it is easy to find that the modified four-factor model has the smallest intercept term. And the Adj-R^2 again proves that the modified four-factor model fits better than the CAPM and the Chinese three-factor model.

5. DISCUSSION

In this paper, we show that textual investor sentiment can be a useful factor in pricing assets and can improve the explanatory power of asset pricing models. Unlike most previous studies, investor sentiment in our study is negatively related to stock excess returns in the current period. Importantly, our textual investor sentiment considers a neutral text and a combination of messages from different styles of stock bars to avoid single sentiment tendencies as much as possible. At present, we only consider texts from the social media source of stock bars and suggest that subsequent studies consider more text data from different sources such as company announcements and professional financial news, and try to construct a composite text-based investor sentiment index such as the BW index.

6. CONCLUSIONS

We extend the Chinese three-factor model to construct the modified four-factor model that includes textual investor sentiment. We use the BU Index, a Chinese internet stock review sentiment index, as a proxy for the textual investor sentiment factor and conduct an empirical analysis of the modified four-factor model. The main findings of this paper are as follows.

First, E/P is a more suitable variable for the value factor than B/M in China. After running regressions with all eligible stocks in China A-shares and GEM, our results support the Chinese three-factor model' s approach of using E/P as a value factor variable.

Second, we used the BU index as a proxy variable for the textual investor sentiment factor and found that the textual investor sentiment factor is an influential factor. The textual investor sentiment factor improves the factor model' s explanatory strength for stock portfolios' excess returns. In addition, we also conducted the GRS test. The results also indicated that the GRS statistic of the four-factor model with the reserve of the textual investor sentiment factor is lower than that of the CAPM and Chinese three-factor models, which means that the modified four-factor performs best among the three models.

Third, the textual investor sentiment factor significantly impacts the excess returns of equity portfolios in the high E/P group, the low E/P group, and the negative sentiment group. In contrast, it has a less significant impact on the portfolios of the other groups. These results remain broadly consistent with the reality of the Chinese stock market, which may also explain the high volatility of the Chinese stock market. Moreover, it also suggests that the textual investor sentiment factor constructed based on textual data of the trading period is not entirely noisy and helps explain the excess returns of the stock portfolios in the current period.

Finally, In this study, only internet stock message board textual data is considered in the textual investor sentiment. Multisource textual data will be focused on in future research, including improving the coverage and size simultaneously.

REFERENCES

Ahern, K. R., & Sosyura, D. (2014). Who writes the news? Corporate press releases during merger negotiations. *The Journal of Finance*, *69*(1), 241-291.

Antweiler, W., & 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. https://doi.org/10.1111/j.1540-6261.2004.00662.x

Arms Jr, R. W. (1989). The Arms Index (TRIN). *Dow Jones-Irwin*.

Asness, C., Frazzini, A., Israel, R., & Moskowitz, T. (2015). Fact, fiction, and value investing. *The Journal of Portfolio Management*, *42*(1), 34-52.

Baker, M., & Wurgler, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. *The Journal of Finance*, *61*(4), 1645-1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x

Bartov, E., Faurel, L., & Mohanram, P. S. (2018). Can Twitter help predict firm-level earnings and stock returns? *The Accounting Review*, *93*(3), 25-57.

Black, F., Jensen, M. C., & Scholes, M. (1972). The capital asset pricing model: Some empirical tests.

Breeden, D. T. (1979). An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics*, *7*(3), 265-296. https://doi.org/https://doi.org/10.1016/0304-405X(79)90016-3

Brown, G. W., & Cliff, M. T. (2004). Investor sentiment and the near-term stock market. *Journal of Empirical Finance*, *11*(1), 1-27.

BU, H., XIE, Z., LI, J.-h., & WU, J.-j. (2018). Investor sentiment extracted from internet stock message boards ank its effect on Chinese stock market. *Journal of Management Science in CHINA*, *21*(04), 86-101.

