Ants that Move the Log: Crashes, Distorted Beliefs, and Social Transmission *

Job Market Paper

Qian Yang†

First Draft: January 6, 2021
This Version: October 24, 2022

Abstract

Have retail investors become the ants that move the log? Social media has proved instrumental for effective coordination that might lead to extreme returns. To study this effect, I construct a novel crash risk measure by estimating ex-ante crash probabilities via logit and machine learning techniques. Stocks with high ex-ante crash risk tend to have lower returns, especially when lagged sentiment is high. Robinhood traders tend to over-buy high crash risk stocks, consistent with the optimal expectations theory (Brunnermeier et al., 2007). By exploiting the staggered first appearances of ticker names on "Wallstreetbets", I document a causal effect of social transmission on crash risk. This effect is significantly more substantial for smaller stocks. To further bolster the finding, I exploit the entire history of Reddit to construct a novel instrument and show that social transmission is likely to cause elevated crash risk on a daily basis.

Keywords: Crash Risk, Left-Tail Risk, Retail Investors, Social Transmission, Cross-Section of Stock Returns, Machine Learning, Robinhood, Wallstreetbets.

*I thank my advisors, Naveen Khanna and Hao Jiang, for their support and encouragement. I thank Ruslan Goyenko (discussant), William Grieser, Ryan Israelsen, Preetesh Kantak (discussant), Donghyun Kim (discussant), Tim Loughran, Sophia Li, Dmitriy Muravyev, Terrance Odean, Tengjia Shu (discussant), Andrei Simonov, Dacheng Xiu, Morad Zekhnini, and participants in Eastern Finance Association Annual Meeting 2022, Australasian Finance and Banking Conference 2021, European Finance Association Doctoral Tutorial 2021, FMA Annual Meeting 2021, 2021 International Risk Management Conference, New Zealand Finance Meeting 2021, 2021 Academy of Behavioral Finance & Economics, 2021 SoFiE Summer School with focus on machine learning, the 17th Annual Conference of the Asia-Pacific Association of Derivatives (APAD) and MSU brown bag for their valuable comments.

†Ph.D. candidate, yangqia8@msu.edu, Eli Broad School of Business, Michigan State University.

1. Introduction

A long-standing point of inquiry in asset pricing and market micro-structure research concerns the role of retail traders. On the one hand, retail traders may generate noise that provides liquidity and incentivizes informed trading, both necessary elements for financial markets to function efficiently (Grossman and Stiglitz, 1980; Kyle, 1985; Black, 1986; Barber and Odean, 2000). On the other hand, correlated sentiment among retail traders can induce modest transitory price impacts that generate limits to arbitrage (Shleifer and Vishny, 1997; Barber et al., 2008, 2009). A few key features of financial markets have likely driven the historical modesty of retail traders' price impact. Specifically, transaction costs have restricted retail trading to a small portion of market volume. Moreover, correlated sentiment among retail traders was mainly confined to herd behavior or everyday exposure to salient events along with inefficient Bayesian updating (e.g., Banerjee, 1992; Bikhchandani et al., 1998; Barber et al., 2021) rather than deliberate coordination.

While these features previously characterized financial markets, recent innovations have dramatically changed the environment for retail traders. For example, Robinhood's advent of commission-free trading in 2015, followed by major online trading platforms such as Charles Schwab, TD Ameritrade, and E-trade in 2019, relaxed retail trading costs considerably. These events partially explain the exponential growth in retail trading, now responsible for as much as 25% of stock market volume (McCrank, 2021). In addition, social media platforms such as Reddit facilitate direct coordination among retail traders. These evolving characteristics are all well represented in the "GameStop" event in 2021, whereby retail traders joined forces to drive up GameStop's stock price by 3,000% to engineer a short squeeze. The GameStop event raises two critical questions. First, has improved coordination introduced the possibility that entertainment motivates many retail traders rather than profit? Second, is the GameStop event an anomaly, or have these features allowed retail traders to become "the ants that move the log," thus potentially altering their role in financial markets?

Gaining an adequate understanding of these questions will likely require considerable theoretical and empirical analysis and, therefore, is well beyond the scope of a single study. Thus, this paper aims to provide an initial systematic exploration of this topic by employing various novel empirical techniques in various settings with granular data on Robinhood trading activity and interactions among retail traders on Reddit. First, I use the standard logit regression to estimate ex-ante crash probabilities, where a "crash" is defined as the log monthly return lower than -20%.[1] Estimating crash risk by a return threshold is informative. According to Beason and Schreindorfer (2022), 80% of the average equity premium is attributable to monthly returns below -10%. However, crashes defined as over -20% monthly return drop constitute only 5% of all stock returns in the CRSP universe from 1996-2021. Thus, predicting ex-ante crash risk is challenging because of the relatively low frequency of crashes, making it hard to construct valid counterfactuals. I employ a novel machine-learning technique that substantially improves the predictive power of low-probability binary outcomes.

Consistently with prior literature (e.g., Jang and Kang, 2019; Atilgan et al., 2020), ex-ante crash risk is negatively correlated with future stock returns. Specifically, a one-standard-deviation increase in crash risk is associated with an approximately 50 bps drop in monthly risk-adjusted returns. The return predictability remains strong conditioning on other tail risk measures (e.g. VaR in Atilgan et al. (2020)). Moreover, when lagged sentiment is high, the overpricing of high crash-risk stocks is more severe. These results are consistent with the predictions in Brunnermeier et al. (2007), where investors underestimate the left-tail probabilities when sentiment is high and thus buy more than the rational amount. Furthermore, consistent with the theory, I document that Robinhood traders disproportionately buy stocks with high ex-ante crash risk. In contrast, institutional investors tend to sell high crash-risk stocks.

It is hard to determine the direction of causality, which perhaps even cuts both ways.

[1]The -20% cutoff is motivated by prior literature (e.g., Jang and Kang, 2019), and I explore alternative return thresholds in Appendix.

That is, are retail traders merely attracted to high-tail-risk stocks? Or are they part of what creates the tail risk? To partially unpack the potential for the latter channel, building on the recent advancement in social transmission theory (Han et al., 2022), I exploit the history of the social media platform Reddit and the first-time appearances of stock tickers on "Wallstreetbets" as a quasi-natural experiment. Specifically, I use a stacked "difference-in-differences" approach (Gormley and Matsa, 2011; Cengiz et al., 2019) to document a causal effect of investors' online conversations on the ex-ante crash risk of stocks. I partially alleviate the possible endogeneity concerns by carefully constructing a match sample and conditioning on a set of characteristics that draw retail attention. The results show that on average the crash risk of stocks increases by approximately 10% within the first three months of appearance on "Wallstreetbets".

Recent work on social transmission (Hu et al., 2021) shows that the online conversations of retail investors on "Wallstreetbets" contain information that possibly drives future stock prices on a daily basis. To bolster the previous results, I build on this work and construct a novel and plausible instrument for investment-related conversations by utilizing the entire history of Reddit posts. Through an instrumental variable estimation approach, I show that a one-standard-deviation increase in online discussions in "Wallstreetbets" is associated with an approximately 2.3% increase in ex-ante crash risk at a daily frequency, where I follow prior literature (e.g., Bollen and Whaley, 2004; Van Buskirk, 2011; Kim and Zhang, 2014; Kim et al., 2016) and use the option implied volatility $SKEW$ as the proxy for crash risk. These results corroborate the previous "difference-in-differences" framework and suggest that retail investors could cause extreme stock returns via efficient herding.

Have retail traders become the ants that move the log? This paper presents a preliminary analysis to address whether we've reached a paradigm shift in the role of retail traders. There are several unique contributions. First, to the best of my knowledge, this is the first study that conducts causal inference on retail influence on crash risk or left-tail risk. Moreover, this paper proposes a new ex-ante crash risk measure via novel methodologies.

The rest of the paper is organized as follows. Section 2 briefly reviews the existing literature. Section 3 explains the construction of ex-ante crash risk and corresponding results for estimating monthly crash probabilities. Section 4 conducts asset pricing tests for crash risk in the cross-section of stock returns. Section 5 discusses the distorted belief mechanism for the negative price of crash risk. Section 6 documents the causal effect of retail conversations on firm crash risk. Section 7 constructs a novel instrument to provide further evidence on the causal effect of social transmission on crash risk. Section 8 conducts robustness tests. Section 9 concludes.

2. Literature Review

This study is related to an extensive list of areas in literature. First and foremost, it concerns the firm-level crash risk. The corporate finance literature studies the determinants of firm crash risk. These determinants are often motivated by managers hoarding bad news (Jin and Myers, 2006). The idea is that the hoarding delays the information transmission such that when it is ultimately released, there is a sudden drop in the price corresponding to the size of the cumulative bad news. Motivated by this theory, the literature has proposed a list of determinants that could endogenously influence crash risk, such as earnings management (Hutton et al., 2009), tax avoidance (Kim et al., 2011), annual report readability (Li, 2008), CSR (Kim et al., 2014), liquidity (Chang et al., 2016), short interest (Callen and Fang, 2015), and governance (Andreou et al., 2016; An and Zhang, 2013). This paper differs from this literature in that it estimates crash risk at a monthly frequency, by utilizing a rich set of conditional information (Chen and Zimmermann, 2021).

In asset pricing, a rich body of literature extracts information from option prices to determine the size of tail risk. For example, Pan (2002) provides theoretical support for the jump-risk premia implied by near-the-money short-dated options that help explain volatility smirk. Xing et al. (2010) studies the relationship between implied volatility smirks and the

cross-section of stock returns. They show that the difference between the implied volatility of out-of-money put options and at-the-money call options shows strong predicting power for future stock returns. Yan (2011) show that jump size proxied by the slope of volatility smile predicts the cross-section of stock returns. The present study uses option information as one set of variables in predicting crashes, thus exploiting a far richer information set.

The third strand of literature on crash risk directly predicts the probability of crashes. Chen et al. (2001) employs cross-sectional regressions to forecast the skewness of daily stock returns. Campbell et al. (2008) use a dynamic logit model to predict distress probabilities for the cross-section of firms. Conrad et al. (2014) show that high distress risk stocks are also likely to become jackpots. They use a logit model to predict the probability of deaths and jackpots. Jang and Kang (2019) exploits a multinomial logit model to jointly predict probabilities of crashes and jackpots at an annual horizon.

This study is also related to the literature on the relationship between investor trading and market efficiency and bubble formation. De Long et al. (1990a), De Long et al. (1990b), and Abreu and Brunnermeier (2003) provide the theoretical support to and empirical evidence of positive feedback traders and their potential impact on market. Retail investors are believed to be "noise traders" that trade too much (Barber and Odean, 2000). Speculative retail traders tend to chase lottery-like stocks, experiencing subsequent negative trading alpha, and affect stock prices accordingly (Han and Kumar, 2013). Recent evidence from "Robinhood Traders" shows that they tend to herd more on extreme past-return stocks, which are more attention-grabbing (Barber et al., 2021), while there is also evidence that mimicking portfolios based on the characteristics of "Robinhood Traders" do not seem to underperform the market, but instead could be a market stabilizing force (Welch, 2020). On the pricing impact of retail trading, Foucault et al. (2011) was one of the first papers that use a quasi-natural experiment to identify the causal effect of retail trading on stock volatility.

Finally, this study is related to the emerging literature that studies the implications and applications of machine learning methodologies in asset pricing. They are mostly concerned

with resolving the "factor zoo" problem (Kozak et al., 2020; Feng et al., 2020; Bianchi et al., 2021; Gu et al., 2020).

3. Data and Estimation of Crash Risk

I use two sets of measures for ex-ante crash risk, one monthly measure, and one daily measure. The monthly measure is the ex-ante probability of stock crashing in a certain month, while the daily measure $SKEW$ is motivated by Xing et al. (2010), and defined as the difference between the implied volatility of out-of-money put option and that of the at-the-money call option.[2] I will start by describing the monthly measure and defer the discussion of the daily measure to Section 7.

3.1. Estimation of Monthly Ex-Ante Crash Risk

I define firm-level crashes as stock monthly log returns lower than -20%. The choice is reasonable in the following sense. Prior literature uses log annual returns of -70% as the cutoff points (Conrad et al., 2014; Jang and Kang, 2019). The unconditional probabilities of crashes defined this way at the annual frequency are roughly 5%. At a monthly frequency, a cutoff point at -20% agrees with this distribution. Thus the universe of stock returns falls into two categories – crashes and otherwise. Then the monthly ex-ante crash risk is defined as follows:

$$CrashRisk_{i,t} = E[P(r_{i,t} < -20\%)|X_{i,t-j}] \tag{1}$$

Where r is the monthly log return. $j \in [1, 2, 3, 4, 5, 6]$ is the months in each training window, or in other words the period we draw conditional information. X is a set of firm-level predictors.

Estimating the ex-ante probabilities of future crashes naturally calls for a logistic regres-

[2]The $SKEW$ measure by Xing et al. (2010) is widely used in the corporate finance literature as a proxy for firm crash risk. See for example...

sion, where the dependent variable is a binary response D_{crash}, where it equals one if the log monthly return is lower than -20%, and zero otherwise. A critical issue arises, however, in forecasting rare events such as crashes. The usual logistic estimator could produce sub-optimal results due to the poor finite sample properties (King and Zeng, 2001). I provide a simple intuition for this argument in Appendix B.1. Though the difficulties and the associated statistical issues in forecasting rare events are rarely studied in economics, the remedy is readily available in machine learning literature. I follow Jiang et al. (2020) and introduce an Ensemble method, "Easy Ensemble" (EEC), that combines random undersampling and bootstrapping (Liu et al., 2008) to supplement the logistic regression approach. A detailed discussion of this technique can be found in Appendix B.2.

To estimate the ex-ante probabilities of a crash, it is essential to conduct out-of-sample procedures. Thus I use a rolling window of 6 months to estimate parameters and fit the following month to produce an OOS estimate of crash risk. With respect to the independent variables, in a slight departure from prior literature, I choose a large set of characteristics that have been shown as return predictors as the independent variables in the estimation process. Specifically, I use variables obtained from Chen and Zimmermann (2021). These are monthly firm-level characteristics that have been shown in the literature as important drivers of future returns, and these variables encompass all variables that were considered as predictors of crashes (Campbell et al., 2008; Conrad et al., 2014; Jang and Kang, 2019).

I limit the data scope to between 1996 and 2020, both to reduce the computation load and to ensure maximum data usage, as some variables are only available from 1996 (for example, option variables). Therefore, with 6-month rolling windows for training, our out-of-sample prediction starts from July 1996 to December 2020, comprised of 294 months. I use CRSP for monthly stock returns. I require common stocks with a share code of 10 or 11 and with prior month-end stock prices greater than \$5 to avoid extreme outliers.

Next, I compare the usual logistic estimator with the EasyEnsemble method in forecasting performances. To illustrate the performance difference, I conduct the following experiment.

For the whole sample, I plot the percentages of real crashes predicted by either model against a decision threshold from zero to one, meaning that at each threshold, all stocks with a predicted probability higher than that would be labeled "crash". The results are shown in Figure 1.

[Fig. 1 about here.]

Note that EasyEnsemble outperforms logistic regression in the low threshold region. This result is desirable because we know that crashes are low-probability events (the unconditional probability of a crash is around 5-6%), and we want the classifier to do well in this region. For example, at the 7% threshold, meaning that we predict all stocks with a probability estimate greater than 7% to crash in the next month, logistic regression is able to capture 72% of all real crashes, while EasyEnsemble is able to capture 85%.

3.2. Summary Statistics

Given the refined estimate of monthly ex-ante crash risk, we can examine its relationship with firm characteristics. In particular, we are interested in the relationship between the risk and the underlying regressors. We summarize the relationship between the machine learning-generated crash risk and the top regressors in Appendix. The summary statistics of both logit-generated ex-ante crash risk and machine learning-generated crash risk, along with all relevant stock characteristics and other data used in later analyses, are presented in Table

[Table 1 about here.]

On top of firm-level crashes, the aggregate probability of a market crash is of great interest to researchers and practitioners alike. Although one can argue that the aggregate stock market crash is systematic, while firm-level crashes are more idiosyncratic in nature, aggregating firm-level crash probabilities might still contain information about the aggregate

crash risk. One possible reason for this logic is that we use a fixed threshold (-20% log return) to define crashes, and thus aggregating these firm-level probabilities contains a systematic component. Therefore, I aggregate monthly firm-level crash risk to the market level by their lagged market capitalizations and plot the series in Figure 2.

[Fig. 2 about here.]

On top of the aggregate crash risk series, I also plot NBER recession periods (NBER, 2021) in the gray shaded areas. Though not immediately clear, the series does contain some information about future possibilities of market crashes, as there are signs of spikes ahead of or during recession periods. Next, we move on to examine the pricing implications of firm-level monthly crash risk.

4. Monthly Crash Risk and Stock Returns

In this section, I examine whether the ex-ante monthly crash risk is priced in the market. I conduct both time-series portfolio analysis and cross-sectional analysis. Prior literature (Conrad et al., 2014; Jang and Kang, 2019; Atilgan et al., 2020) has indicated that crash risk, or left-tail risk, is negatively priced in the market. Though my measure is different in its time frequency and construction, we should expect similar behavior.

4.1. Portfolio Analysis

At the end of each month, I sort stocks into ten decile portfolios based on their estimated ex-ante crash probabilities. Then I compute both value-weighted and equal-weighted excess returns of each portfolio and the hedge portfolio that long high crash risk decile portfolio and short low crash risk decile portfolio. I regress the time series of returns on various asset pricing factors and compute the alpha estimates and their associated T-statistics. The asset pricing models include: CAPM, Fama-French three-factor model (FF3) (Fama and French,

1993), then augmented with a momentum factor (FF4) (Carhart, 1997), Fama-French five-factor model (FF5) (Fama and French, 2015), and then augmented with momentum factor (FF6). To show the consistency of the results and the superiority of the EasyEnsemble method, I show alpha estimates using both logistic regression and EasyEnsemble in Table 2.

[Table 2 about here.]

