Identifying Investors with Sentiment-based Investment Strategies and Predicting their Trading Behavior

by

Sophia Y. Luo

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author .....
Department of Electrical Engineering and Computer Science
May 12, 2020

Certified by .....
Andrew W. Lo
Charles E. and Susan T. Harris Professor, Sloan School of Management
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Identifying Investors with Sentiment-based Investment Strategies and Predicting their Trading Behavior

by

Sophia Y. Luo

Submitted to the Department of Electrical Engineering and Computer Science
on May 12, 2020, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

To the best of our knowledge, there are no algorithms that distinguish different types of investors or predict how investors react to market or non-market events. In this study, we develop a computational approach to investigate investors with sentiment-based investment strategies and predict their trading behavior. We combine a dataset of more than 600,000 brokerage accounts from 2003-2015 with the RavenPack News Analytics dataset. Then, we construct a novel sentiment investor identification mechanism to classify sentiment and non-sentiment investors. Finally, we derive three machine learning models to predict whether a sentiment investor will react to a sentiment event, the reaction magnitude, and the direction of reaction (i.e. buy vs. sell). We select models that are easily interpretable and thus more directly applicable in real-world financial applications. We find that being married and the fraction of positive events in the seven days prior to an event have negative effects on the probability of reaction; whereas, occurring before the financial crisis has a positive effect. On the other hand, the sentiment event with the largest magnitude a week prior to an event and previous sentiment trading behavior have positive effects on reaction magnitude. Finally, being married and previous sentiment trading behavior have negative effects on the probability of buying versus selling, but occurring before the financial crisis has a positive effect.

Thesis Supervisor: Andrew W. Lo
Title: Charles E. and Susan T. Harris Professor, Sloan School of Management

Acknowledgments

First, I want to thank my supervisor, Andrew Lo, for supporting my thesis work from when it first began as a SuperUROP project. Over the past two years, Andrew has been an inspiring mentor. I also want to thank Chi Heem Wong, who has provided valuable advice and guidance on my project. I am very thankful for his consistent mentorship, interesting research ideas, and friendship. His expertise gave me research direction that would have been difficult to develop on my own, and his friendship made the research experience enjoyable. In addition, I want to thank Allen Cheng for his constant support and friendship. He was always there for me whenever I needed another opinion on a research idea or clarification on a statistical concept.

Next, I want to thanks my parents for their unconditional love and support. They have always been there for me, and have offered valuable life advice whenever I needed it most. I will forever appreciate their love and guidance.

Finally, I would like to thank all the friends who have been there for me through the highs and lows of my MIT years. I would not have become who I am today without them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The ability to understand and predict a person's investment decisions can profoundly impact the finance industry. Currently, a large body of literature investigates the relationship between the media and the markets. Many studies also explore the psychology of sentiment trading behavior as well as the impact of media sentiment on stock liquidity, stock returns, market movement, and other market phenomena. However, to the best of our knowledge, there are no algorithms that can distinguish different types of investors and how they may react differently to market or non-market events. Furthermore, as machine learning becomes increasingly more applicable to real-world problems, the need to produce interpretable models also grows. While there is value to using black-box approaches, there exist many real-world machine learning applications that require human-explainability prior to model adoption in practice.

In this study, we develop a computational approach to investigate investors with sentiment-based investment strategies and predict their trading behavior. We combine a dataset of more than 600,000 brokerage accounts from 2003-2015 with the RavenPack News Analytics dataset. Then, we construct a novel sentiment investor identification mechanism to classify sentiment and non-sentiment investors. Finally, we derive three machine learning models to predict whether a sentiment investor will react to a sentiment event, the reaction magnitude, and reaction direction (i.e. buy vs. sell). We select models that are easily interpretable and thus more directly

applicable in real-world financial applications.

In this chapter, we outline the structure of this thesis and summarize our research approach. Then, we provide background on the machine learning and other computational approaches that we use in our study.

1.1 Thesis Structure and Approach

In this study, we investigate the nature of sentiment-based decision-making. We combine a data set of 653,455 brokerage accounts from 2003 to 2015 with the RavenPack News Analytics dataset [1], which covers the same time period. We describe both datasets in depth in Chapter 3. We use the RavenPack-computed sentiment scores, which quantify and aggregate the "positivity" of financial figures, analyst ratings, and opinions expressed in text from media outlets, such as news articles, financial statements, press releases, and other company-relevant publications We focus our study on media sentiments that concern companies based in the United States.

First, we process both datasets such that we can map investor trades from the brokerage accounts data to events captured by Ravenpack (Section 4.1). Given the enormity of the data, one of our contributions is developing a correct and efficient approach to process and merge both datasets. Then, we derive a robust sentiment investor identification mechanism to discover all brokerage accounts that exhibit sentiment trading behavior (Section 4.2). We focus our study on the trading behavior of the identified sentiment investors and derive three predictive models:

1. Using Powell's conjugate direction method, we derive a multivariate logistic regression to predict whether a sentiment investor will react to a given sentiment event (Section 4.3).

2. In the event that there is a reaction, we use a multivariate linear regression to forecast the magnitude of the investor's reaction by predicting the proportion of wealth that the investor will trade (Section 4.4).

3. In addition, we apply Powell's conjugate direction method to derive a multi-

variate logistic regression to predict the direction of the investor's reaction, i.e. buy or sell (Section 4.5).

To derive these predictive models, we generate several samples and use Monte Carlo simulations to derive regression models for each sample. Then, we analyze the mean, standard deviation, and significance of model coefficients.

In Chapter 5, we report and discuss our thesis results. First, we analyze the empirical behavior of our sentiment investor identification mechanism (Section 5.1). Then, we use this mechanism to identify all sentiment investors in our dataset and output summary statistics about their demographic information in Section 5.2. Finally, we report our models that predict whether an investor will react to an event, the magnitude of the reaction, and the direction of the reaction in Sections 5.3-5.5. For each model, we analyze general trends in coefficient magnitudes and signs as well as the individual effects of the variables on the dependent variables. Finally, in Chapter 6, we summarize our key contributions and results, and indicate areas of future work.

We select multivariate linear and logistic regressions as our models of choice mainly due to the interpretability of the two machine learning approaches. Indeed, the design of these algorithms enables us to output useful statistics about model coefficients, accuracy, and predictive power. Additionally, we can structure our experimental design and algorithm results in a way that is reminiscent of economic and financial research papers which enables us to cater our study to a larger audience. We provide the theoretical background to our machine learning approaches in the next section.

1.2 Computational Background

In this section, we explain the machine learning theory of linear and logistic regression. We also provide the mathematical background behind Powell's conjugate direction method. Finally, we explain Monte Carlo simulation.

1.2.1 Linear Regression

Let x be a d-dimensional feature vector and let y be the corresponding output value, which can be any real number. A linear regression model expresses y as a linear function of x. That is, there exists a d-dimensional weight vector w such that

$$y = w_0 + w_1 x_1 + ... + w_d x_d,$$

where w_0 is some constant.

To derive a linear regression model, we compute w, which involves minimizing the mean squared error between the outputted value of the linear model given w and the actual value to output. That is, the cost function $C(w, w_0)$ is

$$C(w, w_0) = \sum_{i=1}^{n} (w^T x^{(i)} + w_0 - y^{(i)})^2,$$

and we want to find the optimal $w*, w_0^*$ such that

$$w^*, w_0^* = \arg\min_{w, w_0} \sum_{i=1}^{n} (w^T x^{(i)} + w_0 - y^{(i)})^2,$$

where n is the number of inputs. To do so, we take the gradient of the cost function, set it to zero, and solve for w and w_0.

We can find the closed-form solution of w and w_0 if we translate the problem into matrix notation. Let X be a matrix with dimension $n \times (d+1)$ such that every row is a different d-dimensional feature input preceded by a 1. That is, the first column of X is a column of ones. In addition, let W be a $(D+1) \times 1$ where the first element is w_0 and the remaining elements correspond to w, and let Y be a $n \times 1$ vector such that the element in row i corresponds with the feature vector in row i of X for all $0 \leq i < n$. The cost function is now defined as

$$C(W) = (XW - Y)^T (XW - Y).$$

Taking the gradient of this cost function, setting it equal to zero, and solving for W^* gives us

$$\begin{aligned}
\nabla_W C(W) &= X^T(XW - Y) + X^T(XW - Y) \\
0 &= 2X^T(XW^* - Y) \\
X^T Y &= X^T X W^* \\
W^* &= (X^T X)^{-1} X^T Y.
\end{aligned}$$

Furthermore, given the linear nature of the model, we can interpret our results easily. For instance, suppose a derived coefficient for a feature f_j is equal to some value w_j. Then, holding all other regressors constant, we expect a w_j unit increase in y given a unit increase in f_j.

A common metric used to measure the overall fit of the linear regression to the data is R-squared, which is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(y^{(i)} - \overline{y})^2},$$

where $y^{(i)}$ is the actual value, $\hat{y}^{(i)}$ is the outputted value from the model, and $\overline{y}$ is the mean $y^{(i)}$ value across all samples i. That is, R^2 is the proportion of the variance in the data explained by the model.

1.2.2 Logistic Regression

Unlike linear regression where the output can be any real number, the output of logistic regression is restricted to being between zero and one inclusive. Thus, logistic regression is commonly used to solve classification problems where the output of the data is binary in nature.

In the logistic regression model, the output is modeled with a logistical function, which is defined as follows:

$$y = \sigma(w^T x + w_0) = \frac{\exp(w^T x + w_0)}{1 + \exp(w^T x + w_0)}.$$

To compute w and w_0, we maximize a likelihood function l:

$$l(w, w_0) = \prod_{i:y^{(i)}=1} \sigma(w^T x^{(i)} + w_0) \prod_{j:y^{(j)}=0} \left(1 - \sigma(w^T x^{(j)} + w_0)\right).$$

Equivalently, we can take the negative log of l and minimize the negative log-likelihood function (L) which is

$$\begin{aligned} L(w, w_0) &= -\sum_{i=1}^{n} \log \left[\sigma(w^T x^{(i)} + w_0)^{y^{(i)}} (1 - \sigma(w^T x^{(i)} + w_0))^{1-y^{(i)}}\right] \\ &= -\sum_{i=1}^{n} \left[y^{(i)} \log \sigma(w^T x^{(i)} + w_0) + (1 - y^{(i)}) \log(1 - \sigma(w^T x^{(i)} + w_0)).\right] \end{aligned}$$

Furthermore, a typical way to interpret the coefficients of a logistic regression model is to reference "log-odds." Suppose a derived coefficient for feature f_j is equal to some value w_j. Then, holding all other features constant, we expect a w_j unit increase in the log-odds of y equalling one to y equalling zero given a unit increase in f_j. Equivalently, given a unit increase in f_j, we expect a less than w_j unit increase in the odds of y equalling one to y equalling zero.

A common metric used to measure the overall fit of the logistic regression to the data is the pseudo R-squared, which is defined as

$$R^2 = 1 - \frac{l(w, w_0)}{l(w_0)},$$

where $l(w, w_0)$ is the log-likelihood of the full model and $l(w_0)$ is the log-likelihood of the model that only includes an intercept. A larger pseudo R-squared value indicates a model with greater log-likelihood. Given the logistic nature of the pseudo R-squared, the actual proportion of variance in the data explained by the model is slightly less than the computed pseudo R^2 value.

1.2.3 Powell's Conjugate Direction Method

Powell's conjugate direction method is an algorithm that finds a local minimum of a function ($f\colon \mathbb{R}^d \to \mathbb{R}$) without taking any derivatives. The procedure is as follows:

1. Select an initial position x_0 (which is d-dimensional) and d directions vectors $p_0, p_1, ...p_{d-1}$.

2. For all $k \in \{0, 1, ..., d-1\}$, let $\alpha_k = \arg\max f(x_k + \alpha_k p_d)$ and let $x_{k+1} = x_k + \alpha_k p_k$.

3. For all $k \in \{0, 1, ..., d-2\}$, set $p_k = p_{k+1}$.

4. Set $p_{d-1} = x_d - x_0$.

5. If $||x_d - x_0|| < \epsilon$, where ϵ is some small convergence criterion, then we are done.

6. Otherwise, set $\alpha_d = \arg\max f(x_d + \alpha_d p_d)$ and set $x_0 = x_d + \delta_d p_d$. Go back to step 1.

In other words, we start from an initial position on the function and with several different direction vectors. The next position is then the linear combination of the directions (i.e. $x_1 = x_0 + \sum_{k=1}^{d} \alpha_k p_k$). The vector $\sum_{k=1}^{d} \alpha_k p_k$ is then added to the set of direction vectors, and the direction vector that contributed the most to x_1, which is the next position, is removed from the set of direction vectors. Then, we set x_0 equal to x_1 and repeat this process until no significant update is made to the current position x_0.

1.2.4 Monte Carlo Simulation

Monte Carlo methods constitute a class of algorithms that involve repeated random sampling to solve problems that might be deterministic in nature. The objective is to generate a large number of random samples from a population and approximate the expected value of the output with the mean of the independent samples.

For example, suppose we want to derive a linear regression model from some data. First, we create N samples of the data, derive run linear regression models from each of these samples. We can think of each sample as a "simulation." Let X_k be the results of the kth simulation, and let μ and σ^2 be the true mean and variance respectively of X_k. Furthermore, let the mean of the Monte Carlo simulations over the N iterations be

$$\hat{\mu}_N = \frac{1}{N} \sum_{k=1}^{N} X_k.$$

By the central limit theorem, the distribution of $\hat{\mu}_N$ converges to a normal distribution with mean μ and variance $n\sigma^2$. Then, we the confidence interval for μ is given by

$$\hat{\mu}_N \pm \frac{cs_N}{\sqrt{N}},$$

where s_N is the sample variance of $\{X_1, ..., X_N\}$ and c is 1.65 for a 90% confidence interval, 1.96 for a 95% confidence interval, and 3.29 for a 99% confidence interval.

Chapter 2

Literature Review

It is widely understood that there is a strong correlation between investor sentiment and the market. Barberis et al. [3] introduce two categories of pervasive market behavior: the underreaction of stock prices to news, such as earnings announcements, and the overreaction of stock prices to a series of good and/or bad news. In their paper, they derive a model of investor sentiment and for how investors form beliefs [3]. In addition, Wurgler and Baker [26] quantify the impact of investor sentiment, and discover that sentiment regularly affects individual firms and the market. Li et al. [13] arrive at a similar result but focus on the relationship between Twitter sentiment and a specific market index. They develop an algorithm to extract words that reflect public sentiment from tweets, predict stock movement on the NASDAQ, and achieve an average accuracy of over 70% [13]. Applying a similar methodology on a different market index, Daniel et al. [6] detect events through sentiment analysis on Twitter data, pinpoint important events that impact companies, and isolate the tweets that influence the Dow Jones Industrial Average (DJIA) index [6]. Ranco et al. [20] narrow their focus on the companies that make up the DJIA, and find that there exist dependencies between Twitter sentiment and abnormal returns during peaks of Twitter volume and several days after such events.

A number of studies also use machine learning algorithms to look into the effect of investor sentiment on stock returns. Huang et al. [10] adopt a linear regression approach and propose a new investor sentiment index to predict the aggregate

stock market. In their study, they prove that this index outperforms many well-known macroeconomic variables and can predict stock returns [10]. They believe that this predictability comes from investors' biased beliefs about future cash flows [10]. Schmeling [22] also uses a linear regression in his investigation of the effect of sentiment on returns. Specifically, he uses long-horizon return regressions, and finds that sentiment has a negative effect on aggregate stock market returns across 18 industrialized countries [22]. He also finds that sentiment negatively affects the returns of value stocks, growth stocks, small stocks, and for different forecasting horizons [22]. In addition, Porshnev et al. [18] adopt a lexicon-based approach to analyze the psychological states of Twitter users, and use support vector machines and neural networks to predict stock market returns on the DJIA and S&P500 [18]. Ren et al. [21] also use support vector machines in their methodology. The authors predict the direction of stock market movement using both financial market data as well as investor psychology-based sentiment features [21]. In particular, they incorporate the day-of-week effect, which is a financial anomaly where average return on Mondays is lower than average return on other days of the week [21]. Furthermore, Jiahong Li et al. [11] propose a long short-term memory neural network model that incorporates investor sentiment and market factors to forecast portfolio performance. Their paper demonstrates the potential of deep learning methodologies in modelling financial time series that intrinsically have a lot noise [11].

On the other hand, some studies find that there is no relationship between sentiment and stock returns. For instance, Neal and Wheatley [15] investigate common measures of individual investor sentiment, including the level of discounts on closed-end funds, net mutual fund redemptions, and the ratio of odd-lot sales to purchases, using first-order autoregressive models. They find that closed-end fund discounts and net mutual fund redemptions are predictive of the difference between small and large firm returns, but do not find evidence that the odd-lot ratio predicts returns [15]. In addition, Kim and Kim [12] investigate whether investor sentiment can forecast stock returns, volatility, and trading volume using a dataset of Yahoo! Finance message board postings. They find no evidence of investor sentiment having an effect on future

stock returns in their intertemporal and cross-sectional regression analyses [12]. They also find no evidence of Internet postings having an effect on volatility and trading volume [12]. Instead, their results suggest that previous stock price movement affects investor sentiment [12]. On the other hand, Chung et al. [5] investigate how the predictive power of investor sentiment on stock returns can vary depending on whether the economy is in an expansion or recession state. Their results suggest that investor sentiment is only a good predictor of returns during periods of economic expansion, and loses significance during periods of economic recession [5].

In addition, several studies focus on the effect of sentiment on stock liquidity, stock volume, stock volatility, and other stock-related aspects of the market. Raissi [19] finds that investor sentiment significantly impacts stock performance and demonstrates that sentiment is the result of factors such as liquidity indicators in the market. Wurgler and Baker [26] arrive at the same conclusion and find that stocks that are difficult to arbitrage or value tend to be the most impacted by sentiment. Agrawal et al. [2] also study the impact of sentiment on liquidity, but focus on the effects of extreme sentiment levels. They demonstrate that extreme sentiment is correlated with higher demand and lower supply of liquidity, causes prices to become more mean-reverting, and leads to narrower spreads [2]. In addition, Tetlock [24] uses content from the *Wall Street Journal* and a vector autoregressive framework to investigate the relationship between media and the market, and finds that media pessimism predicts downward pressure on market prices. The author also finds that low or high pessimism is a predictor of high market trading volume [24]. Furthermore, Chatterjee and Perrizo [4] use Microsoft's Azure Sentiment Analyzer service to study the effect of investor behavior on the market through data-mined tweets, and they find that sentiment affects the volatility of stocks in the market. Wu et al. [25] also make the same conclusion about the relationship between sentiment and stock price volatility. In their paper, they develop a sentiment ontology to conduct sentiment analysis of online posts in stock markets [25]. They use support vector machine and generalized autoregressive conditional heteroscedasticity modeling, and also discover strong correlations between forum sentiment and stock price volatility trends [25].

Trading strategies have also been developed to take advantage of the relationship between sentiment and market movement. Peterson [17] has written a book on how he and his team created a market-neutral social media-based hedge fund that significantly outperformed the S&P 500 during the 2008 financial crisis. In the book, he discusses crowd psychology and patterns between investor sentiment and market movement [17]. In addition, Yang et al. [27] use the Gaussian inverse reinforcement learning method to design an investor sentiment reward-based trading system that only extracts signals that generate either negative or positive market responses. This reward extraction mechanism is based on market returns and market volatility, and back-test results suggest that the sentiment reward-based trading performs better than benchmark strategies on the S&P 500 index and market-based ETFs as well as some other existing news sentiment-based trading signals [27]. Furthermore, Huang et al. [8] use genetic algorithms to derive and optimize a stock selection model from investor sentiment data. They first create a stock scoring model using investment sentiment indicators from behavioral finance literature to compute relative stock rankings [8]. Then, they select the final stocks to form the portfolio from these rankings [8]. A year later, the same authors conduct a comparative study between traditional regression models and evolution-based models [9]. They find that their genetic algorithmic approach significantly outperforms baseline and traditional regression models [9].

Furthermore, there are studies that investigate the effects of sentiment on the markets with some emphasis on the impact of sentiment on investors as well. Pagolu et al. [16] analyze the correlation between Twitter sentiment and stock market movement. They find that positive sentiment about a company encourages people to invest in that company, which increases the company's stock price [16]. Applying a completely different analytical framework, Mohacsy and Lefer [14] utilize a psychodynamic approach to investigate investor sentiment and the markets, concluding that investors tend to react collectively and suggesting that the markets can be quantified as an aggregation of sentiment.

However, we are curious about the extent to which investors react collectively and the nature of individual decision-making in the markets. While the literature

on the correlation between investor sentiment, stock returns, and market movement is plentiful, we find that these studies almost always consolidate and conceptualize investors as a "crowd" or single entity that moves the markets. Furthermore, almost all the current research focuses on what types of stocks and other market properties are likely to be affected by sentiment. As a result, there is a lack of understanding on what types of investors are more likely to be influenced by sentiment, and we are unable to find studies that can predict which individuals are more likely to react to changes in sentiment. However, it is important to understand the individual's decision making process to develop the microeconomics foundations to build more robust macroeconomics models about the markets. This is the main motivation of our study.

Chapter 3

Data

In this chapter, we describe the two datasets that we use in our study: (1) a proprietary brokerage accounts dataset, which is comprised of trade data, monthly positions data, and demographic data, and (2) the RavenPack News Analytics dataset [1]. We introduce the former dataset in Section 3.1 and briefly discuss the latter in Section 3.2. We provide summary statistics and visualizations to contextualize the nature of our data.

3.1 Brokerage Accounts

Our brokerage accounts data set consists of 653,455 anonymous retail accounts, drawn at random, from one of the largest brokerage firms in the United States. Note that we cannot disclose the name of the brokerage firm due to confidentiality agreements.

There are three levels of identification in our data: accounts, customers, and households. Multiple customers can co-own an account, and the brokerage firm has mapped groups of accounts into households based on relationships between customers. For clarity, we visualize the relationship between accounts, customers, and households in Figure 3-1. In this study, we analyze sentiment trading behavior at the individual customer level.

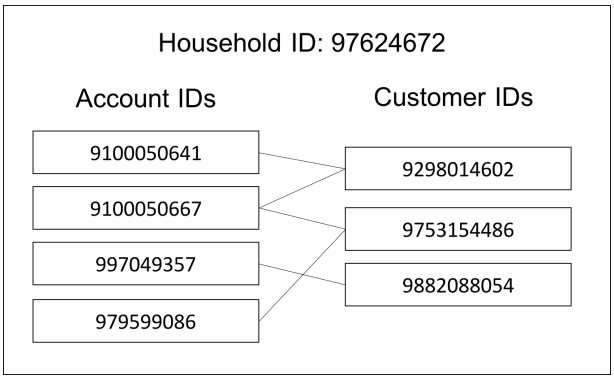

Figure 3-1: Visual representation of the relationship between household, account, and customer IDs.

Each account was active as of December 31st, 2015. Furthermore, the accounts were opened at different times in the past, where some account start dates occur before the earliest recorded activity in our data set. Because our data set includes transactions dating back to January 2003, we artificially set the start dates of these accounts to January 1, 2003.

For each account, the data consist of every trade and the type of asset traded as well as monthly position snapshots. In this paper, we refer to the traded assets by their Committee on Uniform Security Identification Procedures (CUSIP) numbers. We further describe the trades data in Section 3.1.1 and the positions data in Section 3.1.2. The data also consists of self-reported demographics information, which we discuss in Section 3.1.3.

3.1.1 Trades Data

For each account, the trades data is composed of all trades made during the lifetime of the account. A trade is uniquely identified by its timestamp, associated account ID, and CUSIP/ticker. Each trade is also stored with the number of asset units traded, trade commission, and principal amount that is positive or negative if it is a buy or sell, respectively. We provide a complete list of all the data fields and their descriptions in Table A.1.

