Should Retail Investors Listen to Social Media Analysts? Evidence from Text-Implied Beliefs∗

Chukwuma Dim

This version: March 2021

Abstract

Social media is increasingly affecting financial markets, with important implications for market efficiency. This paper uses machine learning to construct beliefs of non-professional social media investment analysts (SMAs) from opinions expressed about individual stocks on social media. On average, SMA beliefs are informative about future stock abnormal returns and earnings surprise. However, there exists important heterogeneity in belief formation ability. Only a small fraction, 10%, of SMAs form beliefs that produce an economically meaningful abnormal return of 56 bps over a 5-day window. SMA characteristics such as specialization, skin in the game, effort, popularity, and disagreement matter for belief formation skill. SMAs herd in belief statements. Herding is less pronounced in bad times and when the consensus is more optimistic, but more pronounced when the consensus is correct ex-post.

Keywords: Social Media, Nonprofessional Analysts, Belief Formation, Investor Skill, Herding, Machine Learning, Textual Analysis, Natural Language Processing

JEL: G11, G12, G14

∗Chukwuma Dim is affiliated with Frankfurt School of Finance & Management, c.dim@fs.de. I received helpful comments and suggestions from Grigory Vilkov, Francesco Sangiorgi, Yigitcan Karabulut, Zacharias Sautner, Michalis Haliassos, Kim Peinjenburg, Nic Schaub, Camelia Kuhnen, Burton Hollifield, Raman Uppal, and Francesco Franzoni. I thank participants of the Frankfurt School of Finance & Management Brownbag Seminar, and the NOVA Finance PhD Pitch Perfect for helpful comments.
1 Introduction

Social networks shape individuals’ expectations and actions as people rely on their network for information. In financial markets, social media intensifies social networks’ role on beliefs due to the ease with which retail investors now interact and share investment ideas. For instance, the recent frenetic trading in the GameStop stock by retail investors, primarily coordinated via social media, arguably led to the 2,000% surge in the GameStop stock price in January 2021 and the near-collapse of a hedge fund, Melvin Capital. At the same time, retail investors, spurred via social media, accounted for 23% of all US equity trading between January and February 2021, making their market footprint as big as that of all hedge funds and mutual funds combined.¹ These and related developments emphasize that the interplay of social media and retail trading poses new challenges for financial markets, which requires a deeper understanding of belief formation on social media and the implications for market efficiency.

A crucial feature of investment social media is the existence of influencers, i.e., nonprofessional **Social Media Investment Analysts (SMAs)**, who publish stock-specific investment research and belief statements that shape the views and actions of many individual investors.² I build on two strands of the theoretical literature to analyze these influencers’ role in financial markets. The model of Pedersen (2021), on the role of social networks on belief formation, implies that the distribution of rational SMAs on social media matters for the overall rationality of financial markets. On the other hand, the model of Bhamra et al. (2021), on psychological distance and belief formation, implies that the views of SMAs on social media can reduce individual investors’ psychological distance from firms, thereby influencing beliefs and investment decisions.³ Two broad questions naturally arise from these theoretical implications: Are SMAs skilled enough to play an informational role? How do they form the beliefs impressed on other investors?

This paper addresses these questions by studying beliefs about individual stocks expressed by SMAs on social media. I provide answers to the following specific questions: Do SMA beliefs

---

¹See, https://www.ft.com/content/7a91e3ea-b9ec-4611-9a03-a8dd3b8bddb5
²Farrell et al. (2020) show that nonprofessional analysts on social media cater to retail investors, while Drake et al. (2019) argue that SMAs’ views on social media reduce the informational value of analyst recommendations.
³Psychological distance is a cognitive separation between oneself and other instances such as persons, events, or times; its dimensions include temporal, spatial, informational, and social distance (Baltatescu, 2014). SMAs’ views on social media could bridge the social and informational distance between individual investors and specific stocks by making the stocks more familiar to investors, impacting beliefs and actions about the stocks.
stated on social media contain value-relevant information? How widespread is SMAs’ ability to form beliefs that provide investment value, i.e., belief formation ability? Which SMA characteristics are associated with belief formation ability? And do SMAs herd in stating their beliefs? The major empirical challenge in studying beliefs about individual stocks is the difficulty in obtaining a reasonably large sample of belief statements. I overcome this challenge by applying Natural Language Processing (NLP) and Machine Learning (ML) techniques to construct SMA beliefs about a large cross-section of stocks from views expressed by SMAs in a popular investment social media, Seeking Alpha.

To be more concrete, since 2018, most SMA opinion articles in Seeking Alpha are accompanied by the author’s explicit belief about a stock as either “Very Bullish”, “Bullish”, “Neutral”, “Bearish”, or “Very Bearish”. I use the sub-set of articles with explicit belief statements to train a relatively simple, yet robust, ML model, namely Support Vector Classifier (similar to Manela and Moreira, 2017), that enables me extract the beliefs implied by the rest of the unlabeled SMA articles in Seeking Alpha dating back to 2004.\(^4\) I then use this large sample of stated and extracted SMA beliefs for the analyses in this paper. Intuitively, the ML algorithm maps words in an SMA’s investment thesis to the SMA’s implicit belief – an approach reminiscent of Stefan Nagel’s recent suggestion to apply ML in constructing proxies of beliefs from textual data to improve upon the limitations of survey data (see, Brunnermeier et al., 2021).

Several factors make Seeking Alpha an ideal setting to study belief formation on social media. First, Seeking Alpha dates back to 2004 and is very popular among retail investors who subscribe to the platform to gain access to SMAs’ views or contribute their opinions. For instance, about 20 million people use Seeking Alpha monthly\(^5\) and roughly 11,000 and 300,000 unique users contributed views and commentaries, respectively, covering about 7,200 firms over my sample period.\(^6\) Second, the goal of Seeking Alpha is to provide opinion and analyses, rather than...
news, primarily through individual investors, instead of journalists, who describe their approach
to stock-picking and portfolio management (Seeking Alpha, 2006). Third, views expressed on
Seeking Alpha are backed by in-depth analyses of an investment thesis, which Seeking Alpha’s
editorial team checks for quality before publication. The second and third features imply that
SMAs in this paper mostly refer to individual investors – relatively more sophisticated than the
average retail investor – who share their beliefs about stocks on investments social media. Unlike
professional analysts, SMAs do not issue investment recommendations; rather, they publish
views that shape the expectations and actions of other investors in their network. SMAs’
testimonials indicate that the main incentives for contributing views on Seeking Alpha include
direct monetary compensation from Seeking Alpha,7 feedback on investment thesis from other
individual investors, and increased visibility and recognition that could result in professional
opportunities.8

I begin the empirical analyses by examining whether, on average, SMA beliefs provide value-
relevant information. I start by aggregating beliefs at the stock level, then examine whether
beliefs predict individual stocks’ abnormal returns (relative to the CAPM, Fama and French
(1993) three-factor model, and Daniel et al. (1997) characteristic-based benchmark) over different
investment horizons. I find that SMA beliefs significantly predict stocks’ future abnormal returns
up to the 3-month investment horizon. For instance, a standard deviation increase in the
bullishness of beliefs predicts an increase in abnormal return of 16 basis points (bps) over a
5-day horizon and 26 (28) bps increase over a 1-month (3-month) horizon. Predictability is
stronger for stocks less likely to be covered by professionals and predominantly traded by retail
investors: low turnover, low price, small size, high book-to-market ratio, and high volatility
stocks. The results are robust to alternative abnormal return benchmarks and to controlling for
contemporaneous news that may drive returns, such as professional analyst recommendations
and sentiment in cash flow news. There is no reversal in predictability even at the 1-year
horizon, which points towards SMA beliefs containing value-relevant information rather than
just sentiment unrelated to fundamentals.

