1. **Introduction**

Because large price volatilities are more likely to be crashes rather than increases in the stock market, investors more focus on the asymmetry of volatility in stock returns. The stock crash denotes the occurrence of extreme negative stock returns in future (Jin and Myers, 2006; Kim et al., 2010, 2011). Stock market crash not only threatens to the stability and development of stock market, but also harms shareholders wealth. The stock crash can result from the simultaneous release of private information by insiders (Romer, 1993; Cao, Coval, and Hirshlerfer, 2002) and the promotion by outsiders (Barlevy and Veronesi, 2003). Both of two sources can include investor sentiments on individual stocks and stock market. Also, the stock crash risk can be attributed to three aspects: the asymmetric information, market fraction and differences of investors’ opinions (Yin and Tian, 2015). The asymmetric information and the differences of investors’ opinions are similar with the differences of investor optimistic and pessimistic sentiments on individual stocks, and market fraction is similar with investor sentiments on market (i.e., market sentiment). Accordingly, this study develops the framework of the effects of investor sentiment on stock crash risks, which considers two sentiment factors, “the differences of investor positive and negative sentiments on individual stocks” and “market sentiment” into the analysis.

Past many studies have demonstrated that investor sentiments significantly affect stock market returns, and some studies among them have empirically found the impacts of investor sentiments on stock return volatilities (Da et al., 2015; Chen et al., 2014; Renault, 2017). Several studies have further analyzed the effects of investor sentiments on stock crash risks and have found the significant evidences (Yin and Tian, 2015; Cui and Zhang, 2019; Xu,et al 2013;Barlevy and Veronesi, 2003;Jang and Kang, 2018). Most of these studies have demonstrated that investor sentiments are positively associated with crash risk of stock price. The positive association can be due to the occultation of negative information in high sentiment period (Cui and Zhang, 2019), sentiment-driven overpricing by noise traders (Jang and Kang, 2018) and the panic expectation of uninformed investors as price declines (Barlevya and Veronesib, 2003). Moreover, the poor quality of financial reports (Yin and Tian, 2015; Cui and Zhang, 2019), high analyst coverage (Xu et al., 2013) and short-sale constraint (Yin and Tian, 2015) all will reinforce this positive relation between investor sentiments and stock crash risks.

过去的文献关于投资人情绪衡量的方法主要有三种。第一、调查基础衡量。Brown and Cliff (2005)以调查的投资人情绪数据为基础，发现了高情绪之后往往伴随着低收益。第二、市场基础衡量。Baker and Wurgler (2007)则以市场数据为基础，通过计算IPO和投资人情绪之间的关系，证明了衡量投资人的情绪是完全可行的；投资人的情绪对个别公司和整个股市有着十分明显、重要和规律的影响，特别是一些难以套利的股票受投资者情绪影响最大。Liao et al. (2011) 采用三项式分布方法来度量管理者羊群行为，并使用主成分分析作为提取市场基础相关变量的复合情绪度量的手段，证明了投资人情绪在共同基金羊群行为中起着重要作用。第三、传统媒体基础衡量。还有一种以传统新闻媒体为基础的投资人情绪衡量方法，Tetlock (2007)利用《华尔街日报》一篇流行专栏的日常内容，来定量衡量媒体与股市互动的性质，发现媒体的高悲观情绪预示着市场价格的下行压力，下行之后市场价格会再回归到基本面，异常高或异常低的悲观情绪预示着更高的市场交易量。Sun et al. (2016) 探讨了高频投资人情绪与股市收益率之间的预测关系，通过对来自新闻联播、互联网新闻和社交媒体等消息来源的文本进行分析计算，证明了标准普尔500指数的回报率是可以预测的，并且认为回报的可预测性很可能是由噪音交易者的交易活动所驱动的。

<**网络或结合金融科技的投资人情绪衡量**>

近年来，越来越多的行为财务学者开始以网络信息**或结合金融科技**为基础衡量投资人的情绪。来自网络的投资人情绪信息应可克服来自于调查基础衡量可能夹杂和答案相关的偏差、市场基础衡量可能有特质及非情绪相关的成分，与传统媒体新闻基础可能混杂因果关系 (Renault, 2017)。同时因为网络沟通的持续变动特质，新闻噪声的问题可被频繁地重新评估。进而，结合金融科技技术可使得用庞大的网络信息以更精确的衡量投资人情绪。

进而,有些文献已利用新闻的文本分析来探讨这个主题。Sun, Najand and Shen (2016)探讨由新闻媒体计算的高频投资人情绪和股票报酬之间的预期关系，结果发现甚至落后报酬存在之下，落后半小时的投资人情绪会最优势的预测日内S&P500指数报酬。Renault (2017)使用一个大众于股市的微博平台，建构由在线大众于分享股市多头或空头看法时使用的字词并形成情绪权重，发现前半小时投资人情绪变动可预测后半小时S&P500指数ETF报酬。然而这些文献的信息多来自单一新闻来源，尚需要多元的信息平台，才能有效结合来自不同投资客群的社会聆听。

However, the direct internet messages published by social platform of online investors are usually shorter and less formal than the indirect internet messages published by financial platform of experts collecting investor information, making the correct classification of optimistic or pessimistic tones difficult. Thus, this study uses the indirect internet messages from experts collecting investor information, which can avoid collecting the ambiguous tone and combine the advantages of dictionary-based approach like traditional media news and machine learning technique from the internet.