As shown in Table 2, when we long top crash risk decile portfolio and short bottom decile portfolio, we produce consistent and significant negative alphas across different asset pricing models, equal-weighted or value-weighted, with T-statistics of magnitude well over 3. Note also that when we compare the results from using logit-generated crash risk and machine learning-generated crash risk, the latter shows superiority in both the magnitude of alpha and the T-statistics. This is a strong piece of evidence that machine learning not only produces consistent results with conventional methods but also demonstrates better forecasting efficacy, as it classifies correctly more actual crashes that contribute to lower returns in the subsequent month.

4.2. Cross-Sectional Regressions

Next, I run Fama-MacBeth cross-sectional regressions (Fama and MacBeth, 1973) following the procedure in Fama and French (2020). Each month, I regress raw stock returns on cross-sectionally standardized lagged firm characteristics. Then I average the coefficients to arrive at the final estimates. The coefficients on characteristics can be directly interpreted as average priced return spread for one standard deviation increase of the corresponding firm risk. I include common risk characteristics such as the natural log of market capitalizations, natural log of book-to-market ratio, asset growth, gross profitability, momentum (prior 11-to-1 month returns), short-term reversal (prior 1-month returns), and my estimated crash probabilities from the Ensemble method. On top of these variables, I control for a set of anomaly characteristics that are shown to be significantly correlated with future stock re-

turns: idiosyncratic volatility, illiquidity (Amihud, 2002), market beta, tail Beta (Kelly and Jiang, 2014), coskewness(Harvey and Siddique, 2000), and net operating assets NOA (Hirshleifer et al., 2004). Bali et al. (2011) proposes a measure MAX that represents investors' preference for lottery-like payoffs. MAX stands for the maximum daily return achieved by each stock in the prior month. To see if the estimated crash risk carries additional information that distinguishes it from MAX, I add the MAX measure as a control variable in the Fama-MacBeth regressions.

Atilgan et al. (2020) also studies the left-tail risk, although their measure is constructed differently. Their "value-at-risk" (VaR) is entirely based on historical returns and is defined as the return conditioning on probability distribution, which differs from our measure that takes return cutoff as given and estimates ex-ante probabilities. To see whether our crash risk contains incremental information about future stock returns than the VaR measure, I include VaR as a control variable. The VaR measure is the negative of 1 percentile daily return of the stock in the past year. I report the regression results in Table 3.

[Table 3 about here.]

Table 3 suggests several points. First, both logit-generated ex-ante crash risk and machine learning-generated crash risk are significantly and negatively correlated with future stock returns, and their magnitudes are very similar to each other. Second, the loadings on crash risk are robust even after controlling for common risk characteristics and go beyond a plethora of tail risk-related variables, including the lottery-payoff proxy MAX (Bali et al., 2011). Third, when our crash risk is not included in the regression, the VaR measure is significantly and negatively correlated with future stock returns, consistent with the results in Atilgan et al. (2020). However, when our crash risk is included in the regression, the loading on VaR becomes insignificant, while our crash risk measure loads negative and significant consistently. This suggests that both our logit-generated and machine learning-generated crash risk measures contain more information than VaR and consequently subsume its effect. Depending on the control variables and the measure we use, a one-standard-deviation

increase in ex-ante monthly crash risk is associated with approximately a 45-51 bps drop in subsequent risk-adjust returns, which translates into -5.47% to -6.12% in annual risk-adjusted returns. These results corroborate the prior literature that ex-ante crash risk is negatively priced, and also provide strong evidence that our crash risk measure contains richer and incremental information than existing crash risk measures.

5. A Possible Economic Mechanism: Distorted Belief

The negative price of crash risk does not agree with rational expectations, as a rational investor would naturally demand a positive risk premium for holding such risk. Prior literature attempts to explain the phenomenon via several arguments. One argument is the limits to arbitrage (Shleifer and Vishny, 1997; Conrad et al., 2014; Jang and Kang, 2019). They show evidence that institutional investors tend to "ride the bubble" as rational speculators, instead of trading against crash risk as rational arbitragers, since high crash risk stocks tend to be small, illiquid, and hence costly to short. The second argument is that investors underestimate the momentum in the left tail (Atilgan et al., 2020), meaning that stocks that crashed the last month may well be highly possible to continue crashing in the subsequent month. Investors somehow fail to understand this dynamic and "bought the dip", which renders the stocks with high crash probabilities overpriced. However, it is unclear why this momentum exists. Moreover, since the VaR measured used Atilgan et al. (2020) is an ex-post measure, it does not answer the question from an investor behavior perspective. A third argument pertains to the observation that stocks with extreme past returns are attention-grabbing, and retail investors have a preference for such stocks (Barber and Odean, 2008; Barber et al., 2021). However, it is reasonable to assume that investors are drawn to extreme past winners, as they might be over-extrapolating past returns. It is nonetheless puzzling why investors should prefer extreme past losers. Moreover, it is difficult to understand why investors should prefer high left-tail stocks. Even if they underestimate the momentum in

the left tail, these are undesirable stocks from a risk-return tradeoff standpoint. In addition, over time, investors should be able to learn from past observations that high crash risk stocks are overpriced, as many of them indeed crashed in the subsequent month.

The literature in behavioral theories provides valuable guidance in terms of investor beliefs and preferences towards crash risk or left tail risk. Two theories, in particular, have clear predictions about investors' attitudes towards the left tail. One is cumulative prospect theory (CPT) by Barberis and Huang (2008). They show that investors with a CPT preference would overweight small probability events. One example is that people would gamble on slim chances of big payoffs, but buy insurance for plane crashes. The implication is that investors with CPT preference should shun high crash risk stocks since they effectively deem those crashes more likely to happen than the true distribution. If all investors have such a preference, high crash risk stocks should be underpriced, and thus produce a positive risk-adjusted return. This prediction does not seem to conform to the empirical observation.

The second theory is the optimal expectations theory (OET) by Brunnermeier et al. (2007). They show that investors may derive anticipatory utility when holding an optimistic subjective belief about stock returns, even though such beliefs prove to be wrong afterward. If investors hold such a belief, they would effectively shift their subjective return distribution to the right when their sentiment is high. The implication is that when sentiment is high, investors with such beliefs tend to think that crashes are less likely than reality, and thus overbuy high crash risk stocks. The pricing implication is that crash risk or left tail risk is overpriced and thus predicts a negative risk-adjusted return.

The evidence presented in this paper and the prior literature for the negative price of crash risk agrees with the optimal expectations theory. To further establish evidence as to whether investors overbuy high crash risk stocks when their sentiment is high, I conducted several tests to provide additional evidence.

5.1. Crash Risk Portfolio Returns and Sentiment

First, I examine the relationship between the crash risk hedge portfolio returns and sentiment. If investors hold optimal expectations, then the loss on the crash risk high-minus-low hedge portfolio would be higher when lagged sentiment is high, since investors' belief distortion would be more severe during such periods.

I follow Baker and Wurgler (2006) and use their sentiment index as a proxy for the market-wide sentiment. In particular, I use the sentiment measure that is orthogonal to macroeconomic indicators to alleviate the impact of market risks. Since their index is available up to the year 2018, my sample is hence limited between July 1996 and December 2018. Then I divide the sample period into two subperiods, where one is the high sentiment period when sentiment is higher than the median value of the whole sample, and another is the low sentiment period. Then I compute the excess returns of the top decile portfolio, the bottom decile portfolio, and the long-short hedge portfolio that long high crash risk stocks and short crash risk stocks, in each of the subperiods. I then compute the differences in these returns between high and low sentiment periods. The results are summarized in Panel A of Table 4.

[Table 4 about here.]

It is immediately clear from the table that the high-crash-risk stocks experience the lowest returns after a high sentiment period when mispricing is most severe, while they do not show negative returns on average after low sentiment months. On the other hand, there is no statistically significant difference between high and low sentiment periods for low-crash-risk stocks. On the whole, a long-short strategy that is long high-crash-risk stocks and short low-crash-risk stocks produces more negative and significant excess returns after high sentiment months. These results are consistent with our hypothesis that when investors are bullish, they are more likely to overbuy high-crash-risk stocks, and thereby the expected returns of these stocks would be lower.

To further examine the relationship between crash risk and sentiment, I run Fama-

MacBeth regressions and panel regressions of stock returns on firm characteristics for high- and low-lag-sentiment months separately. The hypothesis is that the price of crash risk should be more negative immediately after high sentiment months. As before, high sentiment months are defined as those months with lag sentiment higher than the sample median, and low sentiment months are defined as those months with lag sentiment lower than the sample median. The results are reported in the first two columns in Panel B of Table 4.

We can see from the table that when lagged sentiment is high, the coefficient on crash risk is -0.619%, compared to -0.405% when lagged sentiment is low. In other words, the price of crash risk associated with a one-standard-deviation increase in the risk is 21 bps lower when lagged sentiment is high. Though the difference between the two coefficients is not statistically significant (T-statistic of -1.2), the annualized return difference is large at -2.52%. This is another piece of evidence that high crash risk stocks are more overpriced when lagged sentiment is high.

To further assess this phenomenon, I also conduct the following analysis. I define a dummy variable $SentD$, where it equals one if the lagged sentiment is higher than the sample median, and zero otherwise. I first run a panel regression of stock returns on crash risk and other firm characteristics, with firm and time fixed effects. Then I include the $SentD$ variable and interact it with crash risk. The hypothesis is that the interaction term should be significantly negative since when lagged sentiment is high, the overpricing of high crash risk stocks should be more severe. I report the results in Columns (3) and (4) in Panel B of Table 4.

As shown in the table, even after including firm and time-fixed effects, the ex-ante crash risk is consistently priced negatively, albeit with a smaller magnitude. In Column (4), when we interact the sentiment dummy with crash risk, the loading on crash risk is much smaller in magnitude and statistically significant at 5% level, while the coefficient on the interaction term is negative and statistically significant at 1%, with a much higher magnitude. These results are consistent with our hypothesis that when lagged sentiment is high, investors buy

more high crash risk stocks, which causes the overpricing of these stocks even higher, and therefore the subsequent returns turn out to be much lower than in low lagged sentiment periods.

5.2. Trades on Crash Risk

Next, we examine whether some investors are likely to buy high-crash-risk stocks. This hypothesis is the underlying assumption of the previous literature that high-crash-risk stocks are overpriced and is an implication from (Brunnermeier et al., 2007). To explore this hypothesis, I first use Robintrack data to construct a retail trading measure and examine whether they tend to buy high-crash-risk stocks.[3]

As has been extensively discussed in Barber et al. (2021) and Welch (2020), Robintrack data contains hourly stock popularity numbers that are measured by how many users on Robinhood hold a particular stock at a certain hour. Since we cannot observe the number of shares they hold for each stock, and there is no data for the total number of users for each time period, the next best solution is to measure the change in the number of users for each stock. As crash risk is estimated at a monthly frequency, I use month-end numbers of Robinhood users to merge the data. I first construct a log measure for Robinhood trading:

$$Change\,in\,Log(\#User_{i,t}) = \log(\#User_{i,t}) - \log(\#User_{i,t-1}) \tag{2}$$

Then I follow Barber et al. (2021) and construct a percentage change measure for Robinhood trading:

$$\%Change\#User_{i,t} = \#User_{i,t}/\#User_{i,t-1} - 1 \tag{3}$$

Where t is at the monthly frequency to match the frequency of our ex-ante crash risk

[3]Robintrack: `https://www.robintrack.net/`.

measures. The specification is as follows:

$$Robinhood\,Trade_{i,t} = \alpha_0 + \beta \times Crash\,Risk_{i,t} + \sum_p \beta_p Control_{p,i,t-1} + \alpha_i + \lambda_t + \epsilon_{i,t} \quad (4)$$

Where we add firm and time fixed effects to account for unobserved heterogeneity that might be correlated with the error term. The Robinhood sample runs from May 2018 to August 2020. I regress the Robinhood trading measures on both measures of ex-ante crash risk, controlling for the lagged log of the user number and a set of firm characteristics. The results are reported in Columns (1) to (4) of Table 5.

[Table 5 about here.]

The table shows that over the sample period when Robinhood data is available, retail investors on average tend to buy high-crash-risk stocks, consistent with our hypothesis. Importantly, in all specifications, we control for such commonly used lottery characteristics as MAX and MIN (Bali et al., 2011), which are defined as maximum and minimum daily returns of the previous month, and total skewness of the previous month. The coefficient on crash risk is consistently and significantly positive in Robinhood trading tests, meaning that retail preference for high-crash-risk stocks goes beyond the conventional proxies for lottery characteristics defined in the literature (Barberis and Huang, 2008; Bali et al., 2011).

A related question arises as to whether institutional investors would be liquidity providers and act as counterparties since literature has shown that they are reluctant to short the left tail, and would rather ride the bubble. I examine this issue by regressing the change of institutional holdings on the same set of characteristics. The institutional holdings data comes from Thomson Reuters 13F filings data and is defined as the percentage of shares held by institutional investors. The change in the holdings is the difference between the current quarter's holdings and the previous quarter's. The results are shown in Columns (5) and (6) in Table 5.

The results show that there is strong evidence that institutions might be the counterparty of retail investors for crash risk. In sharp contrast to Robinhood trading results, the coefficients on both crash risk measures are negative and statistically significant for institutional trading tests. Taken together, these results support the hypothesis that retail traders derive anticipatory utilities from distorted subjective beliefs. Consistent with the predictions in Brunnermeier et al. (2007), when lagged sentiment is high, investors underestimate left-tail risks and tend to overbuy stocks with high crash risk, which in turn drives up their prices, leading to lower expected returns subsequently. Both the pricing results and retail trading results conform to this theory.

6. Retail Influence on Monthly Crash Risk

Evidence from the previous section shows that retail investors tend to buy high ex-ante crash risk stocks, and this effect is over and beyond the effect of the usual proxies for lottery characteristics. These buying activities could be inconsequential if retail investors are pure "noise traders" (De Long et al., 1990a), as their trades are idiosyncratic and would be canceled out on average. However, when their trades are correlated because of attention or herding, they could forecast subsequent returns (Barber and Odean, 2008; Barber et al., 2021). Social media is instrumental in facilitating herding behavior, as it transmits trading strategies more efficiently. As implied in Han et al. (2022), there is an inherent feedback loop in correlated trading and asset prices. When investors (receivers) take note of other investors' (senders) recent trading success, as demonstrated by their bragging on social media of the high recent returns of their stock picks, they continue to trade in the same direction, thus pushing the stock price even higher. The implication is that regardless of whether investors display a preference for skewness, their trading actions would produce such results and influence stock prices.

There is causal evidence that suggests higher participation by retail investors does induce

higher stock volatility (Foucault et al., 2011). They may be marginal price setters for small stocks (Graham and Kumar, 2006). Retail short sellers predict negative future returns, and they seem to have superior knowledge of small firm fundamentals (Kelley and Tetlock, 2017). Much of the literature focuses on predictive tests, as it is extremely difficult to find ideal settings for the proper identification of causality.

I explore a particular shock to the retail attention and herding channel that might have influenced retail investors' trading behavior, which in turn could drive the change in the crash risk of the underlying stocks.

6.1. The Advent of Wallstreetbets

"Wallstreetbets" is a "Subreddit" on the social media platform "Reddit", and has garnered considerable attention from the investment community largely because of the "GameStop" saga. The Subreddit started in April 2012, and today it has over 12 million subscribers. These subscribers call themselves "degenerates", and frequently exchange trading ideas and post their gains and losses. In a recent study, Hu et al. (2021) shows that conversations on "Wallstreetbets" have information content that predicts next-day returns. A study from a different discipline, Li and Wu (2018) shows that retailers displaying past sales numbers can induce consumers to herd and buy more of the products. These studies suggest that social media as a platform for idea sharing can facilitate more efficient herding. Therefore, it is conceivable that the advent of a highly efficient platform for sharing ideas might affect asset prices, including the crash risk of the underlying stocks, following the results that retail investors exacerbate the overpricing of high-crash-risk stocks.

I examine this issue by tracing back to the origin of "Wallstreetbets" when it was founded in April 2012. I obtain and process all posts from April 2012 when the Subreddit started till December 2020, and find out all stock tickers that were mentioned in these posts.[4] I drop all ticker names that are also common English words, slang, and abbreviations. To illustrate

[4]The complete history of Reddit comments data comes from `https://files.pushshift.io/reddit/comments/`.

the growing community on "Wallstreetbets", I plot the number of posts each month that mention ticker names, and also the number of unique ticker names mentioned each month in Figure 3.

[Fig. 3 about here.]

Panel A of Figure 3 plots the number of posts that mention ticker names on the Subreddit "Wallstreetbets", and Panel B plots the number of unique tickers/firms each month. The time series spans from April 2012 when "Wallstreetbets" was started to December 2020. It shows that the activities on "Wallstreetbets" exploded after the pandemic began in 2020. It also shows the growing breadth of retail investor interests in the number of stocks.

6.2. The Staggered First Appearances of Stock Tickers

Members started to mention stocks in their posts on "Wallstreetbets" on the first day of the Subreddit. According to Han et al. (2022), people are more likely to mention certain stocks if these stocks happen to have high past returns. If other people see these posts, they are more likely to follow suit and trade in the same direction. This could in turn affect the stock returns.

To test this hypothesis, I focus on the seven-month window around each "event", where "event" means a stock ticker appeared for the first time on "Wallstreetbets". Thus there are three months pre-event, and three months post-event. Since conversations about stocks are not exogenous per se, we need a matching strategy and control variables that can offset the endogenous portion of the test. Therefore, I use propensity score matching by running logistic regression. The response is a dummy variable $D_{i,t} = 1$ if a stock i appears on "Wallstreetbets" for the first time at time t, or zero otherwise. The independent variables include lagged market capitalization, prior-month return, asset growth, book-to-market ratio, gross profitability, idiosyncratic risk, illiquidity, MAX, and prior 12-month return, to proxy for the common stock characteristics.