7Seeking Alpha compensates SMAs financially using some criteria that include the number of article-page views. However, the nature of this compensation, which can be as low as $40, suggest that financial compensation is not just the primary motivation of SMAs. For more details on the compensation scheme, see https://seekingalpha.com/page/article_payments
Analyses of daily rebalanced portfolios formed on SMA beliefs suggest that (retail) investors could profit from trading on SMA beliefs. For instance, the portfolio of stocks with the most bullish beliefs have statistically significant alpha ranging between 10 bps and 14 bps per month depending on factor benchmark. On the other hand, the portfolio with the least favorable SMA beliefs has significantly negative alpha, while the difference portfolio has significant and economically sizeable three-factor alpha of roughly 40 bps per month, equivalent to roughly 5% abnormal returns per annum.

To verify that SMAs’ belief predictability of return is indeed due to value-relevant information rather than the price impact of trades generated by belief statements, I examine whether beliefs are informative about future earnings surprise. Because it is unlikely that SMA beliefs affect either analyst forecasts or firms’ earnings, evidence of predictability would support the view that beliefs provide value-relevant information. I find that, indeed, SMA beliefs significantly predict earnings surprise. A standard deviation increase in the bullishness of beliefs predicts an increase in earnings surprise by between 15% and 20% of its standard deviation. This suggests that SMA beliefs’ predictability of returns is more likely due to value-relevant information in beliefs that are not already priced. This result supports the conclusion of Chen et al. (2014) and Kelley and Tetlock (2013), among others, that individual investors produce and trade on value-relevant information.\footnote{SMAs in my paper are also individual investors, since they often disclose investment positions in stocks in Seeking Alpha.}

That SMA beliefs expressed on social media provide value-relevant information is important, because it suggests that the democratization of investment research and the growing popularity of nonprofessional analysts on social media is beneficial for price discovery. However, a complete understanding of the relevance of SMAs for market efficiency requires deeper analyses of how much skill exists among SMAs whose views tend to inform retail trades.\footnote{See Farrell et al. (2020) for evidence that nonprofessionals on social media largely cater to retail investors.} For instance, a desirable outcome would be that the documented predictive ability of SMA beliefs is due to most SMAs being \textit{truly} skilled in producing/processing information that inform correct beliefs. However, it might be that most SMAs are just lucky with only very few being truly skilled.
To rigorously examine the distribution of SMAs’ belief formation ability, I follow recent papers in the professional analyst and delegated management literature (Crane and Crotty, 2020; Harvey and Liu, 2018; Chen et al., 2017) and model SMA belief formation ability as arising from a mixture distribution of multiple skill groups. Because a mixture model uses information in the cross-section of investor performance to reduce noise, it ameliorates the false discovery problem that often arises from the low signal-to-noise feature of abnormal returns. Guided by model selection criteria, I estimate a two-component mixture model where one component comprises the low-skill SMAs and the other the high-skill type. I find that 90% of SMAs belong to the low-skill group. However, roughly 56% of SMAs are skilled enough to generate just positive abnormal returns following belief statements. The average abnormal return of the 90% low-skill SMAs is estimated to be only 6 bps over a 5-day horizon, while that of the 10% high-skill group is a sizeable 56 bps. To provide some context, Crane and Crotty (2020) use a similar setup and estimate the fraction of high-skill professional analysts to be 36%, with roughly 97% being skilled enough to generate just positive abnormal returns. In light of earlier results, it follows that while SMAs as a group appear to add value, at the individual SMA level skill is considerably limited.

Since skill is scarce among SMAs, it helps to know what a skilled SMA looks like – perhaps to reduce the search cost for valuable information on social media. I, therefore, examine how SMA characteristics are associated with skill. Consistent with theoretical results on gains from specialization in information acquisition and cognitive capacity constraints (e.g., Van Nieuwerburgh and Veldkamp, 2010; Hirshleifer et al., 2011), I find that SMAs that specialize in a few industries are 34% more likely to be high-skill. On the other hand, unlike papers that argue that individual investors get better with experience (Seru et al., 2010), I find that SMAs with more experience in Seeking Alpha are 9% less likely to be high-skill. A potential explanation for this result is that more skilled SMAs joined Seeking Alpha over time as the platform became more popular. Furthermore, SMAs with more skin in the game (i.e., those that mostly have an investment position in the stock they express belief on) are 12% more likely to be high-skill, consistent with Campbell et al. (2019) who find that having an investment position in a stock does not impair the credibility of nonprofessional analysts. More popular SMAs also have a
higher probability of being high-skill, suggesting that individuals tend to follow more skilled SMAs on social media.

In the final part of the paper, I examine one plausible channel through which SMAs form beliefs, namely herding, which, depending on its nature, has different implications for market efficiency. Theories of information cascades and reputational herding (e.g., Banerjee, 1992; Scharfstein and Stein, 1990) suggest that SMAs’ herding could take the form of learning from peers to improve one’s forecasts, or simply herding for other reasons unrelated to information. However, although social media is a typical environment for mutual imitation, SMAs may prefer to go against the consensus in order to attract readership which would increase their financial compensation. I, therefore, test for herding and its nature among SMAs using the herding test developed in Welch (2000) for settings where choices are discrete. I find strong evidence that SMAs herd in belief statements. Herding is, however, less pronounced in bad times – during recessions and high market uncertainty – in line with the result of Welch (2000) for professional analysts. On the other hand, contrary to Welch (2000), I find that SMAs herd less when the consensus is more optimistic. A standard deviation increase in consensus optimism is associated with a reduction in herding by 24% of its unconditional value, suggesting that SMAs prefer more attention-grabbing deviation from the consensus potentially to attract readership.

Interestingly, SMAs herd more on consensus that turns out to be correct ex-post. A standard deviation increase in consensus correctness – defined as a Bullish (Bearish) consensus followed by positive (negative) future stock return – is associated with an increase in herding by about 7% of its unconditional value. A possible implication of this result is that SMAs learn fundamental information from the consensus to improve their expectations, which is consistent with models of information-based herding (e.g., Banerjee, 1992). It is equally possible that SMAs independently follow the same fundamental information. Whichever the case, the result suggests that to some extent observational learning among SMAs would likely improve price discovery as investors trade on SMAs’ belief statements.

Put together, my empirical analyses suggest that SMAs as a group tend to improve price discovery, and are not just the mad crowd recent social media events might tempt one to think.
As such, (retail) investors could benefit from SMAs’ average beliefs, even though at the individual SMA level skill is quite limited.

This paper speaks to several strands of research. First, the paper is closely related to the literature that explores whether investors’ opinions stated over the internet are informative about individual stock returns. Chen et al. (2014) show that the fraction of negative words in views expressed on Seeking Alpha predicts stock abnormal returns and earnings surprise, Avery et al. (2016) echo their result using individual investors’ predictions in MotleyFool. Antweiler and Frank (2004) find that messages posted on internet stock message boards predict both return and volatility of a narrow set of stocks. Das and Chen (2007) find similar results for volatility and volume. In contrast to these papers, I go beyond whether views on social media are, on average, informative to investigate the distribution of belief formation ability among nonprofessional analysts on social media. Furthermore, I study how these nonprofessionals form the beliefs impressed on other individuals. These aspects of my work shed new light on modern social media’s role – where the opinion of a few influential individuals can have outsized impacts – on belief formation and price discovery.

A related literature looks at how nonprofessional analysts’ views on social media influence retail investors’ trade, their information environment, and professional analysts’ informational role (Gomez et al., 2020; Farrell et al., 2020; Campbell et al., 2019; Drake et al., 2019). I contribute to this literature by examining whether retail investors should heed nonprofessional analysts in the first place. In so doing, I show that SMAs as a group have value for (retail) investors even though skill may be very limited at the individual SMA level. Furthermore, I document that specific SMA characteristics could help investors identify skilled SMAs.

My paper is related to the literature that studies whether retail investors’ participation in financial markets improves market efficiency or introduces noise (e.g., Boehmer et al., 2020; Seasholes and Zhu, 2010; Barber et al., 2008; Kelley and Tetlock, 2013; Kaniel et al., 2012). In my paper, SMAs are also individual investors as they disclose their investments in Seeking Alpha. In showing that SMA beliefs contain value-relevant information, my paper aligns with papers in this literature that argue that retail investors produce information that improves price
discovery. Unlike the papers in this literature that mostly study retail investors’ transactions, I study beliefs and the distribution of skill among individual investors which offer new insights.