Moreover, the network information of past related studies mostly comes from a single social media platform such like StockTwits (Da et al., 2015; Chen et al., 2014; Renault, 2017), which may be able to glean the emotional response from a specific-type audience. However, we need to get the emotional response from a variety of investor groups to integrate with social listening. Meanwhile, the current literature does not directly capture the sentiment of investors on a stock but only captures the bullish or bearish time series trend of investors on the overall stock market (like Renault, 2017). If the optimistic and pessimistic sentiments of investors on individual stocks can be gotten, the investment decision of choosing the stocks and timing can be carried out. If the strength of the optimistic and pessimistic sentiments of investors on individual stocks can be further gotten through the network information, the network volume of the optimistic and pessimistic sentiments of investors on individual stocks can be used to predict the effects on the stock returns.

我国资本市场自2014年起陆续推出大数据指数基金，指数选股标准除了市值、盈余等传统因子外，还加入了在网络搜索引擎的访问次数及社群媒体的评论正负项得分等大数据因子。随着互联网的不断发展，当前中国国内网民人数已经超过了人口的一半，与传统的信息传播方式相比，互联网使得网民个体之间的沟通具有一种低成本、传播及时和交流广泛的特点，方便了投资者之间的交流，也使得投资者情绪得以更加广泛的传播（孙鲲鹏，2018）。2017年，全球数字经济规模达12. 9万亿美元，其中中国透过网络交易的数字经济总量达27. 2万亿元，位居全球第二。网络平台作为数字经济呈现形式之一，现已是专家或投资者交换观点、情感和知识的重要渠道。基于庞大的数据量，股吧信息可以比较准确、实时地反映出投资者的心理和行为的变化。与调查问卷、档案数据和访谈记录等信息源相比，网络平台数据具有用户基数大、互动性强、涉入性高、客观性高与响应速度快等优势。然而,与网络投资者面对的社交平台相比,专家或记者收集投资者信息的金融平台能够更加精确地区分网络信息中的乐观或悲观语气。与社交平台类似,金融平台仍然拥有大量的数据, 也需要大量的时间和脑力来汇集专家或记者从投资人那里收集到的间接互联网信息。金融科技可以使用网络爬虫技术爬取各网络平台的信息，并通过文本分析准确地提取到和网络信息相关的词汇。并在此文本分析基础上对获得的数据进行分类、汇总和量化。因此我们可以利用金融科技在各网络平台或者新闻媒体中捕捉投资人的情绪并对其进行相关研究。　同时，由于国内股票市场发展的起步较晚，相关机制还不够健全，因此现阶段国内股票市场与已开发国家股市还存在着诸多不同之处，正如当前国内股票市场无法引进熔断机制一样，国外的许多研究成果在国内也并不适用。

近年来陆续已有许多文献使用金融大数据或金融科技技术相关的创新，通过网络社群聆听捕捉大盘股市的投资人情绪信息，进而分析投资人情绪对大盘指数股市报酬波动性或崩盘风险的影响。Das and Chen (2007) 由投票选择出一些结合在一起的分类器算法，开发出一种从股票留言板中提取小投资人情绪的算法。并通过该算法证明了技术部门发布的信息与股票指数水平以及成交量和波动性有关。By analyzing the effect of Raging Bull in the Dow Jones Industrial Average on stock volatilities, the results of Antweiler and Frank (2004) show that the more bullish messages in the current day significantly raise tomorrow’s market volatility. Da et al. (2015) ……….. Johnman, Vanstone and Gepp (2018)…….

仍有少数文章已分析网络社群聆听捕捉投资人对个别股票的情绪信息，并进而影响个股报酬波动性或崩盘风险。孙鲲鹏和肖星（2018）使用周频率的数据探究了互联网社交媒体在投资者情绪传染中的作用以及投资者情绪本身对股价崩盘风险的影响，发现投资者的发帖情绪越乐观，未来股价崩盘风险越高，这种现象源自乐观情绪在股吧的传播扩散，而机构投资者持股会加剧这种效应；社交媒体等互联网技术的发展会助长投资者情绪的蔓延并引起股价崩盘。