The daily nature of the trades data enables us to analyze intra-month trading

behavior that would otherwise be impossible to compute given monthly or quarterly data. In addition, we are able to directly study reactions to daily media events with our trades data, which we would otherwise not be able to if we were given less granular data. Thus, our usage of the trades data distinguishes our study from other financial studies that typically rely on monthly or quarterly data.

3.1.2 Positions Data

For each account, the positions data is comprised of monthly snapshots that record the month-end quantities and prices of each owned security during the lifetime of the account. Each row in the data is uniquely identified by the account ID, the year and month of the snapshot, the associated CUSIP/ticker, and the quantity owned and the price of the CUSIP/ticker at the end of the month. In addition, each row includes the internal asset class assignment associated with each CUSIP/ticker. The possible classes are equities, mutual funds, fixed income securities, cash or cash equivalents, and options. A separate identifier is also included in the data to distinguish cash equities from ETFs within the equities category. We provide the full list of positions data fields and their descriptions in Table A.2.

3.1.3 Demographics Data

The demographics data consists of self-reported information by randomly selected customers. Each row includes the month and year the demographic information was collected from a given customer and their age, income, occupation, marital status, investment knowledge, and investment experience at the time. That is, we have a one-time snapshot of demographic information for each customer. The possible categories that customers could have selected for their investment knowledge and experience are "excellent," "good," "limited," "none," and "decline to report." We report the full list of data fields and their descriptions in Table A.3.

3.2 RavenPack Data

In order to identify sentiment changes, we rely on data from RavenPack News Analytics [1], which aggregates news items and quantifies their sentiment, relevance, topic, novelty, and market impact. Specifically, we use the Equities News Analytics package from the Dow Jones Edition, which analyzes articles from the Dow Jones Newswires, the Wall Street Journal, Barron's, and MarketWatch. At the time of writing, the data spans from January 1, 2000, to November 30, 2019. For this study, we use a subset of the RavenPack data that concerns traded CUSIPs/tickers over the lifetime of the investors' accounts. We present a subset of the fields from the RavenPack dataset and their descriptions in Table A.4.

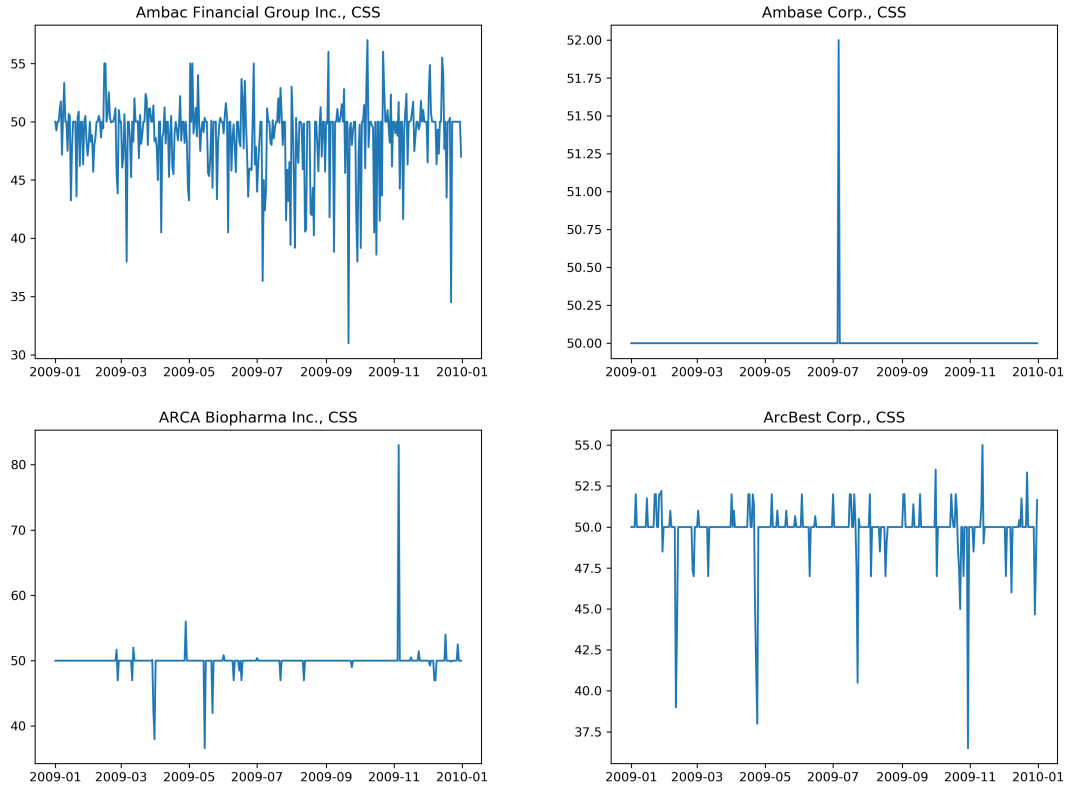

Figure 3-2: Plots of CSS scores over time for Ambac Financial Group Inc., Ambase Corp., ARCA Biopharma Inc., and ArcBest Corp.

In this study, we primarily use the RavenPack Composite Sentiment Score (CSS), which is an aggregate analytic that combines the scores of other RavenPack sentiment classifiers. We plot the CSS scores of four companies over the course of 2009 to

demonstrate that we can analyze the frequency of news events per company over time (Figure 3-2). We specifically select companies that demonstrate a range in news event frequency to illustrate the diversity of news coverage and media sentiment across all companies in the dataset.

The classifiers used to compute the CSS score of a media event are PEQ, BEE, BMQ, BAM, and BCA. PEQ and BEE are trained on articles about global equities and earnings evaluations respectively, and both compute scores based on RavenPack's Traditional Methodology. BMQ is trained on short commentary and editorials on global equity markets; BAM focuses on stories about mergers, acquisitions, and takeover events; and BCA specializes in articles about corporate action announcements. In addition, BMQ, BAM, and BCA compute scores based on RavenPack's Expert Consensus Methodology.

In addition, we present summary statistics of the RavenPack data. We identify companies with valid North American Industry Classification System (NAICS) codes that map to industry names. In Figure 3-3, we plot the number of companies and the number of companies with identifiable industry info per year in the RavenPack dataset. We find that the universe of companies covered by RavenPack is relatively constant over time and that we are unable to identify the industries of a non-negligible number of companies.

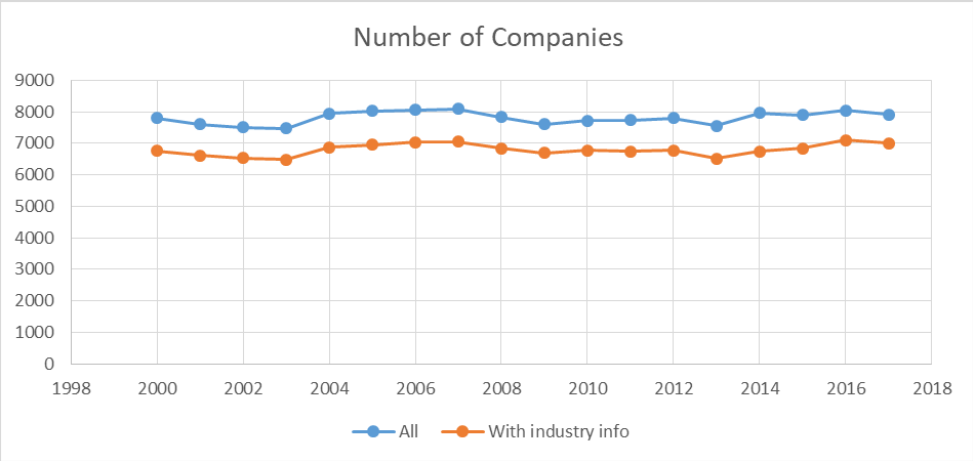

Figure 3-3: The blue line represents the number of companies covered by the RavenPack dataset every year. The orange line represents the number of those companies in the data that have information about their industry.

Finally, we plot the number of companies per year by industry in Figure 3-4. From the plot, we see that a large majority of the companies are from the finance and insurance, manufacturing, and information sectors in decreasing order of coverage.

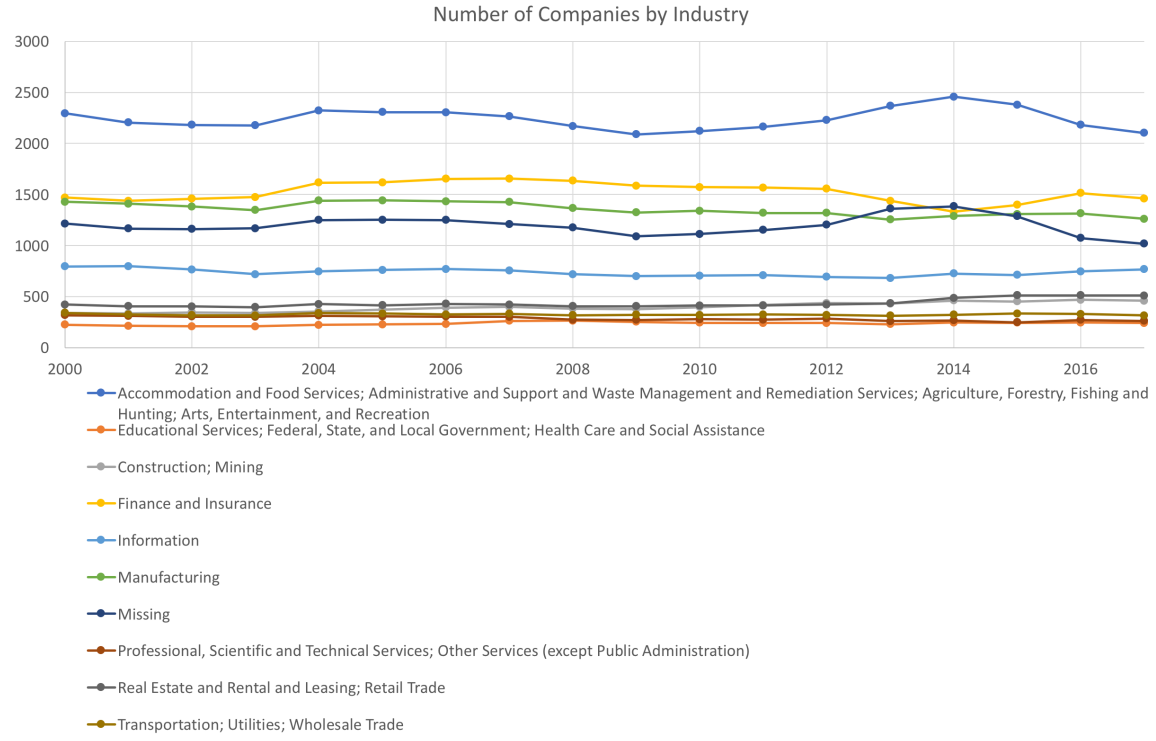

Figure 3-4: The number of companies broken down by industry covered by the RavenPack dataset every year. Note that we aggregated multiple industries together for visual clarity. If a line is an aggregation of multiple industries, the industry names are delimited by semicolons.

Chapter 4

Methods

In this section, we first describe our sentiment event processing step to define key variables and provide contextual information about our data (Section 4.1). Then, we derive a systematic sentiment investor identification method (Section 4.2). Finally, we study the trading behavior of those who we have identified to be sentiment investors and derive three predictive models using machine learning algorithms. Given demographic data, trading history, and nature of the sentiment event(s), we predict whether a sentiment investor will make a trade in reaction to an event (Section 4.3), the magnitude of this reaction (Section 4.4), and the direction of the reaction (Section 4.5).

4.1 Sentiment Event Processing

Generally speaking, we process our data at the CUSIP level. Over time, a company can have different stock ticker symbols, and each ticker symbol can be mapped to one or more CUSIP numbers, which are unique identification numbers that differentiate between securities traded on public markets. Note that we only perform computations on companies that we could find valid CUSIP-to-ticker mappings for.

First, we compute the daily sentiment change across all CUSIPs. RavenPack provides sentiment scores ($S_{j,t}$) for a company that has a particular CUSIP (j) for any given day (t) only if there is at least one new article mentioning the company on

that day [1]. We define a sentiment change ($\Delta_{j,t}$) for a single CUSIP to be the change in sentiment compared to the previous calendar day. That is,

$$\Delta_{j,t} = S_{j,t} - S_{j,t-1}. \tag{4.1}$$

If there is no such article, we assign $\Delta_{j,t} = 0$ by default. Since, the sentiment scores are bounded between -100 and 100, the sentiment changes can range from -200 to 200. In order to reduce noise in the data, we only consider days that have sentiment changes with absolute value of at least 10 and refer to these days as days with "sentiment events." Note that for every day with a RavenPack-identified media event, our construction creates two sentiment events. That is, given a media event on t, we have nonzero $\Delta_{j,t}$ and $\Delta_{j,t+1}$. To avoid double-counting, we keep the former ($\Delta_{j,t}$) and discard the latter ($\Delta_{j,t+1}$).

For our study, we also compute investors' positions during every associated sentiment event. Our brokerage accounts portfolio data is stored at the monthly level, but we need portfolio holdings information at a finer granularity for each investor across the lifetime of the investment account for our analysis. Thus, we calculate the number of shares of a given CUSIP that an investor owns at the time of every event related to that CUSIP. To do this, we use the following algorithm: Given an event that does not occur on the last day of a month, we take each investor's portfolio as of the last day of the previous month and simulate the investor's trading activity by taking all trades the investor makes up to the date of the event. If an event occurs on the last day of a month, we simply take the investor's portfolio snapshot for that month. Note that we only perform our computations for a given account-CUSIP pair if

1. we find a valid CUSIP-to-ticker mapping,

2. the RavenPack dataset provides coverage over the associated entity, and

3. there are sentiment events during the time that the account trades or has a non-zero position of the CUSIP.

Thus, out of the 653,455 brokerage accounts, we are able to use this strategy to compute the quantity of CUSIP shares owned at the time of sentiment events for 376,352 accounts. We conduct the rest of our study on these 376,352 accounts.

4.2 Sentiment Investor Identification Mechanism

We refer to the existing literature when deriving our definition of a sentiment investor. Shleifer et al. [23] derive a model for investor sentiment based on underreactions and overreactions during the period following news announcements. We find this framework to be compelling and applicable. Intuitively, a sentiment investor is someone who makes their trades in response to sentiment events. Equivalently, a non-sentiment investor is someone whose probability of making a trade given an event is approximately equal to the probability of making a trade given no event.

To determine whether someone is a sentiment investor, we first compute the number of days that investor i owns a given CUSIP j. We define the start of the ownership period as the first recorded evidence of the investor owning j (i.e. the first monthly portfolio snapshot that contains j or the first trade made with j). Similarly, the end of the ownership period is the last recorded evidence of the investor owning j. Let N_{ij} be the total number of days during which investor i owns CUSIP j. Note that, using this definition, it is possible that an investor owns zero of a CUSIP during the ownership period of that CUSIP. That is, an investor can exit and re-enter a position on a CUSIP for varying amounts of time at varying frequencies.

During this ownership period, different events concerning j may occur. We construct a rolling window of seven days after each event, and define the total number of these days N_{ij}^{E} to be the reaction period following these events. For instance, if an event occurred on January 1, January 9, and January 11, N_{ij}^{E} would be equal to sixteen. Conversely, we define $N_{ij}^{\overline{E}}$ to be the number of days during the ownership period that are not event reaction days. Furthermore, we define a sentiment trade to be a trade that occurs during an event reaction day.

To determine whether investor i trades CUSIP j on sentiment, we construct the

following difference:

$$p_{ij} = \frac{T_{ij}^{E}}{N_{ij}^{E}} - \frac{T_{ij}^{\overline{E}}}{N_{ij}^{\overline{E}}}, \tag{4.2}$$

where T_{ij}^{E} is the number of trades made by investor i of CUSIP j during all event reaction days and where $T_{ij}^{\overline{E}}$ is the number of trades made during the remaining days of the ownership period. In other words, if investor i trades CUSIP j on sentiment, then p_{ij} should be greater than zero. Note that if $N_{ij}^{E} = 0$, we set $\frac{T_{ij}^{E}}{N_{ij}^{E}} = 0$. Similarly, if $N_{ij}^{\overline{E}} = 0$, we set $\frac{T_{ij}^{\overline{E}}}{N_{ij}^{\overline{E}}} = 0$. Furthermore, $N_{ij} = N_{ij}^{E} + N_{ij}^{\overline{E}}$ and $T_{ij} = T_{ij}^{E} + T_{ij}^{\overline{E}}$. Using our construction, we can also find the total number of sentiment trades investor i makes across all CUSIPs ($T_{i}^{E} = \sum_{j} T_{ij}^{E}$), the total number of non-sentiment trades made ($T_{i}^{\overline{E}} = \sum_{j} T_{ij}^{\overline{E}}$), and the total number of trades made ($T_{i} = \sum_{j} T_{ij}$).

However, it is possible that an investor only trades on sentiment for a few CUSIPs out of their entire portfolio of several hundred CUSIPs. Intuitively, we do not consider such an investor a sentiment investor; we identify sentiment investors as those who trade a significant volume with respect to their other trades in reaction to sentiment events. Measuring the relative volume traded allows us to quantify and capture the magnitude of the reaction relative to investors' general trading behavior. Thus, we construct the following linear combination of p_{ij} values:

$$p_{i} = \sum_{j=1}^{J} w_{ij} p_{ij}, \tag{4.3}$$

where J is the total number of CUSIPs that investor i owns and w_{ij} are weights determined by trade volume. We compute w_{ij} as follows

$$w_{ij} = \frac{\sum_{d=1}^{D} r_{ijd}}{\sum_{j'=1}^{J} \sum_{d=1}^{D} r_{ij'd}}, \tag{4.4}$$

where D is the number of trades made for a given CUSIP and r_{ijd} is the total amount of CUSIP j traded by investor i during trade d. In other words, w_{ij} is the total volume of j ever traded by investor i divided by the total volume traded across all

CUSIPs over the lifetime of the investor's account. Here, volume is defined as the number of shares sold multiplied by the price that the shares were traded at.

Finally, we define sentiment investors as investors who meet the following criteria:

1. $T_i^E \geq 1$ (at least one sentiment trade),

2. $p_i > 0$, and

3. only one person is associated with the account (i.e. no household accounts).

After identifying all sentiment investors in our dataset, we study their demographic composition. In order to determine which groups of investors (G) are more or less likely to react to sentiment events (Δ), we compute the relative prevalence of different groups given an event:

$$\frac{P(G|\Delta)}{P(G)}, \tag{4.5}$$

where

$$P(G|\Delta) = \frac{\text{Number of sentiment investors in } G}{\text{Total number of sentiment investors}}, \tag{4.6}$$

and

$$P(G) = \frac{\text{Number of investors in G}}{\text{Total number of investors}}. \tag{4.7}$$

If the relative prevalence (Equation 4.5) is greater than one, then the group G is more likely to react to sentiment events and thus be sentiment investors compared to other groups. On the other hand, if the relative prevalence is less than one, then G is less likely to be sentiment investors compared to other groups.

We then test our null hypothesis of $P(G|\Delta) = P(G)$ using the two-proportion Z-test. Let $g_\Delta = P(G|\Delta)$ and let $g = P(G)$. Furthermore, let n_Δ be the total number of sentiment investors in G and let n be the total number of investors in G. The test-statistic is then

$$\frac{g_\Delta - g}{x(1-x)\left(\frac{1}{n_\Delta} + \frac{1}{n}\right)}, \tag{4.8}$$

where

$$x = \frac{n_\Delta g_\Delta + ng}{n_\Delta + n}. \tag{4.9}$$

4.3 Predicting Reaction vs. Non-Reaction

For our first model, we predict whether or not a sentiment investor will react to an event. We define a lack of a reaction as a non-reaction. In this section, we discuss how we compute reactions and non-reactions. Then, we describe all features included in our predictive model and explain how we compute each endogenous feature that was not already provided in the dataset. Finally, we discuss our experimental design, including our sample creation and model derivation processes. Note that we disregard all accounts that we identify to exhibit sentiment investing if they are shared by multiple investors. (Refer to Figure 3-1 for more context.) That is, we focus our study on individual sentiment investors, and use "individual sentiment investors" and "sentiment investors" interchangeably for the remainder of this chapter.

4.3.1 Target Variable

For this model, we predict a binary variable for reaction or non-reaction. We define a reaction to a sentiment event as a trade that occurs within seven days after at least one sentiment event. That is, a reaction is a sentiment trade. Given individual sentiment investor i and CUSIP j, let t_{ij}^- and t_{ij}^+ be the days of the first and last sentiment trades respectively. We define i's "window of observation" of j to be 30 days preceding t_{ij}^- to 30 days after t_{ij}^+. If there is a sentiment event that occurs on $t^* \in [t_{ij}^- - 30, t_{ij}^+ + 30]$ but no trade that occurs within seven days after, then we label this data-point as a non-reaction.

4.3.2 Features

We group our features into three main categories: demographic variables, event-specific features, and time-based features. The demographic variables come from the demographics data provided to us by our data vendor (Table A.3). Specifically, we include the following demographic variables in our model:

- Age: For each sentiment investor, we take their reported age and the date of

the record. Then, using the year of this date of record, we compute their age during all reactions and non-reactions.

- Number of dependents: We combine the categories given to us by our vendor into three main groups: has zero dependents, has one to three dependents, and has more than 3 dependents.

- Investment experience: The categories provided by the vendor are "excellent," "good," "limited", "none", and "decline to report." Furthermore, the investor may have never filled out this field. We consider "decline to report" to be missing data.

- Investment knowledge: The investment knowledge data is in the same format as the investment experience data. The provided categories are "excellent," "good," "limited", "none", and "decline to report," and we consider "decline to report" to be equivalent to missing data.

- Marital status: Some provided categories include "single," "married," "divorced," etc. We combine the labels into two main groups: married and not married.

In addition, we compute three event-based features. Let t be the date of a reaction or non-reaction of investor i for CUSIP j. Taking all events that occur within seven days before the reaction/non-reaction, we create two features: (1) $\widetilde{\Delta}$, which is the absolute value of the sentiment change with the largest magnitude and (2) ρ, which is the proportion of positive events. That is,

$$\widetilde{\Delta}_{ijt} = \max_{u \in \{t-7, t-6, \ldots, t\}} |\Delta_{iju}| \tag{4.10}$$

and

$$\rho_{ijt} = \frac{\sum_{u=t-7}^{u=t} \mathbb{1}\{\Delta_{iju} > 0\}}{\sum_{u=t-7}^{u=t} \mathbb{1}\{\Delta_{iju} \neq 0\}}. \tag{4.11}$$

The last event-based feature that we compute for each data-point is the number of

events in the 30 day period leading up to t. Let this feature be η, which is defined as

$$\eta_{ijt} = \sum_{u=t-30}^{t} \mathbb{1}\{\Delta_{iju} \neq 0\}. \tag{4.12}$$

Our last category of variables are time-based features. We compute indicator variables for whether a trade occurs before the financial crisis (before 2007), during the crisis (between 2007 and 2009 inclusive), or after the crisis (after 2009). Finally, we also compute ζ_{ijt}, which is the proportion of volume traded of CUSIP j by investor i prior to t. In other words,

$$\zeta_{ijt} = \frac{\sum_{d=0}^{D_j} r_{ijd}}{\sum_{j'=1}^{J} \sum_{d=0}^{D_{j'}} r_{ij'd}}, \tag{4.13}$$

where r_{ijd} is the total amount of j traded by i during trade d and D_j is the last trade of CUSIP j made before t.

4.3.3 Experimental Design

The nature of the prediction problem lends itself well to a logistic regression model (Section 1.2.2). We label reactions as ones and non-reactions as zeros, and model whether an investor will react to a given event with a multivariate logistic regression that incorporates all previously discussed features.