I also speak to the household finance literature that study individual investors’ beliefs (e.g., Choi and Robertson, 2020; Giglio et al., 2019). This literature largely relies on survey data of beliefs about aggregate outcomes. A recent exception is Bhamra et al. (2021) who extract belief distortions about individual stocks from Finish households’ stock holdings. My main contribution is to use NLP and ML techniques to infer beliefs about a large cross-section of stocks over a relatively long period from textual data, enabling me to study belief formation about stocks on social media. Therefore, my approach is reminiscent of Stefan Nagel’s recent suggestion to deploy innovations such as ML in constructing proxies of beliefs from text to improve upon the limitations of survey data (see, Brunnermeier et al., 2021).

Finally, I contribute to the growing literature that uses ML techniques to extract economic quantities from textual data (e.g., Gu et al., 2020; Chen et al., 2019; Ke et al., 2019; Manela and Moreira, 2017). By leveraging NLP and supervised ML techniques to infer beliefs, I deepen the range of economic research questions these techniques can address.

2 Data

This section describes the data used in this study and the construction of variables. The sample period, January 2005 to December 2019, is determined by the availability of SMAs’ beliefs and opinions in Seeking Alpha.

2.1 Seeking Alpha Data

The analyses in this paper rely on opinions and belief statements of SMAs in Seeking Alpha, a popular investment social media launched in 2004 and used by about 20 million people per month. Any individual can contribute views in Seeking Alpha by submitting an opinion article with extended analyses of the investment thesis (and in more recent times accompanied by an explicit belief statement) on specific stocks. However, the opinion article must scale

\[\text{https://seekingalpha.com/page/about_us}\] Very few opinions were contributed in 2004 after Seeking Alpha’s launch. As a result the analyses in this paper uses data beginning from January 2005.
through Seeking Alpha’s editorial team who checks for quality standards without interfering with the author’s viewpoint. It turns out that most users of Seeking Alpha rather consume and comment on the views of a smaller subset of individuals who do contribute opinion articles and belief statements. Hence, SMAs in this paper refer to this subset of individuals that contribute opinions that shape their followers’ beliefs and actions on Seeking Alpha.

To obtain SMAs’ opinions and stated beliefs from Seeking Alpha, I develop a web-scraping algorithm used to download all opinion articles (and the associated belief statements where available) covering a single US common stock listed on either the NYSE, NASDAQ or AMEX stock exchanges. For each publication, I obtain the SMA’s ID, stock ticker the publication refers to, publication date, and the SMA’s disclosure of investment position in the stock. Furthermore, I retrieve all comments posted in response to the publication by other users of Seeking Alpha. For the comments, I also obtain the author ID and publication date.

In total, I downloaded 257,442 single-ticker opinion articles and 7.3 million comments, contributed by roughly 11,000 SMAs and 300,000 users respective, covering about 7,200 firms over my sample period. SMA belief statement that accompanies each publication is either “Very Bullish”, “Bullish”, “Neutral”, “Bearish”, or “Very Bearish”. However, most publications before 2018 did not explicitly state the SMA’s belief, as it was not part of Seeking Alpha’s requirements at the time. I, therefore, use a machine learning model, described in Section 3, to extract SMAs’ underlying beliefs for the articles without explicit belief statement.

Figures A1 and A2 of the appendix show sample SMA opinion articles where the authors explicitly state their beliefs about a stock as “Bullish” and “Bearish” respectively. From these samples, it can be deduced that individuals that contribute beliefs and opinions to Seeking Alpha are generally more financially literate and sophisticated than the average retail investor.

Footnotes:
12 Most publications include a disclosure section where the author discloses whether he/she has an investment position in the stock being written about. See Figures A1 and A2 for examples of these disclosures. I manually label a randomly selected 5,000 disclosures as either “Long position”, “Short position” or “No position” and then use this labeled sample to train an Support Vector Classifier ML model, as described in Section 3, used to extract the investment position stated in all other disclosures. Given the ease of this particular learning exercise, the trained model achieved an out-of-sample accuracy 99%.
13 Note, these numbers include only publications that cover a single US common stock listed in either the NYSE, NASDAQ or AMEX stock exchange.
2.2 Other data

I obtain stock returns, price and market capitalization data from CRSP, while firm fundamentals data is from Compustat. I compute abnormal return $ABR_{k,t}(h)$ for firm $k$ on day $t$ for investment horizon $h$ relative to three different benchmarks: the Capital Asset Pricing Model (CAPM), Fama and French (1993) three-factor model (FF3), and the Daniel et al. (1997) size/book-to-market/momentum characteristics-based benchmark (SBM). For the CAPM and FF3 benchmarks, betas for each stock are estimated using daily data over the trading-day window $t-272$ to $t-21$, where $t$ is the relevant event day.

Data on professional stock analyst recommendations and forecast of quarterly earnings per share (EPS) are from Institutional Brokers’ Estimate System (IBES). I use analyst recommendations data to compute the number of recommendation upgrades ($Upgrade_{k,t}$) and downgrades ($Downgrade_{k,t}$) for firm $k$ on day $t$. From the unadjusted detail history of analysts’ earnings forecasts, I compute earnings surprise ($SEU_{k,t}$) as the difference between actual EPS and the average forecasts across analysts (consensus estimate) divided by the stock price at the end of the last quarter. To avoid stale forecasts, I use only forecast published within the 30-day period ending one day before the earnings announcement day. Dispersion of earnings forecasts is the standard deviation of the analyst forecasts over the same 30-day window scaled by stock price at the end of the previous quarter.

Finally, I measure sentiment across a comprehensive set of cash flow relevant news events about a stock on a given day using the Event Sentiment Score (ESS) from RavenPack News Analytics. ESS ranges between 0 and 100, where 50 indicates neutral sentiment, values above 50 indicate positive sentiment and values below 50 indicate negative sentiment. I use only news events with a relevance score of at least 75 in order to focus on news events that mostly relate to a specific firm. Finally, for each firm-day pair, I average ESS across all relevant news events and divide by 100. Table A1 of the appendix lists the cash flow relevant news categories used in this paper.
3 Measuring SMA Beliefs

Recent applications of machine learning (ML) and natural language processing (NLP) techniques in finance and economics (e.g., Chen et al., 2019; Ke et al., 2019; Manela and Moreira, 2017) provide evidence that textual data can be used to generate important economic quantities. I build on these papers and use the sub-set of Seeking Alpha opinion articles that include belief statements (labeled samples) to train a supervised ML model used to extract the underlying beliefs for all other articles that do not state the author’s belief. The result is a large sample of stated and extracted beliefs that allows the study of SMAs’ beliefs about individual stocks.

I begin by first processing the raw article text to reduce the vocabulary to a set of terms that would be relevant for the belief classification. The steps involve converting all words to lower case, removing highly frequent words (stopwords), words containing digits, and punctuation. The texts are then lemmatized so that words that represent the same underlying concept are captured by the same word token, while avoiding ambiguity.\textsuperscript{14} The resulting clean text is finally broken into unigrams (single words) and bigrams (consecutive combinations of two words) – both referred to as \textit{n}-grams in this paper.\textsuperscript{15} Afterwards, the labeled samples are converted into an $N \times M$ document-term-matrix $X$, omitting n-grams that occur in less than 3\% of the labeled article corpus. The element in row $n$ and column $m$ of $X$ captures the number of times n-gram $m$ appeared in article $n$. $X$ is a high-dimensional sparse matrix, because its columns comprise all the unique n-grams in the entire labeled articles corpus but only a small fraction of these n-grams appears in each $n$'th article. Finally, term frequency-inverse document frequency (tf-idf) normalization is applied to $X$ such that the entries of the matrix reflect the importance of each $n$-gram to a particular article while adjusting for the fact that some $n$-grams occur frequently in most articles.