綜上所述，第一、孙鲲鹏和肖星（2018）沒有采用文本分析的方法。Avery et. al (2016）文章都仅采用文本分析的内容，但其未采用文本分析的方法来处理。然而使用文本分析的方法赋予每个词频一定的权重，能够使得来自金融论坛的数据获得更加规范的处理，从而增强信息的可读性；同时能够衡量读者破译预期信息的能力，提高信息描述的精确度，能更加精确地衡量投资人的情绪 (Kearney, Liu, 2014；戴德宝，2019)。第二、段江娇等人（2017）在构建指数的过程中只考虑到对个股看涨或者看跌情绪出现的频率。本文在构建情绪变量时，不仅考虑到了投资人情绪出现的频率，而且还赋予词频一定的权重。通过词频和权重，我们可以得到投资人看涨和看跌情绪的情绪密度，能够更精确度量投资人情绪。第三，除了段江娇等人（2017）分析投资人情绪和个股报酬风险之间的关系之外，其他文章都仅分析投资人情绪对股票报酬的影响。本文在此基础之上还研究投资人情绪对个股报酬风险的影响。综合分析个股的报酬和风险可以使投资人的投资行为更加理性，选股更加准确。但是，段江娇等人（2017）是通过分析投资人情绪的一致性而不是情绪的大小来探讨投资人情绪对个股报酬的风险的影响。第四、除了仅Avery et al. (2016)分析投资人对个股乐观及悲观分别的情绪之外，其他文章都仅分析投资人对个股乐观与悲观混合的整体情绪对股票报酬的影响。然而，分析投资人对个股乐观及悲观分别的情绪可更精确衡量投资人情绪对股票报酬的影响。

　　然而，这些文献的投资人情绪信息来源多来自单一网络平台，可能较会捕捉到特定投资客群的投资人情绪信息。也就是说一些基础的金融科技技术虽可于特定平台大量的挖掘信息，但尚需要较成熟的金融科技技术(如网络爬虫技术)等才能广泛于各个网络平台搜集不同投资客群的投资人情绪信息，也才能结合较全面性的网络社群聆听获得较客观多源的网络声量。

细部而言，使用网络爬虫技术爬取各个网络财金平台，本课题间接捕捉投资人对某只股票涨跌看法的信息，形成乐观及悲观情绪的网络声量。我们将正负面情绪的强度整合至金融大数据平台，探讨正负面情绪强度对个股股价报酬的影响进而预测这些股票的报酬。

1. **文獻探討:**

**2.1市场及个股的投资人情绪对股票波动性及崩盘风险影响相关文献**

Past literatures have analyzed the relationship between investor sentiment and stock return volatilities. Using daily frequency data, Da, Engelberg and Gao (2015) find that the Financial and Economic Attitudes Revealed by Search (FEARS) index is strongly related to the transitory component of daily volatility, and it is also correlated with VIX futures returns.

胡昌生，迟阳春（2013）研究了在不同估值水平下投资者情绪对于股票市场波动性影响的差异。我们发现在市场不同的估值阶段投资者情绪对波动性的影响有很大差异。当市场处于高估值期时非理性情绪对波动性有显著影响理性情绪的影响不显著；当市场处于低估值期时，理性情绪对波动性有显著影响，非理性情绪的影响不显著。此外，理性情绪对大盘股波动性的影响更大，非理性情绪则对小盘股波动性的影响大。

Past literatures have analyzed the relationship between investor sentiment and stock crash risk. Only the few study explores the effects of investor sentiments about the stock market on stock crash risk. By using the factor analysis method to combine four investor sentiment variables related to stock market into a comprehensive investor sentiment index, Yin and Tian (2015) examine the association between investor sentiment index and stock crash risk with short-sales constraint conditions. Their results indicate that investor sentiment is positively related with future stock price crash risk and poorer quality of financial report and short-sale constraint both will reinforce this relation. However, most of related studies analyze the effects of investor sentiments about the individual stocks on stock crash risk. Using a large sample of U.S. firms for the period of 1991–2014, Cui and Zhang (2019) examine the impact of investor sentiment on future stock price crash risk. Their results show that firm negative information is more likely to be hidden during high investor sentiment period, thus resulting in greater likelihood of stock price crash risk. The investor sentiment exhibits more pronounced crash risk for the firms with higher leverage ratio, larger default risk and greater analyst forecast dispersion. Using a unique Chinese database of analyst earnings forecasts, Xu, Jiang, Chan and Yi (2013) examine the associations among analyst coverage, analyst optimism and firm-level stock price crash risk. They find that an increase in a firm's analyst coverage results in an increase in stock crash risk. Moreover, this positive association is more pronounced when analysts are more optimistic and tends to be weakened when analysts are non-optimistic. By using equilibrium price function, Barlevy and Veronesi (2003) indicate that the panic behavior of uninformed traders cause the price of stocks to plummet, and then crashes could arise. As prices decline, uninformed traders precipitate a price crash, implying that the magnitude of the crash depends on the fraction of uninformed investors. Jang and Kang (2018) estimate an ex ante probability of crashes of individual stocks and find that stocks with a high probability of crashes produce abnormally low returns. Their results indicate that the worse performance of stocks with high crash probability can come primarily from sentiment-driven overpricing by noise traders and partially from rational speculative bubbles driven by sophisticated traders.

