University of Stavanger

Faculty of Business Administration

MASTER'S THESIS

Study program/ Specialization: Business Administration / Finance	Spring Semester 2021 Open
Writer: Josef Ali Gebory (2022), Erlend Domingo Sæbø (2030)	.... Writer's Signature
Faculty Supervisor Peter Molnar	
Thesis title: **Retail investors and the stock market: Evidence from Robinhood online broker**	
Credits (ECTS): 30	
Key Words • Google Searches • Popularity • Stock returns • Volatility • Trading Volume • Robinhood	Pages.....***....... +enclosure......***..... Stavanger June 15, 2021

Retail investors and the stock market: Evidence from Robinhood online broker

Erlend Domingo Sæbø*   Josef Ali Gebory†

Keywords: Google searches, popularity, stock returns, volatility, trading volume

*UiS Business School, University of Stavanger, Stavanger, Norway
†UiS Business School, University of Stavanger, Stavanger, Norway

ABSTRACT

We study how is the popularity among Robinhood investors related to stock returns, volatility and trading volume for the companies listed on the New York Stock Exchange. Robinhood is an online broker focused on retail costumers, and particularly millenials.

First we investigate which factors can explain popularity among Robinhood investors. We find that for a given week, more popular stocks exhibit positive returns, increased volatility, increased volume and increased ASVI.

We find that during the weeks when stocks are more popular, they exhibit low volatility and high trading volume. Furthermore, high popularity in the current week predicts next week's positive returns, low volatility and low trading volume and positive returns. Our results reveal that popularity is not able to explain contemporary return, but when put together with other control variables it can predict it. We also find that popularity predicts increased trading volume and volatility, as well as the abnormal search volume. Lastly our results showed that popularity are more related to current than future trading activity.

ACKNOWLEDGMENT

Foremost, we would like to express our gratitude to our supervisor Prof. Peter Molnar for the continuous support of our Master thesis, for his immense knowledge, responsiveness, motivation and patience. His guidance helped us a lot, and we could not have imagined a better advisor and mentor for our Master study.

Besides our advisor, we would like to acknowledge our warmest thanks to Prof. Stepan Mikula who made this work possible. His guidance and help carried us through a big part of the study. And we will forever be thankful.

Last but not least, we would like to thank our families for giving birth to us.

Contents

1. Introduction

With trading platforms such as Robinhood (RH) founded in 2013, the barrier for small investors entering the stock market has been lowered, and has therefore led to a surge in new retail investors directly owning stocks, and many investors are doing so for the first time. RH is a trading app which launched in 2015 and it was the first trading platform with zero commission fee. The app represents a certain group of investors, mainly millennials and new entrants to the US stock market. From 2019 to 2020 the amount of RH users went from 10 million to 13 million, and the firm is valued at over 8 billion dollars. Recently, RH enabled the download of large amounts of data showing the number of RH investors holding a particular stock at any given moment. With RHs blessing, the website Robintrack.net (RT) was able to continuously download the data and post it online. This lasted from mid-2018 to mid-2020.

One of the most researched topics within the finance is whether it is possible to predict stock price movements or not. Since the occurrence of new information is random, early research related to the efficient market hypothesis claimed that share prices follow a random path, driven by new information (Fama, 1965). However, it has been long recognized that the stock market is also influenced by investor attention and investor sentiment. Grinblatt and Keloharju (2000) find that stocks which had recently moved dramatically up or down were more often bought by retail investors in the 1990s, and a similar result were found by Welch (2020) who researched RH investors two decades later. Barber and Odean (2008) suggest that the reason to why investors decided to buy such stocks were mainly due to the stocks catching the attention of the investors.

Utilizing the data from RT, containing information regarding how many individual investors holds a specific share within the retail brokerage firm Robinhood, we can analyse a relatively large group of retail investors. Individual RH investors make up an insignificant portion of the market, but collectively they can make a difference. The event of the GameStop stock (GME) which happened in early 2021 is a great example of how the collective force of individual retail investors can move the stock price and impact the stock market (Chohan, 2021). The data downloaded from RT shows that even though some minor (but popular) events occurred in 2020, such as the Cannabis stock being, for a short period, the most held stock, with 244,532 investors holding Aurora Cannabis (ACB). However, most of the investments among the Robinhood investors were not only in the "popular" shares such as Aurora Cannabis, Snapchat or AMD (Welch, 2020). And on top of that most of the Robinhood investors increased their holdings and did not experience margin calls or panic. A comparison to the household-equivalent portfolio of the holdings from accounts at a discount brokerage firm (1991-1996) using Barber and Odean (2000) data and an Robinhood-equialent portfolio showed a 97.1% correlation in investment weight with the Robinhood-equivalent portfolio.

Recently there has been an increase in academic studies considering RH investor. Due to the COVID-19 lockdowns, investors have spent more time in front of their computers and smartphones, leading to higher trading volumes, as reported by Ozik et al. (2020). Barber et al. (2020) find that over the course of five days investors lost as much as 5%, due to herding-related buying, thereby viewing it as being disadvantageous. Welch (2020) showed that RH investors, over a twelve months period, shifted more towards stocks with higher than average trading volumes. The study further indicates that the RH consensus portfolio preformed well in the cross section, which explains why RH investors continued to invest.

Furthermore, Google Trends (GT) tools has often been used by researchers in various fields to identify trends among things such as, petrol prices Molnár and Bašta (2017), unemployment, consumer confidence and car sales Choi and Varian (2012). The use of the GT tool has made it possible to download historical search indexes of keywords, since google records search data for all keywords that reach a certain amount of searches. The availability of data from various resources such as GT, social media and news articles has gradually increased (Tetlock, 2007; Preis et al., 2013, 2010). This has enabled the effective market hypothesis to be examined through more critical ways in later research (Ang and Bekaert, 2007; Malkiel, 2003; Campbell and Yogo, 2006; Cochrane, 2008; Lo and MacKinlay, 1988). Recent studies have had a greater focus on the impact of investor sentiment, in which case Googles search volume index has been recognized as a significant mandate for investor sentiment (Baker and Wurgler, 2006; Barberis et al., 1998).

Researchers have made a few attempts to prognose the future financial market using data retrieved from GT, however the end results have varied. Challet and Ayed (2013) finds that strategies based on completely unrelated keywords do not surpass strategies based on financial keywords. This is achieved by testing the assumption that Google search volume consists of sufficient data to predict future financial index returns. Preis et al. (2013) examines whether it is possible to predict the market movements based on search volumes, and find that it is possible to surpass the market index by creating a strategy based on search volumes. A similar result is found by Moat et al. (2013) who predicts the stock return through the use of Wikipedia page searches. Bijl et al. (2016) find that when transaction costs are included, a trading strategy based on Google search will not continue to be profitable. Kim et al. (2019) show that for the largest Norwegian companies, an increase in Google searches predicts increased trading volume and volatility.

In this paper, we examine how Robinhood Shares traded on the NYSE relate to the relationship between popularity and return, volatility, abnormal search volume and trading volume. We examine how is the popularity among Robinhood investors related to returns, volatility, trading volume and internet searches for company tickers. Second, we examine whether popularity can be predicted through market activity factors. Popularity can neither explain nor predict returns in a univariate model, but becomes significant in a predictive multivariate model. Despite the fact that popularity does not have the ability to predict or explain returns single-handedly within a 8 week horizon, the opposite is true for volatility and trading volume. Our conclusion is in line with what we expected from the market activity.

The following sections are represented as follows. The methodology is described in section 3. Section 4 presents our findings and a discussion of the results. In section 5 we check the robustness of our findings. Section 6 concludes.