I use the tf–idf-normalized matrix $X$ as the features in the machine learning classification task aimed at finding the optimal weights for combining the normalized $n$-gram frequencies.

\textsuperscript{14}Lemmatization groups the inflected forms of a word so they can be analyzed as a single term identified by the word’s lemma, or dictionary form. The alternative is stemming, which reduces inflected words to their word stem. However, stemming has the disadvantage of generating ambiguities where different concepts appear related.

\textsuperscript{15}I experimented with using only single words (unigrams) or unigrams, bigrams and trigrams together. While unigrams produce slightly worse out-of-sample performance than unigrams and bigrams combined, including trigrams to the former yields approximately similar performance.
to give the best out-of-sample (OOS) prediction of SMA beliefs. I adopt the linear Support Vector Classifier (SVC) machine learning algorithm for this exercise because it has been shown to perform well in very high dimensional feature spaces as in this paper (e.g., Chen et al., 2019; Manela and Moreira, 2017; Frankel et al., 2016). Moreover, linear SVC is relatively fast to train on high dimensional spaces and the output can be easily interpreted to understand what word combinations matter for beliefs. Although, penalized logistic regression provides similar features as linear SVC, I find that SVC has better out-of-sample performance which further motivated its use in this paper. Appendix A.1 provides a description of the SVC algorithm.

To apply SVC to my setting, I first collapse the belief labels to three classes, setting the “Very Bullish” and “Bullish” labels to “Bullish”, and setting the “Very Bearish” and “Bearish” labels to “Bearish”; the last label is the “Neutral” label. This reduces the problem of imbalanced data, since the “Very Bullish” and “Very Bearish” belief labels together account for only 3% of the labeled data. As Panel A of Table 1 shows, the belief classes are still highly imbalanced even after the relabeling. To further resolve the problem, I use a conventional ML approach, Synthetic Minority Over-sampling Technique (SMOTE), to obtain a balanced training data by generating synthetic labeled observations for the minority classes.\footnote{SMOTE creates synthetic (not duplicate) samples of the minority classes by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. SMOTE has been shown to outperform alternative data balancing techniques such as random oversampling of the minority classes and undersampling of the majority classes (see, Chawla et al., 2002). I use the default value of k = 5 nearest neighbors.}

Training the linear SVC involves choosing the n-gram weights and regularization parameter $c$ that produce the best OOS belief classification performance. I use grid search over a range of $c$ values, five-fold cross-validation, and OOS accuracy for model selection. The grid search and five-fold cross-validation proceeds by randomly splitting the training data into five equal groups. Then for a given value of $c$ the SVC is trained on folds one to four, while the held-out fold five (test set) is used to compute an OOS accuracy score. The process is repeated until each of the five folds is used once as the test set for the given $c$. The resulting five OOS accuracy scores are averaged and saved and the iteration moves to the next value for $c$. In the end, I select the $c$ parameter value that gives the highest average OOS accuracy score. The resulting SVC classifier achieved an average OOS accuracy of 93%.
As a sanity check, Figure 1 shows the unigrams and bigrams assigned the highest absolute weights by the SVC classifier for the “Bullish” and “Bearish” belief classes. The most important terms for “Bullish belief” include “Game Changer”, “Overvalue”, “Remain Bullish”, “Take Profit”, and “Bode Well”. For “Bearish Belief”, the most important terms include “Add Position”, “Strong Fundamental”, “Buying Opportunity”, “Significant Discount”, and “Under-value”. The picture suggests that the trained SVC model produces intuitive results. Furthermore, bigrams are important for correctly extracting beliefs. Panel A of Table 1 shows the proportion of each belief class extracted by the model.

Finally, using both stated and extracted beliefs, I aggregate individual SMA beliefs about stock \( k \) on day \( t \) by subtracting the number of Bearish beliefs \( NBearish_{k,t} \) from the number of Bullish beliefs \( NBullish_{k,t} \) then normalize by the total number of belief statements \( NBelief_{k,t} \) including Neutral beliefs:

\[
AB_{k,t} = \frac{NBullish_{k,t} - NBearish_{k,t}}{NBelief_{k,t}} \quad (1)
\]
Clearly, the stock-specific aggregate SMA belief \( AB_{k,t} \) ranges between -1 and +1, and is increasing in the bullishness of beliefs. I also consider alternative aggregation measures: 

\[
AB_{1k,t} = \log \left( \frac{1 + NBullish_{k,t}}{1 + NBearish_{k,t}} \right) \quad \text{and} \quad AB_{2k,t} = NBullish_{k,t} - NBearish_{k,t}.
\]

The results are robust to these alternative measures. Figure 2 depicts monthly times series of the cumulative stock market return over the past 12-month period \( R_m \) and the aggregate SMA beliefs averaged across stocks and within each month \( \overline{AB} \). The figure shows that Average SMA belief has rich dynamics and tends to be less bullish when there is a huge drop in the market return over the past year. Regressing \( \overline{AB} \) on a constant and \( R_m \) yields a coefficient of 0.29 \( (t\text{-statistic} = 4.7) \), implying that a standard deviation increase in the past 12-month market return results in average beliefs becoming more bullish by 41.7% of its standard deviation. The main takeaway from the figure is that SMA beliefs are quite similar to beliefs based on survey data in that they tend to be extrapolative, procyclical and, therefore, inconsistent with the implications of rational expectations representative agent models (see e.g., Greenwood and Shleifer, 2014).

Panel B of Table 1 shows summary statistics for the stock-level aggregate SMA beliefs, alongside the abnormal returns for the 5-day (1 week), 21-day (1 month) and 63-day (3 months) investment horizons. The table also shows summary statistics for firms’ market capitalization (Size). The statistics are computed cross-sectionally at each date then averaged over the time-series. The average and median of aggregate beliefs are positive, implying that beliefs are generally more bullish. The average abnormal returns are all positive, while the average market capitalization is $49.6 billion over the sample period.

4 Are SMA Beliefs Informative?

I examine the informativeness of SMA beliefs by looking at whether stock-level aggregate beliefs predict stock abnormal returns in subsection 4.1 and earnings surprise in subsection 4.2.
Figure 2: Average SMA Belief and Past Stock Market Return. The figure shows the time series of the monthly average of aggregate belief ($\bar{AB}$) and the stock market return over the past 12-month period $t - 12$ to $t$ ($R_m$).

Table 1: Summary Statistics of SMA Beliefs, Abnormal Returns and Size. The table reports, in Panel A, the proportion of each belief class in the sub-sample of SMA articles with belief statement; in the sub-sample of extracted beliefs; and finally the entire sample of SMA articles. Panel B shows time series average of cross-sectional summary statistics for the three measures of stock-level aggregate belief ($AB$, $AB1$ and $AB2$); CAPM abnormal returns $ABR(h)$ over horizon $h$ trading days beginning $t + 1$, where $t$ is the belief publication day; and Size measured as stock market capitalization (in millions USD) on day $t$. The sample period is 2005 – 2019.