**2.2 使用网络信息捕捉投资人的情绪**

Recently, the scholars of behavior finance have begun to use the Internet data to instantly measure investor sentiment from network community listening, which should improve the measure of investor sentiments.

By aggregating daily Internet search volume from millions of household concerns, Da et al. (2015) construct a Financial and Economic Attitudes Revealed by Search (FEARS) index as a new measure to reveal market-level investor sentiment. By using the dataset of more than 32 million messages on 91 firms posted on the message board of Yahoo! Finance and a machine learning classification algorithm, Kim and Kim (2014) examine whether the sentiment information of retail investors in posted messages has the predictive power for stock returns, volatility, and trading volume. By comparing the various information sources, content analysis methods, and empirical models, Kearney and Liu (2014) summarize the important findings about how textual sentiment influence on individual, firm-level and market-level behavior and performance. Oliveira et al. (2016）use the statistical measures applied by the labeled messages from StockTwits, like a specialized stock market microblog, to propose the novel and fast approach for creating stock market lexicons. They suppose that the new lexicons are competitive for measuring investor sentiment.

1、Da, Z. , Engelberg, J. , Gao, P. , 2015. The sum of all FEARS: investor sentiment and asset prices. Rev. Financ. Stud. 28 (1), 1–32 .

2、Kim, Kim,2014. Investor sentiment from internet message postings and the predictability of stock returns. J. Econ. Behav. Organ. 107, 708–729 .

3、Kearney, C. , Liu, S. , 2014. Textual sentiment in finance: a survey of methods and models. Int. Rev. Financ. Anal. 33 (3), 171–185 .

4、Oliveira, N., Cortez, P., Areal, N., 2016. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. Decis. Support Syst. 85, 62–73 .

少数国内学者也摘录网络信息以衡量投资人情绪。例如，王靖一 (2018）则收集了和讯网2013年1月到2017年9月1702多万条新闻数据编制了北京大学金融科技情绪指数，数据规模超过了700GB。在应用金融科技来捕捉投资人情绪方面的研究起步较晚，但近年来也取得了一些成果。唐涛提出了当前网络情绪的分析已经超过了传统的分析框架，必须结合金融大数据在分析方法上进行创新，并且指出未来的网络情绪分析方法应该向金融大数据分析的方向发展创新（唐涛，2014）。李金海(2014)根据大数据思想构造了网络情绪的文本挖掘模块，通过计算词汇在文件中的重要性来构造情绪指数。孟雪井等学者曾使用过文本挖掘技术来对中国知网的CSSCI期刊以及新浪微博上的话题进行文本分析，并结合了百度的关键词推荐系统，综合了三大词库来确定我国与投资者情绪相关的网络关键词（孟雪井，2016）。王夫乐等学者通过新浪微博开发平台的开放接口，抓取了微博上2013年3月25日到2016年2月26日的每日数据，并对每一条微博内容都进行了文本情感分析，将情感分为喜悦、惊奇、愤怒、恐惧、悲伤五个维度，并以每一种情绪的微博数量作为权重，以各种情绪的加权均值来代表当日的总情绪（王夫乐，2017）。杨欣等学者利用从百度指数获得的搜索量数据，构建了衡量投资人对于突发事件的关注程度的指标。

**2.3媒体及网络信息基礎的投资人情绪对股價崩盤風險的影响**

Some scholars propose that the usage of the investor sentiment measure from network community listening is more precise to analyze the effect of investor sentiment on stock volatilities or crash risks. They tend to use financial technology to capture the huge amount of information on investor sentiment from the network platforms.

We first summarize the effects of network-based investor sentiment about the financial market on stock index volatilities or crash risks of past studies. Da et al. (2015) use the daily Internet search volume of households to construct a Financial and Economic Attitudes Revealed by Search (FEARS) index, which is strongly related to the daily market volatility index (VIX). They empirically find the increases in FEARS accompany by only temporary increases in market volatility.

We then summarize the effects of media-based investor sentiment for individual stocks on stock volatilities or crash risks of past studies. Johnman, Vanstone and Gepp (2018) use business news published by the Guardian Media Group to examine the statistical and economic effect of positive and negative sentiments on daily excess returns and volatility of the FTSE 100 index. They find that the sentiment from business news doesn't affect the excess index returns but does affect index volatility. However, their further results show that the negative sentiment increases index volatility but positive sentiment reduces it, which is different from the results of other related studies. Furthermore, using the data of media reports from the Internet in Chinese listed firms, Zhu, Wub, Zhang and Yu (2017) examine whether and how the media reports affect the probability of stock price crash. Their empirical results show that the positive media reports reduce the probability of stock price crash, but there is the U-shaped relationship between negative reports and the probability of price crash.