2. Data

The data used in this paper is obtained from RT, GT and Yahoo finance. The data retrieved from Yahoo Finance consists of daily open, close, high, low, adjusted close price, and the trading volume for the companies from the RT list. The GT platform is used to obtain the search volume index (from now on SVI) data. The data collected from RT consists of the number of Robinhood users that hold a particular stock, a timestamp and the security ticker from May 2018 to August 2020. Data from these three sources (RT, GT and Yahoo Finance) are merged based on tickers. All companies that were in the NYSE index from 2018 to 2020 and have complete stock data have been included. Companies with a low search volume on the search words for which Google does not provide any data at all are omitted. Our final sample consists of 3621 companies.

2.1. Google Trends data

Data that individual users enter into the Google Search engine is collected by Google and can be accessed via GT. The scale used in GT is a standardized scale from 0 to 100, where the highest query volume in a given time period for a given region is 100 (Choi and Varian, 2012).

Our analysis is performed using weekly Google Search data. According to Preis et al. (2010) a correlation between GT data and transaction volume of the corresponding shares is observed on a weekly time scale. Furthermore, Bijl et al. (2016) points out that weekly abnormal searches and subsequent stock returns have a significant negative connection.

The GT data is matched with the stock market and Robinhood data, to check if the google Search volume index (from now on SVI) has an impact on the popularity of a stock within the American stock market. The standardized reporting scale from GT is from Sunday to Sunday, while the American stock market is from Monday to Friday. Therefore we defined the financial week from Monday close to Monday close.

In GT you can specify the way you do your search in two ways. (1) Search terms, which show matches for all the terms only in the language the query was made, or (2) topics that share the same concept in any language. Kim et al. (2019) studied the difference between these two search terms defined as: search term (S_t) and business term (B_t). They concluded that models based on the B_t had less explanatory power in comparison to models based on the S_t, they recommend using a simple term. This led us to dropping the business term (B_t)

in our thesis.

The GT tool lets us choose which region we would like to retrieve SVI from. According to Preis et al. (2013) data filtered with the regards of a geographic area, can better explain the movements in this area. Our data from RT consists primarily of daily data from the American stock market, hence we filtered our data to The United States of America.

There are also other several ways of filtering out the data; (1) Arts & Entertainment; (2) Autos & Vehicles; (3) Beauty & Fitness; (4) Books & Literature; (5) Business & Industrial; (6) Computer & Electronics; (7) Finance; (8) Food & Drink; (9) Games; (10) Health; (11) Hobbies & Leisure; (12) Home & Garden; (13) Internet & Telecom; (14) Jobs & Education; (15) Law & Government; (16) News; (17) Online Communities; (18) People & Society; (19) Pets & Animals; (20) Real Estate; (21) Reference; (22) Science; (23) Shopping; (24) Sports; and (25) Travel. The default is "all categories". By using the finance filter, we managed to fix our S_t and B_t even though it gave us a dataset which contained mostly zero values; hence we omitted them. This helps confirm the findings in Bijl et al. (2016) that the finance filter does not provide any further information regarding the terms of predicting stock returns over the unfiltered searches.

Finally, GT makes it possible to filter out which platform you want the informative search to come from: Images, Google Shopping, Web Search and YouTube. We chose to edit our filter to only observe web searches, as web searches were the only platform that gave us the least zero values.

To calculate the abnormal search value index we used raw SVI (from now on ASVI). Because of the dependent variable time period, the raw SVI can not be used directly in the analysis. For example the data from 2019 depends on what data we are using, it could be from 2017 to 2019, or 2019 to 2020. Therefore we need to standardize the data from past history, hence using the method from Da et al. (2011).

The Bijl et al. (2016) method subtracts the average of a time period (in our case 8 weeks) from the weekly SVI and then divides their difference from the standard deviation from that given period:

$$ASVI_t^B = \frac{SVI - \frac{1}{8}\sum_{i=1}^{8} SVI_{t-i}}{\sigma_{SVI,t}} \tag{1}$$

SVI_t the search volume index, and $\sigma_s vi_t$ is the standard deviation of the population period of 8 weeks.

The Da et al. (2011) method subtracts the weekly log SVI from the log of the median SVI from the whole sample:

$$ASVI_t^D = \log SVI_t - \log[Mean(SVI_{t-1}, \ldots, SVI_{t-8}).] \tag{2}$$

Fig.1 shows us how the raw data (before we standardized it) SVI, of the top three companies in the S&P 500 list, and how the illustration changes after we standardized it in two different ways. As you can see from the figure, the standardization makes it way easier to compare between the companies. Further you can see the main difference between the Bijl et al. (2016) and Da et al. (2011), is that the last mentioned often results in very low values, because of its logarithmic form.

2.2. Stock Market data and Robintrack data

Wall Street is home to the two largest stock exchanges in the world, NYSE and NASDAQ, considering total market capitalization. Wall Street also houses four more exchanges such as: The New York mercantile exchange, The New York board of trade, the New York futures exchange, and the former American stock exchange. These six marketplaces offer several financial instruments such as; equities, fixed income products, derivatives products, and equity certificates.

RT offers daily data on all the stocks traded on the Robinhood app between 2018-2020. The data obtained includes: end of day and meanday. The purpose of this website was to show the popularity of certain stocks, and it helped with the transparency of the stock market. This project got taken down in august 2020, hence our historical data between 05.01.2018-08.01.2020. For each stock, we derived the last observation for each day (last UTC).

As an addition to the RT data, we also have a variable called "endofday". This variable will also be conducted as a measure for popularity. This will let us look at a stock's popularity not only based on google searches, but actually on how many retail investors that are in a specific stock- on a specific day.

Our dataset consists of 117 weeks and 4 days. The data downloaded from RT had to be adjusted, as it gave us data from all weekdays including Saturday and Sunday. It also gave us data from American holidays such as: New years day, Martin Luther King Jr Day, Good Friday, memorial day, independence day, Labor day, Thanksgiving day, Presidents day, and Christmas, which all had to be omitted because of zero values. Further we computed the