Carhart, M. M. (1997). On Persistence in Mutual Fund Performance. *The Journal of Finance*, *52*(1), 57-82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x

Chen, H., De, P., Hu, Y. J., & Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, *27*(5), 1367-1403.

Chen, Q., Goldstein, I., & Jiang, W. (2007). Price Informativeness and Investment Sensitivity to Stock Price. *The Review of Financial Studies*, *20*(3), 619-650. https://doi.org/10.1093/rfs/hhl024

Cox, J. C., Ingersoll, J. E., & Ross, S. A. (1985). An Intertemporal General Equilibrium Model of Asset Prices. *Econometrica*, *53*(2), 363-384. https://doi.org/10.2307/1911241

Daniel, K., Hirshleifer, D., & Sun, L. (2019). Short- and Long-Horizon Behavioral Factors. *The Review of Financial Studies*, *33*(4), 1673-1736. https://doi.org/10.1093/rfs/hhz069

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*.

Dougal, C., Engelberg, J., Garcia, D., & Parsons, C. A. (2012). Journalists and the stock market. *The Review of Financial Studies*, *25*(3), 639-679.

Epstein, L. G., & Zin, S. E. (1989). Substitution, risk aversion and the tempoal behavior of comsumption and asset returns: a theoretical framework. *Econometrica*, *57*, 937-969.

Epstein, L. G., & Zin, S. E. (1991). Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis. *Journal of Political Economy*, *99*(2), 263-286. https://doi.org/10.1086/261750

Fama, E. F. (1965). Portfolio Analysis in a Stable Paretian Market. *Management Science*, *11*(3), 404-419. https://doi.org/10.1287/mnsc.11.3.404

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383-417. https://doi.org/10.2307/2325486

Fama, E. F., & French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance*, *47*(2), 427-465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3-56. https://doi.org/10.1016/0304-405x(93)90023-5

Fama, E. F., & French, K. R. (1996). Multifactor Explanations of Asset Pricing Anomalies. *The Journal of Finance*, *51*(1), 55-84. https://doi.org/https://doi.org/10.1111/j.1540-6261.1996.tb05202.x

Fama, E. F., & French, K. R. (2013). *A four-factor model for the size, value, and profitability patterns in stock returns*. University of Chicago

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1-22. https://doi.org/https://doi.org/10.1016/j.jfineco.2014.10.010

Fama, E. F., & MacBeth, J. D. (1973). Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy*, *81*(3), 607-636. https://doi.org/10.1086/260061

Gibbons, M. R., Ross, S. A., & Shanken, J. (1989). A test of the efficiency of a given portfolio. *Econometrica: Journal of the Econometric Society*, 1121-1152.

Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American economic review*, *70*(3), 393-408.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, *50*(4), 1029-1054. https://doi.org/10.2307/1912775

Harris, M., & Raviv, A. (2015). Differences of Opinion Make a Horse Race. *The Review of Financial Studies*, *6*(3), 473-506. https://doi.org/10.1093/rfs/5.3.473

Hirshleifer, D., & Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, *36*(1), 337-386. https://doi.org/https://doi.org/10.1016/j.jacceco.2003.10.002

Holmström, B., & Tirole, J. (2001). LAPM: A Liquidity-Based Asset Pricing Model. *The Journal of Finance*, *56*(5), 1837-1867. https://doi.org/https://doi.org/10.1111/0022-1082.00391

Hou, K., Xiong, W., & Peng, L. (2009). A tale of two anomalies: The implications of investor attention for price and earnings momentum. *Available at SSRN 976394*.

Hou, K., Xue, C., & Zhang, L. (2014). Digesting Anomalies: An Investment Approach. *The Review of Financial Studies*, *28*(3), 650-705. https://doi.org/10.1093/rfs/hhu068

Jegadeesh, N., & Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance*, *48*(1), 65-91. https://doi.org/https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

John Y. Campbell, & John H. Cochrane. (1999). By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior. *Journal of Political Economy*, *107*(2), 205-251. https://doi.org/10.1086/250059

Kelley, E. K., & Tetlock, P. C. (2017). Retail short selling and stock prices. *The Review of Financial Studies*, *30*(3), 801-834.