In Panel A of Table 2, Model 1 indicates that the quantity of positive reports is significantly negatively related to the probability of stock price crash. As expected, positive media reports promote investors’ confidence or reduce the information asymmetry of listed firms over the period of 1 year, thus reducing the probability of stock price crash. However, we cannot differentiate the two effects.

Turning to the negative news tones, the combined results in Models 3 and 4 of Panel A indicate that there is a significant U-shaped relationship between the quantity of negative reports and the probability of stock price crash. When small, the quantity of negative reports is negatively correlated with the probability of stock price crash. It indicates that more negative reports promote information quality more than they damage investors’confidence, thus reducing the probability of stock price crash when the quantity remains low. The degree of information asymmetry accumulates for those firms with fewer negative reports. However, we observe a significant positive relationship between the quantity of negative reports and the probability of stock price crash when the quantity is substantial.

We next summarize the effects of network-based investor sentiment for individual stocks on stock volatilities or crash risks of past studies. Using the search frequency from the Baidu Index to measure the attention of retail investor, Wen, Xu, Ouyang and Kou (2019) examine the effect of the attention of retail investor on stock price crash risk in Chinese listed firms. They find that the firms with higher retail investor attention effectively reduce the information asymmetry and get a lower future stock price crash risk.

The retail investors are the net buyer and the variation of the stock price is positively correlated with the retail investor attention in the short-term. The retail investor attention is positively correlated with the stock price in the short-term. The retail investor attention can enlarge the shareholder base and improve stock liquidity. Ying et al.(2015) use a search frequency index from Baidu.com as a measure of investor attention, and find retail investor attention has a positive influence on the stock return in short-run in China’s stock market. Earnings’ announcements released after the market closes are more likely to contain worse news. When investors pay insufficient attention to listed firms, they may neglect earnings’announcements, resulting in the untimely incorporation of earnings’news into stock prices. In this paper, the more retail investor attention may mean more noise trading in the stock.

Our final sample includes 15,426 firm-year observations representing 2,368 individual firms. The more information the retail investors get, the more difficult and more costly the managers of companies hide the negative news. For a firm with less retail investor attention, retail investors could get less information about the firm and the managers are under low pressure to withhold bad news from the public. As a result, the bad news will accumulate under low retail investor attention leading to greater future crash risk.

Antweiler and Frank (2004) study the effect of Raging Bull about the 45 companies in the Dow Jones Industrial Average on stock volatilities. By using computational linguistics methods, the bullishness of the messages is measured. Their results show that the more bullish messages in the current day significantly raise tomorrow’s market volatility. Antweiler and Frank( 2004) 使用计算机文本分类技术从Yahoo和Raging Bull 股票留言板帖子中提取情绪。研究发现股票当期帖子情绪与收益率显著正相关，当日帖子数与未来一天的收益率显著负相关。同时研究表明，论坛帖子数与同期股价波动显著正相关，而且还能预测未来一天的股价波动。Das and Chen (2007) download the messages from 24 tech-sector stocks in the Morgan Stanley High-Tech (MSH) index posed to the boards. Their results show that there is a strongly positive relationship between the message volume of these stocks and the volatility of MSH index from these stocks, in which high posting volume can be regarded as a proxy of optimistic sentiment. Das and Chen( 2007) 从 Yahoo 在线投资论坛股票留言板提取中小投资者情绪指数，利用网络论坛发帖者的看涨、看跌或看平观点，建立了一个基于Morgan Stanley 高科技股票指数的情绪指数（MSH），研究表明: 市场活跃程度与该情绪指数有很强的相关性。该研究也表明，论坛帖子数与股价波动显著正相关。

段江娇等人（2017）选取东方财富网股吧论坛的个股帖子，使用计算机文本处理技术提取帖子情绪，结合证券分析师对个股的“中性”评级数据，实证研究了我国股票网络论坛的信息含量问题。研究发现:股票当日收益率受当日论坛情绪影响，为显著正相关;股票未来两日收益率与帖子数显著负相关; 股票当日的帖子数显著正向影响当日股价波动，而且能正向影响未来两日的股价波动;当日情绪分歧度越大，未来两日的交易量越大。孙鲲鹏和肖星（2018）使用周频率的数据探究了互联网社交媒体在投资者情绪传染中的作用以及投资者情绪本身对股价崩盘风险的影响，发现投资者的发帖情绪越乐观，未来股价崩盘风险越高，这种现象源自乐观情绪在股吧的传播扩散，而机构投资者持股会加剧这种效应；社交媒体等互联网技术的发展会助长投资者情绪的蔓延并引起股价崩盘。

In sum, the previous studies obtain investor online community listening in the stock market commonly through the direct investor sentiment message published by network community platform of investors, but the caught sentiment message is more crude and noisy. Furthermore, this study objectively collects investor network community listening for individual stocks indirectly through the news of individual stocks published by network financial platform of experts or reporters collecting investor sentiment messages. That is, we assume that the internet-based along with media-based measure of investor sentiment is more precise to analyze the effect of investor sentiment on stock crash risks. Meanwhile, we assume that the simultaneous consideration of investor sentiments on individual stocks and stock market can provide the information on choosing the stocks and the time for the investors. In addition, we propose that the messages of investor sentiment from the multiple network platforms are more various and complete than those from the single network platform of most studies.