Fig. 1. ASVI BIJI ET AL

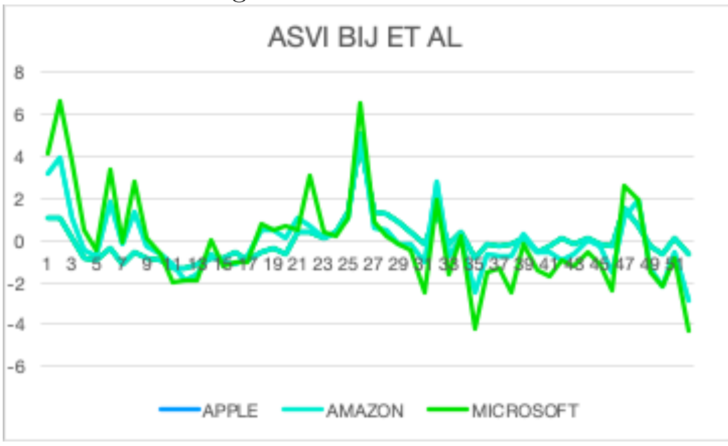


Fig. 2. ASVI DA ET AL

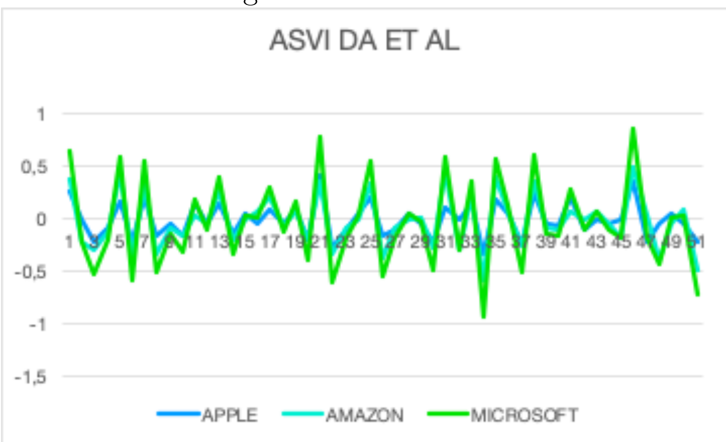
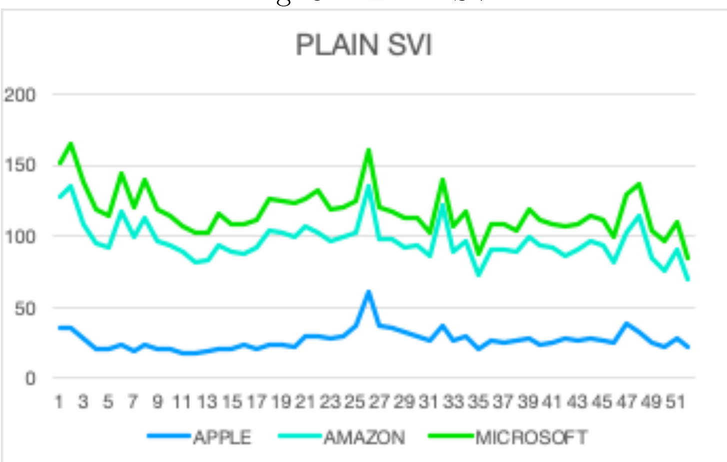


Fig. 3. PLAIN SVI



formula for logarithmic returns:

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right) \tag{3}$$

We downloaded the adjusted closing price directly from Yahoo finance then we merged it with our data from RT, and matched it with the rightful stock. In equation (3) r_t is the raw logarithmic return, P_t is the adjusted stock price for a week, and P_{t-1} is the adjusted price for the previous week.

Next we construct a volume trading variable, since we have daily data on the trading volume TV_t^D, we convert them to weekly TV_t to match with our SVI data. To convert the data, we simply calculate the average trading volume:

$$TV_t = \frac{1}{|S_t|}\sum_{i\in S_t} TV_t^D \tag{4}$$

S_t is the total amount of trading days in a given week, and t and $|S_t|$ is the amount of trading days in a given week.

Further we calculate the abnormal trading volume (from now on ATV), from the formula in Bijl et al. (2016). The ATV_t is calculated in the exact same way as in equation (1), by subtracting the average volume for the 8 week period from the weekly volume, then dividing it by the standard deviation for the 8 week period:

$$ATV_t = \frac{TV_t - \frac{1}{8}\sum_{i=1}^{8} TV_{t-i}}{\sigma_{TV,t}} \tag{5}$$

Where σ_{TV} denote the standard deviation for volume of the whole sample, and TV is the weekly trading volume.

After the abnormal trading volume, we add volatility to our financial variables. Volatility is a measure that makes it possible to evaluate a stock's return over time. Previous studies have found a positive relationship between volatility and future stock returns (French et al., 1987; Banerjee et al., 2007; Bollerslev et al., 2009), hence we will not only add volatility as a control variable in our regression model explaining volume and returns, but as a measure for market activity as well. We measure the volatility by using the (Garman and Klass.1980) volatility estimator adjusted for the opening jump, as explained in Molnar(2012). The variance is calculated by using the information given from the open ($open_t$), high ($high$), low (low_t), close ($close_t$) and adjusted close prices ($radj_t$), and adjusted close price during a given trading day t:

$$Variance_t = \frac{1}{2}(h_t - l_t)^2 - (2\log 2 - 1)c_t^2 + jadj_t^2 \tag{6}$$

From the Equation:

$$\begin{aligned}
c_t &= \log(close_t) - \log(open_t), \\
l_t &= \log(low_t) - \log(open_t), \\
h_t &= \log(high_t) - \log(open_t), \\
j_t &= \log(open_t) - \log(close_{t-1}), \\
r_t &= \log(close_t) - \log(close_{t-1}), \\
jadj_t &= j_t \frac{radj_t}{r_t}.
\end{aligned}$$

Finally, we get the weekly volatility by squaring the average daily variance:

$$Volatility_t = \sqrt{\frac{1}{|S_t|}\sum_{i\in S_t} Variance_t} \tag{7}$$

Lastly we add the Robinhood data and create a variable named "Popularity", which is denoted as "endofday" further up in our thesis. To standardize this variable, we took the logarithmic mean value for 8 weeks and subtracted it from the logarithmic value for a given week:

$$Popularity = log(endofday) - [mean(endofday_{t-1}, endofday_{t-8})] \tag{8}$$

Where endofday is the number of individual investors in a given stock at a given period of time. Popularity is the name of the variable, as we tend to discover how popular a stock is by its unique investors.

2.3. Descriptive Statistics

The summary statistics for the variables generated from our dataset are presented in Table 1. Using the Bijl et al. (2016) formula discussed in section 2.1, we are able to standardize and calculate the variables: popularity, ASVI, ATV and Volatility.Furthermore, the weekly

Garman-class jump-adjusted estimator is used to calculate volatility. Return is calculated as logarithmic returns

Table 1: Descriptive Statistics for all variables

	n	mean	sd	median	min	max	skew	kurtosis
return	$84,053$	-0.115	10.583	0.222	0.040	4.802	-240.832	210.006
Popularity	$84,053$	0.113	0.225	0.050	0.072	0.090	-6.769	5.018
ASVI	$84,053$	0.230	0.693	0.182	0.213	0.641	-2.672	3.858
ATV	$84,053$	-0.049	0.551	-0.081	-0.064	0.404	-6.433	5.660
Volatility	$84,053$	-0.050	0.492	-0.113	-0.090	0.387	-6.928	4.069

In Table 2, we present the correlation between the variables. We can observe that most variables have very low correlations. In particularly it is noted that variable ATV and Volatility has a correlation of 0,58, indicating that both variables share a kind of dependency.

Table 2: Correlation matrix for all variables.

	Return	Popularity	ASVI	ATV	Volatility
Return	1	0.01	-0.02	0.002	-0.16
Popularity	0.01	1	0.16	0.24	0.09
ASVI	-0.02	0.16	1	0.21	0.22
ATV	0.002	0.24	0.21	1	0.58
Volatility	-0.16	0.09	0.22	0.58	1

As suggested by Foster and Viswanathan (1993) the trading volume significantly contributes to higher volatility. This might reflect that high volume trading will affect the price of a given stock.

3. Methodology

In this thesis we investigate the impact popularity has on return, ASVI, volatility, volume, and whether or not it can predict and/or explain these variables by utilizing panel data regressions. In descriptive regressions, the popularity variable is contemporary with the dependent variable. We have two types of regressions: explanatory and predictive. In the last mentioned, we used lagged popularity to investigate whether or not past popularity can predict future returns, volume, ASVI and volatility. The regressions can be found below.