The estimated parameters are then fit to the whole sample to generate fitted values as the propensity score for each stock at each point in time. To match each event, I use the score generated for each "never treated" stock three months prior to the event and find five stocks that have the closest propensity scores to each treated stock.[5]

After the matching process, I follow Gormley and Matsa (2011); Cengiz et al. (2019) and stack each event cohort, where each cohort contains the treated stock and the matched sample. Then I run the following specification:

$$Crash\,Risk_{i,c,t} = \gamma_0 + \beta D_{i,c,t} + \delta_{c,t} + \alpha_{i,c} + \sum_p \beta_p Control_{p,i,t-1} + \epsilon_{i,t} \tag{5}$$

Where $Crash\,Risk_{i,c,t}$ is the estimated crash risk of stock i in cohort c at time t. $D_{i,c,t}$ is a dummy variable that indicates whether a stock i in cohort c is treated at time t. $\delta_{c,t}$ is $Cohort \times Time$ fixed effects. $\alpha_{i,c}$ is $Unit \times Cohort$ fixed effects. Then β is the coefficient of interest that estimates the average treatment effect on the treated stocks. The results are reported in Column (1) and Column (3) of Table 6, where Column (3) adds control variables. The control variables include the natural log of market capitalization, prior-month return, asset growth, gross profitability, illiquidity (Amihud, 2002), MAX (Bali et al., 2011), prior 12-month return, and idiosyncratic risk. Standard errors are clustered at the unit level.

[Table 6 about here.]

When control variables are not included, there is a 1.03 percentage point estimated increase in logit-generated crash risk when a stock is first mentioned on "Wallstreetbets", and the coefficient is highly statistically significant. When control variables are included, the magnitude reduces to approximately 56 bps, and the coefficient remains statistically significant at the 1% level. This corroborates our hypothesis that when a stock was mentioned on social media and subsequently draws more attention that possibly induces more correlated retail trading, which could increase stock crash risk.

[5] "Never treated" means the stock never appears on "Wallstreetbets". This is to ensure the cleanest matching. There are in total 2,276 unique stocks that are never mentioned on "Wallstreetbets".

A critical assumption for the difference-in-differences analysis is the "parallel trend" assumption, where the treated group and the control group should not have significant differences before the event happens. To examine this "parallel-trend" assumption, I conduct a dynamic approach, where instead of examining the coefficient on the treatment dummy, I run the following specification:

$$Crash\ Risk_{i,c,t} = \gamma_0 + \sum_{j=-3}^{+3} \beta_j D_{i,j,c,t} + \delta_{c,t} + \alpha_{i,c} + \sum_{p} \beta_p Control_{p,i,t-1} + \epsilon_{i,t} \qquad (6)$$

Where the dummy variables $D_{i,j,c,t}$ indicate whether a stock i is treated in cohort c at time t, and the distance $j \in [-3, 3]$ from the current month to the treatment month. Month -1 is chosen as the base month that will be omitted from the regression. The results are included in Column (3) and (6) of Table 6.

As shown in the table, the coefficients for the two months before the event are economically and statistically insignificant. On the other hand, the coefficients on the treatment month and the months after the treatment are economically and statistically significant. These results provide strong support to the assumption that there are no significant differences between treatment and control groups before the treatment.

To provide further evidence of the "parallel trend" assumption, I also plotted the coefficients on the dummy variables $D_{i,j,c,t}$ with their 95% confidence intervals in Figure 4.

[Fig. 4 about here.]

The figure provides visual support for the "parallel trend" assumption for our "difference-in-differences" analysis. The dynamic results, together with the static results, provide strong evidence that there is a possible causal effect of increased retail attention on stock crash risk.

6.3. Size and Institutional Ownership

Foucault et al. (2011) show that retail investors have an outsized impact on stock volatility, especially for smaller stocks, where the standard limits to arbitrage argument apply

(Shleifer and Vishny, 1997). Smaller stocks are traded thinly and thus are less liquid. Because of their price tag, they are usually the preferred habitat of retail investors, and thus their institutional holding is usually lower. As a result, their prices can stay distant from their fundamentals for an extended period of time, since rational investors are reluctant to arbitrage for the arbitrage would be costly.

The same argument should apply to crash risk. Prior literature has shown that high crash risk stocks tend to be smaller and more costly to arbitrage (Jang and Kang, 2019). We have also shown in Section 5 that retail investors seem to display a preference for high crash risk stocks possibly because of their distorted beliefs (Brunnermeier et al., 2007). The combination of these factors should lead to a natural hypothesis that retail attention should have an outsized impact on the crash risk of smaller stocks and stocks with lower institutional ownership.

To examine this hypothesis, I divide the universe of stocks into two subgroups based on either lagged size or institutional ownership. Then I define a dummy variable $D_{size/io} = 1$ if the stock is larger than the median or zero otherwise, based on the lagged value of each stock three months prior to each event. In the case of institutional ownership, $D_{size/io} = 1$ if the ratio of institutional ownership for the stock is greater than the median or zero otherwise, based on the lagged value of institutional ownership three months prior to each event. Then I interact $D_{size/io}$ with the $Treated$ dummy variable in the same "stacked difference-in-differences" specification:

$$Crash\,Risk_{i,c,t} = \gamma_0 + \beta_1 D_{i,c,t} + \beta_2 D_{i,c,t} \times D_{size/io} + \delta_{c,t} + \alpha_{i,c} + \sum_p \beta_p Control_{p,i,t-1} + \epsilon_{i,t} \quad (7)$$

I report the results of this specification in Table 7. Columns (1) to (4) report the results of using logit-generated crash risk as the dependent variable, while Columns (5) to (8) use machine learning-generated crash risk as the dependent variable. Columns (1), (3), (5), and (7) only include the treated dummy and the interaction between the treated and the size

dummy or IO dummy. Columns of even numbers add control variables. Standard errors are clustered at the unit level.

[Table 7 about here.]

Consistent with our hypothesis, the coefficient on the interaction term between the treated and size dummy or the IO dummy is negative and economically, and statistically significant. For example, as shown in Column (1) when controls are not included, if the stock is below median size, the first appearance on "Wallstreetbets" increases stock crash risk by 1.5 percentage points, much higher than our baseline estimate of 1.03%. If the stock is above the median size, the effect is much smaller at approximately 57 bps. In column (2) when control variables are included, being a small stock that first appears on 'Wallstreetbets" leads to a 1.04 percentage points increase in crash risk. The interaction term between *Treated* and the size dummy remains significantly negative. The results are consistent when using institutional ownership as the main variable of interest. These results are consistent with prior literature that retail investors have a higher impact on smaller stocks or stocks with a lower level of institutional ownership.

6.4. Supporting Evidence from Trading Volume and Volatility

One necessary assumption for our analysis is that retail investors pile in the stocks that are mentioned on social media. While we do not have individual trading data, there should be a surge in trading volume and volatility (Foucault et al., 2011) around the events. To examine whether this is the case, we re-run the "difference-in-differences" analysis but substitute the dependent variable with trading volume and return volatility, where trading volume is defined as the monthly total volume of shares traded scaled by total shares outstanding, and volatility is defined as daily return volatility of the current month. The results are reported in Table 8.

[Table 8 about here.]

The table shows clearly that there is a significant surge in both trading volume and return volatility in the treated stocks that first appeared on "Wallstreetbets". Moreover, the dynamic tests confirm that there is no evidence that the "parallel trend" assumption is violated. In fact, before the event happens, there is a downward trend for the treated stocks in terms of trading volume and return volatility. This can be more readily shown in Figure 5.

[Fig. 5 about here.]

Taken together, these results support our main analysis that heightened retail attention as a result of social transmission leads to higher ex-ante crash risk. Moreover, there is evidence that retail activities are behind the surge of trading interests in these stocks.

7. Retail Traders and Crash Risk: Daily Evidence

In this section, we approach the main questions using the daily data by exploring the $SKEW$ measure by Xing et al. (2010), which is widely used as a proxy for firm-level crash risk (Bollen and Whaley, 2004; Van Buskirk, 2011; Kim and Zhang, 2014; Kim et al., 2016). It is motivated by the notion that a volatility smirk indicates investors' expectation of a steep decline in the underlying asset value (Bates, 2000).

Using $SKEW$ as a proxy has the following advantages. First, it is available at a daily frequency for stocks that have options traded. Second, it is easy to compute as it only relies on implied volatility. Third, it is ex-ante in nature and thus conforms to our purpose. Formally, $SKEW$ is defined as follows:

$$SKEW_{i,t} = ImpliedVol_{i,t}^{OTM-Put} - ImpliedVol_{i,t}^{ATM-Call} \tag{8}$$

Following Xing et al. (2010), I screen the options based on the following criteria. Days to expiration are between 10 and 60 days. Implied volatilities are between 0.03 and 2. Open

interest must be greater than zero. Option price must be greater than \$0.125. Volume is non-missing. For out-of-money put options, the moneyness is between 0.8 and 0.95. For at-the-money call options, the moneyness is between 0.95 and 1.05. We choose the implied volatility of the put option with moneyness closest to 0.95, and the implied volatility of the call option with moneyness closest to 1 to compute the $SKEW$ measure for the day.

7.1. SKEW and Daily Returns

Xing et al. (2010) show that $SKEW$ is significantly negatively correlated with future weekly returns. To test whether this is the case in the daily frequency and to check whether daily $SKEW$ can be used as a suitable proxy for ex-ante crash risk, we need to examine whether $SKEW$ is significantly negatively correlated with future daily returns.

Therefore I follow Hu et al. (2021) and use the following specification:

$$R_{i,t} = \alpha + \beta SKEW_{i,t-1} + \sum_{p} \beta_p Control_{i,p,t-1} + \lambda_t + \epsilon_{i,t} \tag{9}$$

Where t is at a daily frequency. The control variables include prior day return, prior month-end log of market capitalization, book-to-market ratio, cumulative 19-day returns lagged for 2 days (reversal), cumulative 100-day returns lagged for 21 days (momentum), prior month average trading volume scaled by total shares outstanding (liquidity), and prior month volatility of daily returns. For robustness, I run both Fama-MacBeth regressions and panel regressions and report the results in Panel A of Table 9.

[Table 9 about here.]

Panel A of Table 9 shows that throughout all specifications, the $SKEW$ measure is negatively correlated with future daily stock returns, which is statistically significant at the 1% level. These results corroborate the findings in the prior literature and provide support for using $SKEW$ as a valid proxy for ex-ante crash risk at the daily frequency.

7.2. Retail Trading of SKEW

In section 5, we show that retail investors have a tendency to buy high ex-ante crash risk stocks. To see whether this is also the case in the daily frequency, we again use the trading measure derived from Robintrack to regress on the contemporaneous $SKEW$ measure and the same set of control variables that we used in the previous test. We regress retail trading measures on the contemporaneous $SKEW$ measure instead of the lagged measure because we want to examine retail trading behavior on the "ex-ante" measure of crash risk. The results are reported in Panel B of Table 9.

To control for common market-wide shocks, we follow the prior specifications and include day fixed effects and cluster standard errors at the stock level. From Panel B of Table 9, we see that both regressions using different measures for retail trading load positively and significantly on the contemporaneous $SKEW$, the proxy for ex-ante crash risk measure. These results are consistent with our prior monthly results that retail investors tend to overbuy high crash-risk stocks.

Apparently, these results only report the positive correlation between crash risk and retail trading, while the causality can go both directions, just like in the monthly case. To see whether retail behaviors have a real influence on ex-ante crash risk, we turn to online conversations in "Wallstreetbets" again but follow a different path. We want to examine whether the intensity of daily conversations about certain stocks can have a significantly positive impact on the ex-ante crash risk of these stocks.

7.3. Online Conversations and SKEW: Endogeneity

Apparently, online conversations about stocks are endogenous. As shown in Han et al. (2022), agents receive prominent presentations of other agents' trading strategies, typically represented by high past returns, and thus follow the same strategy, which leads to feedback on the stock returns. Because of this feedback loop, it's impossible to separate the two legs of the circle via the usual regression specifications.

Specifically, consider the following specification, where we regress the $SKEW$ measure on the number of times each stock is mentioned on social media, controlling for a set of stock characteristics.

$$SKEW_{i,t} = \alpha_0 + \beta SocialTransmission_{i,t-1} + \sum_p \beta_p Control_{i,p,t-1} + \lambda_t + \sigma_i + \epsilon_{i,t} \quad (10)$$

In a slight abuse of notation, the $t-1$ in the subscript of "Social Transmission" means the pre-trading hours from 16:30 PM on the previous day to 09:00 AM on the current day, while the $t-1$ in the controls ranges from the previous day to previous month, depending on the variable referred to. In this specification, even when we use the two-way fixed effects estimator, "Social Transmission" is still correlated with the idiosyncratic error term, and thus the estimate of the coefficient β is inconsistent.

7.4. A Plausible Instrument

Let's consider the following scenario. Person A zones away during his long and boring working hours by wandering aimlessly on social media. His/her favorite venue for wandering is Reddit, a popular platform for talking about anything. Each sub-venue specializing in a different topic is called a "Subreddit", a symbol of rich social life in a society. Apart from working, person A spends a tremendous amount of time on hobbies such as football, fishing, and political debates, where he/she posts and comments on the corresponding Subreddits. Apart from all this, person A has developed a keen interest in stock trading, and thus becomes a subscriber of "Wallstreetbets", as he/she can always find interesting ideas for trading there. For person A, Reddit almost satisfies all his/her needs for socializing, and the migration cost is high, plus there is no comparable platform (Chang et al., 2014).

Therefore, person A's activities on "Wallstreetbets" are correlated with his/her activities on other Subreddits. In other words, person A is more likely to post on "Wallstreetbets" if he/she is also posting on other Subreddits. However, it is logical that person A's activities

on other Subreddits have no direct bearing on stock market returns. Such an influence can only be exerted via his/her activities on "Wallstreetbets".

Formally, consider the following specification.

$$WSB\_Posts_{i,t-1} = \alpha_0 + \beta_Z Non\_Finance\_Posts_{i,t-1} + \epsilon_{i,t-1} \tag{11}$$

$$SKEW_{i,t} = \alpha_1 + \beta_X WSB\_Posts_{i,t-1} + \sum_p \beta_p Control_{i,p,t-1} + \lambda_t + u_{i,t} \tag{12}$$

Where the first equation represents the first-stage regression, and the second equation represents the instrumental variable estimation. The subscript i represents stock i and simultaneously all the agents that mention stock i. To operationalize this procedure, we must ensure that the non-finance conversations are truly non-finance related. Therefore, the name of the Subreddit matters.

To ensure that we extract non-finance posts from non-finance "Subreddits", I follow the strategy used in Li et al. (2021). I choose a set of "seed words" and find out 50 words/phrases that are closest in meaning to each seed word. Finally, I choose those "Subreddits" whose title does not contain these keywords. The seed words I choose include: "finance", "stock-market", "stocks", "wall-street", "trading", "forex", "options", "investment", "bond-market", and "bonds".

How to find words/phrases that are similar in meaning to the seed words? Recent advances in computational linguistics offer powerful tools to help solve the problem. First, we want to vectorize the words/phrases into fixed-length vectors. Then we compute the cosine similarity between each pair of vectors to check their distance to each other as a proxy for meaning closeness. To do this, I use the pre-trained word embedding system called "Global Vectors for Word Representation" (GloVe) developed by Pennington et al. (2014). These vectors are trained on the whole corpus of Wikipedia up to 2014 and Gigaword 5 (Parker et al., 2011) on the co-occurrences of words and phrases. I use the 300D version of GloVe, which means that each word/phrase is represented by a 300 dimension vector: $V = [x_1, x_2, ..., x_{300}]$.

Thus the cosine similarity between two words V_1 and V_2 is:

$$CosineSim_{1,2} = \frac{V_1 \cdot V_2}{||V_1|| \cdot ||V_2||} \tag{13}$$

The cosine similarity measure for word vectors ranges between zero and one, with one being the closest meaning.[6] I find out the top 50 most similar words/phrases for each seed word and group them together. Because of duplicates, we end up with a set of 351 keywords that are related to the topic of finance. I then use these keywords to screen all the Subreddits.

7.5. Instrumental Variable Results

With all the data processed, we are ready to construct the instruments. First, we denote a user j's number of posts on "Wallstreetbets" about stock i on day t as $n_{i,j,t}^{WSB}$, and his/her number of posts on non-finance "Subreddits" as $n_{i,j,t}^{nonFin}$. Then stock i's total number of posts on "Wallstreetbets" on day t is $N_{i,t}^{WSB} = \sum_j n_{i,j,t}^{WSB}$. The instrument we construct for this variable would be $N_{i,t}^{nonFin} = \sum_j n_{i,j,t}^{nonFin}$, where the term is summing over all j that have posted on "Wallstreetbets" about stock i on day t.

We proceed to run the regressions of the daily $SKEW$ measure on our main variable of interest – the number of "Wallstreetbets" posts $N_{i,t}^{WSB}$, instrumented by the total number of non-finance posts by the same users $N_{i,t}^{nonFin}$, controlling for the same set of independent variables we use in prior settings. First, we run panel regressions without using the instrument. Then, to test whether there is evidence that the instrument violates the exclusion restriction, I add the instrument into the regression to see whether the instrument is inappropriately excluded. Finally, I run the regression with instrumental variable estimation. The first stage regression is untabulated, but the coefficient on the instrument is 0.049 and statistically significant at the 1% level, and the R^2 is 3.4%. I report the main results in Table 10.

[6]Because all word vectors contain nonnegative numbers, the cosine similarity between any pair of word vectors is nonnegative.

[Table 10 about here.]

The insignificant coefficient on the number of non-finance posts in Column (2) supports the exclusion restriction assumption. The significantly positive coefficient on the number of "Wallstreetbets" posts in the instrumental variable estimation in Column (3) is consistent with our prior results of the "Difference-in-Differences" specification that online conversations among retail investors positively influence the ex-ante crash risk of stocks. A one-standard-deviation increase in the number of "Wallstreetbets" posts is associated with a 15 bps increase in the $SKEW$ measure on average. Since the mean $SKEW$ is 0.065, the 15 bps increase translates into approximately 2.3% increase in ex-ante crash risk on a daily basis.

These results, combined with our prior results on monthly crash risk, support our hypothesis that social media conversations are instrumental in facilitating more efficient herding of individual investors, which in turn drives the increase in the ex-ante crash risk of the underlying stocks.

8. Robustness Tests

In this section, we conduct various alternative tests to examine whether our results are robust.

8.1. Pricing Results Using Crash Risk Defined by Alternative Thresholds

The pricing results for crash risk are robust to the definition of "crash" using other thresholds. I re-run the estimations of ex-ante crash risk and the associated alpha estimates of the hedge portfolios by defining a crash using the following thresholds: log monthly returns of less than -10%, -15%, -25%, and -30%. For brevity, I show only alpha estimates of regressions benchmarking the Fama-French five-factor model plus a momentum factor. In each of these alternative definitions, the crash risk estimated using both logit regression and machine learning method ("EasyEnsemble") consistently produce negative and significant

alphas. Detailed results are presented in Section C.1. These results show further evidence for the superiority of machine learning models.