<table>
<thead>
<tr>
<th><a href="#">Panel A: Proportion of Beliefs</a></th>
<th>Stated Beliefs</th>
<th>Extracted Beliefs</th>
<th>All Beliefs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bullish</td>
<td>0.82</td>
<td>0.63</td>
<td>0.73</td>
</tr>
<tr>
<td>Bearish</td>
<td>0.13</td>
<td>0.26</td>
<td>0.19</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.05</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>N</td>
<td>117,366</td>
<td>119,761</td>
<td>237,127</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><a href="#">Panel B: Aggregate Belief and Abnormal Returns</a></th>
<th>$AB$</th>
<th>$AB1$</th>
<th>$AB2$</th>
<th>$ABR(5)$</th>
<th>$ABR(21)$</th>
<th>$ABR(63)$</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.52</td>
<td>0.38</td>
<td>0.59</td>
<td>0.0006</td>
<td>0.0009</td>
<td>0.0014</td>
<td>49,593</td>
</tr>
<tr>
<td>SD</td>
<td>0.75</td>
<td>0.56</td>
<td>0.90</td>
<td>0.0629</td>
<td>0.1177</td>
<td>0.2078</td>
<td>90,125</td>
</tr>
<tr>
<td>P5</td>
<td>-0.79</td>
<td>-0.56</td>
<td>-0.82</td>
<td>-0.0743</td>
<td>-0.1450</td>
<td>-0.2598</td>
<td>1,051</td>
</tr>
<tr>
<td>P10</td>
<td>-0.64</td>
<td>-0.45</td>
<td>-0.65</td>
<td>-0.0520</td>
<td>-0.1054</td>
<td>-0.1926</td>
<td>1,317</td>
</tr>
<tr>
<td>P25</td>
<td>0.27</td>
<td>0.21</td>
<td>0.31</td>
<td>-0.0237</td>
<td>-0.0500</td>
<td>-0.0955</td>
<td>2,793</td>
</tr>
<tr>
<td>P50</td>
<td>0.89</td>
<td>0.62</td>
<td>0.90</td>
<td>-0.0004</td>
<td>-0.0016</td>
<td>-0.0054</td>
<td>11,604</td>
</tr>
<tr>
<td>P75</td>
<td>0.96</td>
<td>0.67</td>
<td>0.97</td>
<td>0.0228</td>
<td>0.0461</td>
<td>0.0845</td>
<td>54,054</td>
</tr>
<tr>
<td>P90</td>
<td>0.97</td>
<td>0.72</td>
<td>1.11</td>
<td>0.0525</td>
<td>0.1032</td>
<td>0.1886</td>
<td>145,462</td>
</tr>
<tr>
<td>P95</td>
<td>0.97</td>
<td>0.84</td>
<td>1.41</td>
<td>0.0774</td>
<td>0.1512</td>
<td>0.2732</td>
<td>203,237</td>
</tr>
</tbody>
</table>

Electronic copy available at: https://ssrn.com/abstract=3813252
4.1 Evidence from Individual Stock Returns

To examine whether SMA beliefs are informative about stock returns, I employ the following regression specification:

\[
ABR_{k,t+1,t+1+h} = \beta_0 + \beta_1 ABR_{k,t} + X \Gamma + \epsilon_{k,t}
\]  

(2)

where \(ABR_{k,t+1,t+1+h}\) is the future abnormal return of stock \(k\) over horizon \(h \in \{5, 21, 63\}\) trading days. Abnormal returns are computed as described in Section 2.2, and the return calculation horizon starts from day \(t + 1\) to avoid potential complications arising from the time of day when beliefs where published. \(ABR_{k,t}\) is the aggregate belief about stock \(k\) on day \(t\), where \(t\) is the belief publication day in Seeking Alpha. \(X\) captures the following control variables: current abnormal return \((ABR_{k,t})\); past abnormal returns \((ABR_{k,t-1}, ABR_{k,t-2}, \text{ and } ABR_{k,t-h\rightarrow t-3})\); \(Volatility_{k,t}\) measured as the sum of squared daily returns in the calendar month prior to day \(t\); the number of professional stock analysts upgrading \((Upgrade_{k,t})\) and downgrading \((Downgrade_{k,t})\) stock \(k\) on day \(t\); the News Event Sentiment Score \((ESS_{k,t})\); and year-month fixed effects. If there are no upgrades (downgrades) for firm \(k\) on day \(t\), \(Upgrade_{k,t}\) (\(Downgrade_{k,t}\)) is set to zero. If there are no news events for a firm on a given day, \(ESS_{k,t}\) is set to its neutral value of 0.5. All continuous right-hand side variables are standardized to unit variance. The standard errors are clustered by firm and year-month to account for serial correlation, cross-correlation and heteroscedasticity.

Table 2 shows the results of the regression. The column headers indicate the benchmark used to compute abnormal returns: CAPM (Columns 1 and 2); Fama and French (1993) three-factor model, FF3 (Columns 3 and 4); and the Daniel et al. (1997) size/book-to-market/momentum characteristic-based benchmark, SBM (Columns 5 and 6). Coefficients for the control variables are suppressed in the table for brevity; they are shown in Table A2 of the appendix for the 3-month future horizon. Starting with Panel A, it can be seen that SMA beliefs significantly predict future abnormal returns over the 5-day horizon following belief statement. In particular, if beliefs become more bullish by one standard deviation, abnormal returns increase by 16 basis points (bps), with \(t\)-statistic of around 6.8, over the next five trading days. The coefficients
are stable across different abnormal return benchmarks, as well as with and without control variables.

<table>
<thead>
<tr>
<th>Panel A: 5 days</th>
<th>CAPM</th>
<th>FF3</th>
<th>SBM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>(AB_{k,t})</td>
<td>0.0016</td>
<td>0.0016</td>
<td>0.0016</td>
</tr>
<tr>
<td>(6.71)</td>
<td>(6.87)</td>
<td>(6.86)</td>
<td>(7.00)</td>
</tr>
<tr>
<td>Controls</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Obs.</td>
<td>194,490</td>
<td>194,490</td>
<td>194,490</td>
</tr>
<tr>
<td>(R^2) (%)</td>
<td>0.62</td>
<td>0.75</td>
<td>0.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: 1 month</th>
<th>CAPM</th>
<th>FF3</th>
<th>SBM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>(AB_{k,t})</td>
<td>0.0026</td>
<td>0.0025</td>
<td>0.0026</td>
</tr>
<tr>
<td>(4.91)</td>
<td>(4.78)</td>
<td>(5.55)</td>
<td>(5.28)</td>
</tr>
<tr>
<td>Controls</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Obs.</td>
<td>193,839</td>
<td>193,839</td>
<td>193,839</td>
</tr>
<tr>
<td>(R^2) (%)</td>
<td>1.65</td>
<td>1.72</td>
<td>1.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel C: 3 months</th>
<th>CAPM</th>
<th>FF3</th>
<th>SBM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>(AB_{k,t})</td>
<td>0.0028</td>
<td>0.0026</td>
<td>0.0030</td>
</tr>
<tr>
<td>(2.15)</td>
<td>(2.06)</td>
<td>(2.46)</td>
<td>(2.34)</td>
</tr>
<tr>
<td>Controls</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Obs.</td>
<td>192,968</td>
<td>192,968</td>
<td>192,968</td>
</tr>
<tr>
<td>(R^2) (%)</td>
<td>1.98</td>
<td>2.03</td>
<td>1.37</td>
</tr>
</tbody>
</table>

Table 2: SMA Beliefs and Future Stock Abnormal Returns The table reports results for the panel regression of future abnormal returns \(ABR_{k,t+1\rightarrow t+1+h}\) on stock-level aggregate SMA beliefs \(AB_{k,t}\). Abnormal return for firm \(k\) over horizon \(t + 1\) to \(t + 1 + h\), with \(t\) being the belief publication day, is computed relative to either the CAPM, Fama and French (1993) three-factor model (FF3) or a value-weighted portfolio of firms with similar size, book-to-market and momentum characteristics (SBM) as in Daniel et al. (1997). Panel A reports results for horizon \(h = 5\) trading days (1 week), Panel B reports results for \(h = 21\) trading days (1 month), and Panel C results for \(h = 63\) trading days (3 months). All regressions include year-month fixed effects. Other Control variables included in columns (2), (4) and (6) are past abnormal returns: \(ABR_{k,t-1}\); \(ABR_{k,t-2}\); \(ABR_{k,t-h-3\rightarrow t-3}\); \(Volatility_{k,t}\); the number of professional stock analysts upgrading \((Upgrage_{k,t})\) and downgrading \((Downgrade_{k,t})\) stock \(k\) on day \(t\); and the News Event Sentiment Score \((ESS_{k,t})\). All right hand side variables are normalized to unit variance. Values in parentheses are \(t\)-statistics based on standard errors clustered by both firm and year-month. The sample period is 2005 – 2019.