Each investor can trade multiple CUSIPs and have multiple reactions/non-reactions over time for events pertaining to each traded CUSIP. Given that each reaction/non-reaction for each CUSIP is a separate data-point, some investors will inevitably be represented more in the data than others. Thus, we cannot simply run a logistic regression on our entire sample of sentiment investors and all their trades. Instead, we use Monte Carlo simulation (Section 1.2.4). We pseudo-randomly create 1,000 different samples with the following procedure:

Consider the set I of all sentiment investors who we have non-missing data for all previously descried features. (Note that $|I|$ is significantly smaller than the actual number of sentiment investors because not all investors report their demographic

information.) Let $I_{\overline{R}} \subseteq I$ be the set of sentiment investors that have at least one non-reaction. Let $I_R \subseteq I$ be the set of sentiment investors that have at least one reaction. Furthermore, let $I_R^{(B)} \subseteq I_R$ be the set of sentiment investors that have at least one buy reaction and let $I_R^{(S)} \subseteq I_R$ be the set of sentiment investors that have at least one sell reaction.

Our goal is to sample reactions and non-reactions such that each investor is represented in each sample at most once. Furthermore, we want the sample to be "balanced" such that half the data-points are reactions while the other half are non-reactions. We also want an equal number of buy reactions as sell reactions in our samples. To achieve these constraints, we perform the following procedure: First, we select $\min(\lfloor |I|/2 \rfloor, |I_R|, |I_{\overline{R}}|)$ investors uniformly at random from $I_{\overline{R}}$. Let the set of selected investors be $I'_{\overline{R}}$. Then, we select $\min(\lfloor |I|/4 \rfloor, \lfloor |I_R|/2 \rfloor, \lfloor |I_{\overline{R}}|/2 \rfloor)$ investors from $I_R^{(B)} \setminus I'_{\overline{R}}$ uniformly at random. Let the set of selected investors be $I_R^{(B)'}$. We then uniformly at random select $|I_R^{(B)'}|$ investors from $I_R^{(S)} \setminus (I'_{\overline{R}} \cup I_R^{(B)'})$. Let $I_R^{(S)'}$ be the selected investors.

Next, we sample one non-reaction from each investor in $I'_{\overline{R}}$, one buy reaction from each investor in $I_R^{(B)'}$, and one sell reaction from each investor in $I_R^{(S)'}$. Our final sample is then comprised of half non-reaction data-points and half reaction data-points. Furthermore, half of the reaction data-points are buy reactions while the other half are sell reactions. Note that all investors are represented in the sample once. However, because our final sample is only a small fraction of the entire dataset, we create 1,000 of these samples for broader coverage.

Finally, for each sample, we normalize all non-binary variables by subtracting all values of a given variable by the variable mean and dividing the difference by the standard deviation of the variable. The non-binary variables are the features for age during trade, $\widetilde{\Delta}$, and η, and derive multivariate logistic regression models from each of the samples using Powell's conjugate direction method (Section 1.2.3). We also add an intercept to each model. Let V_D, V_E and V_T be the sets of demographic variables, event-based features, and time-based features respectively. We derive regression models on all combinations of V_D, V_E, and V_T to analyze the individual and joint effects

of each category of variables on predicting reactions vs. non-reactions. That is, we have models that involve V_D only, V_E only, V_T only, V_D and V_E, V_D and V_T, V_E and V_T, and all three. We refer to each model according to their names in Table 4.1. For each variable in each model, we take the mean ($\overline{X}$) and standard deviation (σ) across all 1,000 coefficients. We then compute the Z-score for each coefficient with the following equation:

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{N}, \tag{4.14}$$

where $\mu = 0$ and $N = 1000$.

Model Name	Feature Categories
(A)	V_D
(B)	V_E
(C)	V_T
(D)	V_R
(E)	V_D and V_E
(F)	V_D and V_T
(G)	V_D and V_R
(H)	V_E and V_T
(I)	V_E and V_R
(J)	V_T and V_R
(K)	V_D, V_E, and V_T
(L)	V_D, V_E, and V_R
(M)	V_D, V_T, and V_R
(N)	V_E, V_T, and V_R
(O)	V_D, V_E, V_T, and V_R

Table 4.1: Model names and the feature categories they include, where V_D, V_E, V_T, and V_R are the sets of demographic variables, event-based features, time-based features and trade-specific variables respectively. For example, Model (A) only includes the demographic features of V_D; whereas, Model (O) includes all features from V_D, V_E, V_T, and V_R. All feature categories are described in Sections 4.3.2, 4.4.2, and 4.5.1.

We also apply this experimental design on a sub-sample of non-reactions and reactions that occur after 2009, i.e. post-financial crisis. Again, we create 1,000 pseudo-random samples. Instead of I being the set of all sentiment investors, we re-define I to be the set of all sentiment investors who have reactions/non-reactions post-crisis. Similarly, $I_{\overline{R}}$ is the set of sentiment investors that have at least one

non-reaction after 2009, $I_{\overline{R}}^{(B)}$ is the set of sentiment investors that have at least one buy reaction after 2009, and $I_{\overline{R}}^{(S)}$ is the set of sentiment investors that have at least one sell reaction after 2009. We then use the same previously described sampling process to create the 1,000 samples. Finally, we derive multivariate logistic regression models from each of the samples and compute the mean, standard-deviation, and significance of all coefficients using Equation 4.14. Note that we do not include the indicator variables for trades occurring before, during, and after the financial crisis in these post-crisis models due to multicollinearity concerns.

4.4 Predicting Magnitude of Reaction

We measure the magnitude of a reaction (i.e. trade) to a sentiment event with the proportion of wealth traded. In this section, we first discuss how we compute the proportion of wealth traded for each trade. We also describe the features we include in our model. Finally, we present our experimental design.

4.4.1 Target Variable

We use the proportion of wealth traded as a proxy to measure the magnitude of a sentiment investor's reaction to an event. Let ϕ_{it} be the proportion of wealth traded by a given investor i on day t. That is,

$$\phi_{ijt} = \frac{\delta_{ijt}}{W_{it}}, \tag{4.15}$$

where W_{it} is the investor's wealth (Equation 4.16) and δ_{ijt} is the net volume traded of CUSIP j at t (Equation 4.19). If $\delta_{ijt} > 0$, then investor i made a "net buy" at t. If $\delta_{ijt} < 0$, then investor i made a "net sell." Note that our definition controls for day trading by construction.

For each investor, we sum the cash holdings, portfolio value, and net cash gain in

trades per day:

$$W_{it} = h_{it} + P_{it} + \delta_{it} - M_{it}, \tag{4.16}$$

where W_{it} is the wealth of investor i on day t, h is the amount of cash contained in the account, P is the portfolio value (Equation 4.17), δ is the net cash gain from trades (Equation 4.19), and M_{it} is the total trade commission charged on all trades (Equation 4.20).

To calculate P for each investor i on day t we used the following equation

$$P_{it} = \sum_{j}^{J} p_{jt} q_{ijt}, \tag{4.17}$$

where J is the number of CUSIPs that the investor owns at t, p_{jt} is the price of CUSIP j at t, and q_{ijt} is the number of shares that the investor owns of CUSIP j at time t. Note that we cannot simply download stock market data and map CUSIPs with their historic prices because some CUSIPs have been retired, some are thinly traded, and some are penny stocks among other reasons. Thus, we must calculate p_{jt} ourselves. We define p_{jt} as the average of the prices across all trades made for CUSIP j on day t:

$$p_{jt} = \frac{1}{\sum_{i}^{I} \mathbb{1}\{i \text{ traded } j \text{ at } t\}} \sum_{i}^{I} p_{ijt}^{(d)} \mathbb{1}\{i \text{ traded } j \text{ at } t\}, \tag{4.18}$$

where I is the total number of investors and $p^{(d)}$ is the trade price. If there are no trades made at t, we take the most recently available price from fourteen days or less prior. However, we do not want to use stale prices in our computations so if there is no such price, we leave the price as undefined for that day.

We define δ_{it}, which is investor i's net gain in cash from making trades on day t, as

$$\delta_{it} = \sum_{j}^{J} \sum_{d}^{D} r_{ijdt} (\mathbb{1}\{i \text{ sold } j \text{ during } d \text{ at } t\} - \mathbb{1}\{i \text{ bought } j \text{ during } d \text{ at } t\}), \tag{4.19}$$

and define M_{it} as

$$M_{it} = \sum_{j}^{J} \sum_{d}^{D} m_{ijdt}, \tag{4.20}$$

where D is the total number of trades made per CUSIP, r is the total amount transacted, and m is the transactional cost for making the trade.

4.4.2 Features Used

We group our features into three main categories, demographic variables, event-specific features, time-based features, and trade-specific variables. For brevity, we do not describe the demographic variables and event-specific features because they are the same as those introduced in Section 4.3.2.

The category of time-specific features for this model consists of those described in Section 4.3.2: (1) indicator variables for whether the trade occurred before, during, or after the financial crisis and (2) η_{ijt}, which is the proportion of volume traded of CUSIP j by investor i prior to the date t of the current trade. In addition, we define another time-specific feature ψ_{ijt}, which is the proportion of sentiment trades of CUSIP j made by investor i prior to the date t of the current trade. Specifically,

$$\psi_{ijt} = \frac{\sum_{d}^{D_{jt}} r_{ijd} \mathbb{1}\{d \text{ is a sentiment trade}\}}{\sum_{d}^{D_{jt}} r_{ijd}}, \tag{4.21}$$

where r_{ijd} is the total amount of j traded by i during trade d and D_{jt} is number of trades of CUSIP j made before t. Furthermore, let $\psi_{ijt}^{(B)}$ be the proportion of sentiment buys and let $\psi_{ijt}^{(S)}$ be the proportion of sentiment sells. That is,

$$\psi_{ijt}^{(B)} = \frac{\sum_{d}^{D_{jt}} r_{ijd} \mathbb{1}\{d \text{ is a sentiment buy}\}}{\sum_{d}^{D_{jt}} r_{ijd} \mathbb{1}\{d \text{ is a buy}\}} \tag{4.22}$$

and

$$\psi_{ijt}^{(S)} = \frac{\sum_{d}^{D_{jt}} r_{ijd} \mathbb{1}\{d \text{ is a sentiment sell}\}}{\sum_{d}^{D_{jt}} r_{ijd} \mathbb{1}\{d \text{ is a sell}\}}. \tag{4.23}$$

We include $\psi^{(B)}$ and $\psi^{(S)}$ in our model.

Finally, we use trade-specific variables to derive our model. This category of features consists of indicator variables for whether a trade is a buy or sell and for whether the trade is part of a sequence of day trades. Note that we define a day trade as there being more than one trade (buy or sell) made for the same CUSIP on a given day.

4.4.3 Experimental Design

To model the magnitude of a sentiment investor's reaction to a sentiment event, we regress proportion of wealth traded on all features discussed in the previous section using a multivariate linear regression. Furthermore, we use a similar experimental design as discussed in Section 4.3.3.

Again, each investor can trade multiple CUSIPs in reaction to different sentiment events. Because each sentiment trade is an individual data-point, some investors are represented more in the data than others. Thus, we again use Monte Carlo simulation (Section 1.2.4) and pseudo-randomly create 1,000 different samples. As defined in Section 4.3.3, let I_R be the set of sentiment investors who we have non-missing data for all previously descried features. Let $I'_R \subseteq I_R$ be the subset of sentiment investors who we are also able to compute wealth for at each time of trade. Note that $|I'_R| = 601$, so we cannot use the same disjoint sampling technique as we do in Section 4.3.3 given data size constraints. Instead, we randomly sample a trade made from each investor.

Finally, we normalize the variables for age during trade and $\widetilde{\Delta}$, and derive multivariate linear regression models (Section 1.2.1) from each of the samples. As defined in Section 4.3.3, let V_D, V_E, and V_T be the set of all demographic variables, event-based features, and time based features respectively. Furthermore, let V_R be the set of all trade-based features. We derive linear regression models on all combinations of V_D, V_E, V_T, and V_R as described in Section 1.2.1. We refer to each model according to Table 4.1, and we also include an intercept in each model. We compute the significance of each variable coefficient according to Equation 4.14.

We also apply this experimental design on several sub-samples: buys only, sells

only, and trades made after 2009. For the sub-samples that only have buys or only have sells, we do not include the indicator variable for whether the trade is a buy or sell. In addition, we do not include the indicator variables for whether trades occur before, during, or after the financial crisis if the sub-sample only contains data-points that occurred after 2009.

4.5 Predicting Direction of Reaction

We define the direction of reaction (i.e. trade) to a sentiment event as a buy or sell. In this section, we discuss the labels we predict, the features used for the prediction, and the experimental design. Note that the experimental design is most similar to that of Section 4.3.

4.5.1 Target Variable and Features Used

The predicted label is an indicator for whether a reaction was a buy or a sell. We label buys as ones and sells as zeros. Note that every data-point that we consider for this model is a sentiment trade by construction.

The features we use for this model can be broken into three categories: demographic variables, event-based features, and time-based features. We use all variables and features described in Section 4.3.2. For brevity, we do not re-define them. In addition, we include one additional time-based feature ψ_{ijt} (Equation 4.21), which is the proportion of sentiment trades of CUSIP j made by investor i prior to date t of the current trade.

4.5.2 Experimental Design

We regress the direction of a sentiment trade on all features introduced in the previous section using a multivariate logistic regression model (Section 1.2.2). We use the same experimental design as discussed in Section 4.4.3. We pseudo-randomly create 1,000 samples in the same way such that half of the data-points are sentiment buys and the

other half are sentiment sells. Then, we normalize the variables for age during trade and $\widetilde{\Delta}$. Furthermore, using Powell's conjugate direction method (Section 1.2.3), we derive logistic regressions from each of these samples on all combinations of V_D, V_E, and V_T and also include an intercept for each model. For each variable, we take the mean and standard deviation across all 1,000 coefficients and use Equation 4.14 to determine the significance of each variable. Finally, we apply this experimental design on a sub-sample of all trades made after 2009.

Chapter 5

Results and Discussion

In this section, we first discuss the empirical robustness of our sentiment investor identification mechanism in the data (Section 5.1). Then, we report the summary statistics of all accounts that exhibit sentiment investing behavior (Section 5.2). Finally, we discuss the models that we derive to predict whether an investor will react to a given sentiment event (Section 5.3), the magnitude of the reaction (Section 5.4), and the direction of the reaction (Section 5.5).

5.1 Analysis of the Sentiment Investor Identification Mechanism

To evaluate our sentiment investor identification mechanism, we first analyze the distribution of p_{ij} (Equation 4.2), which is the probability that an investor i trades during the period following a sentiment event for a CUSIP j minus the probability that the investor makes a trade when there is no event. We plot a histogram of all p_{ij} values, and observe that the distribution of p_{ij} values is not a normal distribution and had a large peak just below zero (Figure 5-1).

We observe a left skew of the distribution of p_{ji}, and we attribute the large number of negative p_{ij} values to the sparsity of events for most CUSIPs. In other words, CUSIPs with less events have less event reaction days. As a result, trades for such

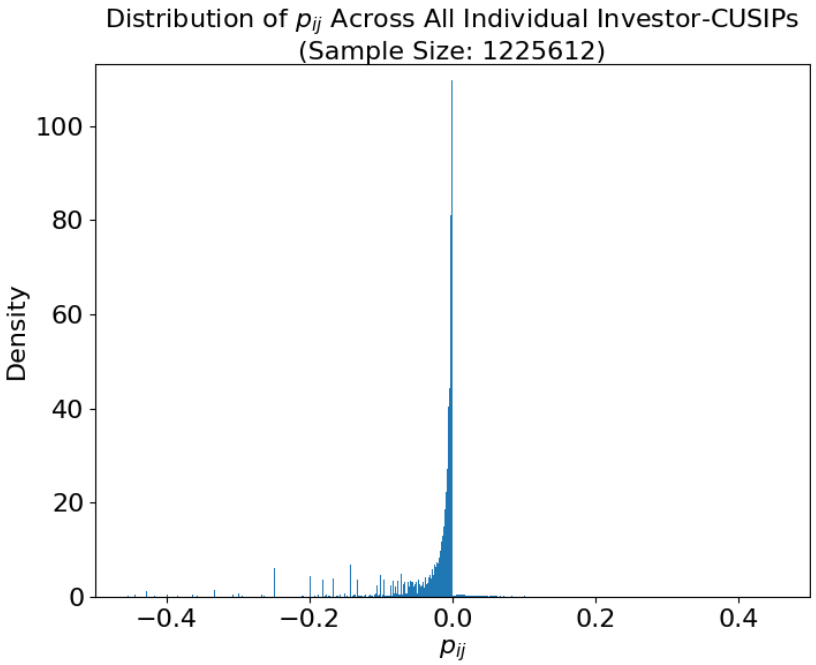

Figure 5-1: Distribution of p_{ij} (Equation 4.2) values across all 1,225,612 Investor-CUSIPs where all investors have individual accounts.

CUSIPs are more likely to occur on non-reaction days, which leads to negative p_{ij} values by construction. To further analyze the distribution of p_{ij}, we filter p_{ij} values based on N_{ij}^E, which is the number of event reaction days for CUSIP j owned by investor i (Figure 5-3). Even after thresholding by N_{ij}^E, the pattern in Figure 5-1 persists (Figure 5-3).

Then, we compute p_i according to Equation 4.3 and plot the values in Figure $5-2$. We observe that the distribution of p_i, which is a linear combination of p_{ij}, is not similar to the distribution of p_{ij}. Thus, we further investigate this discrepancy by separating out the p_{ij} values into three categories: (1) investor i made zero sentiment trades ($T_i^E = 0$), (2) investor i made zero to n sentiment trades exclusive ($T_i^E \in (0, n)$) where n is some threshold, and (3) investor i made at least n sentiment trades ($T_i^E \geq n$) across the lifetime of the account in the dataset.

From Figure 5-4, we observe a truncation of the blue plots at zero. We attribute this observation to the construction of p_i. For each blue data-point, $T_i^E = 0$ which means that $T_{ij}^E = 0$ for all CUSIPs j. That means p_{ij} (Equation 4.2) can be at most

zero for all CUSIPs. If an investor makes any non-sentiment trades, then $p_{ij} < 0$. Thus, we only observe non-positive p_i values in the blue plots. On the other hand, p_i takes on negative, zero, and positive values for groups (2) and (3) in Figure 5-4. Furthermore, we see that the distribution of p_{ij} values for investors in groups (2) and (3) is similar to the distribution of p_i values. This pattern persists as we increase n. We note that p_{ij} can only be nonnegative if there exists event reaction days during investor i's ownership of CUSIP j (i.e. $N_{ij}^E > 0$) and if investor i makes trades during this period (i.e. $T_{ij}^E > 0$). As a result, a large proportion of CUSIPs without events during investor ownership periods skews the distribution of p_{ij}. After isolating this effect, we then see similar distributions between p_{ij} and p_i, as expected.

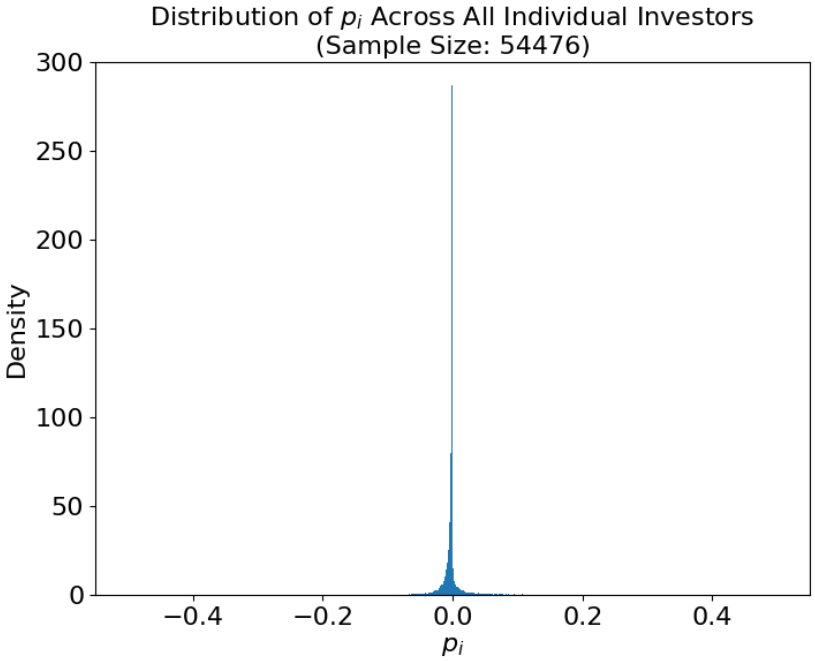

Figure 5-2: Distribution of p_i values (Equation 4.3) across all 54,476 individual investors

Thus, given our observations of the distribution of p_{ij} and p_i in the data, we find p_i to be a compelling metric to use when identifying sentiment investors.

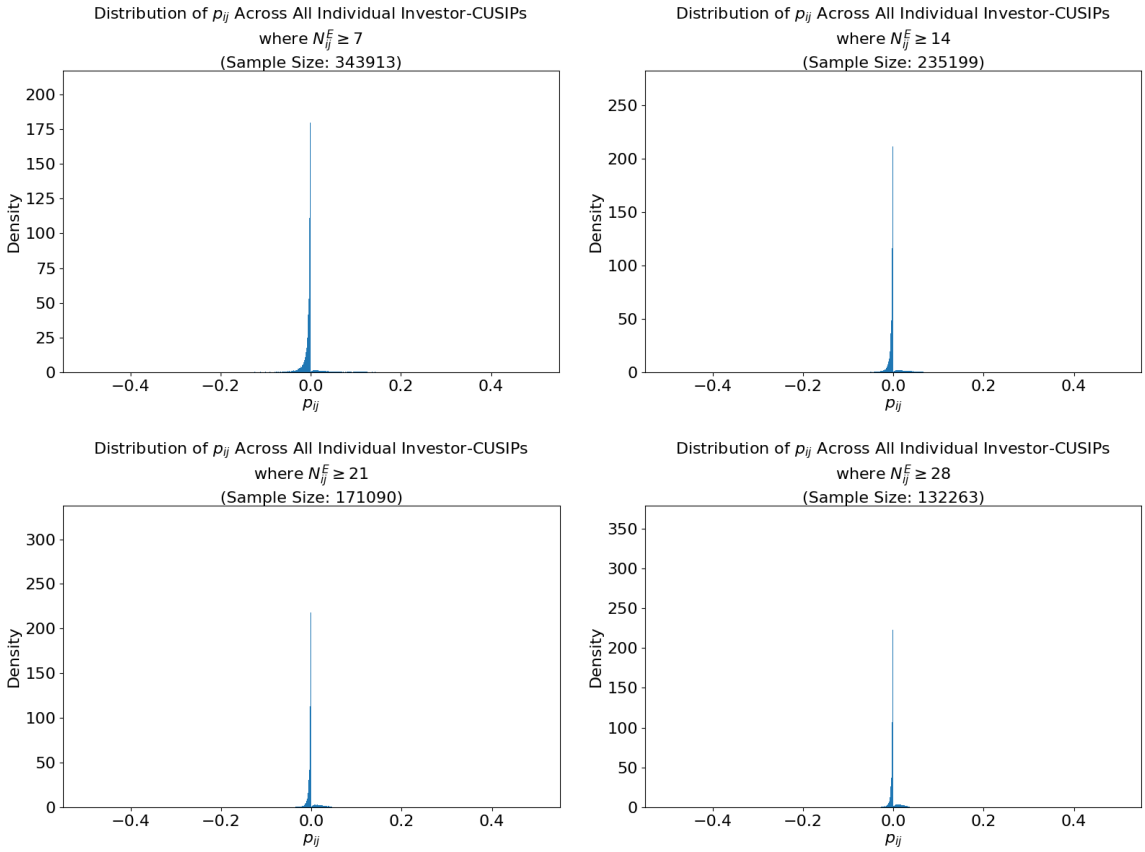

Figure 5-3: Distribution of p_{ij} (Equation 4.2) values across Investor-CUSIP pairs where all investors are individual investors and $N_{ij}^E \geq 7$ (top left), $N_{ij}^E \geq 14$ (top right), $N_{ij}^E \geq 21$ (bottom left), and $N_{ij}^E \geq 28$ (bottom right).