8.2. Retail Influence on Crash Risk Using Other Specifications

Possible causal influence of the online conversations of retail investors can be shown via other specifications. Here we consider several different specifications to demonstrate the robustness of our findings.

8.2.1. Band Widths and Earnings Days

we start by examining whether changing the event windows would make our results go away. Our main test uses a window of [-3,+3], or 7 months in total for each stock's first appearance on "Wallstreetbets". Here, we use the same specifications but change the window to either [-1,+1] or 3 months, or [-2,+2] or 5 months. We report the results in Columns (1), (2), (4), and (5) in Table A.2. Furthermore, there might be a concern that the first appearances of stocks on "Wallstreetbets" are endogenous because these firms' information events might be the underlying reason that prompted the users to talk about these stocks. Therefore, in a separate test, we drop all stock-month observations that happen to be the months of earnings announcements. Arguably, earnings announcements are the biggest information events for individual stocks. After dropping these events, we are left with approximately 77% of our original sample. We then run the same specification as our main test and report the results in Columns (3) and (6) in Table A.2. Throughout all the settings, we find consistent and strong evidence that our results are robust to different event windows. Moreover, the results do not seem to be driven by firms' information events.

8.2.2. First-Year Effects

First, we limit our attention to the narrow window around the first year of "Wallstreetbets", since it is possible that the influence of the Subreddit is particularly prominent during

the early days when there are few other choices of online conversations on investment. I define the period between April 2011 and March 2012 before the advent of "Wallstreetbets" as the pre-period and the period between April 2012 and March 2013 as the post-period. I treat all firms whose ticker names were mentioned on "Wallstreetbets" during the post-period as treated firms. The resulting set of firms whose ticker names were mentioned on "Wallstreetbets" in its first year consists of 236 unique firms.

I use these firms as treated firms that could experience elevated herding in their trading because of the advent of "Wallstreetbets". Then I test the following specification:

$$Crash\,Risk_{i,t} = \gamma_0 + \beta \times Treated + \sum_p \beta_p Control_{p,i,t-1} + \alpha_i + \lambda_t + \epsilon_{i,t} \tag{14}$$

Where $Treated$ is a dummy variable that equals one if the firm's ticker is mentioned in "Wallstreetbets" and if the period is from April 2012 to March 2013 and zero otherwise. Firm and month fixed effects are included, and hence the two dummy variables are absorbed, leaving the interaction intact. The results are reported in Columns (1) and (2) in Table 11.

[Table 11 about here.]

As shown in the table, there is a significant and positive increase in crash risk for the treated stocks whose ticker names were mentioned during the first year of "Wallstreetbets", even after controlling for a group of firm characteristics, including MAX. The magnitude of increase in these specifications is approximately 40 bps or a 6% increase compared to the unconditional mean of stock crash risk.

8.2.3. Treatment in Multiple Periods

I consider the following hypothesis. After a ticker name first appeared on "Wallstreetbets", the attention for the stock is elevated, and subsequently, retail trading follows, which leads to a potential increase in its crash risk. Since we have the full history of "Wallstreetbets" till the end of 2020, we can pinpoint the month when each of the ticker names

was first posted on the Subreddit. There are in total 3507 unique tickers mentioned on "Wallstreetbets", and the months of their first time mentioning scattered across the sample period. In comparison, during the same sample period, there are 5856 unique firms in the CRSP/Compustat universe that satisfy our basic screening.[7]

The assumption here is that once the ticker name is mentioned on "Wallstreetbets" for the first time, the ticker remains treated afterward. I again estimate the following specification:

$$Crash\,Risk_{i,t} = \gamma_0 + \beta \times Treated + \sum_p \beta_p Control_{p,i,t-1} + \alpha_i + \lambda_t + \epsilon_{i,t} \tag{15}$$

Where the treated status begins in the month after the month when the ticker was first mentioned on "Wallstreetbets". Thus the specification contains units that are treated in different time periods. The results are reported in Columns (3) and (4) in Table 11.

Across all specifications, there is a significant and positive coefficient on $Treated$, providing evidence that once a stock is mentioned on "Wallstreetbets", its crash risk becomes elevated. The magnitude is approximately 40 bps or a 6% increase compared to the unconditional mean of ex-ante crash risk.

One obvious concern is endogeneity since retail attention on stocks is not random and could be just reflecting underlying changes of characteristics in stocks. To partially alleviate the concern, I conducted a falsification test, randomly shuffling the months when the stocks were treated. Then I estimate the same specification, except that $Pseudo-Treated$ replaces the dummy variable $Treated$. The results are reported in Table A.3. As shown in the table, the coefficients on "Pseudo-Treated" is virtually indistinguishable from zero. These results lend credibility to the validity of our quasi-natural experiment, supporting our hypothesis that the rise of Reddit contributes to the increased crash risk of the stocks that receive community attention.

[7]Here the screening refers to selecting common stocks with a share code of 10 or 11 with non-missing returns.

8.2.4. Realized Crashes

In our main tests, we used the estimated ex-ante measure of crash risk to test whether the social transmission of investment ideas on investment forums could have a causal impact on the left-tail risk. Arguably the estimated ex-ante measure could contain measurement errors. On the other hand, we might be also interested in directly testing whether social transmission could cause realized crashes. To maintain consistency with the main tests, I run the same specification as the main test except that I replace the dependent variable with a dummy variable $Crash_{i,t}$, which equals one if the stock crashes in the month, or equivalently, its log return is lower than -20% or any of the thresholds. Thus the specification is as follows:

$$Crash_{i,c,t} = \gamma_0 + \sum_{j=-3}^{+3} \beta_j D_{i,j,c,t} + \delta_{c,t} + \alpha_{i,c} + \epsilon_{i,t} \tag{16}$$

I report the results in Table A.4 in Appendix. As the table shows, all the coefficients on the interaction term $Treated$ are positive and statistically significant at the 1% level. These results provide strong evidence that social transmission not only impacts the ex-ante distribution of stock returns but also influences ex-post outcomes.

9. Conclusion

Recent development in financial technology (FinTech) like "Robinhood" has dramatically reduced the hurdle for retail trading. In addition, popular online forums like "Reddit" facilitate more efficient sharing of trading ideas. These innovations can likely amplify the effect of correlated retail trading behaviors. Because of distorted beliefs, retail investors tend to over-buy high crash risk stocks, contributing to the negative price of crash risk. The buying activities and subsequent price reactions formulate a possible feedback loop. The resulting more elevated level of crash risk contributes to exacerbated market volatility, potentially damaging investor welfare.

Future research avenues could further explore social media's role in forming investor beliefs and their subsequent trading behavior. As reflected in the meme stock frenzy, the mass psychology of the online investing community could be influenced without apparent fundamental information, often to the harm of such investors. Studying this interaction between social media conversations and asset prices could help us understand the intricacies of price formation and aid policymakers in their pursuit of protecting potentially novice and vulnerable investor groups.

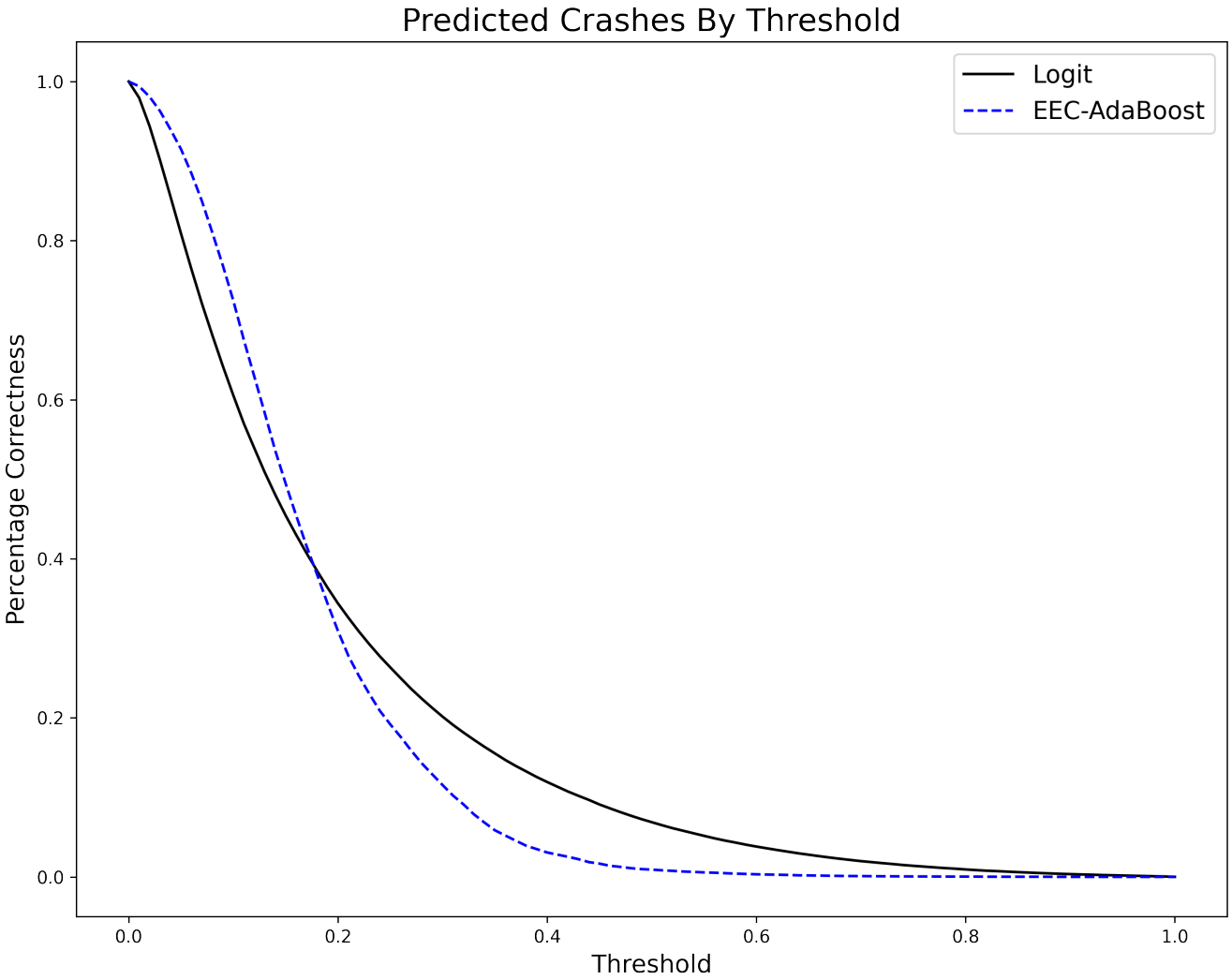

Fig. 1. Out-of-Sample Predicted Crashes by Thresholds. The figure depicts the total percentage of out-of-sample predicted crashes for logistic regression and EasyEnsemble against decision thresholds. The X-axis is the decision threshold from zero to one. The Y-axis gives the percentage of real crashes successfully predicted by either model based on the decision threshold. For example, at the 7% threshold, meaning that we predict all stocks with a probability greater than 7% to crash in the next month, logistic regression is able to catch 72% of all real crashes, while EasyEnsemble is able to catch 85%.

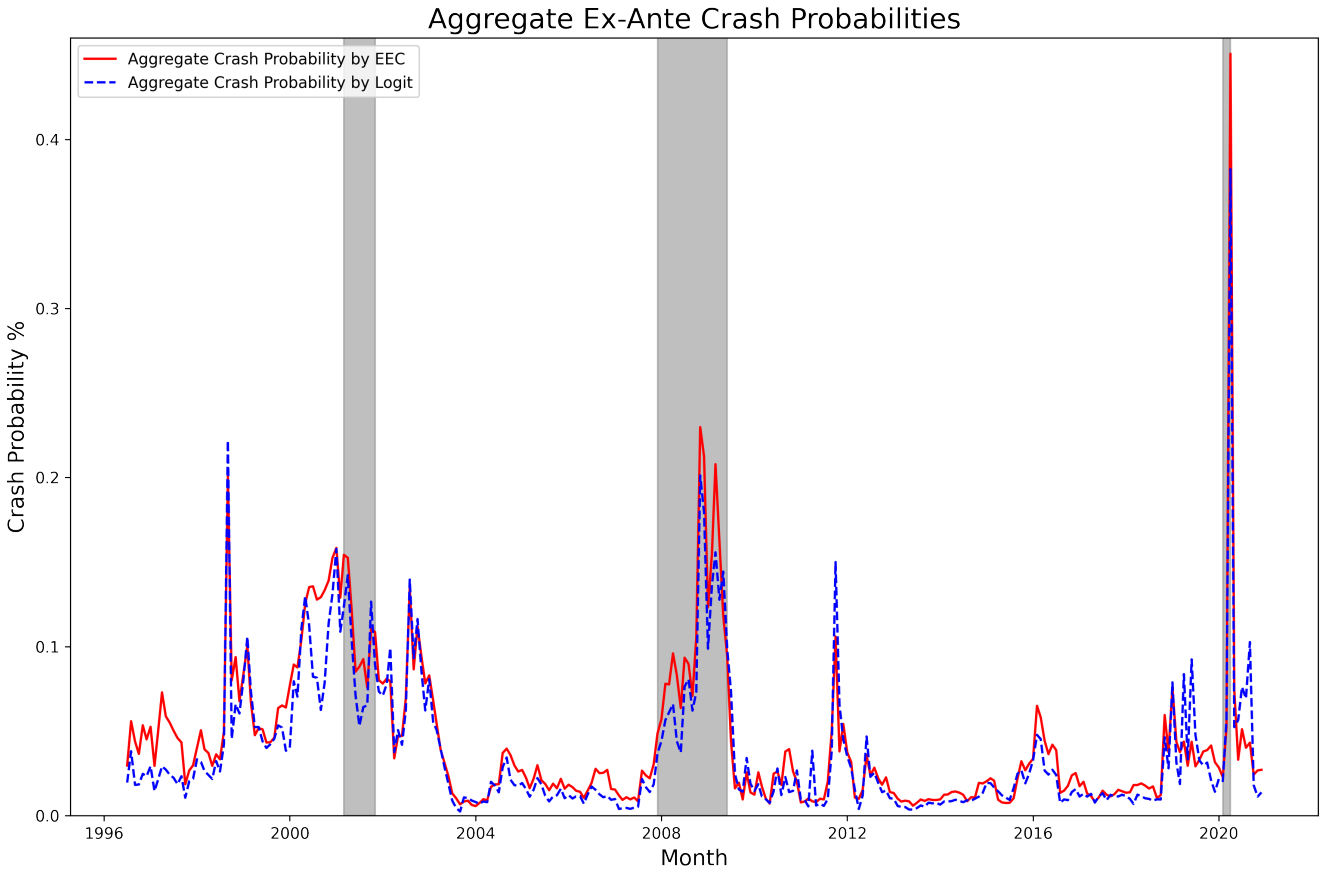

Fig. 2. Aggregate Crash Risk. The figure plots market-wide aggregate ex-ante crash probabilities from 1996 to 2020. The aggregation is done by weighting the monthly ex-ante crash risk of each firm by their lagged market capitalizations as follows:

$$AggCrashRisk_t = \frac{\sum_i MarketCap_{i,t-1} \times CrashRisk_{i,t}}{\sum_i MarketCap_{i,t-1}}$$

The red solid line indicates the aggregate crash probabilities by using the machine learning-generated crash probabilities, while the blue dashed line uses the logit-generated crash probabilities. The gray shaded areas indicate NBER recession periods (NBER, 2021). The time series run from July 1996 to December 2020.

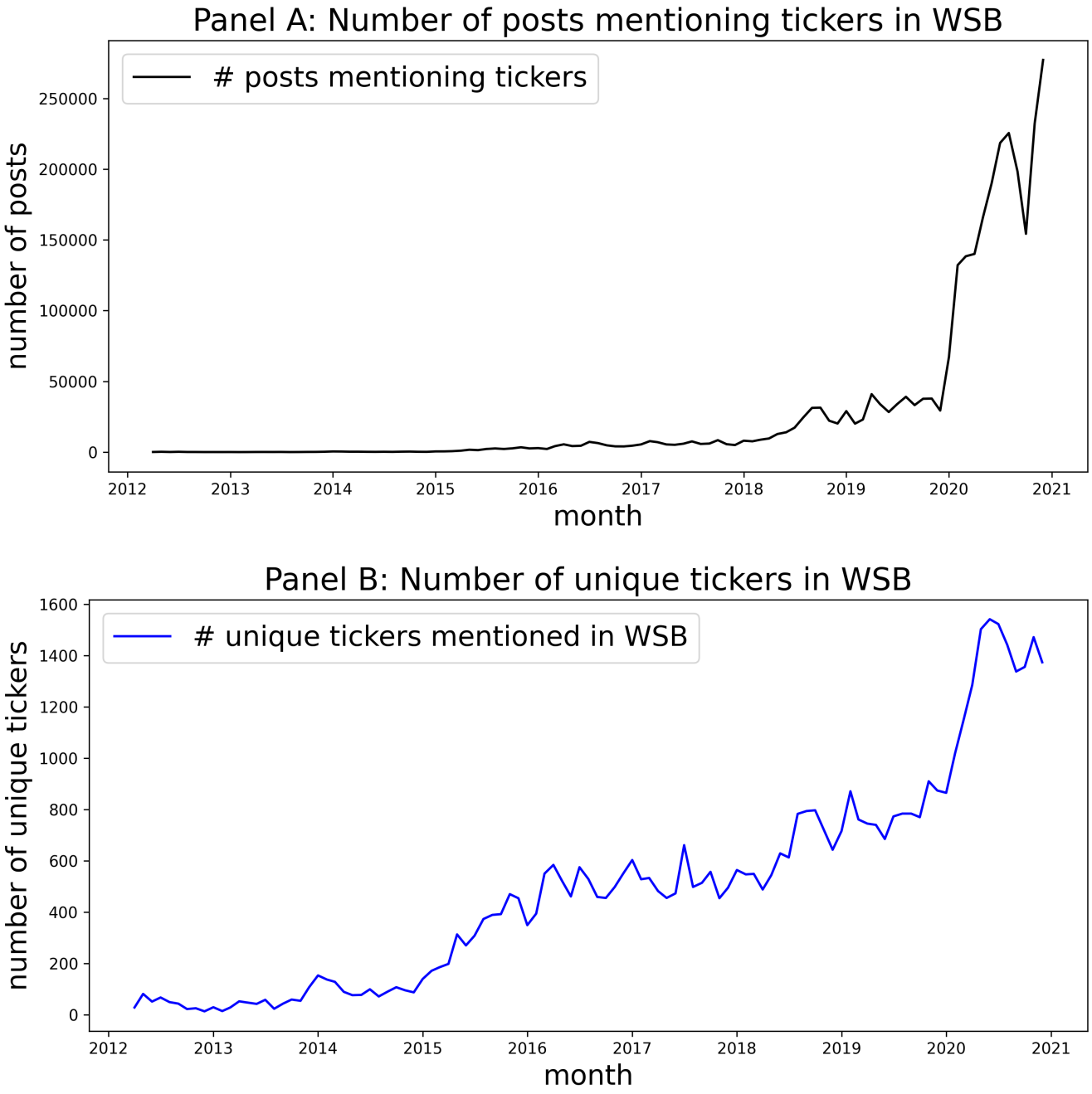

Fig. 3. Monthly Number of Posts and Unique Tickers on "Wallstreetbets". The figure plots the total number of posts each month on the Subreddit "Wallstreetbets" that mention stock ticker names, and also the number of unique ticker names mentioned each month in Panel A and Panel B respectively. For a ticker to be counted, it must not be common English words, slang, or abbreviations. The time series spans from April 2012 when "Wallstreetbets" was established to December 2020.