Panels B and C show results for the 1-month and 3-month future horizons. Interestingly, the predictability does not reverse even after three months. A standard deviation increase in the bullishness of beliefs is associated with approximately 30 bps increase in abnormal returns, with \(t\)-statistic between 2.1 and 2.5, over the next three months following belief statement. In untabulated results, I examine whether investor beliefs predict abnormal returns over the 6-month and 12-month future horizons. While the coefficients in the regressions are insignificant,
I do not find evidence of reversal in predictability which is consistent with investor beliefs containing value-relevant information rather than sentiment unrelated to fundamental or noise.

One could argue that SMA beliefs appear informative because they simply piggyback on other fundamental-relevant news that drive returns. For instance Tetlock et al. (2008) show that the fraction of negative words in the news predict stock returns, while Barber et al. (2001) show that analyst recommendations are informative about future returns. I take care of this concern by directly controlling for these news sources that may drive returns around the same time beliefs are expressed in Columns (2), (4) and (6) of Table 2. In particular, I control for the number of professional stock analysts upgrading and downgrading a given stock, as well as sentiment in cash flow news published about a stock on the same day as belief publication. That the results remain significant after these and other controls that account for momentum and volatility is strong evidence that SMA beliefs are informative about future abnormal returns beyond the information contained in major news outlets and the recommendations of professional analysts. The result suggests that on aggregate SMAs have value for (retail) investors that pay attention to their views.

To drive home the point, I examine the kind of stocks for which SMA beliefs are more informative. The idea is to see if SMAs’ belief informativeness is predominantly on stocks mostly traded by retail investors. I proceed by sorting stocks into deciles on each day based on the following stock characteristics computed as of end of the last calendar month: turnover, price, size, book-to-market ratio, volatility and idiosyncratic volatility. I then repeat regression (2) but now including an interaction of stock aggregate belief $AB_{k,t}$ and the decile rank of the stock on a given characteristic in the specification. Results reported in Table A3 of the appendix confirm that, indeed, SMA beliefs are more informative about difficult to value and hard to arbitrage firms that tend to attract less professional analyst following and more retail trade.\footnote{For some numbers on the kind of stocks retail investors trade, see https://www.nasdaq.com/articles/what-types-of-stocks-do-retail-investors-trade-2019-08-08} In particular, SMA beliefs have stronger predictive power for low turnover firms, low price firms, small firms, high book-to-market ratio firms; firms with high volatility and high idiosyncratic volatility. Besides providing value for retail investors, the result potentially points towards the
informational value of SMAs to the market as whole, given that they provide information on segments of the market where information is generally scarce.

Next, I explore the financial implication of trading in the direction of SMA beliefs using portfolios formed based on beliefs. At the end of day \( t-1 \), I sort stocks into tercile portfolios based on the stock-level aggregate SMA beliefs averaged over the past 30-day period (\( \overline{AB}_{k,t-1} \)). Given that \( \overline{AB}_{k,t} \) ranges between -1 and +1, the tercile portfolios are formed such that portfolio one (“Low”) consists of stocks with the least bullish beliefs, for which \( \overline{AB}_{k,t-1} \leq -0.5 \); portfolio two contains stocks for which \(-0.5 < \overline{AB}_{k,t-1} \leq 0.5 \); and portfolio three (“High”) contains stocks with the most bullish beliefs, \( \overline{AB}_{k,t-1} > 0.5 \). Next, I compute the portfolios’ value-weighted returns on the next trading day \( t \), using the \( t-1 \) normalized market capitalization as weights. Finally, the daily portfolio returns are cumulated to the monthly level.

Figure 3 shows the evolution of $1 invested in the “High” and “Low” portfolios, and for comparison the market portfolio (MKT). Clearly, the portfolio of stocks with the most bullish beliefs (“High”) outperforms both the market and the portfolio perceived less favorably (“Low”) by SMAs. On the other hand, the “Low” portfolio markedly underperforms the market by about 75% as at December 2019. The pictorial evidence is supported by formal tests of the significance of the portfolios’ abnormal returns relative to the CAPM, FF3 and FF3+Momentum (FF4) factor benchmarks. As Table 3 shows, while the “High” portfolio has CAPM (FF4) alpha of 14 (10) bps per month, the “Low” portfolio has significantly negative alpha. Moreover the High minus low (“High - Low”) portfolio has significant and economically sizeable CAPM (FF4) abnormal returns of 43 (37) bps per month, equivalent to 5.1 (4.3) percent per annum.

In untabulated results, I further examine how the belief-based portfolios perform with less frequent portfolio rebalancing. Indeed, rebalancing less frequently sacrifices significant portions of the portfolio abnormal returns. For instance, switching from daily to weekly rebalancing reduces the difference portfolio’s CAPM (FF4) alpha to 34 (27) bps per month, with \( t \)-statistics of 2.21 (1.67) suggesting only marginal significance for the FF4 alpha. Nevertheless, these results indicate that the informativeness of SMA beliefs about future stock returns, beyond systematic risk exposures, is not restricted to specific sample periods. Trading in the direction of average SMA beliefs can generate superior risk-adjusted portfolio returns for investors.
Figure 3: Evolution of Belief-based Portfolios. The figure shows the monthly evolution of $1 invested in the portfolio of stocks with most bullish beliefs (“High”), the portfolio of stocks with the least bullish beliefs (“Low”), and the market portfolio (MKT).

Table 3: Belief-based Portfolio Returns. The table reports average monthly returns and abnormal returns (alpha) of tercile portfolios based on SMA beliefs. P3 (“High”) is the portfolio of stocks with the most bullish beliefs and P1 (“Low”) is portfolio of stocks SMAs view least favorably. High - Low is the long-short portfolio that is long P3 and short P1. Reported in parentheses are t-statistics based on the Newey and West (1987) method. The sample period is 2006 – 2019.

The results so far show that SMA beliefs do predict future (abnormal) returns reliably. It is, however, not quite convincing whether the predictability is because beliefs contain value-relevant information or because SMAs’ views on social media generate trades that drive stock prices in the direction of beliefs. To determine if the predictive ability of beliefs is due to value-relevant information, I follow similar approach as Chen et al. (2014) and Tetlock et al. (2008) and examine if beliefs predict future earnings surprise. The idea behind the analysis is that since it
is implausible that SMA beliefs drive either firm earnings or consensus analyst forecast, evidence that beliefs predict earnings surprise supports the view that SMA beliefs contain value-relevant information.

4.2 Evidence from Earnings Surprise

I use the following regression specification to test whether SMA beliefs are informative about firms’ earnings:

\[ SUE_{k,t} = \beta_0 + \beta_1 AB_{k,t-30\rightarrow t-1} + X \Gamma + \epsilon_{k,t} \]  

where \( t \) and \( k \) denote earnings announcement day and firm respectively, \( SUE_{k,t} \) is earnings surprise computed as quarterly earnings per share minus consensus (average) professional analyst forecast divided by the stock price at the end of last quarter. To avoid stale forecasts, professional analyst forecast data includes only forecasts published/updated within the 30-day period ending one day before earnings announcement. \( AB_{k,t-30\rightarrow t-1} \) is SMA aggregate belief on firm \( k \) averaged over the 30-day period ending one day before earnings announcement. \( X \) captures the following control variables: lag of earnings surprise (\( \text{Lagged}(SEU_{k,t}) \)), price-scaled standard deviation of earnings forecasts (\( \text{Dispersion}_{k,t} \)), cumulative stock return over the past 30-day window (\( R_{k,t-30\rightarrow t-1} \)), log of market value as of end of last quarter (\( \text{Size}_{k,t} \)), log of book-to-market ratio as of the most recent fiscal year end (\( BM_{k,t} \)), news event sentiment score averaged over the past 30-day window (\( ESS_{k,t-30\rightarrow t-1} \)), and year-month and single digit industry SIC code fixed effects. To mitigate the influence of outliers, earnings surprise is winsorized at the 99th percentile. All continuous regressors are normalized to unit-variance, and the standard errors are clustered by firm and year-month to account for serial correlation, cross-correlation and heteroscedasticity.