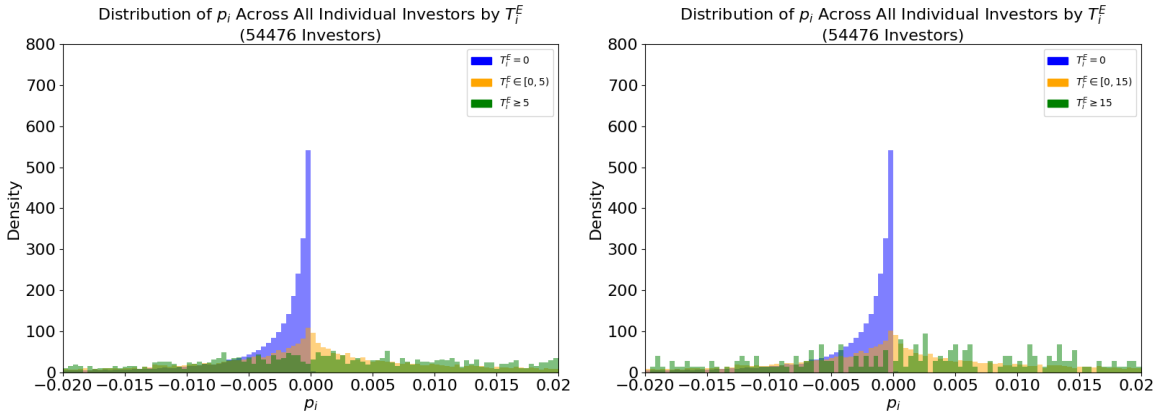

Figure 5-4: p_i values (Equation 4.3) after thresholding for three categories: investor i made zero sentiment trades, investor i made zero to n sentiment trades exclusive, and investor i made at least n sentiment trades across the lifetime of the account in the dataset. We set $n = 5$ on the left and set $n = 15$ on the right. Note that all investors are individual investors. Given that the plots quickly approach zero on both sides of the x axis, we restrict our plots to $p_i \in [-0.02, 0.02]$ for visual convenience.

5.2 Identified Sentiment Investors

Using our criteria, we identify a total of 64,214 accounts, which constitute 9.83% of all accounts, that exhibit sentiment trading behavior. Of the 64,214 sentiment accounts, 55,204 are associated with 40,148 household accounts and 9,010 are individual accounts. We are unable to categorize two accounts as either household or individual accounts, and demographic information is collected from all 9,010 individual sentiment accounts. In the remaining 312,318 accounts, 257,334 accounts are associated with 123,526 household accounts and 54,804 are individual accounts.

In this study, we focus on the sample of 9,010 individual sentiment accounts and 54,804 individual non-sentiment accounts. That is, 14.1% of our sample exhibited sentiment trading behavior. Given that we only study individual accounts, we refer to "individual investor accounts" and "investors" interchangeably.

To better understand our sample of sentiment investors, we report the composition and relative prevalence of the self-reported demographic information of our investors according to Equation 4.5 and compute the significance of the relative prevalence with Equation 4.8. Note that all demographic information was collected at one point in time, and demographic categories were determined by our data vendor. We show the statistics on investors' investment knowledge and investment experience in Tables 5.1 and 5.2, and include remaining demographic statistics (gender, marital status, number of dependents, and occupation group) in Tables A.5-A.8 of Appendix A.

Investment Knowledge	Number of Sentiment Investors	Number of All Investors	Relative Proportion
EXCELLENT	258	1,084	1.69***
			(7.70)
GOOD	879	5,198	1.2***
			(5.17)
LIMITED	1,376	9,661	1.01
			(0.33)
NONE	344	2,679	0.91*
			(-1.69)
Missing	6,153	45,192	0.96***
			(4.92)
Total	9,010	63,814	

Table 5.1: Composition and relative prevalence of sentiment investors' self-reported investment knowledge. "Missing" indicates information that wasn't reported. Note that numbers in parentheses are the standard deviation of coefficients calculated using Equation 4.8. * means significance at the 10% level, ** means significance at the 5% level, and *** means significance at the 1% level.

According to Tables 5.1 and 5.2, sentiment investors are more likely to report that they have excellent or good investment knowledge and experience. They are marginally more likely to report limited investment knowledge and experience and are less likely to report to have no investment knowledge and experience. From Table A.5, we observe that most investors chose not to disclose their gender. Conditioning on the disclosure of gender information, sentiment investors are more likely to be male and less likely to be female. Table A.6 indicates that sentiment investors are more likely to be divorced, married, single, and widowed. In addition, we observe from Table A.7 that sentiment investors are more likely to report information about their number of dependents, and most sentiment investors have zero dependents. On the other hand, they are less likely to be minors, separated, and unmarried. Furthermore, according to Table A.8, sentiment investors are also more likely to be financial professionals, executives and managers, business owners, real estate workers, CPAs, attorneys, retired, skilled laborers (e.g. scientists, government workers, engineers, and paralegals, etc.), self-employed, consultants, and medical professionals and physicians.

They are less likely to be students, in the police force or military, social workers, or disabled.

Investment Experience	Number of Sentiment Investors	Number of All Investors	Relative Proportion
EXCELLENT	377	1,704	1.57***
			(8.07)
GOOD	1,319	6,969	1.34***
			(10.40)
LIMITED	2,019	13,708	1.04**
			(2.00)
NONE	649	4,724	0.97
			(-0.68)
Missing	4,646	36,709	0.90***
			(-10.69)
Total	9,010	63,814	

Table 5.2: Composition and relative proportion of sentiment investors' self-reported investment experience. "Missing" indicates information that wasn't reported. Note that numbers in parentheses are the standard deviation of coefficients calculated using Equation 4.8. * means significance at the 10% level, ** means significance at the 5% level, and *** means significance at the 1% level.

5.3 Reaction vs. No Reaction Model

Here, we introduce our models that predict whether a given sentiment investor will react to a media event. In Section 5.3.1, we discuss the models we derive from running Monte Carlo simulations on event data from all time. We then report models that are trained specifically on post-financial crisis data in Section 5.3.2.

5.3.1 All-Time Model

We report the logistic regression models we derive from running Monte Carlo simulations on reactions and non-reactions for all time in Table 5.3. First, we note that data-points are sampled at the daily level, and these data-points occur across the span of over a decade. Furthermore, they cover all traded CUSIPs during this time

by all sampled sentiment investors. Thus, we expect a nontrivial amount of noise in our data. However, our models that contain two or more variable categories (except for Model (E)) have pseudo-R^2 values that range 0.231-0.369. Noting that pseudo-R^2 values ranging 0.2 to 0.4 represent excellent model fit [7], we conclude that our models have strong predictive power of whether a sentiment investor will react to a given event. In addition, we find that for almost every coefficient in every model, the coefficient's average across all simulations is significant at the 1% level. We attribute the large proportion of significant coefficients to the large number of Monte Carlo simulations that we run.

In Table 5.3, the first three columns are baseline models that each involve one category of variables: demographic variables (V_D), event-based features (V_E), and time-based variables (V_T). The following three columns combine any two of these categories. We observe that that the R^2 values of any two baseline models approximately sum to the R^2 value of the model combining the two categories of variables. For example, the R^2 values of the baseline models for V_D only (Model (A)) and V_E only (Model (B)) are 0.002 and 0.093 respectively, and Model (E) which combines the two has a 0.096 R^2. In the case of Model (H), the R^2 value (0.369) significantly exceeds the sum of the R^2 values of the two baseline models (Model (B): 0.093, Model (C): 0.229). Thus, we conclude that V_D, V_E, and V_T explain largely orthogonal portions of the variance in the data. In addition, the variable categories tend to better fit the data in conjunction with each other.

Next, we analyze the effect of individual variables on the log-odds of reacting to a sentiment event. Beginning with the variables in V_D, we note that the absolute values for the age variable post normalization can be larger than 1 and thus be greater than the other demographic variables, which take on binary values, but have coefficients that are generally on the same order of magnitude. Thus, although the coefficients on age are close to zero, we conclude that differences in age impact the probability of reacting to a given sentiment more than changes in other demographic variables.

Furthermore, after adding V_E to the baseline V_D model (Model (E)), we observe that the magnitudes and signs of the coefficients of the demographic variables tend

to stay consistent. However, there are some exceptions to this pattern. The signs of the coefficients on age and the indicator variable for reporting excellent investment experience change after the addition of event-based features. However, this is unsurprising because the magnitudes of the coefficients in Model (A) were close to zero to begin with, and values within one standard deviation away around the mean exhibit different signs. In addition, the magnitude on the indicator variable for being married significantly changes with the addition of V_E variables. We would expect to see this effect if the other regression variables exhibited stronger positive effects on the log-odds of reacting to a sentiment event such that the effect of being married became more negative. However, this is not the case. At the same time, the standard deviation of the marital status coefficient also increases with the addition of V_E to the model, which indicates that the coefficient of marriage can take a large range of values. Thus, we cannot immediately make many conclusions about the nature of the relationship between marital status and the probability of sentiment reaction.

We now remove the V_D categories and shift our attention to the the baseline V_E model (Model (B)). From the second column of Table 5.3, we observe that the relative magnitudes and signs of the V_E coefficients stay mostly the same. However, we do notice a larger increase in magnitude of $\widetilde{\Delta}$, which is the magnitude of the sentiment event with the largest absolute value in the week leading up to the given reaction/non-reaction. We attribute this effect to omitted variable bias. That is, the variance in the data explained by $\widetilde{\Delta}$ in the baseline model is largely explained away by the demographic variables in Model (E). Furthermore, note that η, which is the number of events in the month prior to the reaction/non-reaction, and $\widetilde{\Delta}$ can have absolute values larger than one post-normalization. On the other hand, ρ, which is the fraction of positive events in the week leading up to the data-point, is constrained to be between zero and one but has coefficients whose values are significantly larger than those on η and $\widetilde{\Delta}$. Noting these facts, we conclude that while it is difficult to evaluate the relative strength of ρ and $\widetilde{\Delta}$, they are all still strong predictors. Holding other predictors constant in Models (B) and (E), a one unit increase in ρ leads to an approximately 1.837 expected decrease in the log-odds of an event reaction. In other

words, a one unit increase in ρ leads to a less than 1.837 decrease in probability of sentiment reaction compared to probability of non-sentiment reaction.

Next, we analyze the baseline V_T model (Model (C)). We observe that the coefficients on the indicator variables for the data-point occurring before and during the financial crisis are positive and consistent in magnitude. Furthermore, the magnitudes of the coefficients are relatively larger than others in the table (especially those on the indicator variable for the data-point occurring before the financial crisis), and have relatively small standard deviations. Thus, we can conclude that the timing of events are strong predictors of the probability that a sentiment investor reacts to the event. In particular, if an event occurs before or during the crisis, there is a 9.008 or 0.632 unit increase respectively in the log-odds of reacting to an event than if the event occurs after the crisis. That is, sentiment investors are more likely to react to sentiment events before and during the financial crisis than after.

We also study the coefficients on ζ, which is the proportion of volume traded of the given CUSIP across an investor's trading history up until the data-point. The sign on these coefficients are consistently negative across all models, but the magnitude of the coefficient in Model (C) is significantly larger than the coefficients on ζ in the other models. We attribute this effect to omitted variable bias. In Model (F), the demographic variables explain away some of the variance captured by ζ, thus decreasing the magnitude of the coefficient from 8.906 to 0.385. In Model (H), even more of the variance is explained away by V_E variables such that the magnitude of the coefficient on ζ decreases from 8.906 to 0.042. That is, in the presence of variables from other categories, the predictive power of ζ significantly decreases. However, despite these smaller coefficient values, the coefficients on ζ in the non-baseline models are still non-trivially positive, i.e. nonzero. Thus, we conclude that ζ is still a good predictor of the probability of sentiment reaction. A sentiment investor is more likely to react to an event concerning a given CUSIP if the investor already has a history of trading that CUSIP.

We now add V_D variables to Model (C) and analyze the results in the third to last column of Table 5.3 (Model (F)). The magnitudes and signs on the coefficients in V_T

remain largely the same except for those of ζ, which we discussed previously. Notably, the magnitudes on the coefficients for the investment knowledge variables, indicator for having excellent investment experience, indicator for marriage, and indicator for having 1-3 dependents are significantly larger in Model (F) than in Model (C). We believe that these variables in particular explain away much of the variance captured by ζ Model (C). We see a similar effect in Model (H) when we add V_E instead of V_D to the baseline V_T model. All magnitudes of the coefficients for variables in V_E increase, particularly those on $\widetilde{\Delta}$. Again, we hypothesize that the variables in V_E explain away data variance previously modeled by ζ. We also attribute the increase in magnitudes of the V_E variables to the possible correlations between the variables in V_E and V_T. For instance, the nature of sentiment events during the financial crisis are with high likelihood different than those that occur pre- and post-crisis.

Finally, we discuss our main contribution of this section, which is the model that involves all categories of variables (Model (K)). We observe that the age of the sentiment investor has a slightly positive (coefficient = 0.023) effect on the log-odds of reacting, and thus a close to zero effect on the probability of sentiment reaction. Similarly, the coefficients on the indicator variables for having 0 or 1-3 dependents (-0.006 and -0.023 respectively) as well as ζ (-0.073) indicate that the number of dependents an investor has and the proportion of volume previously traded by the investor on the given CUSIP have an almost zero effect on the probability of sentiment reaction. On the other hand, the coefficients on the indicators for reporting to have excellent, good, or limited experience as well as the indicators for reporting to have excellent, good, or limited knowledge have larger magnitudes and are consistently negative. In other words, if an investor has limited to excellent investment knowledge and/or experience, we would expect that their probability of reacting to a given event is lower than if they had no investment knowledge and/or experience. Furthermore, if the event occurred before or during the financial crisis, we expect the log-odds of reaction to no reaction to increase by 8.067 or 0.535 respectively. In other words, if the event occurs before 2009, we expect the probability of a reaction to increase by less than 8.067 or 0.535 than if the event occurs after 2009.

		(A)	(B)	(C)	(E)	(F)	(H)	(K)
V_D	Age	0.018***			-0.001	0.058***		0.023***
		(0.063)			(0.066)	(0.085)		(0.105)
	0 dependents	-0.037***			-0.044***	-0.003**		-0.006***
		(0.031)			(0.032)	(0.038)		(0.042)
	1-3 dependents	0.072***			0.037***	0.195***		-0.023
		(0.083)			(0.058)	(0.313)		(0.505)
	Excellent investment	0.028***			-0.018***	0.186***		-0.004
	experience	(0.101)			(0.082)	(0.3)		(0.475)
	Good investment	-0.073***			-0.088***	-0.07***		-0.369***
	experience	(0.159)			(0.159)	(0.28)		(0.35)
	Limited investment	-0.082***			-0.085***	-0.081***		-0.375***
	experience	(0.105)			(0.093)	(0.226)		(0.281)
	Excellent investment	-0.044***			-0.024***	0.012**		-0.188***
	knowledge	(0.093)			(0.085)	(0.19)		(0.232)
	Good investment	-0.156***			-0.076***	-0.373***		-0.171***
	knowledge	(0.165)			(0.154)	(0.234)		(0.288)
	Limited investment	-0.098***			-0.037***	-0.354***		-0.172***
	knowledge	(0.112)			(0.092)	(0.184)		(0.23)
	Married	-0.083***			-0.442***	-7.22***		-7.46***
		(0.074)			(0.13)	(1.324)		(2.49)
V_E	$\widetilde{\Delta}$		-0.506***		-0.039***		-9.251***	-0.339***
			(0.12)		(0.083)		(0.876)	(0.213)
	ρ		-1.839***		-1.837***		-2.601***	-2.598***
			(0.718)		(0.72)		(0.921)	(0.923)
	η		0.026***		0.026***		0.08***	0.081***
			(0.035)		(0.035)		(0.046)	(0.046)
V_T	Is before the			9.008***		7.484***	9.333***	8.067***
	financial crisis			(0.848)		(1.05)	(0.859)	(1.91)
	Is during the			0.632***		0.617***	0.551***	0.535***
	financial crisis			(0.083)		(0.087)	(0.099)	(0.103)
	ζ			-8.906***		-0.385***	-0.042***	-0.073***
				(0.847)		(0.162)	(0.122)	(0.132)
Pseudo R-Squared		0.002***	0.093***	0.229***	0.096***	0.231***	0.369***	0.369***
		(0.001)	(0.05)	(0.007)	(0.05)	(0.007)	(0.073)	(0.073)
No. Observations		2006.826	2006.826	2006.826	2006.826	2006.826	2006.826	2006.826
		(7.096)	(7.096)	(7.096)	(7.096)	(7.096)	(7.096)	(7.096)

Table 5.3: Multivariate logistic regression models for predicting reaction vs. no reaction to a given sentiment event. All symbols and variables are defined in Section 4.3. Note that numbers in parentheses are the standard deviation of coefficients calculated using Equation 4.14. * means significance at the 10% level, ** means significance at the 5% level, and *** means significance at the 1% level. We consider coefficients with a standard deviation of zero to be significant at the 1% level. Models are named according to Table 4.1.

5.3.2 Post-Crisis Model

We also derive logistic regression models on the sub-sample of data-points that occur after 2009, i.e. after the financial crisis, so that our contributions are based on more recent and/or relevant data and thus more directly applicable to real-world applications. We report these models in Table 5.4.

Generally speaking, we observe the same trends and patterns across all variables in all models as those in Table 5.3 of Section 5.3.1. The pseudo R-squared values in the post-crisis models are largely greater than those in the all-time models, which is unsurprising given that there are less data per Monte Carlo simulation due to the post-crisis sampling. However, the models that involve V_T are an exception to this pattern. From Table 5.4, we observe that the baseline V_T model (Model (C)) has a lower pseudo R-squared value (0.201) than its all-time counterpart (0.229). We believe that the smaller R-squared value comes from the fact that we leave out the indicator variables for data-points occurring before and during the crisis for the post-crisis simulations. Furthermore, we omit the variance from pre-crisis data. As a result, Model (C) comprises of only one variable as opposed to the three in the all-time version of the same model. Given the additive nature of the pseudo R-squared values in models with combinations of variable categories, it is therefore unsurprising that models involving V_T have systematically lower pseudo R-squared. Despite this, the models that involve two or more variable categories (except for Model (E)) still achieve pseudo R^2 values ranging from 0.2-0.4. According to McFadden [7], pseudo R^2 values within this range indicate that the corresponding models are excellent fits of the data. Thus, we conclude that our models fit the post-crisis data well.

We now analyze the coefficients on the variables. First, we focus our attention on the demographic variables of V_D. In general, we see similar trends in the relative magnitudes and signs of the coefficients. For instance, like their all-time model counterparts of Section 5.3.1, the coefficients on the age variable are consistently close to zero, and values within one standard deviation of the coefficients can be positive or negative. Similarly, the coefficients on the marital status variable have approxi-

mately the same magnitudes and signs in each column of the all-time and post-crisis tables. We observe similar patterns for the coefficients on the indicator for having no dependents and reporting to have excellent investment knowledge.

On the other hand, there are post-crisis variables with relatively different magnitudes or different signs than those in the all-time models, namely the indicators for having one to three dependents, excellent investment experience, and good investment knowledge. However, the standard deviations on these coefficients in both the all-time and post-crisis models indicate that these coefficients can take values from a relatively large range, and thus account for the discrepancies between these coefficients of the two models.

In addition, there are demographic variables in the post-crisis models with coefficients that share the same signs as those in the all-time models but have consistently larger magnitudes across all models. These variables are the indicators for reporting to have good investment experience, limited investment experience, and limited investment knowledge. Like we did in Section 5.3.1, we conclude that if an investor reports to have good investment experience, limited investment experience, or limited investment knowledge, we generally expect to see a decrease in the log-odds of reacting to not reacting than if the investor reports to have no investment experience or knowledge. We can also conclude that these indicators, which were already relatively strong demographic predictors of reaction vs. non-reaction in the all-time models, have even stronger negative effects on the probability of reacting to not reacting in the post-crisis models.

Next, we discuss the features in V_E and V_T. We see that the coefficients on $\widetilde{\Delta}$ and ρ are similar in both in Table 5.3.1 and Table 5.3.2 and thus exhibit the same trends and relationships with other variables mentioned in Section 5.3.1. In addition, we observe that η is consistently slightly more positive in the post-crisis models than in the all-time models, but otherwise follows the same patterns in both. On the other hand, we observe more discrepancy in the signs of the ζ coefficients. We believe that this effect is caused by the absence of the indicator variables for data-points occurring before or during the crisis in the post-crisis models. Instead of there being

three features in V_T as in the all-time model, the post-crisis model only has one.

Finally, we discuss the main contribution of this sub-section, which is the model that incorporates all feature categories (Model (K)). First, we observe that age, having zero dependents, the number of events in the month before the data-point, and the proportion of volume previously traded by the investor on the given CUSIP have coefficients 0.044, -0.027, 0.099, and 0.072 respectively. Thus, we conclude that these variables have little to no effect on the probability of sentiment reaction. On the other hand, the magnitude on the marital status coefficient is significantly larger (-9.298). That is, if an investor is married, we expect a 9.298 unit decrease in log-odds of reacting to a sentiment event than if the investor were single. Furthermore, from Table 5.4, we can easily see that, out of all the other variables and features, marital status is the strongest predictor of the probability that an investor react to a given event.

Next, we observe that the coefficient on having one to three dependents is 0.317, which implies that the indicator could have a positive effect on the probability of reaction. However, the standard deviation of the coefficient across all Monte Carlo simulations is 0.442. This indicates that we only observe the positive effect on expectation, and values within one standard deviation of 0.317 can be zero and slightly negative. The coefficients on the indicators for having excellent and limited investment experience behave in a similar manner. They indicate that if an investor has excellent or limited investment experience, we would expect a 0.394 unit increase or 0.189 unit decrease respectively in the log-odds of reacting to an event than if the investor were to have no investment experience. However, the standard deviation of these coefficients (0.416 and 0.285 respectively) indicate that this effect could also be zero or slightly nonzero. On the other hand, our results indicate that if an investor has good investment experience, we can expect a 0.423 unit decrease in log-odds of reacting than if the investor were to have no experience.

Finally, we note that the coefficients on the indicators for investment knowledge are consistently smaller in magnitude than those of their investment experience counterparts. Specifically, the indicators for having excellent, good, and limited investment

knowledge are -0.074, 0.031, and -0.19 respectively while the indicators for investment experience are 0.394, -0.423, and 0.189 respectively. Furthermore, we do not observe a similar trend in Table 5.3. Thus, we conclude that investment experience has a stronger average effect on the probability of reacting to an event than investment knowledge is post-financial crisis. We also discuss the effects of $\widetilde{\Delta}$ and ρ on the probability of reacting. The magnitudes of the coefficients indicate that both variables have negative effects on this probability, even after considering values that are within one standard deviation of the coefficients. That is, we expect the magnitude of the sentiment event with the largest absolute value in the week prior to a data-point to have a 0.334 unit decrease on the log-odds of reacting to a sentiment event. We also expect the fraction of positive events over the same period to have a 2.49 unit decrease on the log-odds of reacting to a sentiment event. Thus, we conclude that these two sentiment event features are strong negative predictors of reacting to sentiment relative to other variables in the model.