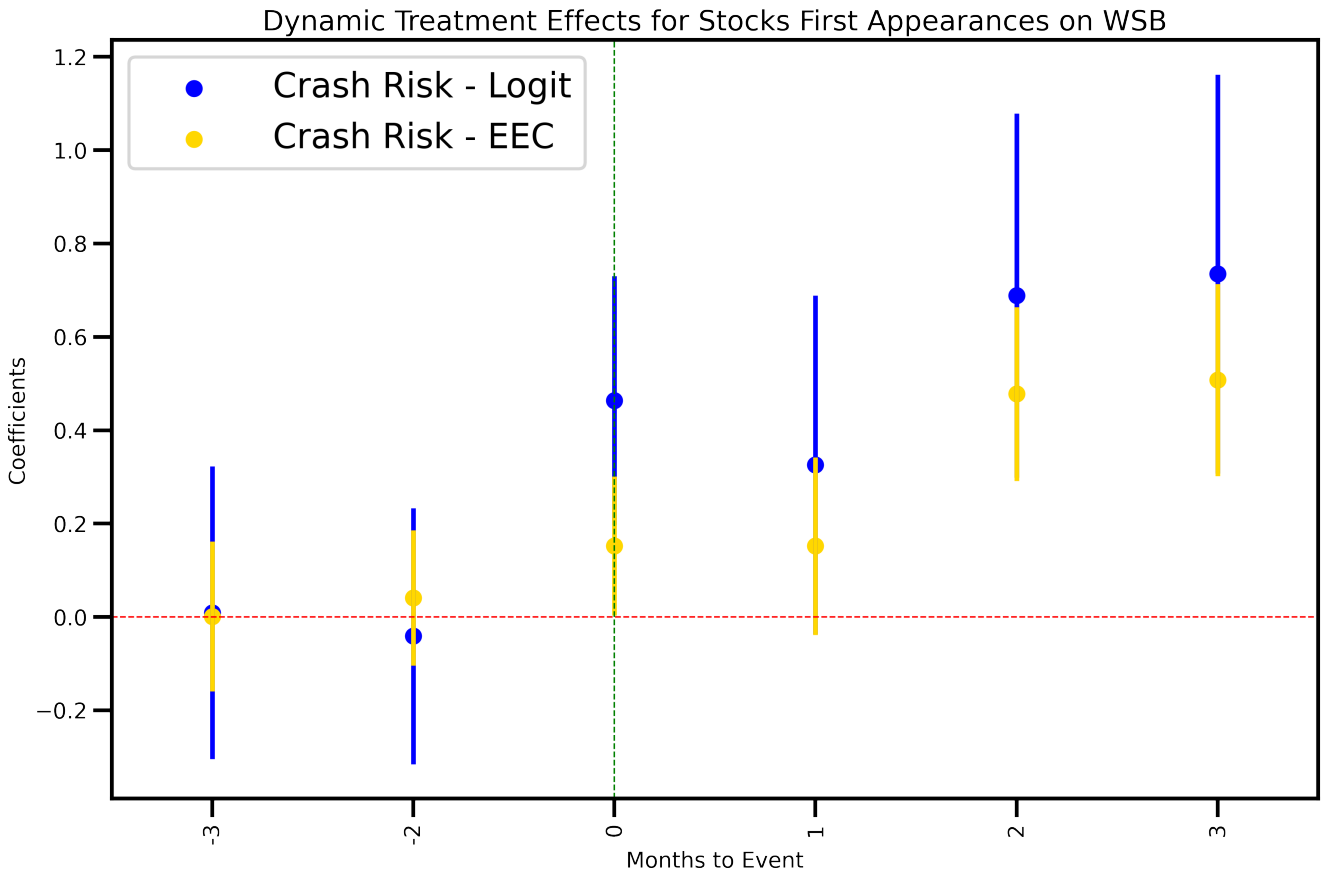

Fig. 4. Dynamic Treatment Effects of the First Appearances of Tickers on "Wallstreetbets". This figure plots the dynamic treatment effects between three months prior to the treatment and three months after the treatment to examine whether the "parallel trend" assumption holds for the "difference-in-differences" analysis on whether the first appearances of stock tickers on "Wallstreetbets" can have a positive and significant effect on stock crash risk. The "difference-in-differences" specification is as follows:

$$Crash\,Risk_{i,c,t} = \gamma_0 + \sum_{j=-3}^{+3} \beta_j D_{i,j,c,t} + \delta_{c,t} + \alpha_{i,c} + \sum_p \beta_p Control_{p,i,t-1} + \epsilon_{i,t}$$

Where the dummy variables $D_{i,j,c,t}$ indicate whether a stock i is treated in cohort c at time t, and the distance $j \in [-3, 3]$ from the current month to the treatment month. Month -1 is chosen to be the base month that will be omitted from the regression. Month 0 is the treatment month, and a green dotted line is plotted for better illustration. The coefficients on the rest of the dummies $D_{i,j,c,t}$ together with their 95% confidence interval bands are then plotted against their respective time periods. The blue markers display results using logit-generated crash risk as the dependent variable, while the golden markers use machine learning-generated crash risk. $Crash\,Risk_{i,c,t}$ is the estimated crash risk of stock i in cohort c at time t. $\delta_{c,t}$ is $Cohort \times Time$ fixed effects. $\alpha_{i,c}$ is $Unit \times Cohort$ fixed effects. Standard errors are clustered at the unit level. The regression results are reported in Column (3) and (6) in Table 6.

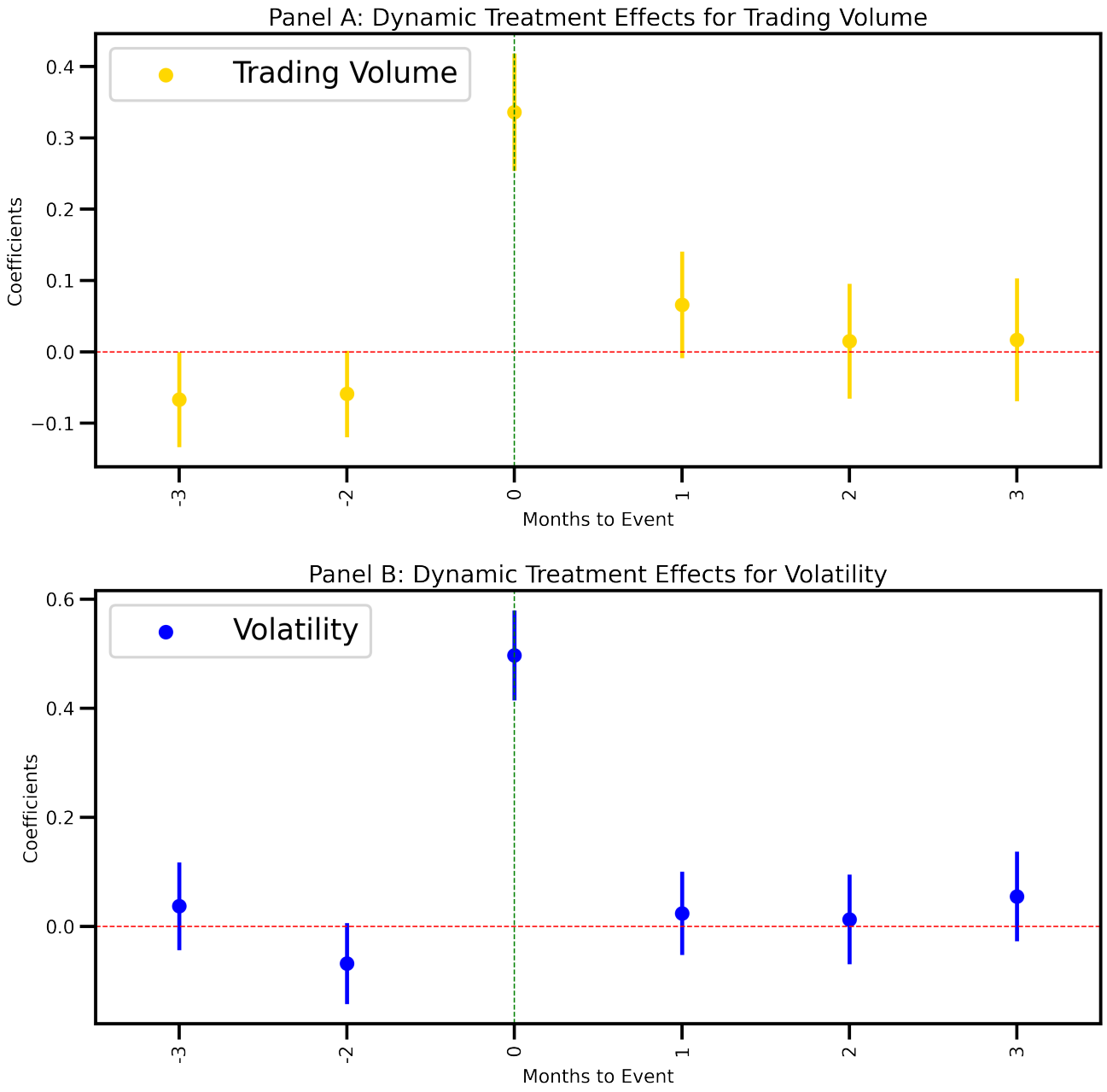

Fig. 5. Dynamic Treatment Effects: Trading Vol & Volatility. This figure plots the dynamic treatment effects between three months prior to the treatment and three months after the treatment to examine whether the "parallel trend" assumption holds for the "difference-in-differences" analysis on whether there is a surge in trading volume and volatility after the first appearances of stock tickers on "Wallstreetbets". The "difference-in-differences" specification is the same specification as in our main test except that we replace the dependent variable with either trading volume or return volatility. Panel A display results on trading volume, while Panel B shows results for return volatility. The regression results are reported in Column (3) and (6) in Table 8.

Table 1: Summary Statistics

This table reports the summary statistics of our main variable ex-ante monthly crash risk and other firm characteristics used later in our analyses. There are two sets of crash risk estimates. $Crash1$ is estimated by logit regression, and $Crash2$ by machine learning (EEC-Adaboost). To differentiate our measure from the left-tail measure VaR in Atilgan et al. (2020), we also include their measure. $VaR1\%$ is defined as the 1 percentile daily return of the stock in the past year, while $VaR5\%$ is the 5 percentile daily return of the stock in the past year. Other variables include the natural log of market capitalizations ($Size$), the natural log of book-to-market ratio, asset growth (ATG), gross profitability (GP), momentum (prior 11-to-1 month returns, MOM), and short-term reversal (prior 1-month returns, $ST-Rev$), idiosyncratic volatility, illiquidity (Amihud, 2002), market beta, tail Beta (Kelly and Jiang, 2014), coskewness(Harvey and Siddique, 2000), MAX (Bali et al., 2011). IO is the institutional ownership for each stock, measured at the quarterly frequency. $UserNum$ is the total number of users for each stock on Robinhood by the end of each month. The sample starts from July 1996 to December 2021, except for Robinhood user numbers where it is limited to between May 2018 and August 2020 due to the data availability of Robintrack (`https://robintrack.net/`).

	Crash1	Crash2	VaR1%	VaR5%	Size	Beta	Log(B/M)	ATG	GP	MOM
count	1,383,264	1,383,264	1,393,933	1,393,933	1,439,823	1,284,824	1,273,512	1,235,657	1,068,246	1,346,720
mean	0.09	0.10	-0.08	-0.05	5.73	1.08	-0.77	0.20	0.32	0.14
std	0.13	0.09	0.05	0.03	2.17	0.85	1.04	3.45	0.39	0.84
1%	0.00	0.00	-0.26	-0.16	1.33	-0.21	-3.75	-0.54	-0.78	-0.86
25%	0.02	0.03	-0.11	-0.07	4.14	0.50	-1.32	-0.03	0.15	-0.23
50%	0.04	0.07	-0.07	-0.04	5.62	0.93	-0.69	0.06	0.30	0.04
75%	0.11	0.14	-0.05	-0.03	7.18	1.48	-0.14	0.19	0.48	0.32
99%	0.64	0.42	0.00	0.00	11.08	3.72	1.72	2.64	1.25	2.89
	ST-Rev	Vol	Skew	TailBeta	Coskew	IdioRisk	Illiq	MaxRet	IO	UserNum
count	1,469,593	1,466,228	1,437,111	951,654	1,356,663	1,435,097	1,345,881	1,440,263	496,204	87,456
mean	0.01	0.03	0.24	0.72	0.22	0.03	4.72	0.08	0.41	3418.46
std	0.20	0.03	1.00	0.58	0.29	0.03	46.24	0.11	0.34	21496.13
1%	-0.43	0.00	-2.64	-0.49	-0.41	0.00	0.00	0.01	0.00	6.00
25%	-0.07	0.02	-0.28	0.37	0.04	0.01	0.00	0.03	0.09	95.00
50%	0.00	0.03	0.20	0.63	0.20	0.02	0.03	0.06	0.36	319.00
75%	0.07	0.04	0.72	0.99	0.37	0.04	0.55	0.10	0.71	1161.00
99%	0.62	0.15	3.25	2.52	1.09	0.15	86.39	0.45	1.09	64691.10

Table 2: Decile High-Minus-Low Portfolio Alphas
This table presents the analysis of portfolios sorted on the ex-ante crash risk measures estimated by both logit and machine learning (EEC-AdaBoost). At the end of each month, stocks are ranked by their ex-ante crash probabilities produced by either logit or machine learning into ten decile portfolios. Then we compute both equal-weighted portfolio returns and value-weighted returns by their lagged market capitalization. The hedge portfolio is long in the top decile ex-ante crash risk portfolio and short in the bottom decile crash risk portfolio. Then the hedge portfolio return series are regressed on risk factor returns from various empirical asset pricing models. The asset pricing models include: CAPM, Fama-French three-factor model (FF3) (Fama and French, 1993), then augmented with a momentum factor (FF4) (Carhart, 1997), Fama-French five-factor model (FF5) (Fama and French, 2015), and then augmented with momentum factor (FF6). Then we report the resulting intercepts (alphas) and their associated T-statistics. The upper panel presents results from using value-weighted portfolio returns, while the lower panel presents equal-weighted results. The left half shows results from using ex-ante crash risk estimated from logistic regressions, and the right from machine learning (EEC-AdaBoost). Standard errors are adjusted using the Newey-West procedure (Newey and West, 1986) with 6 lags.

		Logit		EEC-Adaboost	
	Pricing model	Alpha	T-stat	Alpha	T-stat
Value-weighted	CAPM	-1.852	-3.730	-1.967	-4.393
	FF3	-1.842	-4.440	-1.963	-5.456
	FF4	-1.533	-3.531	-1.775	-4.636
	FF5	-0.874	-2.834	-1.120	-3.947
	FF6	-0.696	-2.263	-1.023	-3.442
Equal-weighted	CAPM	-2.470	-5.571	-2.458	-5.325
	FF3	-2.461	-7.941	-2.452	-7.573
	FF4	-2.106	-7.161	-2.173	-7.005
	FF5	-1.656	-5.637	-1.783	-6.093
	FF6	-1.438	-5.788	-1.614	-5.947

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3: Fama-MacBeth Cross-Sectional Regressions
This table reports Fama-MacBeth cross-sectional regressions of raw returns on ex-ante crash risk and lagged firm characteristics in the spirit of Fama and French (2020). First, we regress monthly stock returns of each month on lagged firm characteristics. Then we average the coefficients and report the associated standard errors. Our main variables of interest are the two ex-ante crash risk measures. One is estimated by logit regression, and the other by machine learning (EEC-Adaboost). Columns (1) and (2) use the logit-generated crash risk as the main variable, while Columns (3) and (4) use the machine learning-generated crash risk. To differentiate our measure from the left-tail measure VaR in Atilgan et al. (2020), I include their measure in Columns (2) and (4) as a control variable, where $VaR1\%$ is defined as the negative of 1 percentile daily return of the stock in the past year. In Column (5), I only include $VaR1\%$ as the sole variable to proxy for left-tail risk to ensure that our results are consistent with Atilgan et al. (2020). Other control variables include the natural log of market capitalizations, the natural log of book-to-market ratio, asset growth, gross profitability, momentum (prior 11-to-1 month returns), and short-term reversal (prior 1-month returns). In Column (3), I add idiosyncratic volatility and illiquidity (Amihud, 2002). In Column (4), I add market beta, tail Beta (Kelly and Jiang, 2014), coskewness(Harvey and Siddique, 2000), net operating assets NOA (Hirshleifer et al., 2004), and MAX (Bali et al., 2011). All independent variables are standardized cross-sectionally each month to be mean zero and standard deviation of unity, such that the coefficients on all the independent variables can be directly read as the percentage increase in average stock returns if the underlying independent variable increase by one standard deviation. Standard errors are adjusted according to Newey-West procedures (Newey and West, 1986) with 6 lags.