然而近年来国内已有一些文献应用金融科技来捕捉投资人情绪对股票报酬波动性或崩盘风险的影响。杨欣 (2014)构建投资人对于突发事件的关注程度的衡量，并以该衡量来研究突发事件关注度对于股数波动的影响。孙鲲鹏和肖星（2018）使用周频率的数据探究了互联网社交媒体在投资者情绪传染中的作用以及投资者情绪本身对股价崩盘风险的影响，发现投资者的发帖情绪越乐观，未来股价崩盘风险越高，这种现象源自乐观情绪在股吧的传播扩散，而机构投资者持股会加剧这种效应；社交媒体等互联网技术的发展会助长投资者情绪的蔓延并引起股价崩盘。

1. **Data Scope and Source**

This study uses the investors thatinvest in the individual stocks of Shanghai 50 Exchange Traded Funds (ETF) and those of Shenzhen 100 ETF in Shanghai and Shenzhen stock markets as our object, and collect the daily emotional information of these investors from the five representatively financial and investing website platforms of the two stock markets. This study uses the daily data from February 2012 to December 2020 as our sample period. Our collecting information includes the news information of the daily investors collected by reporters or experts on the risk appetites, expected returns, opinions, and bullish and bearish views of these ETF stocks in financial website platforms of the two stock markets. The specific content of our collecting news information includes the date, content, individual stocks, characteristic texts, and optimism or pessimism classes initially screened by content and characteristic texts, in order to build a complete database of individual stock sentiments of stock investors.

We collect supporting and non-supporting news of firms’ top executives toward ruling and opposition parties from two popular news groups in Taiwan. the United Daily News Group and Liberty Times, which mainly support the KMT and the DPP, respectively.[[1]](#footnote-1) The market shares of the two newspaper groups are more than 63.95%. Moreover, the United Daily News and Liberty Times Group favor KMT and DPP, respectively, in which KMT-related reports are around 56.58% on the good part, whereas the DPP has around 77.99%. Thus, the news from these two newspapers groups can represent the coverage and unbiasedness toward any parties.

The detailed individual bank-loan contracts and financial ratios are obtained from the bank-loan database of the Taiwan Economic Journal (TEJ). Loan data consist of loan amounts, loan periods, collateral, and fixed or floating interest rate. Shen (2002) first use this loan-transaction database, which other studies employ to examine similar issues. Hence, the quality of data is less of a concern.

Given that firms may not want to reveal their political preference to avoid offending a political party, we collect firms that have released their political preference. The final sample contains 18,718 individual bank-loan contracts from 268 listed in the Taiwan Stock exchange and 54 in the Taipei Stock Exchange.

**4. Construction of the Variables**

**4.1 Dependent variables (Y)**

This study follows the approach of Chen, Hong and Stein (2001) to measure stock crash risk. First, we run the following regressions to obtain the idiosyncratic return of each stock in each week.

 (1)

where  is the stock return of firm  in week  and  is the weighted average market return in week . The lead and lag items of the market are employed in Equation (1) to reduce the deviation because of no synchronous trading (Dimson 1979). We calculate to represent the idiosyncratic return of stock  in week .

We then calculate the following two variables to measure the probability of a stock price crash.

 (2)

 (3)

where  equals the total number of trading weeks of stock  and  is the number of trading weeks with weekly returns higher (lower) than the mean return of the year. We use the variables and to represent the negative return skewness and the down-up return volatility. If is bigger, the stock has a more negative skewness and a larger crash risk.  is a specific return standard deviation of down weeks divided by the one of up weeks.  is bigger means the crash risk become bigger.

**4.2 Independent variables- Sentiment variable about stock market (X1)**

Some studies have used some indicators such as turnovers (TURN) (Jones and Owen, 2002; Baker and Wurgler, 2006) and discount rate of closed-end funds (CEFD) (Lee, Shleifer and Thaler, 1991) to measure the sentiments of investors on market (hereafter, market sentiment of investors). Other studies have used Ratio (i.e., the number of listed company’s going up divided by the number of listed company’s falling down) and Open (i.e., the growth rate of investors’opening accounts) to measure market sentiment of investors. IPO initial return is not a proper measure of investor sentiment although a few studies have considered as a sentiment proxy (Pitter and Welch, 2002; Baker and Wurgler, 2006). This is because there is always a high IPO initial return even if the market is not bullish. Thus, we employ the four market sentiment variables to create a comprehensive market sentiment index using the factor analysis method.