First we check the descriptive model of return measure, here we regress the return against Popularity and the set of control variables:

$$Return_t = \alpha + \beta_1 ATV_t + \beta_2 ASVI_t + \beta_3 Volatility_t + \beta_4 Popularity_t + \epsilon_t \tag{9}$$

Our next descriptive model is motivated by Da et al. (2011) on whether or not ASVI can be used as a proxy to capture investors' attention,hence volume as the dependent variable. We wanted to investigate if this was also the case for both ASVI and popularity. We wanted to see if changes in search and Popularity explain a significant change in trading volume:

$$ATV_t = \alpha + \beta_1 ASVI_t + \beta_2 Return_t + \beta_3 Volatility_t + \beta_4 Popularity_t + \epsilon_t \tag{10}$$

ATV_t, represents the abnormal trading volume for a given time for a specific firm. The β's represents the coefficients for the lagged ATV, ASVI, stock return, volatility and Popularity.

Our next descriptive model investigates if there is a contemporary relationship between the volatility and how many retail investors that are inside a certain stock:

$$Volatility_t = \alpha + \beta_1 ASVI_t + \beta_2 Return + \beta_3 ATV_t + \beta_4 Popularity_t + \epsilon_t \tag{11}$$

$Volatility_t$, is the return volatility at time t for a given firm i, β's are the coefficients for the lagged volatility, ASVI, return, ATV and Popularity.

Next we investigate the contemporary relationship between Popularity and ASVI. We estimate the fourth descriptive model:

$$ASVI_t = \alpha + \beta_1 ATV_t + \beta_2 Return_t + \beta_3 Volatility_t + \beta_4 Popularity_t + \epsilon_t \tag{12}$$

In the descriptive model of Popularity measure, we regress the Popularity against the stock return and the set of control variables mentioned in 2.2. This gave us the following regression model:

$$Popularity_t = \alpha + \beta_1 ASVI_t + \beta_2 Volatility_t + \beta_3 ATV_t + \beta_4 Return + \epsilon_t \tag{13}$$

Where Popularity is the measure for all unique retail investors inside a certain stock, β's are the coefficients for the lagged popularity, ASVI, volatility, Volume and stock return.

The predictive models are built up in a similar way, the only difference is that we only exploited lagged variables in these models. This is done to help us investigate how past information gives us a prediction of future values. The models can be built up in this general form:

$$Y_{t,i} = \alpha_i + \beta_1 controls_{t-1} + \epsilon_t \tag{14}$$

Where $Y_{t,i}$, is the dependent variable for one of the above, at time t for a specific for i, regressed on its lagged value,and on the lagged value of all the other chosen variables

For our regressions we used panel data and consider the firm fixed effects indicated by index i at the intercepts coefficient α .

4. Results

Examining whether the popularity of a stock has a significant effect on the return of a stock, we chose an ordinary least squares regression. Testing for heteroskedasticity and auto-correlation we ran both Breuch-Pagan test and Durbin-Watson test. Our dataset contained both heteroskedasticity and auto-correlation, hence we controlled them and display robust standard errors within all our tables.

The results presented beneath has been standardized with standardization from Da et al. (2011). When all the regular regressions are presented, we conducted robustness checks with different time horizons, this is done to investigate what kind of impact popularity had over time. This is done for all the variables, to get a better view on the predictive power each variable has.

Our first model shows how the results aim to explain the logarithmic returns. Both descriptive and predictive results are presented in Table 3 and Table 4 column 1-4 are the results of a unvaried analysis. Significant results can also be found in the multivariate descriptive analysis column 5. The popularity variable is insignificant in explaining the return both in the univariate and multivariate contemporary models, but becomes significant in the lagged multivariate model. We can conclude that the popularity is insignificant in explaining stock return alone and controlled with other control variables. However when it comes to predicting the stock return, it becomes significant when controlled with other variables. This can indicate that even though there are a lot of unique retail investors inside a certain stock, it does not have a predictive power of one's return as a single factor.

Table 3: Regression results for explanatory model of return. Columns (1)–(4) report results from a single regression to explain returns by various independent variables.Column (5) report results from a multivariate regression. Robust standard errors are reported in brackets. The sample period covers weekly data from 2018 to 2020. The symbols ***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively

	Dependent variable: Return_t				
	(1)	(2)	(3)	(4)	(5)
ATV_t	0.046				2.803^{***}
	(0.176)				(0.171)
ASVI_t		−0.333			0.023
		(0.078)			(0.055)
Volatility_t			-3.447^{***}		-5.283^{***}
			(0.162)		(0.161)
Popularity_t				0.474	−0.110
				(0.500)	(0.457)
R^2	0.00001	0.0005	0.026	0.0001	0.040
Adjusted R^2	−0.00001	0.0005	0.026	0.0001	0.040
Significance levels			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$		

Table 4: Regression results for predictive model of return. Columns (1)–(4) report results from a single regression to predict returns by various independent variables.Column (5) report results from a multivariate regression. Robust standard errors are reported in brackets. The sample period covers weekly data from 2018 to 2020. The symbols ***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively

	Predictive model				
	Return_t				
	(1)	(2)	(3)	(4)	(5)
ATV_{t-1}	-1.372^{***}				-0.687^{***}
	(0.110)				(0.109)
ASVI_{t-1}		-0.669^{***}			-0.444^{***}
		(0.058)			(0.055)
Volatility_{t-1}			-1.929^{***}		-1.506^{***}
			(0.1)		(0.114)
Popularity_{t-1}				0.226	1.838^{***}
				(0.309)	(0.296)
R^2	0.006	0.002	0.008	0.00003	0.011
Adjusted R^2	0.006	0.002	0.008	0.00001	0.011
Significance levels					*p<0.1; **p<0.05; ***p<0.01

Our next regression table presents the movements in the trading volume. Table 5 and Table 6 shows that Popularity is significant at explaining the abnormal trading volume, no matter if it's contemporary or lagged. Popularity is still significant even when controlled with other variables. Even though Popularity has a significant explaining power, most of the variation in trading volume can be explained by the trading volume from the previous week. In accordance with our expectations, the contemporary search volume is significantly correlated with the volume. Kim et al. (2019) find that search volume can both predict and explain trading volume for the Norwegian stock market, which is also the case for the U.S stock market.

Table 5: Regression results for explanatory model of ATV. Columns (1)–(4) report results from a single regression to explain volume by various independent variables.Column (5) report results from a multivariate regression. Robust standard errors are reported in brackets. The sample period covers weekly data from 2018 to 2020. The symbols ***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively

	Explanatory model				
	ATV_t				
	(1)	(2)	(3)	(4)	(5)
$ASVI_t$	0.167***				0.047***
	(0.003)				(0.002)
$Return_t$		0.0001			0.005***
		(0.0004)			(0.0003)
$Volatility_t$			0.654***		0.638***
			(0.006)		(0.005)
$Popularity_t$				0.584***	0.429***
				(0.016)	(0.014)
R^2	0.044	0.00001	0.341	0.057	0.387
Adjusted R^2	0.044	−0.00001	0.341	0.057	0.387
Significance levels			*p<0.1; **p<0.05; ***p<0.01		

Table 6: Regression results for predictive model of ATV. Columns (1)–(4) report results from a single regression to predict volume by various independent variables.Column (5) report results from a multivariate regression. Robust standard errors are reported in brackets. The sample period covers weekly data from 2018 to 2020. The symbols ***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively

	Predictive model				
	ATV_t				
	(1)	(2)	(3)	(4)	(5)
$ASVI_{t-1}$	0.092^{***}				0.053^{***}
	(0.002)				(0.002)
$Return_{t-1}$		-0.002^{***}			-0.001^{***}
		(0.0002)			(0.0002)
$Volatility_{t-1}$			0.288^{***}		0.264^{***}
			(0.006)		(0.006)
$Popularity_{t-1}$				0.231^{***}	0.143^{***}
				(0.013)	(0.011)
R^2	0.017	0.002	0.069	0.010	0.079
Adjusted R^2	0.017	0.002	0.069	0.010	0.079
Significance levels					*p<0.1; **p<0.05; ***p<0.01

Next in table 7 and table 8, we investigate the relationship between our control variables and volatility. In this regressions the stock price volatility is the dependent variable. Popularity is significant at a 99% confidence level, both contemporary and lagged. It also stays significant at a 99% confidence level when augmented with other control variables. The contemporary trading volume is correlated with volatility at a 99% confidence level, both in the univariate and multivariate model, as expected. Therefore an explanatory relationship between the current week's retail investors with stock price volatility seem to exist.