Table 4 shows results of the regression. Column (1) includes only industry and year-month fixed effects as controls, while Columns (2) – (4) progressively include additional sets of control variables. Depending on the set of controls, the coefficient of SMA belief (\( AB_{k,t-30\rightarrow t-1} \)) ranges between 0.06% and 0.08%. To understand the economic magnitude, earnings surprise has a
mean of 0.07%, median of 0.05% and standard deviation of 0.4%. Hence, when SMA beliefs become more bullish by one standard deviation, earnings surprises increase by between 15% and 20% of its standard deviation – an economically sizeable change. That SMA beliefs significantly predict future earnings surprise even after controlling for a bunch of characteristics and cash flow news sentiment strongly support the idea that SMA beliefs expressed on social media contain value-relevant information.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(AB_{k,t-30\rightarrow t-1})</td>
<td>0.0008</td>
<td>0.0006</td>
<td>0.0006</td>
<td>0.0006</td>
</tr>
<tr>
<td></td>
<td>(3.10)</td>
<td>(2.84)</td>
<td>(2.54)</td>
<td>(2.52)</td>
</tr>
<tr>
<td>(Lagged(SEU_{k,t}))</td>
<td>0.0084</td>
<td>0.0084</td>
<td>0.0084</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(9.48)</td>
<td>(9.55)</td>
<td>(9.55)</td>
<td></td>
</tr>
<tr>
<td>Dispersion_{k,t}</td>
<td>-0.0016</td>
<td>-0.0016</td>
<td>-0.0016</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-2.85)</td>
<td>(-2.82)</td>
<td>(-2.82)</td>
<td></td>
</tr>
<tr>
<td>(R_{k,t-30\rightarrow t-1})</td>
<td>0.0015</td>
<td>0.0015</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.81)</td>
<td>(3.80)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size_{k,t}</td>
<td>-0.0002</td>
<td>-0.0002</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-0.77)</td>
<td>(-0.83)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BM_{k,t}</td>
<td>-0.0003</td>
<td>-0.0003</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-1.35)</td>
<td>(-1.35)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(ESS_{k,t-30\rightarrow t-1})</td>
<td></td>
<td></td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.06)</td>
</tr>
<tr>
<td>(R^2) (%)</td>
<td>2.22</td>
<td>15.75</td>
<td>16.12</td>
<td>16.12</td>
</tr>
<tr>
<td>Obs.</td>
<td>19,083</td>
<td>18,716</td>
<td>18,716</td>
<td>18,716</td>
</tr>
</tbody>
</table>

Table 4: SMA Beliefs and Earnings Surprise. The table reports results for the panel regression of price-scaled earnings surprise \((SEU_{k,t})\) on stock-level SMA beliefs averaged over the 30-day window from \(t-30\) to \(t-1\) \((AB_{k,t-30\rightarrow t-1})\), where \(t\) is the earnings announcement day. \(Lagged(SEU_{k,t})\) is the lag of price-scaled earnings surprise. Dispersion_{k,t} is the price-scaled standard deviation of earnings forecasts. \(R_{k,t-30\rightarrow t-1}\) is cumulative return over the 30 days window from \(t-30\) to \(t-1\). Size_{k,t} is log of market value as at June of the previous calendar year; \(BM_{k,t}\) is log of book-to-market ratio as of the most recent fiscal year end. \(ESS_{k,t-30\rightarrow t-1}\) is the News Event Sentiment Score averaged over the 30 days window from \(t-30\) to \(t-1\). Regressors are normalized to unit-variance, and all regressions include year-month and single digit industry SIC code fixed effects. Values in parentheses are \(t\)-statistics based on standard errors clustered by both firm and year-month. The sample period is 2005 – 2019.

Overall, the results of this section suggest that earlier results on stock returns are unlikely to be driven purely by the price impact of trades following SMAs’ belief statements; it, therefore, reinforces the argument that SMAs on aggregate have value for (retail) investors. It is, however, important to note that this conclusion does not easily extend beyond the group of nonprofessionals considered as SMAs in this paper. In particular, SMAs do not include the mostly unsophisticated crowd in some segments of social media who express views that are not backed by some extended fundamental analyses.
While it is reassuring to find that SMA beliefs do provide value-relevant information, it is still unclear whether the average informativeness of beliefs arises from luck or because these SMAs have the ability to consistently form correct beliefs. Put differently, are SMAs’ beliefs informative because the vast majority of SMAs are skilled or is the result driven by a tiny fraction of SMAs skilled in forming correct beliefs? Understanding the distribution of skill among SMAs would shed more light on how (retail) investors should approach SMAs’ views, as well as on SMAs’ role on market efficiency.

5 Modeling Belief Formation Ability as a Mixture Distribution

To understand the distribution of SMAs’ belief formation ability, I follow recent literature that studies the skill of financial professionals with the aim of avoiding false discovery (Crane and Crotty, 2020; Harvey and Liu, 2018; Chen et al., 2017) and model belief-formation ability as arising from a mixture distribution of multiple skill groups. Modeling SMA skill as a mixture distribution allows me avoid common pitfalls that arise from the low signal-to-noise feature of estimated abnormal returns – the common measure of unobservable skill. Noise in estimated abnormal returns can result in conventional individual significance tests misjudging good luck for skill or bad luck for lack of skill. Such tests also suffer from low power that makes it difficult to separate skill from luck. On the other hand, the mixture distribution model can use information from the cross-section of SMA performance to reduce noise and is not impeded by low test power.

The approach in this paper closely follows Crane and Crotty (2020). I begin by assuming that there is an unknown number $J$ of skill groups. For each group $j \in \{0, 1, 2, ..., J\}$, there is a fraction $\pi_j$ of SMAs with true belief formation ability, captured by abnormal returns, centered on $\mu_j$. The dispersion of true abnormal returns for SMAs in group $j$ is driven by variation in true ability, say, investor specific traits. Let $\alpha_i = \mu_j + \omega_i$ denote true belief formation ability of SMA

---

18Belief formation skill could arise from SMAs’ experience, expertise, information-processing capabilities, or literacy levels.

19Papers such as Bajgrowicz and Scaillet (2012) and Barras et al. (2010) use a different approach, the False Discovery Rate (FDR) method, to differentiate luck from skill. However, Andrikogiannopoulou and Papakonstantinou (2019) show that the FDR method suffers from significant power problems due to the low signal-to-noise ratio in returns data.
\( \omega_i \) captures individual-specific traits and is normally distributed with zero mean and variance \( \sigma_j^2 \). On the other hand, estimated ability, \( \hat{\alpha}_i \), is measured with noise, \( e_i \), which is assumed to be independent of \( \omega_i \) and normally distributed with zero mean and variance \( s_i^2 \) (i.e., \( s_i \) is the standard error of estimated alpha). Thus, the estimated abnormal performance of an SMA is \( \hat{\alpha}_i = \mu_j + \omega_i + e_i \). Setting \( J = 2 \), following model selection criteria to be discussed shortly, the specifications boil down to a two-component distribution for SMA \( i \)'s belief formation ability with the following density function:

\[
f(\hat{\alpha}_i) = \pi_0 \cdot \phi(\hat{\alpha}_i; \mu_0, \sigma_{i,0}) + \pi_1 \cdot \phi(\hat{\alpha}_i; \mu_1, \sigma_{i,1})
\]

(4)

where \( \phi(\hat{\alpha}_i; \mu_j, \sigma_{i,j}) \) is the normal density function with mean \( \mu_j \) and variance \( \sigma_{i,j}^2 = \sigma_j^2 + s_i^2 \) evaluated at \( \hat{\alpha}_i \).\(^{20}\) The log likelihood function \( L \) for a sample of \( N \) estimated SMA belief-formation ability is

\[
L(\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_N|s_1, s_2, ..., s_N, \Theta) = \sum_{i=1}^{N} \log(f(\hat{\alpha}_i))
\]

(5)

where the parameter set \( \Theta = \{\pi_0, \pi_1, \mu_0, \mu_1, \sigma_0, \sigma_1\} \) is estimated via maximum likelihood subject to the constraints that \( 0 \leq \pi_0 \leq 1, \pi_1 = 1 - \pi_0 \), and \( \sigma_j \geq 0 \). The distribution is ordered such that \( \mu_0 \leq \mu_1 \), which implies that SMAs in skill group \( j = 1 \) have higher belief formation ability than those in group \( j = 0 \).