		(A)	(B)	(C)	(E)	(F)	(H)	(K)
V_D	Age	-0.035***			-0.038***	-0.006**		0.044***
		(0.084)			(0.09)	(0.096)		(0.114)
	0 dependents	-0.022***			-0.037***	0.003**		-0.027***
		(0.04)			(0.042)	(0.047)		(0.051)
	1-3 dependents	0.127***			0.124***	0.136***		0.317***
		(0.128)			(0.152)	(0.297)		(0.442)
	Excellent investment	0.18***			0.178***	0.217***		0.394***
	experience	(0.148)			(0.173)	(0.293)		(0.416)
	Good investment	-0.298***			-0.226***	-0.417***		-0.423***
	experience	(0.246)			(0.282)	(0.352)		(0.393)
	Limited investment	-0.204***			-0.128***	-0.226***		-0.189***
	experience	(0.164)			(0.184)	(0.284)		(0.285)
	Excellent investment	-0.102***			-0.086***	-0.081**		-0.074***
	knowledge	(0.151)			(0.155)	(0.242)		(0.241)
	Good investment	-0.107***			-0.15***	-0.089***		0.031**
	knowledge	(0.247)			(0.291)	(0.317)		(0.359)
	Limited investment	-0.138***			-0.197***	-0.231***		-0.19**
	knowledge	(0.174)			(0.209)	(0.245)		(0.257)
	Married	0.059***			-0.323***	-6.54***		-9.298***
		(0.135)			(0.182)	(1.735)		(3.336)
V_E	$\widetilde{\Delta}$		-0.508***		-0.238***		-9.653***	-0.334***
			(0.104)		(0.183)		(0.499)	(0.227)
	ρ		-1.76***		-1.755***		-2.507***	-2.49***
			(0.387)		(0.392)		(0.486)	(0.493)
	η		0.096***		0.099***		0.097***	0.099***
			(0.041)		(0.041)		(0.05)	(0.05)
V_T	ζ			-10.214***		0.459***	0.094***	0.072***
				(0.12)		(0.22)	(0.158)	(0.166)
Pseudo R-Squared		0.005***	0.095***	0.201***	0.099***	0.205***	0.346***	0.35***
		(0.002)	(0.046)	(0.008)	(0.046)	(0.008)	(0.065)	(0.065)
No. Observations		1330.288	1330.288	1330.288	1330.288	1330.288	1330.288	1330.288
		(6.265)	(6.265)	(6.265)	(6.265)	(6.265)	(6.265)	(6.265)

Table 5.4: Multivariate logistic regression models for predicting reaction vs. no reaction to a given sentiment event post-financial crisis. All symbols and variables are defined in Section 4.3. Note that numbers in parentheses are the standard deviation of coefficients calculated using Equation 4.14. * means significance at the 10% level, ** means significance at the 5% level, and *** means significance at the 1% level. We consider coefficients with a standard deviation of zero to be significant at the 1% level. Models are named according to Table 4.1.

5.4 Magnitude of Reaction Model

In this section, we discuss the models we derive to predict the magnitude of reaction to sentiment events. First, we discuss the model that we derive from running Monte Carlo simulations on both buy and sell trade data for all time (Section 5.4.1). Then, we report our model results from the sub-samples involving only buy reactions (Section 5.4.3) and only sell reactions (Section 5.4.4). Finally, we discuss the model that we derive from running Monte Carlo simulation on trades that occurred after the financial crisis (Section 5.4.2).

5.4.1 All-Time Model for All Trades

We report the models we derive from running Monte Carlo simulations on all trades for all time in Table 5.5. Given the nature of the data, we expect a large amount of noise in our results. The granularity of the data-points is at the daily level that spans over a little over a decade, and data-points cover all traded CUSIPs during this time by all sampled sentiment investors. However, our models that contain three or more variable categories (with the exception of Model (C)), have R^2 values that range 0.089-0.102. That is, our models are able explain 8.9%-10.2% of the variance in the data, which are nontrivial percentages. Furthermore, almost all of the coefficients across all models are significant at the 1% level. We attribute this large proportion of significant coefficients to the large number of Monte Carlo simulations we run.

The first four columns in the table are baseline models that each involve one category of variables: demographic variables (V_D), event-based features (V_E), time-specific features (V_T), and trade-specific features (V_R). These models exhibit R^2 values of 0.013, 0.012, 0.075, and 0.003 respectively. Furthermore, from the following six columns in Table 5.5, models that combine any two of these categories have R^2 values that are approximately equal to the sum the R^2 values of the two corresponding baseline models. For instance, Model (H) (column 8) has an R^2 value of 0.088 which is slightly greater than $0.012+0.075 = 0.087$. We see a similar additive effect in the R^2 values for the models that involve any three of the variable categories (columns 11-14)

as well as all four of the categories (column 15). Thus, we conclude that V_D, V_E, V_T, and V_R are largely orthogonal to each other with respect to explaining variance in the data.

Next, we discuss the effect of individual variables on the proportion of wealth traded in reaction to an event. Beginning with the variables in V_D, we observe that the coefficient on the age during a trade (0.307) has the largest magnitude and thus has the largest effect on the proportion of wealth traded compared to other variables in the baseline demographic model (coefficients range -0.083 to 0.033 inclusive). Noting that the absolute value of age can be larger than one after normalization, the magnitude of the coefficient on age is about ten times larger than the coefficients on the other demographic features, which are binary variables. This effect persists when adding V_T (Model ((F)), V_R (Model (G)), or both V_T and V_R (Model (M)) to the regression. However, adding V_E by itself or in combination with the other categories of variables removes this effect. We believe that we may see this pattern due to omitted variable bias. That is, $\widetilde{\Delta}$ explains much of the variance previously captured by age. Specifically, the magnitude of the coefficients on $\widetilde{\Delta}$ (which is a variable in V_E) across all models is consistently positive and relatively large in magnitude, ranging from 0.196 to 0.318. Though significantly smaller in magnitude, the coefficients on ρ are also consistently positive and equal to either 0.015 or 0.016.

In other words, the magnitude of the sentiment event with the largest absolute value across recent events is a key predictor of the proportion of wealth traded. Furthermore, a one unit increase in the proportion of positive events in the seven-day period leading up to a reaction raises the expected proportion of wealth traded during that reaction by about 0.016. In the absence of the features in V_E, the age of the sentiment investor at the time of a given sentiment event becomes a significantly more important predictor of reaction magnitude.

In general, the coefficients on the other demographic variables are all close to zero. The coefficients on the indicator variables for having zero dependents and having one to three dependents are slightly negative. Due to multicollinearity concerns, we leave out the indicator variable for having more than three dependents. Thus, we

can conclude that if a sentiment investor has three or less dependents, the expected proportion of wealth traded in reaction to an event is slightly less than if the sentiment investor has more than three dependents. Similarly, the coefficients on the indicator variable for being married are also slightly negative and equal to -0.004. That is, if a sentiment investor is married, we expect the investor to trade a proportion of wealth that is -0.004 units less than if the investor were single.

We analyze the indicator variables for reported investment knowledge and experience in tandem. Note that the categories for both groups of indicator variables are excellent, good, limited, and none. Looking at the coefficients across each row, we can see that the signs of the coefficients are largely consistent for each indicator variable. The indicator variable for having good investment experience and the indicator variable for having limited investment knowledge do not have coefficients with consistent signs. However, the magnitudes of these coefficients are close to zero and their standard deviations are large enough such that values within one standard deviation away from the mean do not necessarily share the same sign.

Although the coefficients on all the variables across all models are relatively small and close to zero, we can still observe some interesting patterns. For instance, the coefficients on the indicator variable for having excellent investment experience are consistently negative and those that correspond with the indicator variable for having limited investment experience are consistently positive. Furthermore, if we add and subtract these coefficients by one standard deviation, the signs do not change. Thus, we expect sentiment investors who report to have excellent investment experience to trade a smaller proportion of their wealth compared to those who report to have no investment experience. In addition, we expect sentiment investors who report to have limited investment experience to trade a larger proportion of their wealth compared to those who report to have no investment experience. By transitivity, we can conclude that those who report to have excellent investment experience will trade a smaller proportion of their wealth in reaction to a sentiment event compared to those who report to have limited experience.

While the coefficients for having excellent investment experience are negative, the

coefficients for having excellent investment knowledge are consistently positive even after adding/subtracting by one standard deviation. This suggests that sentiment investors who report to have excellent investment experience do not necessarily report to have excellent investment knowledge, and vice versa. The inconsistency raises interesting psychological and behavioral economic questions as we would expect people who have excellent investment experience to also have excellent investment knowledge, and vice versa.

Furthermore, upon close inspection of Models (F), (K), (M), and (O), we find that the magnitudes of the coefficients for having excellent investment experience and good investment knowledge are consistently smaller than those in other columns, i.e. models without V_T. On the other hand, the magnitudes of the coefficients for having limited investment experience are consistently larger in models with V_T than those without. We attribute this pattern to omitted variable bias. That is, the variables in V_T explain away some of the variance captured by the indicator for having limited investment experience.

We also study the individual effects of V_T variables on reaction magnitude, and note that the values within one standard deviation of the coefficients on η, $\psi^{(B)}$, and $\psi^{(S)}$ are consistently positive, and the coefficients range from 0.021-0.023, 0.126-0.131, and 0.012-0.018 respectively. In other words, η, which is the proportion of volume traded of a given CUSIP across all other CUSIPs before the current trade, has a small positive but significantly non-zero effect on the proportion of wealth traded in reaction to a sentiment event. Thus, we conclude that if a sentiment investor has a history of trading a given CUSIP, the magnitude of their reaction to an event concerning the CUSIP is larger than if the investor never traded the CUSIP before. We draw the same conclusion from observing the coefficients on $\psi^{(B)}$ and $\psi^{(S)}$. That is, if an investor has a history of trading a given CUSIP on sentiment, they will have a larger reaction to the current event than if they had never traded the CUSIP on sentiment before. We also observe that the coefficients on $\psi^{(B)}$ are an order of magnitude larger than those on $\psi^{(S)}$. We attribute this phenomenon to the fact that, in order to sell a CUSIP on sentiment, an investor must first own some shares of that CUSIP before the

event. Noting that most sentiment-based strategies tend to have shorter investment horizons, the likelihood of owning CUSIP shares before an event is relatively low. Thus, the proportion of buys that were made on sentiment up until the current trade is a better predictor of reaction magnitude than is the proportion of sells that were made on sentiment.

We now analyze the coefficients on the indicator variables for the timing of the reaction. Due to multicollinearity concerns, we leave out the indicator variable for the reaction occurring after the financial crisis and include the variables for the reaction occurring before and during the financial crisis. From the table, we observe that the coefficients on the variable for the trade occurring during the financial crisis are consistently negative and almost zero in magnitude. That is, if an investor reacts to an event that occurs during the financial crisis, we expect to observe a slightly smaller proportion of wealth traded compared to if the investor were to react to an event post-financial crisis.

We also analyze the coefficients on the indicator variable for the reaction occurring before the financial crisis. Upon initial inspection, the coefficients seem inconsistent in magnitude and sign. However, a clear pattern emerges if we study pairs of models where one model doesn't include V_R and the other one is the same model with V_R included. The pairs are as follows: Models (F) and (M), Models (H) and (N), and Models (K) and (O). For each pair, we observe that the coefficients on the indicator for pre-crisis reaction and their standard deviations are equal across both elements. For example, the coefficient is 0.003 with a standard deviation of 0.006 for Models (H) and (N). From this observation, we can conclude that the variables in V_T are essentially orthogonal in nature to those in V_R. Introducing V_R to a model that contains V_T does not impact the predictive power of the pre-crisis indicator variable on the magnitude of a sentiment reaction.

We study the nature of the coefficients on the variables in V_R. The coefficients on the indicator variable for whether a reaction is a buy are not consistent in sign but are all close to zero, with the exception of the coefficient in the V_R only model, which is a couple orders of magnitude larger. Given the low R^2 value of the model (0.003)

and the fact that there are only two variables in V_R, we can attribute the inconsistent coefficient value to omitted variable bias. On the other hand, the coefficients on the indicator variable for whether a reaction is a day trade are consistently close to zero.

Finally, we discuss our main contribution of this section, which is the model that involves all categories of variables (Model (O)). Although the R^2 value of this model indicates that we can explain 10.2% of the variance in the data, the magnitudes of most of the coefficients are quite small. In general, the investor's age, reporting of having good or limited experience, reporting of having excellent or good investment knowledge, proportion traded of the given CUSIP across all previously traded CUSIPs, proportion of sells of the CUSIP made on sentiment, and the fraction of positive events in the week prior to the current data-point have slightly positive effects on the log-odds of reacting to a sentiment event. On the other hand, having three or less dependents, reporting to have excellent investment experience or limited investment knowledge, and being married as well as the event occurring before 2009 have slightly negative effects on the log-odds of reacting to a sentiment event. The coefficients on $\widetilde{\Delta}$ and $\phi^{(B)}$ are relatively larger in magnitude (0.256 and 0.13 respectively) compared the other coefficients. Thus, we can conclude that the magnitude of the sentiment event with the largest absolute value in the week prior to the data-point and the proportion of buys of the CUSIP made on sentiment have the strongest average effects on the log-odds of reacting to a sentiment event. left=1in,right=1in,top=1in,bottom=1in

		(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)	(N)	(O)
V_D	Age	0.307***				0.005***	0.246***	0.324***				0.002***	0.005***	0.262***		0.003***
		(0.028)				(0.007)	(0.028)	(0.029)				(0.006)	(0.007)	(0.029)		(0.006)
	0 dependents	-0.004***				-0.003***	-0.009***	-0.003***				-0.008***	-0.002***	-0.008***		-0.007***
		(0.003)				(0.003)	(0.003)	(0.003)				(0.003)	(0.004)	(0.003)		(0.003)
	1-3 dependents	-0.083***				-0.081***	-0.064***	-0.082***				-0.062***	-0.08***	-0.063***		-0.061***
		(0.027)				(0.027)	(0.027)	(0.027)				(0.027)	(0.027)	(0.027)		(0.027)
	Excellent investment experience	-0.073***				-0.073***	-0.059***	-0.071***				-0.059***	-0.072***	-0.057***		-0.058***
		(0.027)				(0.027)	(0.027)	(0.027)				(0.027)	(0.028)	(0.027)		(0.027)
	Good investment experience	-0.006***				-0.009***	0.022***	-0.006***				0.019***	-0.008***	0.021***		0.019***
		(0.025)				(0.025)	(0.024)	(0.025)				(0.024)	(0.025)	(0.024)		(0.024)
	Limited investment experience	0.013***				0.013***	0.029***	0.012***				0.029***	0.012***	0.029***		0.028***
		(0.008)				(0.008)	(0.008)	(0.008)				(0.008)	(0.008)	(0.008)		(0.008)
	Excellent investment knowledge	0.022***				0.022***	0.03***	0.021***				0.029***	0.021***	0.029***		0.028***
		(0.006)				(0.006)	(0.006)	(0.006)				(0.006)	(0.006)	(0.006)		(0.006)
	Good investment knowledge	0.033***				0.037***	0.016***	0.033***				0.021***	0.038***	0.017***		0.023***
		(0.022)				(0.022)	(0.021)	(0.022)				(0.021)	(0.022)	(0.021)		(0.021)
	Limited investment knowledge	-0.001***				0.001**	-0.008***	-0.001**				-0.006***	0.001***	-0.007***		-0.005***
		(0.008)				(0.009)	(0.008)	(0.008)				(0.008)	(0.009)	(0.008)		(0.008)
	Married	-0.004***				-0.004***	-0.004***	-0.004***				-0.004***	-0.005***	-0.004***		-0.004***
		(0.002)				(0.002)	(0.002)	(0.002)				(0.002)	(0.002)	(0.002)		(0.002)
V_E	$\widetilde{\Delta}$		0.238***			0.301***			0.196***	0.256***		0.241***	0.318***		0.213***	0.256***
			(0.004)			(0.029)			(0.004)	(0.007)		(0.029)	(0.03)		(0.008)	(0.03)
	ρ		0.015***			0.016***			0.016***	0.015***		0.016***	0.015***		0.016***	0.016***
			(0.002)			(0.002)			(0.002)	(0.002)		(0.002)	(0.002)		(0.002)	(0.002)
V_T	Is before the financial crisis			0.199***			-0.008***		0.003***		0.216***	-0.007***		-0.008***	0.003***	-0.007***
				(0.003)			(0.006)		(0.006)		(0.007)	(0.007)		(0.006)	(0.006)	(0.007)
	Is during the financial crisis			-0.018***			-0.019***		-0.015***		-0.018***	-0.016***		-0.02***	-0.015***	-0.017***
				(0.004)			(0.005)		(0.004)		(0.004)	(0.005)		(0.005)	(0.004)	(0.005)
	ζ			0.022***			0.021***		0.023***		0.022***	0.023***		0.021***	0.023***	0.023***
				(0.005)			(0.005)		(0.005)		(0.005)	(0.005)		(0.005)	(0.005)	(0.005)
	$\psi^{(B)}$			0.127***			0.126***		0.131***		0.127***	0.13***		0.126***	0.131***	0.13***
				(0.008)			(0.007)		(0.007		(0.007	(0.007)		(0.007)	(0.007)	(0.007)
	$\psi^{(S)}$			0.023***			0.025***		0.021***		0.027***	0.022***		0.031***	0.024***	0.028***
				(0.012)			(0.013)		(0.012		(0.018	(0.013)		(0.018)	(0.018)	(0.018)
V_R	Is buy				0.259***			-0.004***		0.006***	0.007***		-0.003***	0.01***	0.003***	0.005***
					(0.007)			(0.006)		(0.007)	(0.006)		(0.007)	(0.006)	(0.006)	(0.006)
	Is day trade				0.005***			0.004***		0.006***	-0.001***		0.005***	-0.002***	-0.001**	-0.002***
					(0.006)			(0.007)		(0.006)	(0.009)		(0.007)	(0.009)	(0.009)	(0.009)
R-squared		0.013***	0.012***	0.075***	0.003***	0.025***	0.087***	0.016***	0.088***	0.015***	0.078***	0.099***	0.028***	0.089***	0.09***	0.102***
		(0.004)	(0.003)	(0.006)	(0.002)	(0.005)	(0.006)	(0.004)	(0.007)	(0.003)	(0.006)	(0.007)	(0.005)	(0.006)	(0.007)	(0.007)
No. Observations		601	601	601	601	601	601	601	601	601	601	601	601	601	601	601
		(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)

Table 5.5: Multivariate linear regression models for predicting the magnitude of reaction to a given sentiment event. All symbols and variables are defined in Section 4.4. Note that numbers in parentheses are the standard deviation of coefficients calculated using Equation 4.14. * means significance at the 10% level, ** means significance at the 5% level, and *** means significance at the 1% level. We consider coefficients with a standard deviation of zero to be significant at the 1% level. Models are named according to Table 4.1.

5.4.2 Post-Crisis Model for All Trades

We derive logistic regression models on the sub-sample of post-financial crisis data-points. In this way, we may contribute models based on more recent and potentially more relevant data to real-world applications. We report these models in Table 5.6.

We observe generally the same trends and patterns across most variables as we do in Table 5.5 of Section 5.4.1. The R-squared values in the post-crisis models are consistently greater than those in the all-time models, and the model that contains all variable categories achieves a 0.135 R^2 value. That is, our most comprehensive model can explain 13.5% of the variance in the post-crisis data. The consistently larger R^2 values are unsurprising because there are less samples per Monte Carlo simulation in the post-crisis model.

Next, we analyze the coefficients on the individual variables across all models. First, we notice that in both the all-time and post-crisis models, the magnitudes of the coefficients on $\widetilde{\Delta}$ are the largest relative to others, and those on $\psi^{(B)}$ are the second largest. Thus, we conclude that $\widetilde{\Delta}$ has the strongest effect on reaction magnitude and is followed by $\psi^{(B)}$. That is, the score of the sentiment event with the largest value in the week preceding a reaction has the largest effect on the proportion of wealth traded in reaction to an event. Given an increase in this score, we can expect an increase in reaction magnitude. Similarly, the proportion of buys made on sentiment before the reaction has the second largest effect on the proportion of wealth traded in reaction to an event. Given a unit increase in this proportion, we expect an increase in reaction magnitude.

On the other hand, the indicator for day trading and $\psi^{(S)}$ exhibit somewhat dissimilar trends in the post-crisis models than they do in the all-time models. While both have coefficients that are consistently approximately zero in both types of models, they do not share the same signs. Specifically, for all models that do not involve the full set of variable categories, both the indicator for day trading and $\psi^{(S)}$ have slightly negative coefficients in the post-crisis models and slightly positive coefficients in the all-time models. However, because their coefficients are close to zero, they

have the same relative strength of predictability of reaction magnitude in both post-crisis and all-time data. In addition, the magnitudes on the coefficients of $\psi^{(S)}$ are consistently lower than those of $\psi^{(B)}$. That is, in both all-time and post-crisis data, previous sentiment buying history has a stronger effect on reaction magnitude than previous sentiment selling history does.

The remaining variables exhibit the same trends in both all-time and post-crisis models. The coefficients on ρ, ζ, and the indicator for a reaction being a buy are all close to zero and slightly positive. Given unit increases in any of these three variables, we expect to see a slight increase – if any – in proportion of wealth traded in reaction to an event. The demographic variables also exhibit similar behavior between the all-time and post-crisis models. They all have coefficients approximately equal to zero. The variables for age, having limited investment experience, having excellent investment knowledge, and having good investment knowledge have coefficients that are slightly positive. Meanwhile, the coefficients on the variables for number of dependents, being married, having excellent investment experience, and limited investment knowledge are slightly negative.

Finally, we discuss Model (O), which involves all feature categories. The variable with the strongest effect on reaction magnitude is $\widetilde{\Delta}$, which has the coefficient with the largest magnitude (0.275). If there is a unit increase in $\widetilde{\Delta}$ we expect a 0.275 unit increase in the proportion traded in reaction to a sentiment event. The variable with the next strongest effect if $\psi^{(B)}$ with a coefficient of 0.152. On the other hand, the coefficient of $\psi^{(S)}$ is 0.021, which is an order of magnitude less than that of $\psi^{(B)}$. Given a unit increase in $\psi^{(B)}$, we expect to see a 0.152 increase in reaction magnitude; whereas, we expect a 0.021, or close to zero, unit increase in reaction magnitude if there is a unit increase in $\psi^{(S)}$. Thus, we can conclude that that previous sentiment buying history is more predictive of current reaction magnitude than previous sentiment selling history is.

The coefficients on the remaining variables are all approximately zero. We observe that the coefficients on ρ and ζ are slightly positive (0.006 and 0.001 respectively) while those on the indicators for buying and day trading are slightly negative

(-0.022 and -0.051 respectively). Furthermore, the coefficient for age is slightly positive (0.009), while the coefficients for the number of dependents (zero dependents: -0.016, one to three dependents: -0.091) and being married (-0.009) are slightly negative. We also investigate the signs on the coefficients for the investment experience and knowledge variables. If an investor reports to have excellent or limited investment experience, we expect a 0.066 or 0.04 unit decrease in reaction magnitude. On the other hand, if the investor reports to have good investment experience, limited experience, excellent knowledge, or good knowledge, we expect a 0.034, 0.077, 0.04, or 0.024 unit increase in reaction magnitude.