Table 7: Regression results for explanatory model of volatility. Columns (1)–(4) report results from a single regression to explain Volatility by various independent variables.Column (5) report results from a multivariate regression. Robust standard errors are reported in brackets. The sample period covers weekly data from 2018 to 2020. The symbols ***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively

	Explanatory model				
	Volatility$_t$				
	(1)	(2)	(3)	(4)	(5)
ASVI$_t$	0.155***				0.073***
	(0.003)				(0.002)
Return$_t$		$-$0.007***			$-$0.007***
		(0.0003)			(0.0002)
ATV$_t$			0.521***		0.515***
			(0.01)		(0.009)
Popularity$_t$				0.201***	$-$0.133***
				(0.01)	(0.01)
R^2	0.048	0.026	0.341	0.008	0.379
Adjusted R^2	0.048	0.026	0.341	0.008	0.379
Significance levels					*p<0.1; **p<0.05; ***p<0.01

Table 8: Regression results for predictive model of volatility. Columns (1)–(4) report results from a single regression to predict Volatility by various independent variables.Column (5) report results from a multivariate regression. Robust standard errors are reported in brackets. The sample period covers weekly data from 2018 to 2020. The symbols ***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively

	Predictive model				
	Volatility_t				
	(1)	(2)	(3)	(4)	(5)
ASVI_{t-1}	0.086^{***}				0.058^{***}
	(0.002)				(0.002)
Return_{t-1}		-0.006^{***}			-0.007^{***}
		(0.0002)			(0.0002)
ATV_{t-1}			0.235^{***}		0.260^{***}
			(0.005)		(0.005)
Popularity_{t-1}				-0.080^{***}	-0.278^{***}
				(0.009)	(0.01)
R^2	0.019	0.023	0.077	0.001	0.128
Adjusted R^2	0.019	0.023	0.077	0.001	0.128
Significance levels					*p<0.1; **p<0.05; ***p<0.01

Our next model aims to explain the relationship between our dependent variable search volume. Also in table 9 and table 10 the Popularity variable shows a significant relationship at a 99% confidence level, both at univariate, multivariate, contemporary and lagged. Therefore we can conclude that there is an explaining and predictive power between the search volume and the popularity.

Table 9: Regression results for explanatory model of ASVI. Columns (1)–(4) report results from a single regression to explain ASVI by various independent variables.Column (5) report results from a multivariate regression. Robust standard errors are reported in brackets. The sample period covers weekly data from 2018 to 2020. The symbols ***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively

	Explanatory model				
	ASVI_t				
	(1)	(2)	(3)	(4)	(5)
ATV_t	0.265***				0.113***
	(0.005)				(0.005)
Return_t		−0.001			0.0001***
		(0.0002)			(0.0002)
Volatility_t			0.308***		0.218***
			(0.004)		(0.005)
Popularity_t				0.507***	0.397***
				(0.011)	(0.01)
R^2	0.044	0.0005	0.048	0.027	0.074
Adjusted R^2	0.044	0.0005	0.048	0.027	0.074
Significance levels					*p<0.1; **p<0.05; ***p<0.01

Table 10: Regression results for predictive model of ASVI. Columns (1)–(5) report results from a single regression to predict ASVI by various independent variables.Column (6) report results from a multivariate regression. Robust standard errors are reported in brackets. The sample period covers weekly data from 2018 to 2020. The symbols ***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively

	Predictive model				
	ASVI_t				
	(1)	(2)	(3)	(4)	(5)
ATV_{t-1}	0.212***				0.134***
	(0.004)				(0.004)
Return_{t-1}		−0.0001*			−0.0002
		(0.0002)			(0.0002)
Volatility_{t-1}			0.211***		0.131***
			(0.004)		(0.005)
Popularity_{t-1}				0.339***	0.191***
				(0.009)	(0.010)
R^2	0.032	0.00000	0.024	0.014	0.041
Adjusted R^2	0.032	−0.00001	0.024	0.014	0.041
Significance levels			*p<0.1; **p<0.05; ***p<0.01		

Finally, table 11 and table 12 presents the results for Popularity as a dependent variable. What we are most interested to look at in this table is the return. We want to look at how return might explain and/or predict future and present popularity. The R^2 for the return both in univariate and multivariate, and both for lagged and contemporary, is quite high. Even though the coefficients are very low, return becomes significant in explaining the popularity when combined with other market factors. Return is actually better at predicting popularity than explaining it, this signifies that the sentiment and attention for companies trading on the NYSE are captured by the logarithmic return.

Table 11: Regression results for explanatory model of Popularity. Columns (1)–(4) report results from a single regression to explain Popularity by various independent variables.Column (5) report results from a multivariate regression. Robust standard errors are reported in brackets. The sample period covers weekly data from 2018 to 2020. The symbols ***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively

	Explanatory model				
	Popularity_t				
	(1)	(2)	(3)	(4)	(5)
ATV_t	0.097^{***}				0.108^{***}
	(0.002)				(0.002)
ASVI_t		0.053^{***}			0.042^{***}
		(0.0008)			(0.0006)
Volatility_t			0.042^{***}		-0.041^{***}
			(0.002)		(0.001)
Return_t				0.0002	-0.00005^{*}
				(0.00015)	(0.00013)
R^2	0.057	0.027	0.008	0.0001	0.076
Adjusted R^2	0.057	0.027	0.008	0.0001	0.076
Significance levels					$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 12: Regression results for predictive model of Popularity. Columns (1)–(4) report results from a single regression to explain Popularity by various independent variables.Column (5) report results from a multivariate regression. Robust standard errors are reported in brackets. The sample period covers weekly data from 2018 to 2020. The symbols ***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively

	Explanatory model				
	Popularity$_t$				
	(1)	(2)	(3)	(4)	(5)
ATV$_{t-1}$	0.110***				0.095***
	(0.001)				(0.001)
ASVI$_{t-1}$		0.053***			0.034***
		(0.0006)			(0.0005)
Volatility$_{t-1}$			0.057***		-0.005^{***}
			(0.001)		(0.001)
Return$_{t-1}$				0.0001***	-0.0004^{***}
				(0.00009)	(0.00008)
R^2	0.081	0.034	0.017	0.00001	0.084
Adjusted R^2	0.081	0.034	0.017	-0.00000	0.084
Significance levels				*p<0.1; **p<0.05; ***p<0.01	

We can sum all these tables up, and conclude that Popularity can tell us more about both future and present trading activity for the US stock market.

5. Robustness checks

In this section we conduct several robustness checks to see if our hypothesis still holds after different types of time horizons. First we check if our results are robust with the respect of the model from Da et al. (2011) where we change the time lapse from 8 weeks, 16 weeks, and 32 weeks.