Kumar, A., & Lee, C. M. (2006). Retail investor sentiment and return comovements. *The Journal of Finance*, *61*(5), 2451-2486.

LeDoux, J. (2012). Rethinking the emotional brain. *Neuron*, *73*(4), 653-676.

Lee, C. M., Shleifer, A., & Thaler, R. H. (1991). Investor sentiment and the closed-end fund puzzle. *The Journal of Finance*, *46*(1), 75-109.

Lee, W. Y., Jiang, C. X., & Indro, D. C. (2002). Stock market volatility, excess returns, and the role of investor sentiment. *Journal of Banking & Finance*, *26*(12), 2277-2299.

Lintner, J. (1965a). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *The Review of Economics and Statistics*, *47*, 13-47.

Lintner, J. (1965b). Security Prices, Risk, and Maximal Gains From Diversification. *The Journal of Finance*, *20*(4), 587-615. https://doi.org/10.2307/2977249

Liu, J., Stambaugh, R. F., & Yuan, Y. (2019). Size and value in China. *Journal of Financial Economics*, *134*(1), 48-69. https://doi.org/https://doi.org/10.1016/j.jfineco.2019.03.008

Long, J. B. D., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise Trader Risk in Financial Markets. *Journal of Political Economy*, *98*(4), 703-738. https://doi.org/10.1086/261703

Lucas, R. E., & Stokey, N. (1987). Money and Interest in a Cash-In-Advance Economy. *Econometrica*, *55*, 491-513.

Markowitz, H. (1952). The Utility of Wealth. *Journal of Political Economy*, *60*(2), 151-158. https://doi.org/10.1086/257177

Mehra, R., & Prescott, E. C. (1985). The equity premium: A puzzle. *Journal of Monetary Economics*, *15*(2), 145-161.

Merton, R. C. (1973). An Intertemporal Capital Asset Pricing Model. *Econometrica*, *41*(5), 867-887. https://doi.org/10.2307/1913811

Mossin, J. (1966). Equilibrium in a Capital Asset Market. *Econometrica*, *34*(4), 768-783. https://doi.org/10.2307/1910098

Newey, W. K., & West, K. D. (1987). Hypothesis Testing with Efficient Method of Moments Estimation. *International Economic Review*, *28*(3), 777-787. https://doi.org/10.2307/2526578

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, *108*(1), 1-28. https://doi.org/10.1016/j.jfineco.2013.01.003

Ross, S. A. (1976). Options and Efficiency*. *The Quarterly Journal of Economics*, *90*(1), 75-89. https://doi.org/10.2307/1886087

Shanken, J. (2015). On the Estimation of Beta-Pricing Models. *The Review of Financial Studies*, *5*(1), 1-33. https://doi.org/10.1093/rfs/5.1.1

Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance*, *19*(3), 425-442. https://doi.org/10.2307/2977928

Shefrin, H. (2001). Behavioral corporate finance. *Journal of Applied Corporate Finance*, *14(3)*, 113-126.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and Trades: the Information Content of Stock Microblogs. *European Financial Management*, *20*(5), 926-957. https://doi.org/10.1111/j.1468-036x.2013.12007.x

Stambaugh, R. F., & Yuan, Y. (2016). Mispricing Factors. *The Review of Financial Studies*, *30*(4), 1270-1315.

https://doi.org/10.1093/rfs/hhw107

Weil, P. (1989). The equity premium puzzle and the risk-free rate puzzle. *Journal of Monetary Economics*, *24*(3), 401-421.

Zhou, G. (2018). Measuring Investor Sentiment. *Annual Review of Financial Economics*, *10*(1), 239-259. https://doi.org/10.1146/annurev-financial-110217-022725

Received: July 21, 2023 Revised: July 25, 2023 Accepted: July 28, 2023