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Returns in %				
Crash Risk (Logit)	-0.491***	-0.453***			
	(0.080)	(0.077)			
Crash Risk (EEC)			-0.507***	-0.459***	
			(0.097)	(0.086)	
VaR1%		-0.123		-0.097	-0.246***
		(0.082)		(0.074)	(0.083)
Controls	YES	YES	YES	YES	YES
Observations	545,367	545,290	545,367	545,290	564,466
R-squared	0.083	0.086	0.083	0.085	0.084
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 4: Sentiment and Crash Risk Returns

This table presents the relationship between the price of crash risk and sentiment. Panel A reports the value-weighted portfolio excess returns for high-crash-risk, low-crash-risk, and long-short hedge portfolios in both high sentiment and low sentiment periods and their differences. Market-wide sentiment is defined by Baker and Wurgler (2006). Our sample is limited between July 1996 and December 2018 because of the availability of the index. High and low sentiment periods are defined as either above or below median sentiment over the sample period. Panel B reports regression results. In Columns (1) and (2), we run Fama-MacBeth cross-sectional regressions of stock returns on crash risk and lagged firm characteristics for high- and low-lag-sentiment periods separately. Control variables include all the firm characteristics used in Table 3. Standard errors are estimated according to Newey-West procedure (Newey and West, 1986) with 6 lags. In Columns (3) and (4), we run panel regressions of stock returns on the same independent variables as the previous specification, with firm and time fixed effects. In Column (4), we add a dummy variable *SentD*, where it equals one if the lagged sentiment is higher than the sample median, and zero otherwise. We interact *SentD* with crash risk, and hence the coefficient on the interaction term can be interpreted as the incremental price of crash risk when lagged sentiment is high. All independent variables are standardized cross-sectionally each month to be mean zero and standard deviation of unity. Standard errors are clustered at the firm level.

Panel A: Portfolio Excess Returns and Sentiment

	High Sent	Low Sent	High-Low
Low crash risk	0.597*	0.812**	-0.215
	(0.314)	(0.309)	0.436
High crash risk	-1.849*	0.879	-2.728**
	(1.008)	(0.880)	(1.280)
Long-short	-2.446**	0.067	-2.513**
	(0.943)	(0.709)	(1.144)

Panel B: Price of Crash Risk and Sentiment

	(1)	(2)	(3)	(4)
	FMB		Panel	
VARIABLES	Low Sent	High Sent	Return	Return
Crash Risk	-0.405***	-0.619***	-0.335***	-0.135**
	(0.108)	(0.141)	(0.050)	(0.062)
SentmentD×Crash Risk				-0.374***
				(0.063)
Controls	YES	YES	YES	YES
Observations	240,805	269,577	545,227	510,260
R-squared	0.078	0.085	0.168	0.159

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5: Investor Trading and Crash Risk

This table presents results from regressing Robinhood user trading measures and institutional trading measures on crash risk, controlling for other firm characteristics. The first Robinhood user trading measure is the monthly change of the natural log of user numbers holding a particular stock, where the user numbers are from the online brokerage Robinhood (Robintrack). The second Robinhood user trading measure is the percentage change in the number of users over the previous month. The institutional trading measure is the quarterly change in the ratio of institutional holding for each stock. We regress all of these trading measures on the contemporaneous crash risk measures constructed from both logit regressions and the machine learning method (EEC-AdaBoost). Columns (1) to (4) add lagged log of the number of users as a control variable. For all specifications, the control variables include the natural log of market capitalization, the natural log of book-to-market ratio, asset growth, gross profitability, momentum, short-term reversal, MAX and MIN (Bali et al., 2011), defined as the highest and lowest daily returns of the previous month, total skewness of daily returns in the previous month, illiquidity (Amihud, 2002), and Fama-French three-factor betas estimated from the previous month. Firm and Time fixed effects are included, and robust standard errors are included in parentheses.

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Change in Log(User)		User%Change		IO Change	
Crash Risk (Logit)	0.093***		0.154***		-0.026***	
	(0.010)		(0.020)		(0.002)	
Crash Risk (EEC)		0.104***		0.156***		-0.013***
		(0.016)		(0.028)		(0.003)
Controls	YES	YES	YES	YES	YES	YES
Observations	63,692	63,692	63,692	63,692	375,339	375,339
R-squared	0.241	0.240	0.191	0.190	0.500	0.500
Firm & Time FE	YES	YES	YES	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Table 6: First Appearances of Stock Tickers on "Wallstreebets" and Crash Risk
This table reports results from a "stacked difference-in-differences" approach (Gormley and Matsa, 2011) that examines the effect of first appearances of stocks tickers on "Wallstreetbets" on their ex-ante crash risk. Columns (1) to (3) use logit-generated crash risk as the dependent variable, while Columns (4) to (6) use machine learning-generated crash risk. "Wallstreetbets" was started in April 2012. From the beginning of "Wallstreetbets" to the end of 2020, we find all the stock tickers that are ever mentioned in the Subreddit and the first month they were mentioned. We then define each of these instances as one event and each of the stocks as a treated stock. We match each treated stock with five control stocks from the pool of "never treated" stocks via propensity score matching based on lagged characteristics three months prior to each event. Then the "cohorts" containing treated and control observations are stacked together and the following specification is run:

$$Crash\,Risk_{i,c,t} = \gamma_0 + \beta D_{i,c,t} + \delta_{c,t} + \alpha_{i,c} + \textstyle\sum_p \beta_p Control_{p,i,t-1} + \epsilon_{i,t}$$

Where $Crash\,Risk_{i,c,t}$ is the estimated crash risk of stock i in cohort c at time t. $D_{i,c,t}$ is a dummy variable that indicates whether a stock i in cohort c is treated at time t. $\delta_{c,t}$ is $Cohort \times Time$ fixed effects. $\alpha_{i,c}$ is $Unit \times Cohort$ fixed effects. Then β is the coefficient of interest that estimates the average treatment effect on the treated stocks. The results are reported in Columns (1), (2), (4), and (5), where Columns (2) and (5) add control variables. The control variables include the natural log of market capitalization, prior-month return, asset growth, gross profitability, illiquidity, MAX (Bali et al., 2011), prior 12-month return, and idiosyncratic risk. Columns (3) and (6) examine the dynamic treatment effects around the events. Standard errors are clustered at the unit level.

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Crash Risk (Logit)			Crash Risk (EEC)		
Treated	1.032***	0.560***		0.674***	0.303***	
	(0.103)	(0.129)		(0.054)	(0.064)	
Month -3			0.009			0.001
			(0.160)			(0.082)
Month -2			-0.041			0.041
			(0.140)			(0.074)
Month 0			0.464***			0.152**
			(0.136)			(0.076)
Month +1			0.326*			0.152
			(0.185)			(0.097)
Month +2			0.689***			0.478***
			(0.199)			(0.095)
Month +3			0.735***			0.508***
			(0.218)			(0.105)
Observations	208,502	125,734	125,734	208,502	125,734	125,734
R-squared	0.874	0.909	0.909	0.921	0.946	0.946
Cohort×Units FE	YES	YES	YES	YES	YES	YES
Cohort×Month FE	YES	YES	YES	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Table 7: First Appearances of Stock Tickers on "Wallstreebets" and Crash Risk: Size & IO

This table reports results from a "stacked difference-in-differences" approach that examines whether the effect of first appearances of stocks tickers on "Wallstreetbets" on their ex-ante crash risk differs because of size or level of institutional ownership. "Wallstreetbets" was started in April 2012. From the beginning of "Wallstreetbets" to the end of 2020, we find all the stock tickers that are ever mentioned in the Subreddit and the first month they were mentioned. We then define each of these instances as one event and each of the stocks as a treated stock. We match each treated stock with five control stocks from the pool of "never treated" stocks via propensity score matching based on lagged characteristics three months prior to each event. Then the "cohorts" containing treated and control observations are stacked together and the following specification is run:

$$Crash\,Risk_{i,c,t} = \gamma_0 + \beta_1 D_{i,c,t} + \beta_2 D_{i,c,t} \times D_{size} + \delta_{c,t} + \alpha_{i,c} + \sum_p \beta_p Control_{p,i,t-1} + \epsilon_{i,t}$$

Where $Crash\,Risk_{i,c,t}$ is the estimated crash risk of stock i in cohort c at time t. $D_{i,c,t}$ is a dummy variable that indicates whether a stock i in cohort c is treated at time t. $D_{size/io}$ is a dummy variable that equals one if the stock is larger than the median or zero otherwise, based on the lagged value of each stock three months prior to each event. In the case of institutional ownership, $D_{size/io} = 1$ if the ratio of institutional ownership for the stock is greater than the median or zero otherwise. $\delta_{c,t}$ is $Cohort \times Time$ fixed effects. $\alpha_{i,c}$ is $Unit \times Cohort$ fixed effects. Then β_2 is the coefficient of interest that estimates the difference in average treatment effect on the treated stocks if the stocks belong to the large stock subgroup. Column (2) adds control variables that include the natural log of market capitalization, prior-month return, asset growth, gross profitability, illiquidity (Amihud, 2002), MAX (Bali et al., 2011), prior 12-month return, and idiosyncratic risk. Standard errors are clustered at the unit level to account for possible duplicate observations.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Crash Risk (Logit)				Crash Risk (EEC)			
Treated	1.501***	1.038***	1.539***	0.988***	1.060***	0.434***	1.019***	0.422***
	(0.182)	(0.326)	(0.177)	(0.310)	(0.090)	(0.142)	(0.090)	(0.145)
Treated×D_{size}	-0.930***	-0.743**			-0.766***	-0.202		
	(0.205)	(0.343)			(0.104)	(0.153)		
Treated×D_{io}			-1.082***	-0.689**			-0.735***	-0.191
			(0.202)	(0.330)			(0.102)	(0.155)
Controls	NO	YES	NO	YES	NO	YES	NO	YES
Observations	208,502	125,734	208,502	125,734	208,502	125,734	208,502	125,734
R-squared	0.874	0.909	0.874	0.909	0.921	0.946	0.921	0.946
Cohort×Units FE	YES	YES	YES	YES	YES	YES	YES	YES
Cohort×Month FE	YES	YES	YES	YES	YES	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Table 8: First Appearances of Stock Tickers on "Wallstreebets": Trading Vol & Volatility
This table reports results from a "stacked difference-in-differences" approach (Gormley and Matsa, 2011) that examines the effect of first appearances of stocks tickers on "Wallstreetbets" on their trading volume and volatility. Columns (1) to (3) use trading volume as the dependent variable, while Columns (4) to (6) use return volatility. "Wallstreetbets" was started in April 2012. From the beginning of "Wallstreetbets" to the end of 2020, we find all the stock tickers that are ever mentioned in the Subreddit and the first month they were mentioned. We then define each of these instances as one event and each of the stocks as a treated stock. We match each treated stock with five control stocks from the pool of "never treated" stocks via propensity score matching based on lagged characteristics three months prior to each event. Then the "cohorts" containing treated and control observations are stacked together and the following specification is run:

$$TradingVol_{i,c,t} = \gamma_0 + \beta D_{i,c,t} + \delta_{c,t} + \alpha_{i,c} + \sum_p \beta_p Control_{p,i,t-1} + \epsilon_{i,t}$$

Where $TradingVol_{i,c,t}$ is the trading volume of stock i in cohort c at time t. $D_{i,c,t}$ is a dummy variable that indicates whether a stock i in cohort c is treated at time t. $\delta_{c,t}$ is $Cohort \times Time$ fixed effects. $\alpha_{i,c}$ is $Unit \times Cohort$ fixed effects. Then β is the coefficient of interest that estimates the average treatment effect on the treated stocks. The results are reported in Columns (1), (2), (4), and (5), where Columns (2) and (5) add control variables. The control variables include the natural log of market capitalization, prior-month return, asset growth, gross profitability, illiquidity, MAX (Bali et al., 2011), prior 12-month return, and idiosyncratic risk. Columns (3) and (6) examine the dynamic treatment effects around the events. Standard errors are clustered at the unit level.

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Trading Volume			Volatility		
Treated	0.227***	0.146***		0.260***	0.162***	
	(0.026)	(0.032)		(0.021)	(0.026)	
Month -3			-0.067*			-0.037
			(0.034)			(0.041)
Month -2			-0.059*			-0.068*
			(0.031)			(0.038)
Month 0			0.336***			0.497***
			(0.042)			(0.042)
Month +1			0.066*			0.024
			(0.038)			(0.039)
Month +2			0.015			0.013
			(0.041)			(0.042)
Month +3			-0.017			-0.055
			(0.044)			(0.042)
Observations	209,478	125,748	125,748	212,961	125,748	125,748
R-squared	0.931	0.954	0.954	0.790	0.842	0.843
Cohort×Units FE	YES	YES	YES	YES	YES	YES
Cohort×Month FE	YES	YES	YES	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Table 9: Daily Returns, Retail Trading, and Crash Risk ($SKEW$)
This table examines the relationship between daily returns and lagged $SKEW$ measure, and the relationship between Robinhood user trading and the contemporaneous $SKEW$ measure. Panel A reports regressions of daily stock returns on lagged $SKEW$ measure as a proxy for crash risk in the daily frequency. The $SKEW$ measure follows Xing et al. (2010):

$$SKEW_{i,t} = ImpliedVol_{i,t}^{OTM-Put} - ImpliedVol_{i,t}^{ATM-Call}$$

The option data is from Option Metrics. We screen the option data based on the following conditions. Days to expiration are between 10 and 60 days. Implied volatilities are between 0.03 and 2. Open interest must be greater than zero. Option price must be greater than \$0.125. Volume is non-missing. For out-of-money put options, the moneyness is between 0.8 and 0.95. For at-the-money call options, the moneyness is between 0.95 and 1.05. We choose the implied volatility of the put option with moneyness closest to 0.95, and the implied volatility of the call option with moneyness closest to 1 to compute the $SKEW$ measure for the day. Columns (1) and (2) report Fama-MacBeth cross-sectional regressions, while Columns (3) and (4) report panel regressions. The control variables include prior day return, prior month-end log of market capitalization, book-to-market ratio, cumulative 19-day returns lagged for 2 days (reversal), cumulative 100-day returns lagged for 21 days (momentum), prior month average trading volume scaled by total shares outstanding (liquidity), and prior month volatility of daily returns. For panel regressions, we include day fixed effects, and standard errors are clustered at the stock level. Panel B reports panel regressions of Robinhood user trading measures on contemporaneous $SKEW$ measure as a proxy for ex-ante crash risk and control variables. The trading measures include the change in the log of user numbers and the percentage change of user numbers from the previous day.

	(1)	(2)	(3)	(4)
Panel A: Daily Stock Returns and Crash Risk ($SKEW$)				
VARIABLES	FMB		Panel	
Lag Option SKEW	-0.001***	-0.002***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Controls	NO	YES	NO	YES
Observations	2,071,209	2,010,815	2,071,209	2,010,815
R-squared	0.003	0.072	0.199	0.201
Panel B: Robinhood User Trading and Crash Risk ($SKEW$)				
VARIABLES	Change in Log(Robinhood Users)		% Change in Robinhood Users	
Option SKEW	0.001**		0.001**	
	(0.000)		(0.001)	
Controls	YES		YES	
Observations	703,614		862,423	
R-squared	0.011		0.003	
Note:			*p<0.1; **p<0.05; ***p<0.01	

Table 10: Instrumental Variable Estimation: "WSB" Posts and Crash Risk ($SKEW$)
This table reports the results of regressing the daily $SKEW$ measure on the number of "Wallstreetbets" posts, controlling for other stock characteristics. Column (1) reports a panel regression of $SKEW$ on the number of "Wallstreetbets" posts. Column (2) adds the proposed instrument "number of non-finance posts" to test the exclusion restriction. Column (3) reports the result of instrumental variable estimation. Denote a user j's number of posts on "Wallstreetbets" about stock i on day t as $n^{WSB}_{i,j,t}$, and his/her number of posts on non-finance "Subreddits" as $n^{nonFin}_{i,j,t}$. Then stock i's total number of posts on "Wallstreetbets" on day t is $N^{WSB}_{i,t} = \sum_j n^{WSB}_{i,j,t}$. The instrument we construct for this variable would be $N^{nonFin}_{i,t} = \sum_j n^{nonFin}_{i,j,t}$, where the term is summing over all j that have posted on "Wallstreetbets" about stock i on day t. The IV specification is as follows:

$$N^{WSB}_{i,t-1} = \alpha_0 + \beta_Z N^{nonFin}_{i,t-1} + \epsilon_{i,t-1}$$
$$SKEW_{i,t} = \alpha_1 + \beta_X N^{WSB}_{i,t-1} + \sum_p \beta_p Control_{i,p,t-1} + \lambda_t + u_{i,t}$$

The $t-1$ subscripts on the number of posts refer to the time period of 16:30 PM the previous day to 9:00 AM on day t. The first stage regression of $N^{WSB}_{i,t}$ on $N^{nonFin}_{i,t}$ produces a coefficient of 0.049, statistically significant at the 1% level, and a R^2 of 3.4%, which dispels the weak instrument concern. In all specifications, the control variables include prior day return, prior month-end log of market capitalization, book-to-market ratio, cumulative 19-day returns lagged for 2 days (reversal), cumulative 100-day returns lagged for 21 days (momentum), prior month average trading volume scaled by total shares outstanding (liquidity), and prior month volatility of daily returns. We include day fixed effects, and standard errors are clustered at the stock level.