 (4)

**4.3 Independent variables- Sentiment variables about individual stocks (X2)**

This study establishes the variables of investor sentiment about individual stocks on the following five steps.

**Step 1**: We use both general-class and financial-class sentimental dictionary as the seeds of text characteristics. Simplified Chinese Emotional Polarity Dictionary of Taiwan University is considered as our general-class sentimental dictionary, and Chinese Financial Emotion Dictionary of Xiamen University is used as our financial-class sentimental dictionary. The two dictionaries belong to the representative and authoritative dictionaries of investor sentiments used in Chinese characters. We mix the two dictionaries together to get the seeds of our text characteristics.

**Step 2**: We look for the network news with extremely positive and negative scores as potential text characteristics by using the seeds of text characteristics. First, the seeds of text characteristics are thrown into all network news of our sample firms. Next, the term frequency-inverse document frequency (*TF-IDF*) method is used to calculate and sort the scores of the characteristic texts of positive and negative sentiments for each news document. The definition of a *TF-IDF* score for *jth* term from optimistic and pessimistic characteristic texts in *kth* news is shown as follows:



 (5)

Where =.  means the number of the news’ documents for *jth* term of optimistic characteristic texts, and *N* is the number of totalnews’ documents in this study. Then,  is the number of *jth* term of optimistic characteristic texts in each news document. The definitions of and  of pessimistic characteristic texts are the similar with and . As the frequency of the term in the news document is more and the specific term number of all news’ documents is less, there is higher score of *TF-IDF*. If the score for optimistic characteristic texts in *kth* news is higher than that for pessimistic characteristic texts in the same news, the news is divided by optimistic class; otherwise, it is divided by pessimistic class. The characteristic texts for the news of the former 5% scores are chosen as optimistic characteristic texts in this study, and those for the news of the latter 5% scores are regarded as pessimistic characteristic texts.

**Step 3: Weight of characteristic texts of investor sentiment**

After matching the terms of our news documents with the seeds of our text characteristics, their weights should be decided in the linguistic field. By extending the sentiment weight of *TF-IDF* method computed in Oliveira et al. (2016), we revise it as the weights of specific-term (***j\****) in all the optimistic and pessimistic texts of all the related news’ documents. We take the weight of specific-term (***j\****) in the optimistic text for example to compute in the following:

 (7)

Where the subscript*s i, t, d*, and *k* are firm *i*th on year *t*, date *d* in news *k*th;  is the sum of sentiment score of specific-term (***j\****) in the optimistic characteristic texts of the related news’ documents. When  is lower, specific-term (***j\****) is less in the optimistic class; and when  is higher, its term frequency is more. Thus, the weight of specific-term (***j\****) in the optimistic texts, which indicates the ***j\**** term is more important in the collection of optimistic news’ documents.

-------------------------------------------------------------------------------------------------------

**Step 4:** Out-of-sample classification accuracy and the comparison between dictionary-base and machine learning

To assess the accuracy of classification for optimistic (or pessimistic) sentiment based on TF-IDF method, we select a list of optimistic and pessimistic messages for our 150 firms in 2020. We select the messages as positive for *kth* news in the testing samples if the score in *kth* news with optimistic characteristic texts is higher than that with pessimistic characteristic texts. Conversely, the messages are selected as negative. The selection can be regarded as real sentiment of each message. We use the same pre-processing techniques and messages as for our training samples (i.e., 2013-2019).

Then, we calculate a sentiment score for each message. The optimistic (or pessimistic) sentiment score is calculated as the average weight () of the terms present in the message for optimistic (or pessimistic) sentiment. If the pessimistic sentiment score is larger than the optimistic sentiment score, the sentiment score in this message belongs to pessimistic message. We compare the real sentiment of investors with the sentiment score in these messages. To consider out-of-sample classification accuracy for various classes, the proportion of correct classification of dropping unclassified messages (CC) and the proportion of correct classification of each class are computed.

Next, in terms of machine learning algorithm, we use maximum entropy proposed by . We use the algorithm of maximum entropy because the maximum entropy exhibits better results than naïve Bayes and support vector machines (Renault, 2017). Taking maximum entropy for example, the probability of document k belongs to class L given a weight vector v is the following:



Where is a set of m features that appear in a document.