5.1. *Different Time horizons*

This subsection displays the model results for different time lapse. We calculated three different time horizons: 8 weeks, 16 weeks and 32 weeks which will be presented respectively.

These different time horizons are all calculated in the same way. We used the formula from equation (2), and changed the mean variation between 8, 16 and 32 weeks. This gave us the ability to see whether the explanatory and predictive power for popularity was as robust as in the results section. If the explanatory and predictive power still holds for a longer time horizon, we can conclude that the results are robust.

The tables below will display the explanatory and predictive power of Popularity, when it comes to Volatility, return, trading volume and ASVI. The models will only show the estimated coefficients, along with robust standard errors and the R^2.The significance will also here be shown by the "*".

In table 13 and 14 ; Sensitivity to Popularity standardization: Returns as the dependent variable. Based on the formula from Da et al. (2011), the results of the standardization in the last 8, 6 and 32 weeks have been reported. Univariate regressions with lagged or contemporary popularity have been reported. The tables are set up in the order in which the complete descriptive model with the simultaneous independent variables comes first. The table is then followed by second predictive model with lagged values. The robust standard errors are reported in parentheses. Weekly data from the period 2018 to 2020 are covered in the sample period. The symbols,***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 13: Sensitivity to Popularity standardization: Return as the dependent variable.

	Explanatory models:					
	Return_t					
	8 weeks		16 weeks		32 weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
Popularity_t	0.474	−0.110	0.936***	0.418	1.101***	0.781***
	(0.500)	(0.457)	(0.324)	(0.282)	(0.210)	(0.179)
ASVI_t		0.023		0.037		0.004
		(0.054)		(0.051)		(0.050)
ATV_t		2.803***		2.629***		2.321***
		(0.171)		(0.172)		(0.165)
Volatility_t		−5.283***		−4.798***		−4.186***
		(0.161)		(0.146)		(0.142)
R^2	0.0001	0.040	0.001	0.039	0.002	0.032
Adjusted R^2	0.0001	0.040	0.001	0.039	0.002	0.032
Significance levels						*p<0.1; **p<0.05; ***p<0.01

Table 14: Sensitivity to Popularity standardization: Return as the dependent variable. Lagged.

	Predictive models:					
	Return$_t$					
	8 weeks		16 weeks		32 weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
Popularity$_{t-1}$	0.226	1.838***	0.709***	2.066***	0.929***	2.017
	(0.308)	(0.295)	(0.221)	(0.208)	(0.154)	(0.149)
ASVI$_{t-1}$		−0.444***		−0.378***		−0.365***
		(0.055)		(0.058)		(0.052)
ATV$_{t-1}$		−0.687***		−0.661***		−0.665***
		(0.108)		(0.110)		(0.112)
Volatility$_{t-1}$		−1.506***		−1.122***		−0.596***
		(0.114)		(0.107)		(0.107)
R^2	0.00003	0.011	0.0004	0.010	0.001	0.007
Adjusted R^2	0.00001	0.011	0.0004	0.010	0.001	0.007
Significance levels						*p<0.1; **p<0.05; ***p<0.01

In Table 15 and 16: Sensitivity to Popularity standardization: ASVI as the dependent variable. Based on Da et al. (2011), the results of the standardization in the last 8, 6 and 32 weeks have been reported. Univariate regressions with lagged or contemporary popularity have been reported. The tables are set up in the order in which the complete descriptive model with the simultaneous independent variables comes first. The table is then followed by second predictive model with lagged values. The robust standard errors are reported in parentheses. Weekly data from the period 2018 to 2020 are covered in the sample period. The symbols,***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively.

levels, respectively.

Table 15: Sensitivity to Popularity standardization: ASVI as the dependent variable.

	Explanatory models:					
	$ASVI_t$					
	8 weeks		16 weeks		32 weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
$Popularity_t$	-0.227^{***}	0.397^{***}	-0.201^{***}	0.396^{***}	-0.207^{***}	0.411^{***}
	(0.011)	(0.010)	(0.009)	(0.008)	(0.007)	(0.007)
$Volatility_t$		0.218^{***}		0.227^{***}		0.218^{***}
		(0.005)		(0.005)		(0.004)
ATV_t		0.113^{***}		0.142^{***}		0.167^{***}
		(0.005)		(0.005)		(0.005)
$Return_t$		0.0001^{***}		0.0002^{***}		0.00002^{***}
		(0.0002)		(0.0002)		(0.0002)
R^2	0.001	0.074	0.0003	0.106	0.010	0.140
Adjusted R^2	0.001	0.074	0.0003	0.106	0.010	0.140
Significance levels					*p<0.1; **p<0.05; ***p<0.01	

Table 16: Sensitivity to Popularity standardization: ASVI as the dependent variable. Lagged

	Predictive models:					
	$ASVI_t$					
	8 weeks		16 weeks		32 weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
$Popularity_{t-1}$	-0.080^{***}	0.191^{***}	0.028^{***}	0.255^{***}	0.133^{***}	0.333^{***}
	(0.009)	(0.011)	(0.006)	(0.008)	(0.005)	(0.007)
$Volatility_{t-1}$		0.131^{***}		0.164^{***}		0.160^{***}
		(0.005)		(0.005)		(0.005)
ATV_{t-1}		0.134^{***}		0.156^{***}		0.177^{***}
		(0.005)		(0.005)		(0.005)
$Return_{t-1}$		−0.0002		−0.0002		−0.0001
		(0.0002)		(0.0002)		(0.0002)
R^2	0.001	0.041	0.0003	0.075	0.010	0.116
Adjusted R^2	0.001	0.041	0.0003	0.075	0.010	0.116
Significance levels				*p<0.1; **p<0.05; ***p<0.01		

In Table 17 and 18; Sensitivity to Popularity standardization: Volume as the dependent variable. Based on Da et al. (2011), the results of the standardization in the last 8, 6 and 32 weeks have been reported. Univariate regressions with lagged or contemporary popularity have been reported. The tables are set up in the order in which the complete descriptive model with the simultaneous independent variables comes first. The table is then followed by second predictive model with lagged values. The robust standard errors are reported in parentheses. Weekly data from the period 2018 to 2020 are covered in the sample period. The symbols,***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 17: Sensitivity to Popularity standardization: Volume as the dependent variable.

	Explanatory modele:					
	Volume_t					
	8 weeks		16 weeks		32 weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
Popularity_t	0.275***	0.429***	0.223***	0.404***	.215***	0.381***
	(0.016)	(0.014)	(0.012)	(0.010)	(0.009)	(0.008)
Volatility_t		0.638***		0.609***		0.606***
		(0.003)		(0.003)		(0.003)
ASVI_t		0.047***		0.059***		0.073***
		(0.002)		(0.002)		(0.002)
Return_t		0.005***		0.005***		0.005***
		(0.0001)		(0.0001)		(0.0002)
R^2	0.001	0.387	0.0003	0.410	0.010	0.423
Adjusted R^2	0.001	0.387	0.0003	0.410	0.010	0.423
Significance levels				*p<0.1; **p<0.05; ***p<0.01		

Table 18: Sensitivity to Popularity standardization: Volume as the dependent variable. Lagged

	Predictive modelse:					
	Volume$_t$					
	8 weeks		16 weeks		32 weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
Popularity$_{t-1}$	-0.080^{***}	0.143^{***}	0.028	0.240	0.133^{***}	0.306^{***}
	(0.013)	(0.012)	(0.009)	(0.008)	(0.007)	(0.006)
Volatility$_{t-1}$		0.264^{***}		0.305^{***}		0.314^{***}
		(0.006)		(0.005)		(0.005)
ASVI$_{t-1}$		0.053^{***}		0.076^{***}		0.094^{***}
		(0.002)		(0.003)		(0.003)
Return$_{t-1}$		-0.001^{***}		-0.001^{***}		-0.0002
		(0.0002)		(0.0002)		(0.0002)
R^2	0.001	0.079	0.0003	0.136	0.010	0.180
Adjusted R^2	0.001	0.079	0.0003	0.136	0.010	0.180
Significance levels						*p<0.1; **p<0.05; ***p<0.01

In Table 19 and 20: Sensitivity to Popularity standardization: Volatility as the dependent variable. Based on Da et al. (2011), the results of the standardization in the last 8, 6 and 32 weeks have been reported. Univariate regressions with lagged or contemporary popularity have been reported. The tables are set up in the order in which the complete descriptive model with the simultaneous independent variables comes first. The table is then followed by second predictive model with lagged values. The robust standard errors are reported in parentheses. Weekly data from the period 2018 to 2020 are covered in the sample period. The symbols,***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 19: Sensitivity to Popularity standardization: Volatility as the dependent variable.