		(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)	(N)	(O)
V_D	Age	0.324***				0.007***	0.255***	0.355***				0.007***	0.01***		0.282***	0.009***
		(0.032)				(0.008)	(0.033)	(0.033)				(0.008)	(0.008)		(0.034)	(0.008)
	0 dependents	-0.016***				-0.016***	-0.018***	-0.015***				-0.017***	-0.014***		-0.017***	-0.016***
		(0.004)				(0.004)	(0.004)	(0.004)				(0.004)	(0.004)		(0.004)	(0.004)
	1-3 dependents	-0.108***				-0.107***	-0.088***	-0.114***				-0.086***	-0.112***		-0.093***	-0.091***
		(0.03)				(0.031)	(0.031)	(0.03)				(0.032)	(0.031)		(0.031)	(0.032)
	Excellent investment experience	-0.074***				-0.073***	-0.06***	-0.083***				-0.058***	-0.082***		-0.067***	-0.066***
		(0.03)				(0.03)	(0.031)	(0.03)				(0.032)	(0.031)		(0.032)	(0.032)
	Good investment experience	0				0	0.036***	0.01***				0.036***	0.009***		0.035***	0.034***
		(0.016)				(0.016)	(0.016)	(0.016)				(0.016)	(0.016)		(0.016)	(0.016)
	Limited investment experience	0.057***				0.056***	0.077***	0.063***				0.077***	0.062***		0.077***	0.077***
		(0.009)				(0.009)	(0.009)	(0.009)				(0.009)	(0.009)		(0.009)	(0.009)
	Excellent investment knowledge	0.021***				0.02***	0.04***	0.027***				0.038***	0.026***		0.042***	0.04***
		(0.008)				(0.008)	(0.008)	(0.008)				(0.008)	(0.008)		(0.008)	(0.008)
	Good investment knowledge	0.028***				0.03***	0.018***	0.031***				0.02***	0.034***		0.022***	0.024***
		(0.014)				(0.014)	(0.014)	(0.014)				(0.014)	(0.014)		(0.014)	(0.014)
	Limited investment knowledge	-0.043***				-0.042***	-0.044***	-0.043***				-0.043***	-0.041***		-0.041***	-0.04***
		(0.01)				(0.011)	(0.01)	(0.01)				(0.011)	(0.011)		(0.01)	(0.011)
	Married	-0.01***				-0.01***	-0.008***	-0.012***				-0.008***	-0.012***		-0.01***	-0.009***
		(0.002)				(0.002)	(0.002)	(0.002)				(0.002)	(0.002)		(0.002)	(0.002)
V_E	$\widetilde{\Delta}$		0.232***			0.319***			0.195***	0.258***		0.249***	0.348***	0.216***		0.275***
			(0.005)			(0.034)			(0.005)	(0.007)		(0.035)	(0.035)	(0.008)		(0.036)
	ρ		0.005***			0.004***			0.007***	0.004***		0.006***	0.003***	0.006***		0.006***
			(0.002)			(0.002)			(0.002)	(0.002)		(0.002)	(0.002)	(0.001)		(0.001)
V_T	ζ			0.199***			0		0.01***		0.221***	0.002***		0.011***	-0.001***	0.001***
				(0.003)			(0.009)		(0.008)		(0.007)	(0.01)		(0.007)	(0.009)	(0.009)
	$\psi^{(B)}$			0.157***			0.159***		0.159***		0.148***	0.161***		0.15***	0.15***	0.152***
				(0.008)			(0.007)		(0.008)		(0.008)	(0.007)		(0.008)	(0.007)	(0.007)
	$\psi^{(S)}$			-0.111***			-0.153***		-0.113***		0.069***	-0.157***		0.065***	0.026***	0.021***
				(0.026)			(0.028)		(0.026)		(0.032)	(0.029)		(0.032)	(0.032)	(0.032)
V_R	Is buy				0.264***			0.001***		0.014***	-0.019***		0.003***	-0.02***	-0.02***	-0.022***
					(0.006)			(0.009)		(0.008)	(0.007)		(0.01)	(0.007)	(0.008)	(0.008)
	Is day trade				-0.039***			-0.045***		-0.039***	-0.051***		-0.045***	-0.051***	-0.052***	-0.051***
					(0.007)			(0.008)		(0.007)	(0.01)		(0.008)	(0.01)	(0.01)	(0.01)
R-squared		0.038***	0.003***	0.083***	0.017***	0.04***	0.119***	0.059***	0.087***	0.019***	0.095***	0.122***	0.06***	0.098***	0.132***	0.135***
		(0.006)	(0.002)	(0.007)	(0.005)	(0.006)	(0.01)	(0.009)	(0.007)	(0.005)	(0.008)	(0.01)	(0.009)	(0.008)	(0.011)	(0.011)
No. Observations		328	328	328	328	328	328	328	328	328	328	328	328	328	328	328
		(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)

Table 5.6: Multivariate linear regression models for predicting the magnitude of reaction to a given sentiment event post-financial crisis. All symbols and variables are defined in Section 4.4. Note that numbers in parentheses are the standard deviation of coefficients calculated using Equation 4.14. * means significance at the 10% level, ** means significance at the 5% level, and *** means significance at the 1% level. We consider coefficients with a standard deviation of zero to be significant at the 1% level. Models are named according to Table 4.1.

5.4.3 All-Time Model for Buys

We also run Monte Carlo simulations on data that comprises of only buys in order to isolate the effects of our variables on the magnitude of reaction when investors make sentiment buys. The models are reported in Table 5.7.

First, we notice that the R^2 values of the buy-only models are consistently greater than those of the models that are trained on both buy and sell data. Specifically, the model that contains all variable categories boasts an R^2 value of 0.197 which is almost double that of the model in Section 5.4.1. However, we believe that these higher R^2 values come from the fact that there are only 158 observations per Monte Carlo simulation in the buy-only models, which is more than a third less than the 601 observations modeled in Section 5.4.1. Thus, instead of analyzing the overall fit of each buy-only model, we focus our attention instead on the relative predictability of each variable.

We observe some similar trends and patterns in the buy-only models as those in the all-trades models of Section 5.4.1. For instance, we notice that the magnitude on the coefficients for marital status, number of dependents, age are essentially zero in both the buy-only and all-trades models. While the coefficients on each of these variables may not necessarily be consistent across all buy-only models, the standard deviations indicate that values within one standard deviation away can be both positive and negative. This effect is most notable in the coefficients on the indicators for investors' number of dependents. Additionally, some of the coefficients for age in the models in Table 5.7 are not always close to zero. In general, when V_E variables are not included in the model, the coefficients on the age variable are an order of magnitude larger than those of the models with V_E. We attribute this to omitted variable bias. That is, in the absence of V_E variables, age can explain away some of variance in the data. However, most of this variance is explained away by event-based features when they are added to the model. Thus, we conclude that both marital status and age have approximately no effect on the magnitude of sentiment reaction.

Next, we discuss the relative predictability of the indicators on investment expe-

rience and knowledge. These coefficients are similar in sign to those in the all-trades models, but differ in value by an order of magnitude. From Table 5.7, we observe that the coefficients on the indicator for having excellent and good experience are consistently slightly negative across all models. On the other hand, the coefficients for the indicator of having limited investment experience are generally slightly positive. However, there are cases where the signs on these coefficients are negative. In such cases, the magnitudes on the coefficients also tend to be smaller. Thus, we conclude that if an investor has good to excellent experience, we expect a lower proportion of wealth used to make buys compared to if the investor were to have no investment experience. On the other hand, if an investor has limited investment experience, we expect a higher proportion of wealth traded than if the investor had no investment experience. Furthermore, the coefficients on the indicators for investment knowledge are consistently close to zero. In most of the models, we observe slightly positive coefficients on the indicator for having good investment knowledge and slightly negative coefficients on the indicator for having limited investment knowledge. However, for all three investment knowledge indicators, values within one standard deviation of their coefficients are generally positive, negative, and close to zero. Thus, we conclude that investment knowledge has little to no effect on the magnitude of sentiment reaction.

We now analyze the coefficients for the variables in V_E, V_T, and V_R in Table 5.7. Though most trends are similar to those in Table 5.5, we notice that the variables with the strongest average effects on reaction magnitude are not all the same. First, we observe that the coefficients on ρ, $\psi^{(B)}$, and $\psi^{(S)}$ are consistently slightly positive across all models. This indicates that given an increase in the fraction of positive events in the week prior to a reaction or an increase in the proportion bought/sold on sentiment previously, we would expect to see some increase in the proportion of wealth used to make a buy. We conclude that these three variables have weak positive effects on reaction magnitude. We also observe that the predictive power of $\psi^{(B)}$ is stronger than that of $\psi^{(S)}$. In other words, sentiment buying history is more predictive of the buy reaction magnitude than sentiment selling history is. Furthermore, the coefficients on the indicator for reactions occurring before the financial crisis are

generally slightly negative, while those on the indicator for reactions occurring during the financial crisis are generally positive. This means that, if a reaction occurs before the crisis, we expect a decrease in the proportion of wealth used to make buys than if the reaction occurs after the crisis. On the other hand,if a reaction occurs during the crisis, we expect an increase in the proportion of wealth used to make buys than if the reaction occurs after the crisis.

In addition, the coefficients on the indicator for a reaction to consist of day trading are either slightly negative or relatively strongly positive, which significantly differs from the trend of approximately being zero across all models in Table 5.4.1. Specifically, the coefficients are positive in the models that also include V_T and negative otherwise. Thus, we believe that making a day trade is somewhat correlated with the variables in V_T and are thus jointly predictive of sentiment reaction magnitude. In fact, when the coefficients on day trading are positive, their magnitudes are consistently largest out of all the other coefficients. We conclude that, with the presence of V_T, day trading has a strong average effect on magnitude reaction. Furthermore, the magnitude of the coefficients on ζ are consistently positive across all models. When V_T is not included in the models, the coefficients on ζ are second largest out of all the other variable coefficients. Otherwise, they are third largest. Thus, we conclude that the number of events in the month leading up to a reaction have weak positive effects on sentiment reaction magnitude. From Table 5.7, we also observe that the coefficients on $\widetilde{\Delta}$ are consistently positive and generally have the largest magnitudes out of the other coefficients. Keeping all other variables constant, we expect a unit increase in $\widetilde{\Delta}$ to lead to an increase in the proportion of wealth used to make buys. Thus, we conclude that, sentiment in the week prior to a reaction has the strongest average effect on sentiment reaction magnitude when the reaction is a buy.

Finally, we analyze the comprehensive model that is comprised of all variable categories (Model (O)). First we note that the variables with the strongest average effects on sentiment magnitude reaction for buys are $\widetilde{\Delta}$, ζ, and the indicator for day trading. Specifically, given a unit increase in these coefficients, we expect to see a 0.327, 0.125, and 0.337 unit increase respectively in the proportion of wealth used to

make a buy in reaction to an event. In addition, both $\widetilde{\Delta}$ and ζ are variables that can have absolute values larger than one, while the indicator for day trading is constrained to being zero or one. Thus, we conclude that the relative predictability of $\widetilde{\Delta}$ and ζ are stronger than their coefficients may initially suggest.

On the other hand, the other coefficients in the model have coefficients closer to zero. The coefficients for the age and number of dependents are slightly negative while those on $\psi^{(B)}$ and $\psi^{(S)}$ are slightly positive. In addition, the coefficients for having excellent or good investment experience are slightly negative while those for having limited investment experience are slightly positive. The opposite is true for investment knowledge: The coefficients for having excellent or good investment knowledge are slightly positive while those for having limited investment knowledge are slightly negative. In other words, holding other variables constant, if an investor has excellent experience, good experience, or limited knowledge, we expect a slight decrease in the sentiment reaction magnitude compared to if the investor had no experience or knowledge; whereas, if an investor has excellent knowledge, good knowledge, or limited experience, we expect a slight increase. Finally, the coefficient on the indicator for a reaction occurring before the financial crisis is slightly negative while the one on the indicator for a reaction occurring after the financial crisis is slightly positive. That is, if a reaction occurs before the financial crisis, we expect a slight decrease in the proportion of wealth used to buy on sentiment than if the reaction occurs after or during the financial crisis. Similarly, if a reaction occurs during the financial crisis we expect a slight increase in sentiment buy reaction magnitude than if the reaction occurs before or after the crisis.

		(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)	(N)	(O)
V_D	Age	0.317***				-0.011***	0.292***	0.345***				-0.016***	-0.012***	0.33***		-0.018***
		(0.011)				(0.007)	(0.011)	(0.016)				(0.007)	(0.007)	(0.018)		(0.007)
	0 dependents	-0.011***				-0.008***	-0.023***	-0.01***				-0.018***	-0.006***	-0.02***		-0.016***
		(0.005)				(0.005)	(0.005)	(0.005)				(0.005)	(0.005)	(0.005)		(0.005)
	1-3 dependents	-0.062***				-0.058***	-0.098***	-0.056***				-0.093***	-0.054***	-0.09***		-0.088***
		(0.005)				(0.006)	(0.005)	(0.005)				(0.005)	(0.005)	(0.005)		(0.005)
	Excellent investment experience	-0.08***				-0.075***	-0.111***	-0.072***				-0.107***	-0.069***	-0.101***		-0.099***
		(0.005)				(0.006)	(0.005)	(0.005)				(0.006)	(0.005)	(0.005)		(0.006)
	Good investment experience	-0.086***				-0.104***	-0.032***	-0.085***				-0.06***	-0.103***	-0.035***		-0.062***
		(0.013)				(0.015)	(0.015)	(0.014)				(0.017)	(0.015)	(0.016)		(0.017)
	Limited investment experience	0.009***				-0.001	0.044***	0.005***				0.032***	-0.003***	0.038***		0.028***
		(0.013)				(0.014)	(0.013)	(0.013)				(0.014)	(0.014)	(0.013)		(0.014)
	Excellent investment knowledge	0.006***				0.004***	0.043***	-0.001				0.039***	-0.001**	0.035***		0.032***
		(0.011)				(0.011)	(0.01)	(0.011)				(0.01)	(0.011)	(0.011)		(0.011)
	Good investment knowledge	0.11***				0.116***	0.06***	0.108***				0.071***	0.115***	0.058***		0.07***
		(0.017)				(0.017)	(0.017)	(0.017)				(0.017)	(0.017)	(0.017)		(0.017)
	Limited investment knowledge	0.008***				0.013***	-0.039***	0.006***				-0.035***	0.011***	-0.042***		-0.037***
		(0.017)				(0.018)	(0.018)	(0.017)				(0.019)	(0.018)	(0.018)		(0.019)
	Married	0.001***				-0.001***	0.003***	0.001***				0.001***	-0.001***	0.003***		0.001***
		(0.002)				(0.002)	(0.002)	(0.002)				(0.002)	(0.002)	(0.002)		(0.002)
V_E	$\widetilde{\Delta}$		0.251***			0.319***			0.184***	0.277***		0.295***	0.339***		0.223***	0.327***
			(0.004)			(0.012)			(0.005)	(0.013)		(0.012)	(0.018)		(0.016)	(0.02)
	ρ		0.028***			0.028***			0.03***	0.027***		0.03***	0.027***		0.029***	0.029***
			(0.004)			(0.004)			(0.003)	(0.004)		(0.003)	(0.004)		(0.003)	(0.003)
V_T	Is before the financial crisis			0.182***			-0.063***		-0.01***		0.229***	-0.062***		-0.063***	-0.011***	-0.062***
				(0.003)			(0.014)		(0.007)		(0.015)	(0.014)		(0.014)	(0.006)	(0.014)
	Is during the financial crisis			0.019***			0.018***		0.024***		0.018***	0.023***		0.017***	0.023***	0.022***
				(0.006)			(0.007)		(0.006)		(0.006)	(0.007)		(0.007)	(0.006)	(0.007)
	ζ			0.122***			0.129***		0.121***		0.122***	0.126***		0.128***	0.121***	0.125***
				(0.007)			(0.007)		(0.007)		(0.007)	(0.007)		(0.008)	(0.007)	(0.008)
	$\psi^{(B)}$			0.061***			0.073***		0.076***		0.063***	0.089***		0.075***	0.078***	0.09***
				(0.014)			(0.014)		(0.014)		(0.013)	(0.015)		(0.014)	(0.014)	(0.014)
	$\psi^{(S)}$			0.076***			0.084***		0.06***		0.079***	0.069***		0.087***	0.063***	0.072***
				(0.02)			(0.02)		(0.02)		(0.019)	(0.021)		(0.019)	(0.019)	(0.02)
V_R	Is day trade				0.284***			-0.016***		-0.003***	0.413***		-0.016***	0.225***	0.479***	0.337***
					(0.012)			(0.014)		(0.006)	(0.072)		(0.014)	(0.084)	(0.07)	(0.085)
R-squared		0.026***	0.029***	0.124***	0.005***	0.054***	0.162***	0.03***	0.157***	0.032***	0.133***	0.192***	0.056***	0.168***	0.164***	0.197***
		(0.003)	(0.008)	(0.01)	(0.003)	(0.008)	(0.01)	(0.005)	(0.012)	(0.008)	(0.01)	(0.011)	(0.009)	(0.01)	(0.011)	(0.011)
No. Observations		158	158	158	158	158	158	158	158	158	158	158	158	158	158	158
		(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)

Table 5.7: Multivariate linear regression models for predicting the magnitude of reaction to a given sentiment event if the reaction is a buy. All symbols and variables are defined in Section 4.4. Note that numbers in parentheses are the standard deviation of coefficients calculated using Equation 4.14. * means significance at the 10% level, ** means significance at the 5% level, and *** means significance at the 1% level. We consider coefficients with a standard deviation of zero to be significant at the 1% level. Models are named according to Table 4.1.

5.4.4 All-Time Model for Sells

Next, we derive linear regression models on the sub-sample of data that comprises of only sells in order to isolate the effects of our variables on sell reaction magnitude when investors make sells. We report the models in Table 5.8, and compare the results with those in Table 5.7 of Section 5.4.3.

The trends in the sell-only model are similar to those in the buy-only model with a couple of noticeable differences. In general, the sell-only trends more closely resemble those of the models with both buys and sells. This is unsurprising because there are more sells than buys in the data. We also observe that the R^2 values of the sell-only models are consistently less than those of the buy-only models but are consistently greater than those of the models that contain both buys and sells. However, we believe that this trend comes from the fact that there are 520 observations in the sell-only models, 158 in the buy-only models, and 601 in the models that comprise of both. The more observations there are, the more variance there is to be explained by model coefficients.

First, we begin our discussion with variables that have similar coefficients in both the buy-only and sell-only models. The coefficients on $\widetilde{\Delta}$ and ρ are positive and have similar magnitudes across both model types. We can thus draw the same conclusions about both the effects of these two variables as we do in Section 5.4.3. Specifically, $\widetilde{\Delta}$ has a relatively strong effect and ρ has a relatively weak effect on reaction magnitude in both the buy-only and sell-only samples.

Furthermore, the coefficients on the demographic variables of the sell-only models are consistently close to zero, which we also observe to be the case in the buy-only models. In addition, the signs on the coefficients for age, number of dependents, having excellent investment experience, having limited investment experience, and having excellent investment knowledge are consistent between the two. That is, these variables have the same direction of effect (i.e. positive or negative) on buy reactions as they do on sell reactions. However, this is not the case for the other demographic variables. The coefficients for the indicators of having good investment experience and

limited investment knowledge are slightly negative in the buy-only models, but are slightly positive in the sell-only models. Similarly, the coefficients for the indicator of having good investment knowledge are slightly positive in the buy-only models and slightly negative in the sell-only ones. The coefficient signs on the indicator for being married are inconsistent across each model for both the buy-only and sell-only simulations. However, the coefficients for marital status are negative more often in the sell-only ones. In addition, while the signs on the coefficients for the pre-crisis indicator are similar for both buys and sells, they are different for the during-crisis indicator. Specifically, the indicator for a reaction occurring during the financial crisis has a consistently slightly positive effect in the buy-only models but a consistently slightly negative effect in the sell-only models.

We also observe variables that share the same signs as those in the buy-only models, but have significantly different relative magnitudes. The coefficients on ζ and the indicator for day trading exhibit this trend. In the buy-only models, these two variables have coefficients that are positive and generally greater than 0.1, which are relatively large values compared to other coefficients. However, both these variables lose an order of magnitude in the sell-only models such that they have slightly positive coefficients that are close to zero.

In addition, $\psi^{(B)}$ and $\psi^{(S)}$ exhibit the same pattern. Their coefficients are slightly positive but close to zero in the buy-only models. Furthermore, the coefficients on $\psi^{(B)}$ are generally greater than those on $\psi^{(S)}$. On the other hand, in the sell models, the magnitudes on the coefficients for both variables are positive and an order of magnitude larger than those in the buy models. Furthermore, the coefficients on $\psi^{(S)}$ are consistently greater than those on $\psi^{(B)}$. Thus, we conclude that previous sentiment buying and selling behavior are more predictive of reaction magnitude when the investor is making a sell versus a buy. In addition, past sentiment selling behavior is more predictive of current sentiment selling behavior, while past sentiment buying behavior is more predictive of current sentiment buying behavior.

Finally, we discuss the main contribution of this section, which is the model that involves all variable categories (Model (O)). Most of the coefficients on the variables

are close to zero. The ones that are furthest away from zero are those on Δ (0.23), $\psi^{(B)}$ (0.153), and $\psi^{(S)}$ (2.274). This suggests that these three variables have the strongest effects on reaction magnitude for sells relative to the other model features. Furthermore, this implies that previous sentiment selling history has the largest positive effect on the proportion of wealth sold in reaction to an event. Specifically, if there is a unit increase in $\psi^{(S)}$, we expect a 2.274 unit increase in the sell reaction magnitude.

Similarly, if there is a unit increase in $\widetilde{\Delta}$ or $\psi^{(B)}$, we expect a 0.23 or 0.153 unit increase in proportion of wealth sold. In addition, given that $\psi^{(B)}$ and $\psi^{(S)}$ are constrained to be between zero and one but $\widetilde{\Delta}$ can take on values greater than one, we can infer that the relative predictive power of $\widetilde{\Delta}$. That is, if $\psi^{(B)}$ also had a coefficient of 0.23 instead of 0.153, we would expect a larger change in reaction magnitude given a change in $\widetilde{\Delta}$ than in $\psi^{(B)}$. Thus, the sentiment score of the event with the largest magnitude in the week prior to a reaction has a relatively strong effect on reaction magnitude than $\psi^{(B)}$ does.

The coefficients on the remaining variables are close to zero. Specifically, the coefficients on age, ρ, and ζ are slightly positive while those on the indicators for having one to three dependents, being married, occurring before the crisis, and occurring during the crisis are slightly negative. In addition, the coefficients on having good or limited investment experience are slightly positive; whereas, those on having good or limited investment knowledge are slightly negative. The opposite is true for having excellent investment experience and knowledge. That is, the coefficient on having excellent investment experience is slightly negative, and the coefficient on having excellent investment knowledge is slightly positive. Thus, if an investor has good experience, limited experience, or excellent knowledge, expect to see a slight increase in sell reaction magnitude. On the other hand, if an investor has good knowledge, limited knowledge, or excellent experience, we expect to see a slight decrease in sell reaction magnitude.