Specifically, to compare the classification accuracy of dictionary-base with that of machine learning, we use the machine learning techniques to estimate the sentiment score of individual messages that belongs to bull (or bear) class. The similar approach can be used to get out-of-sample classification accuracy based on machine learning techniques.

|  |  |  |  |
| --- | --- | --- | --- |
|  | CC (%) | CCoptimistic(%) | CCpessimistic(%) |
| TF-IDF |  |  |  |
| Machine learning |  |  |  |

**Step 5: The intensity of investor sentiment about individual stocks**

This step calculates the annual “intensity of investor sentiment” toward the *IOS* (*IOSit*) and *IPS* (*IPSit*) for *i*th firm at year *t*. To derive *IOSit* or *IPSit*, we mainly select the characteristic texts that investors are optimistic in the news for *i*th firm at year *t*. Taking investor optimistic sentiment for an example, we calculate *IOSit* by summing the daily intensity *IOSdk,it* as the sum of the term frequency  multiplied by weight in each firm at each year. Then, we calculate *IPSit* by summing daily influences of *IPSdk,it*. A similar approach is applied to calculate *IPSit*.

; (8)

, (9)

where  and  are the term frequencies of the *j*th characteristic text of investor optimistic and pessimistic sentiments for the *i*th firm on date *d* in the *k*th news;  and  are the weights of the *j*th characteristic texts of the optimistic and pessimistic sentiments. A higher intensity of *IOSit* or *IPSit* means a stronger optimistic or pessimistic sentiment of investors for firm *ith* on year *t*.

In addition, the maximum values of positive and negative sentiment variables are unlimited since the number of newsdocument *K* and number of characteristic texts *J* are unlimited. Thus, the concept of uniform distribution is used to standardize the nonzero-intensity values in which the minimum and maximum are between 1% and 100%, respectively.[[2]](#footnote-2) However, we still regard the zero-intensity value as zero. Thus, a high intensity means a strong investor sentiment.

To decompose investor sentiments about stock market and pure individual stocks, we regress *IOS* and *IPS* on respectively to obtain error term, which can be regarded as investor sentiments about pure individual stocks. The process is as follows:

  (10)

**

** (11)

where  denotes investor sentiment about stock market and is defined by section 4.2; and  and  are the adjusted IOS and IPS measures in which we control investor sentiments about stock market, which denote the adjusted investor optimistic and pessimistic sentiments about pure individual stocks, respectively. Thus, the higher the  and  are, the stronger the investor optimistic and pessimistic sentiments about pure individual stocks are.

**5. Econometric method**

**5.1 Regression model**

We first examine how investor optimistic and pessimistic sentiments about individual stocks affect the probability of stock price crash. Our model is specified as follows:

, (12)

where, the subscripts *i*, and *t* denote the *i*th stock at year *t*. The dependent variable *Loan* is the loan contract. The variables *IPTKMT* and *IPTDPP* are the strengths of a firm’s party tendencies toward supporting the KMT and DPP in our sample period, respectively.We have two types of proxies for loan contracts, namely, loan rate and non-loan rate terms. Loan rate is the spread that a borrower pays based on points over a risk-free rate. The non-loan rate terms comprise LoanPeriod, LoanSize, NumLend, LoanSecu, and DNEW. LoanPeriod is the number of loan days; LoanSize is the loan amounts over book value of assets; NumLend is the total number of lenders to a firm during a given year; LoanSecu is an index ranging from 0 to 3 with 0 denoting no collateral, 1 for a credit collateral, 2 for a security collateral, and 3 for a real estate collateral. The DNEW is a dummy variable that denotes the loans in the first year of the contract.

**Firm** is the vector of the control variables that comprises six firm characteristic variables; namely, LnAsset, Leverage, Tobin's *Q*, Tangibility, Profitability, and Z-score (Graham, Li, and Qiu, 2008). LnAsset is the log of total firm assets; Leverage is the ratio of the long-term debt plus debt in current liabilities divided by total assets; Tobin’s *Q* is the market value of equity’s ratio plus the book value of debt divided by total assets; Tangibility is the ratio of the net property, plant, and equipment divided by total assets; modified Z-score of Altman (1968).[[3]](#footnote-3) All firm characteristic variables are estimated one year prior to the initial loan year.

**OtherLoanTerm** is the vector that comprises five loan characteristic variables that is different from loan contract; namely, LoanPeriod, LoanSize, LoanSecu, Rating, and DFix. Rating is credit-risk rating of a firm ranging from 1 to 9, where a high score denotes high default probabilities. The DFix is a dummy variable that equals one if the type of loan has a fixed interest rate and zero otherwise. The variables in this study are defined in Table 2. Our model uses industry, bank, and year dummies in the regression to control for fixed effects.

 (13)

1. The United Daily News Group include United Daily News, Economic Daily News, United Evening News. [↑](#footnote-ref-1)
2. For non-zero-intensity values, our standardized method is [99\*(*x* – *x*min)/(*x*max-*x*min) +1], where *x* is the intensity value. For zero-intensity values, we still use zero. [↑](#footnote-ref-2)
3. The modified Z-score is (1.2\*Working capital + 1.4\*Retained earnings + 3.3\*EBIT + .999\*Sales)/Total assets. (See Altman, 1968). [↑](#footnote-ref-3)