VARIABLES	(1) Panel	(2) Panel	(3) IV
Number of "Wallstreetbets" Posts	0.070***	0.067***	0.193***
	(0.019)	(0.018)	(0.035)
Number of Non-Finance Posts		0.005	
		(0.004)	
Controls	YES	YES	YES
Observations	2,655,209	2,655,209	2,655,209
R-squared	0.089	0.089	0.042
Day FE	YES	YES	YES
Firm Cluster	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Table 11: Wallstreebets and Crash Risk: Alternative Settings
This table examines the impact of the advent of "Wallstreetbets" on ex-ante crash risk. Columns (1) and (2) examine the effect of stock ticker appearances in the first year of "Wallstreetbets", which was founded in April 2012. The dependent variable is the crash risk, while the independent variable of interest is $Treated$, which is a dummy variable that equals one if the firm's ticker is mentioned in "Wallstreetbets" and if the period is from April 2012 to March 2013, and zero otherwise. Columns (3) and (4) examine the "continued" effect of stock ticker appearances on "Wallstreetbets". The independent variable of interest is again $Treated$, but now it equals one if the firm's ticker is first mentioned in "Wallstreetbets" in a particular month and for all the period after the month when the ticker is first mentioned. Thus the treated stocks have different time periods for their treated status. The sample period for Test 2 starts from January 2012 to December 2020. In all specifications, we control for the natural log of market capitalization, the natural log of book-to-market ratio, asset growth, gross profitability, momentum, short-term reversal, idiosyncratic risk, illiquidity (Amihud, 2002), MAX (Bali et al., 2011), defined as the highest daily returns of the previous month, market beta, tail beta (Kelly and Jiang, 2014), coskewness (Harvey and Siddique, 2000), and net operating assets (Hirshleifer et al., 2004). Firm and Time fixed effects are included, and standard errors are clustered at the firm level.

	(1)	(2)	(3)	(4)
	Dependent Var: Crash Risk			
	Setting 1		Setting 2	
VARIABLES	logit	EEC	logit	EEC
Treated	0.008***	0.004***	0.004***	0.004***
	(0.002)	(0.001)	(0.001)	(0.001)
Controls	YES	YES	YES	YES
Observations	51,842	51,842	211,984	211,984
R-squared	0.677	0.787	0.691	0.814
Firm & Time FE	YES	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix A. Selected Variable Definitions

ATG	= asset growth over the previous year
$Book\,Value\,of\,Equity$	= $SEQ + TXDITC - Perferred$, preferred is $PSTKRV$, or $PSTKL$, or $PSTK$, whichever is first available.
$Crash\,Risk$	= predicted monthly ex-ante probability of a stock crash, where a crash is defined as the log monthly return less than a threshold. The main results use -20% as the threshold
GP	= gross profitability, equals $(REVT - COGS)/AT$
$IdioRisk$	= Daily residual volatility obtained by regressing the previous month's daily excess returns on the market factor returns
$Illiquidity$	= monthly mean of daily absolute return over price times volume of that day, see Amihud (2002)
MOM	= Prior eleven-to-one month return
NOA	= $net\_operating\_assets/lag\_AT$
$SKEW$	= difference between the implied volatility of out-of-money put option and that of the at-the-money call option, see Xing et al. (2010)

Appendix B. In Praise of Machine Learning

B.1. The Black Swan Problem

When predicting rare events, the usual logistic estimator could produce biased estimates due to the poor finite sample properties (King and Zeng, 2001). A simple intuition can be illustrated as follows.

The cost of misclassifying crashes as "normal" cases is far higher than misclassifying "normal" cases as crashes. In the first scenario, investors would be faced with huge unexpected losses, whereas the second would be analogous to giving up average returns. Thus,

the cost of misclassification is asymmetric. The loss function in a generic logistic regression is not cost-sensitive, meaning that it treats each observation equally.

Formally, for logistic regression, its loss function is log loss, or cross-entropy, as represented by Equation 17.

$$logLoss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{17}$$

Now we separate the two classes and denote the sizes of them as N_{normal} and N_{crash}, where N_{normal} denotes the number of "normal" observations, and N_{crash} denotes the number of crashes. Then the log loss function can be written as:

$$logLoss = -\frac{1}{N_{normal} + N_{crash}}[\sum_{i=1}^{N_{normal}} \log(p_i^{normal})] - \frac{1}{N_{normal} + N_{crash}}[\sum_{i=1}^{N_{crash}} \log(p_i^{crash})] \tag{18}$$

Where the first term refers to the log loss of classifying "normal" cases, and the second term refers to "crashes".

Now consider the "imbalanced sample" case, where $N_{normal} >> N_{crash}$. In the limit, fix N_{crash} and let $N_{normal}/N_{crash} \to \infty$. Then the second term of Equation 18 tends to zero, and effectively we are only minimizing the log loss on the "normal" cases. King and Zeng (2001) shows that in a finite sample, using generic logistic regression on an imbalanced sample, or "rare event classification" problems, would produce biased coefficients and underestimate the probability of rare events.

B.2. EasyEnsemble Estimation

"Easy Ensemble" (EEC) is conducted as follows. In each rolling training window of 6 months, I randomly sample a subset of normal observations (non-crash) and pair them with the crash observations, thus ensuring an equal number of observations between crashes and normal observations. This sample is thereby deemed "balanced". I fit an estimator on this

sample and save the parameters. I repeat this process 50 times by independently (with replacement) sampling 50 subsets from the non-crash observations, hence constructing 50 bootstrapped and balanced samples and their associated saved parameters. An Ensemble is built upon these results and arrives at a final estimate. Since each bootstrapped sample is balanced, the cost of misclassifying crashes is given sufficient attention, contributing to a better estimate.

EasyEnsemble method requires setting a base estimator from which it builds the Ensemble. I use Adaptive Boosting (Freund and Schapire, 1997), or AdaBoost, as the base estimator. Boosting is an Ensemble method that converts a group of weak learners to a strong one (Zhou, 2012). In AdaBoost, each iteration of the algorithm dynamically adapts to the falsely classified instances of the last iteration. This has been shown to produce superior forecasting performance.[8]

A relevant concern over any resampling technique is that such technology would invariably change prior distribution before training. In our case, we are undersampling the "normal" cases to match the number of "crashes", resulting in a distribution of 50-50. The probabilities produced in such a system would reflect the new distribution, thus making further inferences about these probabilities less realistic. To match realistic prior distributions, I use the cross-validated classifier calibration technique (Zadrozny and Elkan, 2001, 2002; Platt et al., 1999; Niculescu-Mizil and Caruana, 2005) to bring the probability estimates back to realistic priors.[9]

[8]EasyEnsemble is a flexible algorithm that allows a large set of estimators. In untabulated results, using other base estimators produce similar results.

[9]The calibration is done via cross-validation. Each training sample, in our case, 6-month rolling data, is randomly split into training and validation sets. The base estimator (EasyEnsemble) is first trained on the training set and then calibrated to fit the validation set to obtain a probability estimate. Then the probabilities are averaged across each of the calibrated estimators for predicting the test set. Please refer to the referenced papers for technical details.

B.3. Forecasting Performance

To evaluate the forecasting performance of EasyEnsemble against logistic regression, I choose two commonly used metrics from the machine learning literature: ROC-AUC and Average Precision (McClish, 1989; Brodersen et al., 2010; Yue et al., 2007; He and Garcia, 2009). ROC-AUC is the area under the curve of ROC. Average precision is the weighted mean of precisions at each threshold, where the weight is the increase in recall from the prior threshold. The two metrics are threshold-free, meaning they measure the model's overall performance regardless of the decision threshold.

We choose these threshold-free metrics because we are uncertain about the true prior distribution of crashes since they are relatively rare. The threshold-free feature gives us tremendous flexibility in choosing the best model.

Since there are 294 forecasting windows, we have a time series of the metrics above. Thus we are able to compare the mean metrics between logistic regression and EasyEnsemble and compute standard errors and T-statistics. EasyEnsemble has a mean ROC of 0.775, while logit has a mean ROC of 0.759. The difference is 0.016 and is statistically significant at 1%.

B.4. Interpretation

EasyEnsemble, unlike logistic regression, does not produce readily interpretable coefficients. To pinpoint what variables are significant predictors of crashes, I follow Jiang et al. (2020) and compute Spearman's rank correlations between each variable and the estimated crash probabilities.

Specifically, I compute the rank correlations between each of the 204 variables and crash probabilities each month. Then I average each variable's rank correlations across time and save the time-series mean. To examine which variables are important predictors, I take the absolute values of these correlations and rank them based on the absolute values. The higher the absolute value, the more important the variable is. Since all variables are standardized before entering into machine learning algorithms, their levels of importance are directly

comparable. I plot the top 20 most important variables with their absolute rank correlations with crash risk in Figure A.1.

[Fig. A.1 about here.]

It can be seen from the figure that from the bottom to the top of the bar chart, or from the most important to the 20th most important: idiosyncratic volatility (Ali et al., 2003), idiosyncratic volatility by Fama-French three-factor model (IdioVol3F) (Ang et al., 2006), idiosycratic risk by CAPM (IdioRisk), bid-ask spread (Amihud and Mendelson, 1986), Maximum daily return over last month (MaxRet) (Bali et al., 2011), analyst forecasted earnings (FEPS) (Cen et al., 2006), 52-week high (High52) (George and Hwang, 2004), market cap (size), quarterly ROA (roaq) (Balakrishnan et al., 2010), ROE (Haugen and Baker, 1996), cash flow to market (CF) (Lakonishok et al., 1994), firm age (Barry and Brown, 1984), governance (Gompers et al., 2003), net payout yield (Boudoukh et al., 2007), net external financing (Bradshaw et al., 2006), share turnover volatility (std_turn) (Chordia et al., 2007), earnings forecast to price (SFE) (Elgers et al., 2001), predicted analyst forecast error (PredictedFE) (Frankel and Lee, 1998), revenue growth rank (Lakonishok et al., 1994), and off-season reversal years 16 to 20 (Heston and Sadka, 2008).

Many of these variables make intuitive sense. For example, idiosyncratic risk, 52-week high, and MaxRet have been shown to have a strong and negative correlation with future stock returns. Apart from the fact that these predictors are important in forecasting the cross-section of stock returns, their forecasting power naturally extends to stock crashes.

Appendix C. Other Results

C.1. Robustness: Alpha Estimates by Alternative Definitions of Crash Risk

This section presents the alpha estimates by defining a crash using the following thresholds: log monthly returns of less than -10%, -15%, -25%, and -30%. For brevity, I show only

alpha estimates of benchmarking the Fama-French five-factor model plus a momentum factor. The alpha estimates and their associated T-statistics for both logit and "EasyEnsemble" are presented in Table A.1.

[Table A.1 about here.]

C.2. Robustness: Changing Bandwidths and Dropping Earnings Months

This subsection reports the results of robustness tests for our main results. First, we use the same specifications but change the window to either [-1,+1] or 3 months, or [-2,+2] or 5 months. We report the results in Columns (1), (2), (4), and (5) in Table A.2. Second, we drop all stock-month observations that happen to be the months of earnings announcements, then run the same specification as our main test and report the results in Columns (3) and (6) in Table A.2.

[Table A.2 about here.]

C.3. Falsification Tests

To partially alleviate the endogeneity concern that online conversation of stock tickers is not random, I design a falsification test as follows. I randomly shuffle the months when the stocks become treated. Then I estimate the same specification as in Columns (3) and (4) of Table 11, except that the dummy variable $Treated$ is replaced by $Pseudo - Treated$. The results are reported in Table A.3.

[Table A.3 about here.]

C.4. Realized Crashes

Here we are interested in testing whether social transmission could directly cause realized crashes. To maintain consistency with the main tests, I run the same specification as the

main test except that I replace the dependent variable with a dummy variable $Crash_{i,t}$, which equals one if the stock crashes in the month, or equivalently, its log return is lower than -20% or any of the thresholds (namely, -10%, -15%, -25%, and -30%). Thus the specification is as follows:

$$Crash_{i,c,t} = \gamma_0 + \sum_{j=-3}^{+3} \beta_j D_{i,j,c,t} + \delta_{c,t} + \alpha_{i,c} + \epsilon_{i,t} \tag{19}$$

The results are reported in Table A.4.

[Table A.4 about here.]

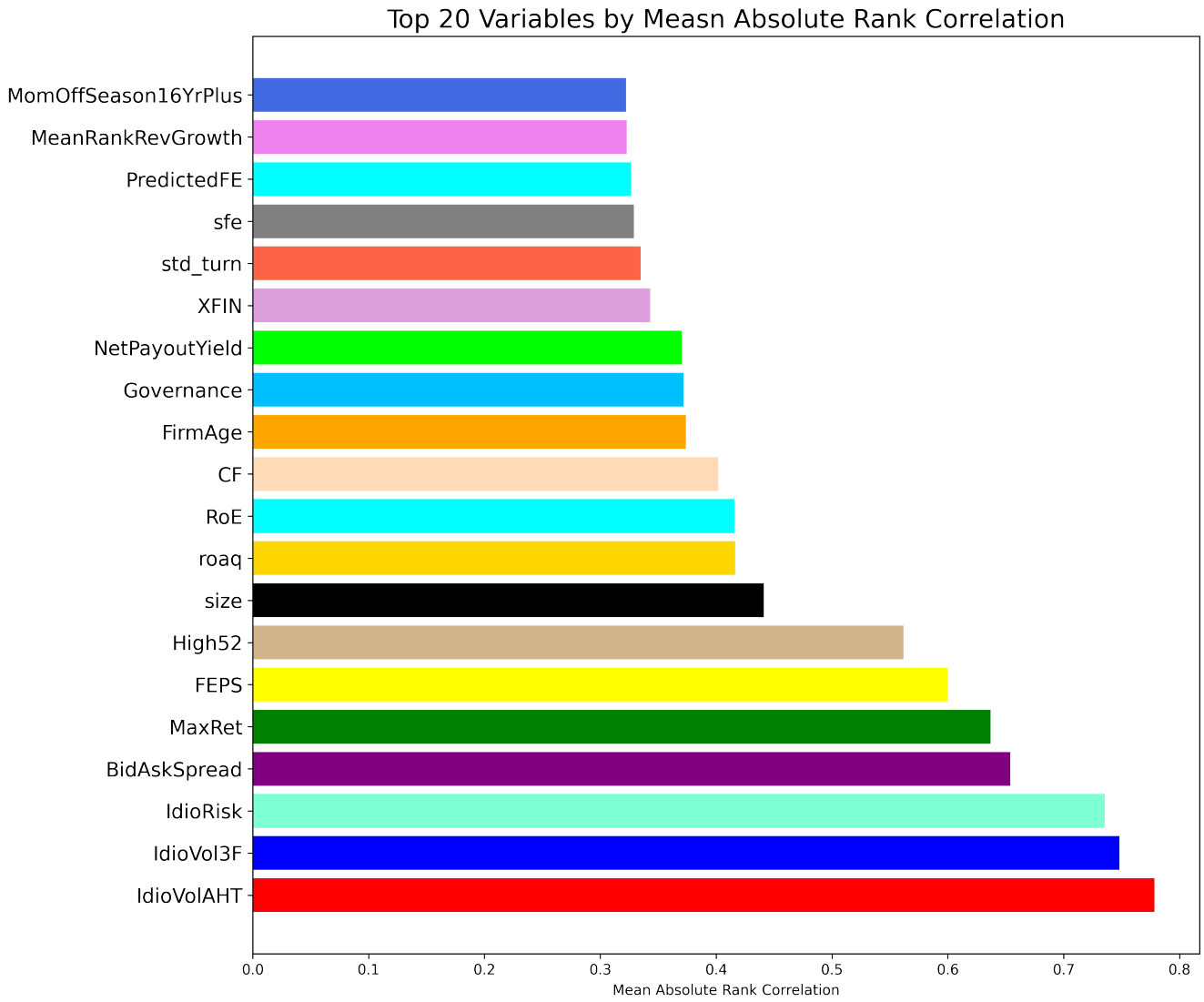

Fig. A.1. Top 20 Variables with Highest Absolute Rank Correlations with Crash Risk. The figure depicts the top 20 most important variables with their absolute rank correlations with crash risk. Specifically, I compute the rank correlations between each of the 204 variables and crash probabilities each month. Then I average each variable's rank correlations across the 294 test windows and save the time-series mean. To examine which variables are important predictors, I take the absolute values of these correlations and rank them based on the absolute values. The higher the absolute value, the more important the variable is. Since all variables are standardized before entering into machine learning algorithms, their levels of importance are directly comparable. From top to bottom of the bar chart, the 20th most important variable is off-season reversal years 16 to 20 (Heston and Sadka, 2008), and the most important variable is idiosyncratic volatility (Ali et al., 2003).

Table A.1: Decile High-Minus-Low Alphas: Alternative Definitions
This table presents the high-minus-low long-short zero-cost strategy alpha estimates and their associated T-statistics for both logistic regression and EasyEnsemble, by alternative definitions of crash risk. Crashes are defined using the following thresholds: log monthly returns of less than -10%, -15%, -25%, and -30%. Then the probabilities of crashes are estimated using these thresholds. At the end of each month, stocks are ranked by their ex-ante crash probabilities produced by either logit or EasyEnsemble into ten decile portfolios. Then the hedge portfolio return series are regressed on risk factor returns. The asset pricing model is the Fama-French five-factor model (Fama and French, 2015) augmented with momentum factor (FF6). For each of the thresholds, the upper panel presents results from value-weighted portfolios, while the lower panel presents equal-weighted results. The left half shows results from using crash risk estimated from logistic regressions, and the right from EasyEnsemble. T-statistics are included and standard errors are adjusted using the Newey-West (Newey and West, 1986) procedure with 6 lags.

		Logit		EEC-AdaBoost	
Threshold	Weighting	Alpha	T-stat	Alpha	T-stat
$log(ret) < -10\%$	value	-0.405	-1.291	-1.164	-3.989
	equal	-1.637	-6.467	-1.783	-6.466
$log(ret) < -15\%$	value	-0.855	-2.920	-1.249	-4.059
	equal	-1.601	-6.704	-1.758	-6.615
$log(ret) < -25\%$	value	-0.825	-2.764	-1.157	-3.920
	equal	-1.475	-5.816	-1.716	-6.358
$log(ret) < -30\%$	value	-0.751	-2.444	-1.047	-3.714
	equal	-1.444	-5.544	-1.603	-6.120

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.2: Robustness: Changing Band Widths and Dropping Earnings Months
This table reports the results of robustness tests that examine whether our main results hold if we change the length of the event windows or drop earnings months. The specification follows Table 6. Columns (1) and (4) use an event window of [-1,+1] or 3 months. Columns (2) and (5) use an event window of [-2,+2] or 5 months. Columns (3) and (6) drop all observations that report quarterly earnings during that month. Columns (1) – (3) use logit-generated crash risk as the dependent variable, while Columns (4) – (6) use machine learning-generated crash risk. Control variables include the natural log of market capitalization, prior-month return, asset growth, gross profitability, illiquidity (Amihud, 2002), MAX (Bali et al., 2011), prior 12-month return, and idiosyncratic risk. Standard errors are clustered at the unit level to account for possible duplicate observations.