	Explanatory models:					
	Volatility$_t$					
	8 weeks		16 weeks		32 weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
Popularity$_t$	0.201***	−0.133***	0.206***	−0.133***	0.242***	−0.080***
	(0.010)	(0.012)	(0.007)	(0.009)	(0.005)	(0.007)
ASVI$_t$		0.073***		0.083***		0.078***
		(0.002)		(0.002)		(0.002)
ATV$_t$		0.515***		0.535***		0.500***
		(0.003)		(0.003)		(0.003)
Return$_t$		−0.007***		−0.008***		−0.007***
		(0.0001)		(0.0001)		(0.0001)
R^2	0.008	0.379	0.013	0.387	0.031	0.373
Adjusted R^2	0.008	0.379	0.013	0.387	0.031	0.373

Significance levels *p<0.1; **p<0.05; ***p<0.01

Table 20: Sensitivity to Popularity standardization: Volatility as the dependent variable. lagged

	Predictive models:					
	Volatility$_t$					
	8 weeks		16 weeks		32 weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
Popularity$_{t-1}$	-0.080^{***}	-0.278^{***}	0.028^{***}	-0.194^{***}	0.133^{***}	-0.079^{***}
	(0.009)	(0.010)	(0.007)	(0.005)	(0.005)	(0.005)
ASVI$_{t-1}$		0.058^{***}		0.076^{***}		0.076^{***}
		(0.002)		(0.002)		(0.002)
ATV$_{t-1}$		0.260^{***}		0.307^{***}		0.293^{***}
		(0.003)		(0.003)		(0.003)
Return$_{t-1}$		-0.007^{***}		-0.008^{***}		-0.008^{***}
		(0.0001)		(0.0001)		(0.0001)
R^2	0.001	0.128	0.0003	0.164	0.010	0.166
Adjusted R^2	0.001	0.128	0.0003	0.164	0.010	0.166
Significance levels						*p<0.1; **p<0.05; ***p<0.01

In Table 21 and 22 ; Sensitivity to Popularity standardization: Popularity as the dependent variable. Based on Da et al. (2011), the results of the standardization in the last 8, 6 and 32 weeks have been reported. Univariate regressions with lagged or contemporary popularity have been reported. The tables are set up in the order in which the complete descriptive model with the simultaneous independent variables comes first. The table is then followed by second predictive model with lagged values. The robust standard errors are reported in parentheses. Weekly data from the period 2018 to 2020 are covered in the sample period. The symbols,***, ** and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 21: Sensitivity to Popularity standardization: Popularity as the dependent variable.

	Explanatory models:					
	Popularity$_t$					
	8 weeks		16 weeks		32 weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
Return$_t$	0.0002	-0.00005^{*}	0.001	0.0003^{*}	0.001	0.001^{***}
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Volatility$_t$		-0.041^{***}		-0.060^{***}		-0.056^{***}
		(0.002)		(0.002)		(0.003)
ASVI$_t$		0.042^{***}		0.066^{***}		0.102^{***}
		(0.001)		(0.001)		(0.002)
ATV$_t$		0.108^{***}		0.161^{***}		0.217^{***}
		(0.002)		(0.002)		(0.002)
R^2	0.0001	0.076	0.001	0.113	0.002	0.169
Adjusted R^2	0.0001	0.076	0.001	0.113	0.002	0.169
Significance levels						*p<0.1; **p<0.05; ***p<0.01

Table 22:

	Predictive models:					
	Popularity					
	8 weeks		16 weeks		32 weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
Return_{t-1}	0.0001	-0.0004^{***}	0.001^{***}	-0.00004	0.002^{***}	0.0005^{***}
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Volatility_{t-1}		-0.005^{***}		-0.009^{***}		0.012^{***}
		(0.002)		(0.002)		(0.003)
ASVI_{t-1}		0.034^{***}		0.060^{***}		0.102^{***}
		(0.001)		(0.001)		(0.002)
ATV_{t-1}		0.095^{***}		0.138^{***}		0.185^{***}
		(0.002)		(0.002)		(0.002)
R^2	0.00001	0.084	0.001	0.120	0.002	0.179
Adjusted R^2	-0.00000	0.084	0.001	0.120	0.002	0.179

Significance levels *p<0.1; **p<0.05; ***p<0.01

6. Conclusion

The motivation behind this research was to investigate whether Popularity can explain and/or predict the stock market activity in the U.S. Our main focus was to investigate whether a correlation between Popularity and stock return existed.

Unfortunately we saw that Popularity could not explain or predict return when we conducted a univariate model. However, when put together with control variables, Popularity becomes significant in predicting but not explaining stock return. The results from our robustness checks, tells us that when the time horizon is extended, popularity becomes more significant in explaining and predicting stock returns. We can also see that Popularity has a strong explanation power when it comes to ASVI, and ASVI has a strong explanation power when we conducted the contemporary multivariate model of return. This is also the results of earlier studies (Da, Engelberg, Gao,2011,Bijl, Kringhaug, Molnar, Sandvik, 2016). Our findings help with extending the predictive and explaining power of ASVI, as well as adding the new market factor popularity. If RT starts downloading daily data again, people should definitely consider adding the popularity variable for future research.

Further, Popularity can also predict and explain volume and volatility. This can indicate that retail investors used the daily data from RT, along with other available information when putting their money in the stock market.

The result that popularity could predict more or less all of the market factors, may be one of the main reasons why the RT project got taken off air and shut down. Its explaining and predictive power could have made it easy for future investors to benefit from.

7. References

References

Ang, A., Bekaert, G., 2007. Stock return predictability: Is it there? The Review of Financial Studies 20, 651–707.

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

Banerjee, P. S., Doran, J. S., Peterson, D. R., 2007. Implied volatility and future portfolio returns. Journal of Banking & Finance 31, 3183–3199.

Barber, B. M., Huang, X., Odean, T., Schwarz, C., 2020. Attention induced trading and returns: Evidence from robinhood users. Available at SSRN 3715077 .

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.

Barberis, N., Shleifer, A., Vishny, R., 1998. A model of investor sentiment. Journal of financial economics 49, 307–343.

Bijl, L., Kringhaug, G., Molnár, P., Sandvik, E., 2016. Google searches and stock returns. International Review of Financial Analysis 45, 150–156.

Bollerslev, T., Tauchen, G., Zhou, H., 2009. Expected stock returns and variance risk premia. The Review of Financial Studies 22, 4463–4492.