		(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)	(N)	(O)
V_D	Age	0.291***				0.01***	0.223***	0.302***				0.011***	0.011***		0.237***	0.012***
		(0.021)				(0.006)	(0.022)	(0.022)				(0.006)	(0.006)		(0.023)	(0.006)
	0 dependents	0***				0.001***	-0.002***	0.001***				-0.001***	0.001***		-0.001***	0***
		(0.003)				(0.003)	(0.003)	(0.003)				(0.003)	(0.003)		(0.003)	(0.003)
	1-3 dependents	-0.07***				-0.068***	-0.048***	-0.07***				-0.046***	-0.068***		-0.047***	-0.045***
		(0.02)				(0.02)	(0.02)	(0.02)				(0.02)	(0.021)		(0.02)	(0.021)
	Excellent investment	-0.063***				-0.063***	-0.045***	-0.062***				-0.045***	-0.062***		-0.044***	-0.044***
	experience	(0.02)				(0.02)	(0.02)	(0.02)				(0.02)	(0.02)		(0.02)	(0.02)
	Good investment	0.025***				0.023***	0.061***	0.025***				0.057***	0.023***		0.061***	0.057***
	experience	(0.014)				(0.014)	(0.015)	(0.014)				(0.014)	(0.014)		(0.015)	(0.014)
	Limited investment	0.009***				0.009***	0.032***	0.008***				0.033***	0.009***		0.032***	0.033***
	experience	(0.007)				(0.007)	(0.007)	(0.007)				(0.007)	(0.007)		(0.007)	(0.007)
	Excellent investment	0.025***				0.024***	0.037***	0.025***				0.036***	0.024***		0.037***	0.036***
	knowledge	(0.006)				(0.006)	(0.006)	(0.006)				(0.006)	(0.006)		(0.006)	(0.006)
	Good investment	0.009***				0.013***	-0.015***	0.01***				-0.01***	0.014***		-0.014***	-0.009***
	knowledge	(0.011)				(0.011)	(0.011)	(0.011)				(0.011)	(0.011)		(0.011)	(0.011)
	Limited investment	0***				0.001***	-0.007***	0.001***				-0.005***	0.002***		-0.006***	-0.004***
	knowledge	(0.007)				(0.007)	(0.007)	(0.007)				(0.007)	(0.007)		(0.007)	(0.007)
	Married	-0.006***				-0.006***	-0.007***	-0.006***				-0.007***	-0.007***		-0.007***	-0.007***
		(0.001)				(0.001)	(0.001)	(0.001)				(0.001)	(0.001)		(0.001)	(0.001)
V_E	$\widetilde{\Delta}$		0.236***			0.285***			0.196***	0.249***		0.216***	0.296***	0.211***		0.23***
			(0.004)			(0.023)			(0.004)	(0.007)		(0.023)	(0.024)	(0.008)		(0.024)
	ρ		0.011***			0.01***			0.011***	0.011***		0.011***	0.01***	0.011***		0.011***
			(0.002)			(0.002)			(0.002)	(0.002)		(0.002)	(0.002)	(0.002)		(0.002)
V_T	Is before the			0.202***			-0.004***		0.013***		0.217***	-0.004***		0.013***	-0.004***	-0.003***
	financial crisis			(0.003)			(0.007)		(0.006)		(0.007)	(0.007)		(0.006)	(0.007)	(0.007)
	Is during the			-0.025***			-0.029***		-0.023***		-0.025***	-0.027***		-0.023***	-0.029***	-0.027***
	financial crisis			(0.004)			(0.005)		(0.005)		(0.004)	(0.005)		(0.005)	(0.005)	(0.005)
	ζ			0.008***			0.005***		0.009***		0.008***	0.006***		0.009***	0.005***	0.006***
				(0.004)			(0.004)		(0.004)		(0.004)	(0.004)		(0.004)	(0.004)	(0.004)
	$\psi^{(B)}$			0.148***			0.15***		0.151***		0.148***	0.153***		0.151***	0.15***	0.153***
				(0.006)			(0.006)		(0.006)		(0.006)	(0.006)		(0.006)	(0.006)	(0.006)
	$\psi^{(S)}$			2.038***			2.037***		2.138***		2.165***	2.127***		2.272***	2.175***	2.274***
				(0.062)			(0.077)		(0.064)		(0.076)	(0.078)		(0.072)	(0.089)	(0.085)
V_R	Is day trade				0.254***			0**		0.012***	-0.003***		0	-0.007***	-0.003***	-0.007***
					(0.006)			(0.007)		(0.006)	(0.005)		(0.007)	(0.005)	(0.005)	(0.005)
R-squared		0.014***	0.007***	0.098***	0.001***	0.02***	0.111***	0.015***	0.106***	0.008***	0.1***	0.119***	0.021***	0.108***	0.113***	0.121***
		(0.003)	(0.002)	(0.006)	(0.001)	(0.003)	(0.007)	(0.003)	(0.007)	(0.002)	(0.006)	(0.007)	(0.004)	(0.006)	(0.007)	(0.007)

Table 5.8: Multivariate linear regression models for predicting the magnitude of reaction to a given sentiment event if the reaction is a sell. All symbols and variables are defined in Section 4.4. Note that numbers in parentheses are the standard deviation of coefficients calculated using Equation 4.14. * means significance at the 10% level, ** means significance at the 5% level, and *** means significance at the 1% level. We consider coefficients with a standard deviation of zero to be significant at the 1% level. Models are named according to Table 4.1.

5.5 Direction of Reaction Model

Conditioning on an sentiment investor reacting to an event, we predict the direction of that reaction, i.e. whether that investor will make a buy or a sell. In Section 5.5.1, we introduce the models we derive from running Monte Carlo simulations on sentiment trades from all time. In Section 5.5.2, we discuss the models that we train specifically on post-financial crisis sentiment trades. Note that buys are labeled as ones and sells are labeled as zeros in our models. Thus, coefficients that are positive indicate that increases in corresponding variable implies an increase in the log-odds of buying to selling.

5.5.1 All-Time Model

We report the models we derive from running Monte Carlo simulation on all sentiment trades for all time in Table 5.9. Again, given the granularity of our data, we expect a large amount of noise to construe our results. Unlike the previous model results that we compute across all time (Sections 5.3.1 and 5.4.1), the pseudo R-squared that we achieve in predicting direction reaction is relatively lower. The models that contain two or more variable categories (with the exception of Model (E)) have R^2 values that range from 0.05-0.055. Furthermore, almost all of the coefficients across all the models are significant at the 1% level. Again, we attribute this large proportion of significant coefficients to the large number of Monte Carlo simulations that we run.

We provide a similar discussion about our model results as we do in Section 5.3.1. In Table 5.9, the first three columns are the baseline models (Models (A)-(C)), the following three columns are models that involve any two of the variable categories, and the last column is the final model that involves all three variable categories. Again, we observe an additive effect across the pseudo R^2 values. That is, the R^2 values of models that combine different variable categories are approximately equal to the sum of the R^2 values of the individual baseline models of the selected categories. Thus, we conclude that V_D, V_E, and V_T are largely orthogonal to each other with respect to explaining the variance in the sentiment trade data.

Next, we discuss the effect of individual variables on the direction of event reaction. We start with the variables in V_D, and observe that the coefficient on being married has the largest magnitude and thus has the largest effect on whether a sentiment investor makes a buy or a sell compared to the other variables in the baseline demographic model. The variable with the next largest magnitude is the indicator on whether the investor reported to have good investment knowledge. We observe that the magnitudes of both these variable coefficients are consistently the two largest across all demographic variables in all models in Table 5.9. Thus, we conclude that marital status and reporting to have good investment knowledge have the strongest demographic effect on reaction direction.

Furthermore, we note that the coefficients for the indicators for number of dependents change sign across the different models. This is unsurprising because, in all the models, the magnitudes of these coefficients are consistently close to zero and values within one standard deviation of each coefficient can have different signs. In addition, the coefficients on age have similar magnitudes and are consistently positive. As a result, we conclude that age and number of dependents do not have strong effects on reaction direction.

We also discuss the indicators for reported investment experience and knowledge. First, we note that the coefficients on the indicators for excellent and limited investment experience and excellent knowledge do not have consistent signs across all models. We attribute this phenomenon to the fact that the coefficients have small magnitudes and values within one standard deviation of each coefficient across all models can have different signs. Thus, it is unsurprising to see the coefficient means have inconsistent signs. On the other hand, the coefficients on the indicators for having good investment experience and good and limited investment knowledge are consistently positive and relatively larger in magnitude. Thus, if an investor reports to have good investment experience, good investment knowledge, or limited investment knowledge, we expect to see an decrease in log-odds of buying compared to if the investor reports to have no investment experience or knowledge. Furthermore, we conclude that these indicators have relatively strong effects on reaction direction.

We now shift our attention to the variables in V_E variables. First, we note that the coefficients on ρ, which is the fraction of positive events in the seven days prior to the current data-points, are consistently negative and range from -0.105 to -0.133 across the different models. This implies that if there is a one unit increase in ρ, then we expect to see 0.105-0.133 unit decrease in log-odds of buying to selling. That is, a unit increase in ρ corresponds with a less than 0.105-0.133 unit decrease in the probability of buying as opposed to selling. On the other hand, the magnitudes and signs of $\widetilde{\Delta}$ across all models are inconsistent. In Models (B) and (H), $\widetilde{\Delta}$ has coefficients are -0.256 and -1.343 respectively. On the other hand, the coefficients are 0.012 and 0.103 in Models (E) and (K) respectively. We attribute this to the possible correlation between $\widetilde{\Delta}$ and other variables introduced into the regression. Furthermore, the models with positive coefficients for $\widetilde{\Delta}$ have standard deviations such that values within one standard deviation of the coefficient could be negative. Thus, we cannot draw any immediate conclusions about how the probability of buying as opposed to selling is affected by the event score with the largest score magnitude in the seven days prior to a given reaction.

Next, we analyze the variables in V_T. We note that the signs and magnitudes on the coefficients for the indicators on the reaction occurring before and during the financial crisis are consistent across all models. Specifically, the coefficients for the indicator of occurring before the financial crisis range from 2.38-2.46 and the coefficients for the indicator of occurring during the financial crisis range from -0.156 to -0.13. The signs and relative magnitudes of these coefficients persist for values within one standard deviation of the coefficients. Thus, we can conclude the following: If a reaction occurs before the financial crisis, we expect a 2.38-2.46 unit increase in the log-odds of buying versus selling compared to if the reaction occurs after the financial crisis. If a reaction occurs during the financial crisis, we can expect a 0.13-0.156 unit decrease in the log-odds of buying versus selling compared to if the reaction occurs after the crisis.

We also discuss ψ, which is the variable for the proportion of all volume traded on sentiment by the investor up until the current data-point. From Table 5.9, we

see that the magnitude and sign of the coefficients for ψ are consistent across all models, and range from -1.216 to -1.195. Again, we see that even when taking values one standard deviation away from each coefficient on ψ, the signs of the coefficients are still negative and the absolute value of the coefficients are still greater than one. Thus, holding other variables constant, we can expect that the log-odds of buying to selling decreasing by a little over one unit given a unit increase in ψ. That is, if an investor has a history of trading a larger proportion of volume on sentiment, then the probability that they buy in reaction to an event is lower compared to the probability that they sell. This finding is especially interesting since, as mentioned in Section 5.4.1, the pre-requisite of selling in reaction to an event involves owning some shares of the CUSIP before the event. On the other hand, there is no pre-requisite to buying in reaction to an event. Thus, there may be some inherent selection bias in the data.

The last variable to analyze in V_T is ζ, which is the proportion of volume traded of the given CUSIP across an investor's trading history up until the data-point. While the signs on the coefficients on ζ across all the models are consistently negative, values one standard deviation from the mean can have different signs. Furthermore, the magnitudes of the coefficients in the non-baseline models range from -0.056-0.104 and are thus close to zero in value. Thus, we conclude that the effect of ζ on the log-odds of buying to selling is slightly negative to zero.

Finally, we discuss our main contribution of this section, which is the model that comprises of all categories of variables (Model (K)). The investor's age and reporting of having limited investment experience have slightly positive (0.05 and 0.067 respectively) to zero effect on the log-odds of buying to selling. On the other handing, having three or less dependents (coefficients: -0.024 and -0.035), reporting to have excellent investment knowledge (-0.014), and the proportion traded of the CUSIP by an investor before the current data-point (-0.056) have slightly negative to no effect on the log-odds of buying to selling. Meanwhile, the coefficients for the following variables are ten times greater: indicators for reporting to have good experience (0.14), good knowledge (0.351), limited knowledge (0.207), and excellent experience (-0.168); the features for the fraction of positive events (-0.105) and the

event score with the largest magnitude (0.103); and the indicator for the data-point occurring during the financial crisis (-0.156). Finally, the coefficients with the largest magnitude in this model are the investor's marital status (-1.481) and history of sentiment trading (-1.195) as well as the indicator for the reaction occurring before the financial crisis (2.422). Thus, we conclude that marital status, sentiment trading history, and an reaction occurring before the financial crisis have the strongest effects on the direction of an event reaction.

		(A)	(B)	(C)	(E)	(F)	(H)	(K)
V_D	Age	0.036***			0.033***	0.056***		0.05***
		(0.071)			(0.069)	(0.071)		(0.071)
	0 dependents	0.008***			0.013***	-0.028***		-0.024***
		(0.03)			(0.03)	(0.034)		(0.033)
	1-3 dependents	0.124***			0.122***	-0.002		-0.035
		(0.16)			(0.163)	(0.182)		(0.174)
	Excellent investment	0.026***			0.034***	-0.142***		-0.168
	experience	(0.168)			(0.17)	(0.183)		(0.177)
	Good investment	0.134***			0.142***	0.138***		0.14***
	experience	(0.276)			(0.261)	(0.291)		(0.288)
	Limited investment	-0.005			-0.005	0.064***		0.067***
	experience	(0.153)			(0.147)	(0.171)		(0.17)
	Excellent investment	0.004			-0.001	-0.016***		-0.014**
	knowledge	(0.13)			(0.126)	(0.141)		(0.141)
	Good investment	0.284***			0.247***	0.361***		0.351***
	knowledge	(0.272)			(0.269)	(0.29)		(0.286)
	Limited investment	0.178***			0.149***	0.216***		0.207***
	knowledge	(0.153)			(0.153)	(0.174)		(0.173)
	Married	-0.471***			-0.428***	-1.57***		-1.481***
		(0.21)			(0.233)	(0.255)		(0.254)
V_E	$\widetilde{\Delta}$		-0.256***		0.012**		-1.343***	0.103***
			(0.045)		(0.126)		(0.102)	(0.148)
	ρ		-0.133***		-0.129***		-0.11***	-0.105***
			(0.037)		(0.037)		(0.037)	(0.038)
V_T	Is before the			2.408***		2.46***	2.38***	2.422***
	financial crisis			(0.179)		(0.178)	(0.177)	(0.178)
	Is during the			-0.155***		-0.13***	-0.141***	-0.156***
	financial crisis			(0.08)		(0.087)	(0.084)	(0.085)
	ζ			-1.393***		0.104***	-0.049***	-0.056***
				(0.092)		(0.15)	(0.109)	(0.113)
	ψ			-1.21***		-1.216***	-1.199***	-1.195***
				(0.143)		(0.145)	(0.139)	(0.145)
Pseudo R-Squared		0.004***	0.003***	0.048***	0.007***	0.053***	0.05***	0.055***
		(0.002)	(0.002)	(0.006)	(0.002)	(0.006)	(0.006)	(0.006)
No. Observations		1095	1095	1095	1095	1095	1095	1095
		(0)	(0)	(0)	(0)	(0)	(0)	(0)

Table 5.9: Multivariate logistic regression models for predicting the direction of reaction to a given sentiment event. All symbols and variables are defined in Section 4.4. Note that numbers in parentheses are the standard deviation of coefficients calculated using Equation 4.14. * means significance at the 10% level, ** means significance at the 5% level, and *** means significance at the 1% level. We consider coefficients with a standard deviation of zero to be significant at the 1% level. Models are named according to Table 4.1.

5.5.2 Post-Crisis Model

In addition, we derive logistic regression models on the sub-sample of data with reactions occurring after the financial crisis. We report these models in Table 5.10.

In general, we see similar trends across all variables as those in Table 5.9 of Section 5.5.1, and the magnitudes of the coefficients are generally larger in the post-crisis models than they are in the all-time models. Furthermore, the pseudo R-squared values in the post-crisis models are consistently greater than their all-time counterparts. Notably, the model that contains all categories of variables achieves a 0.101 pseudo R-squared value, which is double that of the comprehensive all-time model. This increase in pseudo R-squared may come from the smaller number of data-points in the post-crisis model (post-crisis: 672 observations, all-time: 1,095 observations). With less data-points, there is generally less variance to be explained. Nevertheless, we observe a doubling in pseudo R-squared given a less than 50% decrease in the number of observations per Monte Carlo simulation as well as the removal of two relatively predictive features (the indicators for the data-point occurring before and during the crisis). Thus, we can conclude that our post-crisis model fits post-crisis data relatively more than the all-time model fits data from all time.

Now, we investigate the trends in the coefficients of the variables. First, we notice that the signs and magnitudes of the coefficients on the indicators for the number of dependents, the proportion traded of a given CUSIP previously, and the sentiment score of the event with the largest magnitude in the week leading up to the event are similar to those in Table 5.9. Specifically, the number of dependents have a close to zero effect and the proportion previously traded has a slightly negative to zero effect on the probability of reacting to an event. In addition, the signs and magnitudes of the coefficients on Δ are not consistent across all models, but they match those in Table 5.9. Again, we attribute these inconsistencies with the correlations between $\widetilde{\Delta}$ and other variables introduced to each model.

From Table 5.10, we also observe that the coefficients on the indicator for marital status and the feature for the proportion of volume traded on sentiment previously

share the same signs as their counterparts in the all-time models. In addition, the magnitudes on these coefficients are consistently larger in the post-crisis models than the all-time ones. This implies that these two variables explain more of the variance and thus have stronger effect on reaction direction in the post-crisis data. Specifically, if an investor is married, we expect to see a 2.334 unit decrease in log-odds of buying to selling. That is, if an investor is married, we expect to see a less than 2.334 unit decrease in the probability of buying compared to the probability of selling. Furthermore, given a unit increase in the proportion of volume traded on sentiment previously, we expect to see a 1.957 unit decrease in the log-odds of buying to selling, or a less than 1.957 unit decrease in the probability of buying compared to the probability of selling.

Next, we analyze the indicators for self-reported levels of investment experience and knowledge. Across all post-crisis models, the magnitudes of the coefficients on these indicators are either similar or consistently larger than those in the all-time models. Specifically, the indicators for excellent investment experience, limited investment experience, and good investment knowledge share similar coefficients. Meanwhile, the coefficients on the indicators for good investment experience, excellent investment knowledge, and limited investment knowledge are consistently greater in magnitude and thus explain more of the variance in the post-crisis models than that of the all-time models. We can also make similar statistical interpretations of the coefficients for investment experience and knowledge as we do in Section 5.5.1. If an investor has excellent experience, we expect a slightly negative decrease in the log-odds of buying to selling than if the investor were to have no investment experience. On the other hand, we do not expect to observe a significant change in log-odds of buying if the investor were to have limited experience as opposed to no experience. However, if the investor has good investment experience or limited to excellent investment knowledge, we expect a positive increase in log-odds of buying to selling in reaction to an event compared to if the investor were to have no investment experience or knowledge.

Finally, we discuss Model (K) which is comprised of all variables. First, we observe that the number of dependents and the magnitude of the event with the largest

absolute value in the week prior to a reaction have approximately no effect on the probability of buying versus selling. On the other hand, age has a slightly positive coefficient (0.116). Noting that age is one of a few variables that can take negative values and values greater than one, the coefficient indicates that age is relatively predictive of the odds of buying, especially given that variables with smaller absolute values have even smaller coefficients.

Meanwhile, the coefficients on ρ and ζ indicate that the fraction of positive events in the week prior to the reaction and the proportion traded of the CUSIP previously have negative effects on the probability of buying. Specifically, if there is a unit increase in ρ or ζ, we expect a 0.167 or 0.128 unit decrease in the log-odds of buying to selling. We observe that ψ, which is the proportion traded on sentiment previously, has significantly stronger negative effect. If there is a unit increase in ψ, we expect a 1.957 unit decrease in the log-odds of buying to selling. That is, we expect a less than 1.957 unit decrease in the probability of buying versus selling given a unit increase in ψ. Thus, we conclude that an investor's history of sentiment trading has a stronger negative effect on reaction direction than the investor's general CUSIP-specific trading history and the positivity of the events prior to the reaction do.

Furthermore, we analyze the predictive power of investment knowledge and experience on reaction direction after the financial crisis. We observe that, if an investor has excellent investment experience, we can expect a 0.245 unit decrease in log-odds of buying to selling than if the investor were to have no investment experience. On the other hand, if the investor has good investment experience, we expect a 0.539 unit increase in the log-odds of buying to selling than if the investor had no investment experience. However, having limited investment experience has little to no effect on the probability of buying versus selling compared to having no investment experience. Furthermore, if the investor reports to have limited, good, or excellent investment knowledge, we expect to see a 0.370, 0.317, and 0.178 unit increase in log-odds of buying to selling than if the investor had no investment knowledge. In other words, holding all other variables constant, an investor who reports to have limited to excellent investment knowledge is more likely to buy in reaction to an event than sell

compared to an investor who reports to have no investment experience.

		(A)	(B)	(C)	(E)	(F)	(H)	(K)
V_D	Age	0.007**			-0.005*	0.123***		0.116***
		(0.082)			(0.084)	(0.093)		(0.089)
	0 dependents	0.035***			0.038***	-0.026***		-0.026***
		(0.041)			(0.042)	(0.046)		(0.045)
	1-3 dependents	0.140***			0.126***	-0.040***		0.039***
		(0.148)			(0.152)	(0.341)		(0.166)
	Excellent investment	-0.053***			-0.050***	-0.331***		-0.245***
	experience	(0.162)			(0.165)	(0.333)		(0.180)
	Good investment	0.465***			0.468***	0.476***		0.539***
	experience	(0.330)			(0.333)	(0.405)		(0.366)
	Limited investment	0.002			0.005	0.032***		0.098***
	experience	(0.191)			(0.184)	(0.242)		(0.202)
	Excellent investment	0.226***			0.220***	0.133***		0.178***
	knowledge	(0.151)			(0.145)	(0.189)		(0.170)
	Good investment	0.253***			0.233***	0.358***		0.317***
	knowledge	(0.333)			(0.343)	(0.404)		(0.376)
	Limited investment	0.308***			0.303***	0.395***		0.370***
	knowledge	(0.173)			(0.185)	(0.251)		(0.217)
	Married	-0.542***			-0.489***	-2.308***		-2.334***
		(0.225)			(0.253)	(0.475)		(0.274)
V_E	$\widetilde{\Delta}$		-0.200***		0.076***		-1.913***	0.069***
			(0.073)		(0.148)		(0.171)	(0.189)
	ρ		-0.238***		-0.235***		-0.170***	-0.167***
			(0.051)		(0.051)		(0.052)	(0.053)
	ζ			-2.034***		0.065***	-0.183***	-0.128***
				(0.147)		(0.203)	(0.185)	(0.173)
	ψ			-1.981***		-1.986***	-1.943***	-1.957***
				(0.224)		(0.231)	(0.228)	(0.240)
Pseudo R-Squared		0.011***	0.011***	0.085***	0.021***	0.096***	0.090***	0.101***
		(0.003)	(0.004)	(0.011)	(0.005)	(0.011)	(0.011)	(0.011)
No. Observations		672	672	672	672	672	672	672
		(0)	(0)	(0)	(0)	(0)	(0)	(0)

Table 5.10: Multivariate logistic regression models for predicting the direction of reaction to a given sentiment event post-financial crisis. All symbols and variables are defined in Section 4.5. Note that numbers in parentheses are the standard deviation of coefficients calculated using Equation 4.14. * means significance at the 10% level, ** means significance at the 5% level, and *** means significance at the 1% level. We consider coefficients with a standard deviation of zero to be significant at the 1% level. Models are named according to Table 4.1.

Chapter 6

Conclusion

In this study, we develop a scalable computational approach to investigate financial decision-making of individual sentiment investors. Specifically, we focus on investigating sentiment-based investment strategies and predicting trading behavior. In addition, we derive a robust sentiment investor classification system and produce predictive models that are not only able to fit our data well but also easily interpretable and thus more directly applicable in real-world financial applications.

In this chapter, we summarize our key results and contributions (Section 6.1). Then, we identify areas of future work to further investigate and better understand sentiment-based financial decision-making (Section 6.2).

6.1 Key Results and Contributions

Given the enormity of the data, one of our contributions is developing a correct and efficient approach to process and merge the brokerage and Ravenpack datasets. Furthermore, we develop a novel, robust sentiment investor identification mechanism to discover all brokerage accounts that exhibit sentiment trading behavior. This mechanimsm involves computing the difference between the probability that an investor trades a certain CUSIP in the days following sentiment events and the probability than the investor trades the CUSIP during other times. For each investor i, we take the trade-volume weighted average of these differences across all CUSIPs traded

by i, and refer to the final result as p_i. If we observe at least one sentiment trade, $p_i 0$, and only one investor managing the investment account, then i is an individual sentiment investor.