	(1)	(2)	(3)	(4)	(5)	(6)
	Crash Risk (Logit)			Crash Risk (EEC)		
VARIABLES	3 Months	5 Months	No Earnings	3 Months	5 Months	No Earnings
Treated	0.362***	0.500***	0.529***	0.144*	0.241***	0.321***
	(0.140)	(0.126)	(0.165)	(0.076)	(0.066)	(0.083)
Observations	53,962	89,975	92,347	53,962	89,975	92,347
R-squared	0.943	0.923	0.921	0.966	0.955	0.954
Cohort×Units FE	YES	YES	YES	YES	YES	YES
Cohort×Month FE	YES	YES	YES	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.3: Wallstreetbets and Crash Risk: Falsification

This table conducts a falsification or placebo test on the impact of the appearances of stock tickers on "Wallstreetbets" on the stock crash risk. "Wallstreetbets" was founded in April 2012. The dependent variable is the crash risk estimated by both logit regression and machine learning method (EEC-AdaBoost), while the independent variable of interest is $Pseudo - Treated$, which is a dummy variable that equals one if the firm's ticker is first mentioned on "Wallstreetbets" in a particular month and if the period is after the "pseudo-month" when the ticker is first mentioned. The pseudo-treated status is generated by randomly shuffling the months when the treated stocks are first mentioned on "Wallstreetbets". The sample period starts from January 2012 to December 2020. We control for the natural log of market capitalization, the natural log of book-to-market ratio, asset growth, gross profitability, momentum, short-term reversal, idiosyncratic risk, illiquidity (Amihud, 2002), MAX (Bali et al., 2011), defined as the highest daily returns of the previous month, market beta, tail beta (Kelly and Jiang, 2014), coskewness (Harvey and Siddique, 2000), and net operating assets (Hirshleifer et al., 2004). Firm and time fixed effects are included, and standard errors are clustered at the firm level.

	(1)	(2)
	Dependent Var: Crash Risk	
VARIABLES	Logit	EEC
Pseudo-Treated	-0.000	0.000
	(0.001)	(0.001)
Controls	YES	YES
Observations	211,984	211,984
R-squared	0.691	0.814
Firm & Time FE	YES	YES
Note:		*p<0.1; **p<0.05; ***p<0.01

Table A.4: Wallstreetbets Conversations on Realized Crashes

This table reports results from a "stacked difference-in-differences" approach (Gormley and Matsa, 2011) that examines the effect of first appearances of stocks tickers on "Wallstreetbets" on whether they will experience realized crashes. The dependent variable is a dummy variable $Crash_{i,t}$, which equals one if the stock crashes in the current month, or equivalently, its log return is lower than -20% or any of the thresholds. From Column (1) to (5), the threshold varies from -10% to -30%. "Wallstreetbets" was started in April 2012. From the beginning of "Wallstreetbets" to the end of 2020, we find all the stock tickers that are ever mentioned in the Subreddit and the first month they were mentioned. We then define each of these instances as one event and each of the stocks as a treated stock. We match each treated stock with five control stocks from the pool of "never treated" stocks via propensity score matching based on lagged characteristics three months prior to each event. Then the "cohorts" containing treated and control observations are stacked together and the following specification is run:

$$Crash_{i,c,t} = \gamma_0 + \sum_{j=-3}^{+3} \beta_j D_{i,j,c,t} + \delta_{c,t} + \alpha_{i,c} + \epsilon_{i,t}$$

Where $D_{i,c,t}$ is a dummy variable that indicates whether a stock i in cohort c is treated at time t. $\delta_{c,t}$ is $Cohort \times Time$ fixed effects. $\alpha_{i,c}$ is $Unit \times Cohort$ fixed effects. Then β is the coefficient of interest that estimates the average treatment effect on the treated stocks. Standard errors are clustered at the unit level.

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Crash10	Crash15	Crash20	Crash25	Crash30
Treated	0.015***	0.010***	0.008***	0.008***	0.011***
	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
Constant	0.170***	0.110***	0.075***	0.052***	0.035***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	215,770	215,770	215,770	215,770	215,770
R-squared	0.550	0.552	0.548	0.547	0.541
Cohort×Units FE	YES	YES	YES	YES	YES
Cohort×Month FE	YES	YES	YES	YES	YES

Note: *p<0.1; **p<0.05; ***p<0.01

References

Abreu, D., Brunnermeier, M. K., 2003. Bubbles and crashes. Econometrica 71, 173–204.

Ali, A., Hwang, L.-S., Trombley, M. A., 2003. Arbitrage risk and the book-to-market anomaly. Journal of Financial Economics 69, 355–373.

Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. Journal of financial markets 5, 31–56.

Amihud, Y., Mendelson, H., 1986. Asset pricing and the bid-ask spread. Journal of financial Economics 17, 223–249.

An, H., Zhang, T., 2013. Stock price synchronicity, crash risk, and institutional investors. Journal of Corporate Finance 21, 1–15.

Andreou, P. C., Antoniou, C., Horton, J., Louca, C., 2016. Corporate governance and firm-specific stock price crashes. European Financial Management 22, 916–956.

Ang, A., Hodrick, R. J., Xing, Y., Zhang, X., 2006. The cross-section of volatility and expected returns. The journal of finance 61, 259–299.

Atilgan, Y., Bali, T. G., Demirtas, K. O., Gunaydin, A. D., 2020. Left-tail momentum: Underreaction to bad news, costly arbitrage and equity returns. Journal of Financial Economics 135, 725–753.

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. The journal of Finance 61, 1645–1680.

Balakrishnan, K., Bartov, E., Faurel, L., 2010. Post loss/profit announcement drift. Journal of Accounting and Economics 50, 20–41.

Bali, T. G., Cakici, N., Whitelaw, R. F., 2011. Maxing out: Stocks as lotteries and the cross-section of expected returns. Journal of financial economics 99, 427–446.

Banerjee, A. V., 1992. A simple model of herd behavior. Quarterly Journal of Economics 107, 797–817.

Barber, B. M., Huang, X., Odean, T., Schwarz, C., 2021. Attention induced trading and returns: Evidence from robinhood users. Journal of Finance, forthcoming .

Barber, B. M., Odean, T., 2000. Trading is hazardous to your wealth: The common stock investment performance of individual investors. The journal of Finance 55, 773–806.

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

Barber, B. M., Odean, T., Zhu, N., 2008. Do retail trades move markets? The Review of Financial Studies 22, 151–186.

Barber, B. M., Odean, T., Zhu, N., 2009. Systematic noise. Journal of Financial Markets 12, 547–569.

Barberis, N., Huang, M., 2008. Stocks as lotteries: The implications of probability weighting for security prices. American Economic Review 98, 2066–2100.

Barry, C. B., Brown, S. J., 1984. Differential information and the small firm effect. Journal of financial economics 13, 283–294.

Bates, D. S., 2000. Post-'87 crash fears in the s&p 500 futures option market. Journal of econometrics 94, 181–238.

Beason, T., Schreindorfer, D., 2022. Dissecting the equity premium. Journal of Political Economy 130, 2203–2222.

Bianchi, D., Büchner, M., Tamoni, A., 2021. Bond risk premiums with machine learning. The Review of Financial Studies 34, 1046–1089.

Bikhchandani, S., Hirshleifer, D., Welch, I., 1998. Learning from the behavior of others: Conformity, fads, and informational cascades. Journal of Economic Perspectives 12, 151–170.

Black, F., 1986. Noise. The journal of finance 41, 528–543.

Bollen, N. P., Whaley, R. E., 2004. Does net buying pressure affect the shape of implied volatility functions? The Journal of Finance 59, 711–753.

Boudoukh, J., Michaely, R., Richardson, M., Roberts, M. R., 2007. On the importance of measuring payout yield: Implications for empirical asset pricing. The Journal of Finance 62, 877–915.

Bradshaw, M. T., Richardson, S. A., Sloan, R. G., 2006. The relation between corporate financing activities, analysts' forecasts and stock returns. Journal of accounting and economics 42, 53–85.

Brodersen, K. H., Ong, C. S., Stephan, K. E., Buhmann, J. M., 2010. The balanced accuracy and its posterior distribution. In: *2010 20th international conference on pattern recognition*, IEEE, pp. 3121–3124.

Brunnermeier, M. K., Gollier, C., Parker, J. A., 2007. Optimal beliefs, asset prices, and the preference for skewed returns. American Economic Review 97, 159–165.

Callen, J. L., Fang, X., 2015. Short interest and stock price crash risk. Journal of Banking & Finance 60, 181–194.

Campbell, J. Y., Hilscher, J., Szilagyi, J., 2008. In search of distress risk. The Journal of Finance 63, 2899–2939.

Carhart, M. M., 1997. On persistence in mutual fund performance. The Journal of finance 52, 57–82.

Cen, L., Wei, J., Zhang, J., 2006. Forecasted earnings per share and the cross section of expected stock returns. Tech. rep., Working Paper, Hong Kong University of Science & Technology.

Cengiz, D., Dube, A., Lindner, A., Zipperer, B., 2019. The effect of minimum wages on low-wage jobs. The Quarterly Journal of Economics 134, 1405–1454.

Chang, I.-C., Liu, C.-C., Chen, K., 2014. The push, pull and mooring effects in virtual migration for social networking sites. Information Systems Journal 24, 323–346.

Chang, X. S., Chen, Y., Zolotoy, L., 2016. Stock liquidity and stock price crash risk. Journal of Financial and Quantitative Analysis (JFQA), Forthcoming .

Chen, A. Y., Zimmermann, T., 2021. Open source cross-sectional asset pricing. Critical Finance Review, Forthcoming .

Chen, J., Hong, H., Stein, J. C., 2001. Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. Journal of financial Economics 61, 345–381.

Chordia, T., Huh, S.-W., Subrahmanyam, A., 2007. The cross-section of expected trading activity. The Review of Financial Studies 20, 709–740.

Conrad, J., Kapadia, N., Xing, Y., 2014. Death and jackpot: Why do individual investors hold overpriced stocks? Journal of Financial Economics 113, 455–475.

De Long, J. B., Shleifer, A., Summers, L. H., Waldmann, R. J., 1990a. Noise trader risk in financial markets. Journal of political Economy 98, 703–738.

De Long, J. B., Shleifer, A., Summers, L. H., Waldmann, R. J., 1990b. Positive feedback investment strategies and destabilizing rational speculation. the Journal of Finance 45, 379–395.

Elgers, P. T., Lo, M. H., Pfeiffer Jr, R. J., 2001. Delayed security price adjustments to financial analysts' forecasts of annual earnings. The Accounting Review 76, 613–632.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of .

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

Fama, E. F., French, K. R., 2020. Comparing cross-section and time-series factor models. The Review of Financial Studies 33, 1891–1926.

Fama, E. F., MacBeth, J. D., 1973. Risk, return, and equilibrium: Empirical tests. Journal of political economy 81, 607–636.

Feng, G., Giglio, S., Xiu, D., 2020. Taming the factor zoo: A test of new factors. The Journal of Finance 75, 1327–1370.

Foucault, T., Sraer, D., Thesmar, D. J., 2011. Individual investors and volatility. The Journal of Finance 66, 1369–1406.

Frankel, R., Lee, C. M., 1998. Accounting valuation, market expectation, and cross-sectional stock returns. Journal of Accounting and economics 25, 283–319.

Freund, Y., Schapire, R. E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 55, 119–139.

George, T. J., Hwang, C.-Y., 2004. The 52-week high and momentum investing. The Journal of Finance 59, 2145–2176.

Gompers, P., Ishii, J., Metrick, A., 2003. Corporate governance and equity prices. The quarterly journal of economics 118, 107–156.

Gormley, T. A., Matsa, D. A., 2011. Growing out of trouble? corporate responses to liability risk. The Review of Financial Studies 24, 2781–2821.

Graham, J. R., Kumar, A., 2006. Do dividend clienteles exist? evidence on dividend preferences of retail investors. The Journal of Finance 61, 1305–1336.

Grossman, S. J., Stiglitz, J. E., 1980. On the impossibility of informationally efficient markets. The American economic review 70, 393–408.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.

Han, B., Hirshleifer, D., Walden, J., 2022. Social transmission bias and investor behavior. Journal of Financial and Quantitative Analysis 57, 390–412.

Han, B., Kumar, A., 2013. Speculative retail trading and asset prices. Journal of Financial and Quantitative Analysis 48, 377–404.

Harvey, C. R., Siddique, A., 2000. Conditional skewness in asset pricing tests. Journal of Finance 55, 1263–1295.

Haugen, R. A., Baker, N. L., 1996. Commonality in the determinants of expected stock returns. Journal of financial economics 41, 401–439.

He, H., Garcia, E. A., 2009. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering 21, 1263–1284.

Heston, S. L., Sadka, R., 2008. Seasonality in the cross-section of stock returns. Journal of Financial Economics 87, 418–445.

Hirshleifer, D., Hou, K., Teoh, S. H., Zhang, Y., 2004. Do investors overvalue firms with bloated balance sheets? Journal of Accounting and Economics 38, 297–331.

Hu, D., Jones, C. M., Zhang, V., Zhang, X., 2021. The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery. Available at SSRN 3807655 .

Hutton, A. P., Marcus, A. J., Tehranian, H., 2009. Opaque financial reports, r2, and crash risk. Journal of financial Economics 94, 67–86.

Jang, J., Kang, J., 2019. Probability of price crashes, rational speculative bubbles, and the cross-section of stock returns. Journal of Financial Economics 132, 222–247.

Jiang, H., Khanna, N., Yang, Q., Zhou, J., 2020. The cyber risk premium. Available at SSRN: https://ssrn.com/abstract=3637142 or http://dx.doi.org/10.2139/ssrn.3637142 .

Jin, L., Myers, S. C., 2006. R2 around the world: New theory and new tests. Journal of financial Economics 79, 257–292.

Kelley, E. K., Tetlock, P. C., 2017. Retail short selling and stock prices. The Review of Financial Studies 30, 801–834.

Kelly, B., Jiang, H., 2014. Tail risk and asset prices. The Review of Financial Studies 27, 2841–2871.

Kim, J.-B., Li, L., Lu, L. Y., Yu, Y., 2016. Financial statement comparability and expected crash risk. Journal of Accounting and Economics 61, 294–312.

Kim, J.-B., Li, Y., Zhang, L., 2011. Corporate tax avoidance and stock price crash risk: Firm-level analysis. Journal of Financial Economics 100, 639–662.

Kim, J.-B., Zhang, L., 2014. Financial reporting opacity and expected crash risk: Evidence from implied volatility smirks. Contemporary Accounting Research 31, 851–875.

Kim, Y., Li, H., Li, S., 2014. Corporate social responsibility and stock price crash risk. Journal of Banking & Finance 43, 1–13.

King, G., Zeng, L., 2001. Logistic regression in rare events data. Political analysis 9, 137–163.

Kozak, S., Nagel, S., Santosh, S., 2020. Shrinking the cross-section. Journal of Financial Economics 135, 271–292.

Kyle, A. S., 1985. Continuous auctions and insider trading. Econometrica: Journal of the Econometric Society pp. 1315–1335.

Lakonishok, J., Shleifer, A., Vishny, R. W., 1994. Contrarian investment, extrapolation, and risk. The journal of finance 49, 1541–1578.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. Journal of Accounting and economics 45, 221–247.

Li, K., Mai, F., Shen, R., Yan, X., 2021. Measuring corporate culture using machine learning. The Review of Financial Studies 34, 3265–3315.

Li, X., Wu, L., 2018. Herding and social media word-of-mouth: Evidence from groupon. Forthcoming at MISQ .

Liu, X.-Y., Wu, J., Zhou, Z.-H., 2008. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39, 539–550.

McClish, D. K., 1989. Analyzing a portion of the roc curve. Medical Decision Making 9, 190–195.

McCrank, J., 2021. Factbox: The u.s. retail trading frenzy in numbers. Thomson Reuters, URL: `https://www.reuters.com/article/us-retail-trading-numbers-idUSKBN29Y2PW`.

NBER, 2021. Us business cycle expansions and contractions.

Newey, W. K., West, K. D., 1986. A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Niculescu-Mizil, A., Caruana, R., 2005. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632.

Pan, J., 2002. The jump-risk premia implicit in options: Evidence from an integrated time-series study. Journal of financial economics 63, 3–50.

Parker, R., Graff, D., Kong, J., Chen, K., Maeda, K., 2011. English gigaword fifth edition, 2011. Linguistic Data Consortium, Philadelphia, PA, USA .

Pennington, J., Socher, R., Manning, C. D., 2014. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Platt, J., et al., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 10, 61–74.

Shleifer, A., Vishny, R. W., 1997. The limits of arbitrage. The Journal of finance 52, 35–55.

Van Buskirk, A., 2011. Volatility skew, earnings announcements, and the predictability of crashes. Earnings Announcements, and the Predictability of Crashes (April 28, 2011) .

Welch, I., 2020. Retail raw: Wisdom of the robinhood crowd and the covid crisis. Tech. rep., National Bureau of Economic Research.

Xing, Y., Zhang, X., Zhao, R., 2010. What does the individual option volatility smirk tell us about future equity returns? Journal of Financial and Quantitative Analysis pp. 641–662.

Yan, S., 2011. Jump risk, stock returns, and slope of implied volatility smile. Journal of Financial Economics 99, 216–233.

Yue, Y., Finley, T., Radlinski, F., Joachims, T., 2007. A support vector method for optimizing average precision. In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 271–278.

Zadrozny, B., Elkan, C., 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: *Icml*, Citeseer, vol. 1, pp. 609–616.

Zadrozny, B., Elkan, C., 2002. Transforming classifier scores into accurate multiclass probability estimates. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699.

Zhou, Z.-H., 2012. Ensemble methods: foundations and algorithms. CRC press.