Campbell, J. Y., Yogo, M., 2006. Efficient tests of stock return predictability. Journal of financial economics 81, 27–60.

Challet, D., Ayed, A. B. H., 2013. Predicting financial markets with google trends and not so random keywords. arXiv preprint arXiv:1307.4643 .

Chohan, U. W., 2021. Counter-hegemonic finance: The gamestop short squeeze. Available at SSRN .

Choi, H., Varian, H., 2012. Predicting the present with google trends. Economic record 88, 2–9.

Cochrane, J. H., 2008. The dog that did not bark: A defense of return predictability. The Review of Financial Studies 21, 1533–1575.

Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. The Journal of Finance 66, 1461–1499.

Fama, E. F., 1965. The behavior of stock-market prices. The journal of Business 38, 34–105.

Foster, F. D., Viswanathan, S., 1993. Variations in trading volume, return volatility, and trading costs: Evidence on recent price formation models. The Journal of Finance 48, 187–211.

French, K. R., Schwert, G. W., Stambaugh, R. F., 1987. Expected stock returns and volatility. Journal of financial Economics 19, 3–29.

Grinblatt, M., Keloharju, M., 2000. The investment behavior and performance of various investor types: a study of finland's unique data set. Journal of financial economics 55, 43–67.

Kim, N., Lučivjanská, K., Molnár, P., Villa, R., 2019. Google searches and stock market activity: Evidence from norway. Finance Research Letters 28, 208–220.

Lo, A. W., MacKinlay, A. C., 1988. Stock market prices do not follow random walks: Evidence from a simple specification test. The review of financial studies 1, 41–66.

Malkiel, B. G., 2003. The efficient market hypothesis and its critics. Journal of economic perspectives 17, 59–82.

Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., Preis, T., 2013. Quantifying wikipedia usage patterns before stock market moves. Scientific reports 3, 1–5.

Molnár, P., Bašta, M., 2017. Google searches and gasoline prices. In: *2017 14th International Conference on the European Energy Market (EEM)*, IEEE, pp. 1–5.

Ozik, G., Sadka, R., Shen, S., 2020. Flattening the illiquidity curve: Retail trading during the covid-19 lockdown. Available at SSRN 3663970 .

Preis, T., Moat, H. S., Stanley, H. E., 2013. Quantifying trading behavior in financial markets using google trends. Scientific reports 3, 1–6.

Preis, T., Reith, D., Stanley, H. E., 2010. Complex dynamics of our economic life on different scales: insights from search engine query data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368, 5707–5719.

Tetlock, P. C., 2007. Giving content to investor sentiment: The role of media in the stock market. The Journal of finance 62, 1139–1168.

Welch, I., 2020. The wisdom of the robinhood crowd. NBER Working Paper .

Appendix A. Explanatory Model

Table 23: Sensitivity to the time horizon standardization, with Return as the dependent variable. Univariate regression models between Popularity and return are presented in columns (1),(3),(5),(7),(9),(11) and (13). While multivariate regression models between all the control variables are presented in columns (2),(4),(6),(8),(10),(12) and (14)

			Return			
	1 week		2 weeks		3 weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
	4 weeks		5 weeks		6 weeks	
	(7)	(8)	(9)	(10)	(11)	(12)
	7 weeks					
	(13)	(14)				
Return_{t-1}	-0.059^{***}	-0.065^{***}	-0.043^{***}	-0.056^{***}	-0.043^{***}	-0.064^{***}
	-0.042^{***}	-0.067^{***}	-0.042^{***}	-0.069^{***}	-0.040^{***}	-0.069^{***}
	-0.040^{***}	-0.070^{***}				
	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
	(0.003)	(0.003)				
Popularity	-1.589^{***}	-2.924^{***}	-1.208^{***}	-2.255^{***}	-0.543^{**}	-1.372^{***}
	-0.118	-0.916^{***}	0.227	-0.485^{**}	0.453^{**}	-0.239
	0.484^{***}	-0.157				
	(0.380)	(0.386)	(0.287)	(0.290)	(0.244)	(0.246)
	(0.216)	(0.218)	(0.197)	(0.199)	(0.183)	(0.185)
	(0.171)	(0.174)				
Volatility		-4.439^{***}		-5.247^{***}		-5.719^{***}
		-5.800^{***}		-5.761^{***}		-5.781^{***}
		-5.669^{***}				
		(0.118)		(0.103)		(0.099)
		(0.097)		(0.095)		(0.093)
		(0.092)				

Continuation of Table 23

	Return					
	1 week		2 weeks		3 weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
	4 weeks		5 weeks		6 weeks	
	(7)	(8)	(9)	(10)	(11)	(12)
	7 weeks					
	(13)	(14)				
ASVI		0.169***		0.231***		0.131**
		0.133**		0.076		0.089
		0.069				
		(0.064)		(0.059)		(0.057)
		(0.056)		(0.055)		(0.055)
		(0.054)				
ATV		3.392***		3.435***		3.366***
		3.301***		3.161***		3.114***
		2.991***				
		(0.107)		(0.092)		(0.088)
		(0.086)		(0.085)		(0.084)
		(0.083)				
R^2	0.004	0.030	0.002	0.036	0.002	0.042
	0.002	0.045	0.002	0.045	0.002	0.046
	0.002	0.045				
Adjusted R^2	0.004	0.030	0.002	0.036	0.002	0.042
	0.002	0.044	0.002	0.045	0.002	0.046
	0.002	0.045				

Significance levels

*p<0.1; **p<0.05; ***p<0.01

Appendix B. Explanatory Model

Table 24: Sensitivity to the time horizon standardization, with Popularity as the dependent variable. Univariate regression models between Popularity and return are presented in columns (1),(3),(5),(7),(9) and (11). While multivariate regression models between all the control variables are presented in columns (2),(4),(6),(8),(10) and (12).

	Popularity					
	1 week		2 weeks		3 weeks	
	(1)	(2)	(3)	(4)	(5)	(6)
	4 weeks		5 weeks		6 weeks	
	(7)	(8)	(9)	(10)	(11)	(12)
Popularity_{t-1}	0.471***	0.494***	0.581***	0.593***	0.689***	0.694***
	0.736***	0.736***	0.775***	0.771***	0.803***	0.796***
	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Return	−0.00003	−0.0001**	−0.00004	−0.00001	0.00004	0.0001***
	0.0001	0.0002***	0.0001**	0.0002***	0.0001***	0.0003***
	(0.00004)	(0.00003)	(0.00003)	(0.00003)	(0.00003)	(0.00003)
	(0.00003)	(0.00003)	(0.00003)	(0.00003)	(0.00003)	(0.00003)
Volatility		0.019***		0.021***		0.024***
	0.024***		0.025***		0.025***	
		(0.001)		(0.001)		(0.001)
	(0.001)		(0.001)		(0.001)	
ASVI		0.007***		0.009***		0.008***
	0.009***		0.009***		0.009***	
		(0.001)		(0.001)		(0.001)
	(0.001)		(0.001)		(0.001)	
ATV		0.051***		0.051***		0.049***
	0.048***		0.048***		0.047***	
		(0.001)		(0.001)		(0.001)
	(0.001)		(0.001)		(0.001)	
R^2	0.313	0.393	0.448	0.513	0.594	0.645
	0.654	0.696	0.705	0.742	0.741	0.773
Adjusted R^2	0.313	0.393	0.448	0.513	0.594	0.645
	0.654	0.696	0.705	0.742	0.741	0.773

Significance levels

*p<0.1; **p<0.05; ***p<0.01