Using this sentiment investor identification mechanism, we identify a total of 64,214 accounts, which constitute 9.83% of all accounts, that exhibit sentiment trading behavior. Of the 64,214 sentiment accounts, 55,204 are associated with 40,148 household accounts and 9,010 are individual accounts. In this study, we focus on the sample of 9,010 individual sentiment accounts and 54,804 individual non-sentiment accounts. That is, 14.1% of our sample exhibits sentiment trading behavior.

We also mention some key points from our investigation of the demographic information of our sample. We observe that sentiment investors are more likely to report that they have excellent or good investment knowledge and experience. They are marginally more likely to report limited investment knowledge and experience and are less likely to report to have no investment knowledge and experience. Sentiment investors are also more likely to be financial professionals, executives and managers, business owners, real estate workers, and retired and less likely to be students, in the police or military, social workers, or disabled

Given investors' demographic information, investment history, and sentiment event data, we also derive three models that can predict whether a sentiment investor will react to a sentiment event, the magnitude of the reaction, and the direction of the reaction. First, we summarize our main conclusions from the logistic regression models that predict whether a sentiment investor will react to a given sentiment event. Our most comprehensive model achieves a 0.369 pseudo R^2 value. Noting that pseudo R^2 values ranging 0.2 to 0.4 represent excellent model fit [7], we conclude that our models have strong predictive power of whether a sentiment investor will react to a given event. We observe that age, number of dependents, and the proportion of volume previously traded by the investor on the given CUSIP have close to zero effect on the probability of sentiment reaction. From our results, we also learn that if an investor has limited to excellent investment knowledge and/or experience, we can expect that their probability of reacting to a given event is lower than if they had no investment

knowledge and/or experience. Furthermore, if the event occurred before or during the financial crisis, we expect the log-odds of reaction to no reaction to increase compared to if the event occurred after 2009.

We also derive logistic regression models on the sub-sample of data-points that occur after the financial crisis, so that our contributions are based on more recent data. The most comprehensive model achieves a 0.35 pseudo R-squared value. Generally speaking, we observe the same trends and patterns across all variables as those in the previous model. For instance, age continues to have a close to zero effect on the probability of reacting to an event. Furthermore, there are demographic variables in the post-crisis models with coefficients that share the same signs as those in the all-time models but have consistently larger magnitudes. These include the indicators for reporting to have good investment experience, limited investment experience, and limited investment knowledge. In addition, the variable that has the strongest positive effect on the probability of reacting to an event out of all other variables is marital status. On the other hand, the variables that have the strongest negative effect are the fraction of positive events and the sentiment score of the event with the largest absolute value in the week prior to the given data-point.

Second, we derive linear regression models to predict the magnitude of reaction, i.e. the proportion of wealth traded, given that the investor has already reacted to a given sentiment event. First, we derive multivariate linear regressions from data that include all sentiment trades from all time. The most comprehensive model achieves a 0.102 R-squared value. Furthermore, the magnitude of the sentiment event with the largest absolute value in the week prior to the data-point and the proportion of buys of the CUSIP made on sentiment previously have the strongest average effects on reaction magnitude. We also find that the proportion of buys that were made on sentiment up until the current trade is a better predictor of reaction magnitude than is the proportion of sells that were made on sentiment. Furthermore, our results indicate that if an investor reacts to an event that occurs during the financial crisis, we expect to observe a slightly smaller proportion of wealth traded compared to if the investor were to react to an event post-financial crisis. Finally, our results suggest

that if a sentiment investor has a history of trading a given CUSIP, the magnitude of their reaction to an event concerning the CUSIP is larger than if the investor never traded the CUSIP before.

We then derive multivariate linear regressions on only post-crisis data. The most comprehensive model achieves a 0.135 R-squared value. The variables with the strongest effects on reaction magnitude in post-crisis data are the same as those in all-time data. We find that the score of the sentiment event with the largest value in the week preceding a reaction has the largest effect on the proportion of wealth traded in reaction to an event. The proportion of buys made on sentiment before the reaction has the second largest effect. In general, the coefficients in this model tend to follow the same trends as those in the all-time model. However, there are some variables, such as the indicator for day trading, that exhibit somewhat dissimilar trends in the post-crisis models than they do in the all-time models. Nonetheless, given the similar coefficient magnitudes of these variables across both model types, we conclude that they have the same relative predictive strength of reaction magnitude in both post-crisis and all-time data.

We also derive linear regressions from data that only involve buys. Our most comprehensive model achieves a 0.197 R-squared value. We observe some similar trends and pattern in the buy-only models as those in the all-trades models. For instance, we find that sentiment in the week prior to a reaction has the strongest average effect on sentiment buy reaction magnitude. We also conclude that sentiment buying history is more predictive of the buy reaction magnitude than sentiment selling history is. Furthermore, our results suggest that the effect of marital status, number of dependents, and age are essentially zero. However, the effects of investment experience and knowledge on reaction magnitude have a relatively stronger effect on reaction magnitude in buy-only data than in data that includes all trades. Finally, we find that if a reaction occurs before the crisis, we should expect a decrease in the proportion of wealth used to make buys than if the reaction occurs after the crisis. On the other hand, if a reaction occurs during the crisis, we should expect an increase in the proportion of wealth used to make buys than if the reaction occurs after the crisis.

Then, we derive linear regression models from data that only includes sells in order to predict sell reaction magnitude. Our most comprehensive model achieves a 0.121 R-squared value. There are some similar trends between the buy-only and sell-only models. For instance, the magnitude of the sentiment event with the largest absolute value in the week prior to the reaction has a strong effect on the proportion of wealth traded. However, there are some noticeable differences between the two models. For example, the coefficients for the indicators of having good investment experience and limited investment knowledge are slightly negative in the buy-only models, but are slightly positive in the sell-only models. From our results, we also see that the number of events in the month prior to the reaction as well as the indicator for day trading have relatively large effects on reaction magnitude in buy-only data, but lose their effects by an order of magnitude in the sell-only data. Furthermore, we observe that past sentiment selling behavior is more predictive of current sentiment selling behavior, while past sentiment buying behavior is more predictive of current sentiment buying behavior.

Third and finally, we derive logistic regression models to predict the direction of reaction, i.e. buy or sell. First, we train our models on data from all time, and our most comprehensive model achieves a 0.055 pseudo R-squared value. We note that the sentiment score for the event with the largest magnitude in the week prior to the event has a weak effect on reaction direction, which is not the case in our previous models that predict reaction/non-reaction and reaction magnitude. Furthermore, out of all the demographic variables, marital status has the largest effect on sentiment trade direction. The demographic variable with the next largest effect is the indicator for whether the investor reports to have good investment knowledge. On the other hand, demographic variables likes age and the number of dependents have weak to zero effects on reaction direction. Other variables that have relatively strong effects on trade direction include the indicator for a reaction occurring before the crisis as well as the proportion of all volume traded on sentiment previously. Specifically, if a reaction occurs before the crisis, we expect an increase in the log-odds of buying to selling compared to if the reaction occurred after the crisis. Furthermore, our results

indicate that if an investor has a history of trading a larger proportion of volume on sentiment, then the probability that they buy in reaction to an event is lower compared to the probability that they sell.

We also derive logistic regression models on the sub-sample of data with reactions occurring after the financial crisis. Our most comprehensive model achieves a pseudo R-squared value of 0.101, which is double that of the comprehensive all-time model. In general, we see similar trends across all variables as those in the all-time data. For instance, the signs and magnitudes of the coefficients on the indicators for the number of dependents, the proportion traded of a given CUSIP previously, and the sentiment score of the event with the largest magnitude in the week leading up to the event are similar to those of the all-time model. We also observe that marital status and the proportion of volume traded on sentiment previously explain more of the variance and thus have stronger effects on reaction direction in the post-crisis data than they do in the all-time data. Furthermore, our results indicate that the fraction of positive events in the week prior to the reaction and the proportion traded of the CUSIP previously have negative effects on the probability of buying. We also conclude that an investor's history of sentiment trading has a stronger negative effect on reaction direction than the investor's general CUSIP-specific trading history and the positivity of the events prior to the reaction do.

6.2 Future Work

In this section, we draw inspiration from the studies described in Chapter 2 and point out areas of future work.

First, a notable area of future work involves deriving the same models from different media sentiment datasets. We could use Twitter data as a measure of sentiment (e.g. [11], [6], [18], [16]) as well as Yahoo! Finance forums (e.g. [12]). Then, we could compare model fit and the predictability of different model variables on Ravenpack news media data, Twitter data, and Yahoo! Finance forum data. Furthermore, we could look into combining the three datasets together to develop more generalizable

models derived from a more comprehensive and diverse dataset.

Another area of future work involves experimenting with different stock- and market-specific features in our models. For instance, we could incorporate the day-of-the-week effect discussed by Ren et al. [21]. Both Chatterjee and Perrizo [4] and Kim and Kim [12] find strong correlations between stock volatility and investor sentiment, so we could also include measures of stock and market volatility in our models. In addition, Kim and Kim [12] discuss the impact of previous stock movement on investor sentiment, so we could look into including features that measure previous stock movement in our models as well. Furthermore, Chung et al. [5] indicate that they only discover a relationship between investor sentiment and stock returns during times of economic expansion and not during economic recessions. While we do already include similar features in our models (the indicators for data-points occurring before, during, or after the crisis), we could look into more granularly defining times of booms versus busts and incorporating such features instead.

Finally, one potential area of future work involves experimenting with different machine learning model approaches. While our linear and logistic regressions do fit our data relatively well, it would be valuable to compare the performance of these models against other machine learning algorithms. We could look into using decision tree models and experimenting with using support vector machines and neural network architectures like Porshnev et al. [18], Jiahong Li et al. [11], and Ren et al. [21] do. Indeed, Jiahong Li et al. [11] demonstrate the potential of deep learning in finance applications, and a notable area of future work could involve experimenting with different deep learning architectures. However, given that another one of our focuses is to produce interpretable models that are usable in real-world applications, an area of future work would be to investigate the trade-off between the potential improvements performance and interpretability between these models.

Appendix A

Tables

Key	Description	Format	Example
trade_date	Date of trade	YYYYMMDDD	20110819
buy_sell	Indicator for the trade being a buy or sell	string	B
principal	Principal amount traded, i.e. volume traded	double	2098.0
quantity	Number of asset units traded	integer	100
tcommission	Trade commission	double	8.95
cusip_nr	CUSIP number	9-character alphanumeric	26613Q106
ticker_symbol	Ticker symbol	string	DFT
item_issue_id	Item issue identification number	integer	1482888359
product_code	Product code	string	RET
product_grplvl1	Identifier for top-level security type	string	EQUITY
product_grplvl2	Identifier for mid-level security type	string	EQUITY
product_grplvl3	Identifier for bottom-level security type	string	EQUITY
acid_key	Account identification number	integer	9127986000

Table A.1: All fields in the trades data.

Key	Description	Format	Example
month	Month of snapshot	YYYYMM	200903
item_issue_id	Item issue identification number	integer	577414143
settle_qty	Quantity of shares held in security	integer	503.2907
ticker_symbol	Ticker symbol	string	C
cusip_num	CUSIP number	9-character alphanumeric	172967101
issue_price	Closing price listed on the exchange on the last market day of the month	double	2.53
product_code	Product code	string	COM
product_grplvl1	Identifier for top-level security type	string	EQUITY
product_grplvl2	Identifier for mid-level security type	string	EQUITY
product_grplvl3	Identifier for bottom-level security type	string	EQUITY
acid_key	Account identification number	integer	9127986000

Table A.2: All fields in the positions data.

Key	Description	Format	Example
month_last_record	Month the data was collected	YYYYMM	201512
cust_age	Age	integer	70
dmsa_martl_stat_cd	Marital status	string	MARRIED
cust_depndt_qy	Number of dependents	integer	0
ps_gndr_cd	Gender	string	M
actual_curr_occup_tx	Occupation	string	SKI INSTRUCTOR
dmsa_curr_occup_tx	Occupation category	string	EDUCATION
indiv_annl_incm_am	Individual annual income category	integer	100
fin_tot_nwrth_am	Total financial worth category	integer	1000
invst_knldg_cd	Investment knowledge category	string	L
invst_exprc_cd	Investment experience category	string	G
acctid_key	Account identification number	integer	999969455
custid_key	Customer identification number	integer	9893700864

Table A.3: All fields in the demographics data.

Key	Description	Format	Example
BAM	Score computed by the BAM classifier	integer	50
BCA	Score computed by the BCA classifier	integer	50
BEE	Score computed by the BEE classifier	integer	50
BMQ	Score computed by the BMQ classifier	integer	50
CATEGORY	Tag to label a particular type of news event	string	executive-resignation
COMPANY	The entity's ISO code and stock ticker	string	US/AAPL
COUNTRY_CODE	ISO-3166 country code	string	US
CSS	Score computed from the PEQ, BEE, BMQ, BCA, and BAM classifiers	int	50
ENTITY_TYPE	5 entity types: COMP (Company), ORGA (Organization), CURR (Currency), CMDT (Commodity), PLCE (Place)	string	ORGA
PEQ	Score computed by the PEQ classifier	int	50
TIMESTAMP_UTC	Timestamp of the media event	timestamp	2015-01-02 19:35:26.454

Table A.4: Subset of fields from the RavenPack data.

Gender	Number of Sentiment Investors	Number of All Investors	Relative Proportion
F	171	1,559	0.78***
			(-3.18)
M	356	2,317	1.09
			(1.51)
Missing	8,483	59,938	1.0
			(0.84)
Total	9,010	63,814	

Table A.5: Composition and relative prevalence of sentiment investors' gender. "Missing" indicates information that wasn't reported.

Marital Status	Number of Sentiment Investors	Number of All Investors	Relative Proportion
DIVORCED	309	1,717	1.27***
			(3.99)
MARRIED	2,167	12,344	1.24***
			(10.47)
MINOR	2	16	0.89
			(-0.16)
SEPARATED	2	29	0.49
			(-1.00)
SINGLE	1,384	8,458	1.16***
			(5.47)
UNMARRIED	4	29	0.98
			(-0.04)
WIDOWED	106	703	1.07
			(0.63)
Missing	5,036	40,518	0.88***
			(-13.95)
Total	9,010	63,814	

Table A.6: Composition and relative prevalence of sentiment investors' marital status. "Missing" indicates information that wasn't reported. Note that numbers in parentheses are the standard deviation of coefficients calculated using Equation 4.8. * means significance at the 10% level, ** means significance at the 5% level, and *** means significance at the 1% level.

Number of Dependents	Number of Sentiment Investors	Number of All Investors	Relative Proportion
0	3,850	25,158	1.08***
			(6.00)
1	525	3,126	1.19***
			(3.78)
2	515	2,954	1.23***
			(4.53)
3	239	1,322	1.28***
			(3.56)
4	90	487	1.31***
			(2.36)
5	29	139	1.48*
			(1.93)
6	10	27	2.62***
			(2.71)
7	2	8	1.77
			(0.73)
8	2	3	4.72*
			(1.88)
9	0	1	0.0
			(-0.38)
10	1	3	2.36
			(0.77)
Missing	3,747	30,586	0.87***
			(-11.29)
Total	9,010	63,814	

Table A.7: Composition and relative prevalence of sentiment investors' number of dependents. "Missing" indicates information that wasn't reported. Note that numbers in parentheses are the standard deviation of coefficients calculated using Equation 4.8. * means significance at the 10% level, ** means significance at the 5% level, and *** means significance at the 1% level.

Table A.8: Composition and relative prevalence of sentiment investors' occupation group. Note that this table spans multiple pages.

Occupation Group	Number of Sentiment Investors	Number of All Investors	Relative Proportion
PILOT	22	101	1.54*
			(1.86)
FINANCIAL	186	931	1.41***
			(4.38)
EXECUTIVE	276	1,398	1.4***
			(5.17)
OWNER	100	515	1.38***
			(2.94)
REALESTATE	54	288	1.33*
			(1.92)
CPA	78	429	1.29*
			(2.07)
ATTORNEY	107	591	1.28***
			(2.38)
RETIRED	893	5,011	1.26***
			(6.70)
SCIENTIST	41	235	1.24
			(1.26)
GOVERNMENT	7	40	1.24
			(0.53)
ENGINEER	81	476	1.21
			(1.56)
SKILLLABOR	295	1,740	1.2***
			(2.95)
MANAGER	345	2,069	1.18***
			(2.91)
PARALEGAL	15	91	1.17
			(0.56)
SELFEMPLOYED	180	1,101	1.16*
			(1.84)
CONSULTANT	90	565	1.13
			(1.07)
S-SKILLEDOFFICE	91	575	1.12
			(1.02)
MEDICAL	94	605	1.1
			(0.87)
PHYSICIAN	87	558	1.1
			(0.86)

MARKETING	216	1,404	1.09
			(1.19)
COMPUTER	154	998	1.09
			(1.03)
CLERGY	5	33	1.07
			(0.15)
WHITE-COLLAR	254	1,730	1.04
			(0.59)
UNEMPLOYED	208	1,474	1.0
			(-0.01)
PROFESSIONAL	313	2,261	0.98
			(-0.33)
EDUCATION	113	835	0.96
			(-0.43)
HOMEMAKER	104	777	0.95
			(-0.51)
SECRETARY	57	437	0.92
			(-0.56)
MINOR	2	16	0.89
			(-0.16)
ARTIST	36	287	0.89
			(-0.67)
STUDENT	19	154	0.87
			(-0.56)
POLICE-MILITARY	26	213	0.86
			(-0.70)
SOCIALWORKER	2	24	0.59
			(-0.72)
DISABLED	0	2	0.0
			(-0.53)
Missing	4,459	35,850	0.88***
			(-11.96)
Total	9,010	63,814	

Table A.8: Composition and relative prevalence of sentiment investors' occupation group. "Missing" indicates information that wasn't reported. Note that numbers in parentheses are the standard deviation of coefficients calculated using Equation 4.8. * means significance at the 10% level, ** means significance at the 5% level, and *** means significance at the 1% level.

Bibliography

[1] Peter Ager Hafez and Junqiang Xie. Web news analytics enhance stock portfolio returns. *SSRN Electronic Journal*, 01 2014. doi: 10.2139/ssrn.2423362.

[2] Shreyash Agrawal, Pablo D. Azar, Andrew W. Lo, and Taranjit Singh. Momentum, mean-reversion, and social media: Evidence from stocktwits and twitter. *The Journal of Portfolio Management*, 44:85–95, 07 2018.

[3] Nicholas Barberis, Andrei Shleifer, and Robert W. Vishny. A Model of Investor Sentiment. NBER Working Papers 5926, National Bureau of Economic Research, Inc, February 1997. URL `https://ideas.repec.org/p/nbr/nberwo/5926.html`.

[4] A. Chatterjee and W. Perrizo. Investor classification and sentiment analysis. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1177–1180, 2016.

[5] San-Lin Chung, Chi-Hsiou Hung, and Chung-Ying Yeh. When does investor sentiment predict stock returns? *Journal of Empirical Finance*, 19(2):217 – 240, 2012. ISSN 0927-5398. doi: https://doi.org/10.1016/j.jempfin.2012.01.002. URL `http://www.sciencedirect.com/science/article/pii/S0927539812000035`.

[6] Mariana Daniel, Rui Ferreira Neves, and Nuno Horta. Company event popularity for financial markets using twitter and sentiment analysis. *Expert Systems with Applications*, 71, 11 2016.

[7] Thomas A. Domencich and Daniel McFadden. Urban travel demand: A behavioral analysis. 1977.

[8] C. Huang, C. Chang, B. R. Chang, and T. Hsieh. A genetic-based stock selection model using investor sentiment indicators. In *2011 IEEE International Conference on Granular Computing*, pages 262–267, 2011.

[9] C. Huang, T. Hsieh, B. R. Chang, and C. Chang. A comparative study of regression and evolution-based stock selection models for investor sentiment. In *2012 Third International Conference on Innovations in Bio-Inspired Computing and Applications*, pages 73–78, 2012.

[10] Dashan Huang, Fuwei Jiang, Jun Tu, and Guofu Zhou. Investor Sentiment Aligned: A Powerful Predictor of Stock Returns. *The Review of Financial Studies*, 28(3):791–837, 10 2014. ISSN 0893-9454. doi: 10.1093/rfs/hhu080. URL https://doi.org/10.1093/rfs/hhu080.

[11] Jiahong Li, Hui Bu, and Junjie Wu. Sentiment-aware stock market prediction: A deep learning method. In *2017 International Conference on Service Systems and Service Management*, pages 1–6, 2017.

[12] Soon-Ho Kim and Dongcheol Kim. Investor sentiment from internet message postings and the predictability of stock returns. *Journal of Economic Behavior Organization*, 107:708 – 729, 2014. ISSN 0167-2681. doi: https://doi.org/10.1016/j.jebo.2014.04.015. URL http://www.sciencedirect.com/science/article/pii/S0167268114001206. Empirical Behavioral Finance.

[13] Bing Li, Keith C.C. Chan, Carol Ou, and Sun Ruifeng. Discovering public sentiment in social media for predicting stock movement of publicly listed companies. *Inf. Syst.*, 69(C):81–92, September 2017. ISSN 0306-4379. doi: 10.1016/j.is.2016.10.001. URL https://doi.org/10.1016/j.is.2016.10.001.

[14] Ildiko Mohacsy and Heidi Lefer. Money and sentiment: A psychodynamic approach to behavioral finance. *The Journal of the American Academy of Psychoanalysis and Dynamic Psychiatry*, 35:455–75, 02 2007.

[15] Robert Neal and Simon M. Wheatley. Do measures of investor sentiment predict returns? *The Journal of Financial and Quantitative Analysis*, 33(4):523–547, 1998. ISSN 00221090, 17566916. URL http://www.jstor.org/stable/2331130.

[16] Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, and Babita Majhi. Sentiment analysis of twitter data for predicting stock market movements. *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 1345–1350, 2016.

[17] Richard L Peterson. *Trading on sentiment: the power of minds over markets*. John Wiley & Sons, 2016. ISBN 9781119219149. doi: 10.1002/9781119219149.

[18] A. Porshnev, I. Redkin, and A. Shevchenko. Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 440–444, 2013.

[19] Nizar Raissi. Role of investor sentiment in financial markets: an explanation by behavioural finance approach. *Int. J. Accounting and Finance*, Vol. 5:362–401, 04 2016.

[20] Gabriele Ranco, Darko Aleksovski, Guido Caldarelli, Miha Grčar, and Igor Mozetič. The effects of twitter sentiment on stock price returns. *PloS one*, 10(9):e0138441, 2015.

[21] R. Ren, D. D. Wu, and T. Liu. Forecasting stock market movement direction using sentiment analysis and support vector machine. *IEEE Systems Journal*, 13(1):760–770, 2019.

[22] Maik Schmeling. Investor sentiment and stock returns: Some international evidence. *Journal of Empirical Finance*, 16(3):394 – 408, 2009. ISSN 0927-5398. doi: https://doi.org/10.1016/j.jempfin.2009.01.002. URL http://www.sciencedirect.com/science/article/pii/S0927539809000048.

[23] Andrei Shleifer, Nicholas Barberis, and Robert Vishny. A model of investor sentiment. *Journal of Financial Economics*, 49:307–343, 02 1998.

[24] Paul Tetlock. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62:1139–1168, 02 2007. doi: 10.2139/ssrn.685145.

[25] D. D. Wu, L. Zheng, and D. L. Olson. A decision support approach for online stock forum sentiment analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(8):1077–1087, 2014.

[26] Jeffrey A. Wurgler and Malcolm Baker. Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21:129–152, 02 2007.

[27] Steve Y. Yang, Yangyang Yu, and Saud Almahdi. An investor sentiment reward-based trading system using gaussian inverse reinforcement learning algorithm. *Expert Systems with Applications*, 114:388 – 401, 2018. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2018.07.056. URL http://www.sciencedirect.com/science/article/pii/S0957417418304810.