I use the bootstrap-based likelihood ratio test and the Bayesian information criterion (BIC) to select the number of components in the mixture model. To perform the likelihood ratio test, let \( M_0 \) and \( M_1 \) be two models, with \( M_0 \) being a constrained version of \( M_1 \). I estimate the parameter sets \( \Theta_0 \) and \( \Theta_1 \) for models \( M_0 \) and \( M_1 \) respectively, and compute the likelihood ratio

Assuming that the component distributions are normal allows for easy interpretability of the model parameters. For instance, one could view true skill, \( \alpha_j \), as the sum of several random investor characteristics, which approaches normal distribution under the central limit theorem. More so, even though component distributions are assumed normal, the composite distribution is not necessarily normally distributed, but is rather fit to best characterize the data.

\(^{20}\)Electronic copy available at: https://ssrn.com/abstract=3813252
test statistic (LRTS):

\[ LRTS = 2(L(\hat{\Theta}_1) - L(\hat{\Theta}_0)) \]  

The distribution of LRTS under the null of \( M_0 \) being the true model is obtained via bootstrapping. I simulate 1,000 samples using the estimated parameter set \( \hat{\Theta}_0 \). Models \( M_0, M_1 \) and the LRTS is estimated for each bootstrap sample, yielding a distribution of LRTS. The null hypothesis that \( M_0 \) is the true model is rejected if the empirical LRTS exceeds the 95% percentile of the bootstrapped distribution.

To use the BIC for model selection, I compute the sample-size-adjusted BIC (Sclove, 1987; Schwarz, 1978) for a model with \( c \) components as

\[ BIC_c = -2 L_c(\Theta) + p_c \log((N + 2)/24) \]  

where \( L_c(\Theta) \) and \( p_c \) are the log-likelihood function and number of estimated parameters in the model, and \( N \) is the number of observations. The model with the lowest BIC value is selected as better characterizing the data.

To take the mixture model to the data, I use 5-day abnormal return relative to the CAPM as the measure of estimated SMA belief formation ability.\(^{21}\) I proceed by first computing abnormal returns \( ABR_{it}^k \) over window \( t+1 \) to \( t+6 \) trading days for each belief statement by SMA \( i \) about stock \( k \) on day \( t \). As before, the return calculation window starts from \( t+1 \) to avoid potential complications arising from the time of day when belief statements where published. \( ABR_{it}^k \) is then signed by pre-multiplying it by +1 for bullish beliefs (long position) and -1 for bearish beliefs (short positions). Neutral beliefs are excluded because they do not provide clear investment signal.

Figure 4 shows the distribution of the signed abnormal returns \( ABR_{it}^k \) following bullish and bearish belief statements. The respective abnormal returns are on average positive which

\(^{21}\)I use 5-days abnormal returns because results in the section 4.1 show that return predictability of beliefs decline with horizon. The results are robust to using alternative benchmarks, such as the Fama and French (1993) three-factor model and the Daniel et al. (1997) characteristics-based benchmark, for computing abnormal returns.
compares favorably with results in the professional analyst recommendation literature (Crane and Crotty, 2020; Loh and Stulz, 2011); the histogram further shows that abnormal returns following bearish statements have more mass on the right, consistent with earlier results (e.g., Womack, 1996; Stickel, 1995) that sell recommendations tend to be more informative. Finally, I aggregate the abnormal returns to the SMA level by calculating the average across all belief statements, \( n_i \), by SMA \( i \) as follows:

\[
\overline{ABR}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} ABR^k_i
\]  

(8)

The main analysis uses only SMAs with at least five belief statements over the sample period, that is \( n_i \geq 5 \). \( \overline{ABR}_i \) is an SMA’s estimated belief formation ability, \( \hat{\alpha}_i \). Its standard error, \( s_i \) is calculated by clustering on belief statement date, to account for across-belief statement correlation among belief statements on the same date, and stock level, to account for across-SMA correlation in belief statements on the same stock. Summary statistics for \( \overline{ABR}_i \) and \( s_i \) are captured in Table 5, which shows that average estimated SMA ability is 27 basis points (bps), with a median of 11 bps. While these numbers seem economically meaningful, the standard errors \( s_i \), with an average (median) of 140 (97) bps suggest that the SMA-specific abnormal returns are measured with substantial noise. The reported skewness (1.5) and kurtosis (24.0) suggest that the abnormal returns is not normally distributed, while the Kolmogorov-Smirnov test strongly rejects normality at the 1% significance level. The implication is that standard significance tests based on normality can lead to incorrect inference about SMA skill, which validates the application of a mixture distribution model to isolate true belief formation skill among SMAs.

5.1 How Many SMAs Have Belief Formation Ability?

The parameters of the mixture model are first estimated with the restriction that \( \mu_0 = 0 \) and \( \sigma_0 = 0 \). This implies that the true belief formation ability of SMAs in component \( j = 0 \) is zero and the variation in their estimated performance is principally due to luck. \( \pi_0 \) therefore
Figure 4: Histogram of Abnormal Returns for Bullish and Bearish Beliefs. The figure shows the distribution of the signed abnormal stock returns $ABR_i^k$ following bullish and bearish belief statements by SMAs.

<table>
<thead>
<tr>
<th></th>
<th>$ABR_i^k$</th>
<th>$s_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.0027</td>
<td>0.0140</td>
</tr>
<tr>
<td>SD</td>
<td>0.0269</td>
<td>0.0165</td>
</tr>
<tr>
<td>P10</td>
<td>-0.0189</td>
<td>0.0031</td>
</tr>
<tr>
<td>P25</td>
<td>-0.0066</td>
<td>0.0054</td>
</tr>
<tr>
<td>P50</td>
<td>0.0011</td>
<td>0.0097</td>
</tr>
<tr>
<td>P75</td>
<td>0.0099</td>
<td>0.0167</td>
</tr>
<tr>
<td>P90</td>
<td>0.0251</td>
<td>0.0285</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.4879</td>
<td>5.8086</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>24.009</td>
<td>62.985</td>
</tr>
</tbody>
</table>

Obs. 4,145
Fraction positive 0.55
K-S p-value 0.00

Table 5: Summary Statistics for SMAs’ Abnormal Returns. The table reports summary statistics for the average 5-day abnormal returns, $ABR_i^k$, and its standard error, $s_i$, following SMA belief statements. Abnormal returns for belief statements result from buying stocks with bullish belief and selling stocks with bearish belief. The benchmark return for each event is based on CAPM. Abnormal returns are aggregated to the SMA level by estimating the average across all of an SMA’s belief statements. Standard errors, $s_i$, are clustered by both recommendation date, to account for across-publication correlation for publications on the same date, and by the underlying stock, to account for across-SMA correlation in beliefs on the same stock. The reported Fraction positive is the fraction of the SMA cross-section with positive abnormal return. The K-S p-value is the p-value of a Kolmogorov-Smirnov test of the null hypothesis that the demeaned cross-sectional distribution of $ABR_i$ is normally distributed. The sample period is 2005 – 2019.

captures the fraction of SMAs with zero belief formation ability, while $\pi_1$ captures the fraction of skilled SMAs.
Table 6 shows results for the constrained two-component mixture model. Panel A shows the estimated parameters for Component 0, i.e. unskilled group, in Column (1) and Component 1, i.e. skilled group, in Column (2). A large fraction, 75.5%, of SMAs is estimated to be unskilled, while 24.5% is estimated to have belief formation skill. The estimated average abnormal return for the skilled group is 39 bps, with a large dispersion of 260 bps. The large dispersion suggest that some SMAs in Component 1 actually have negative belief formation skill, implying that such SMAs form beliefs that generate negative abnormal returns. Combining the fraction of unskilled SMAs of Component 0 and those in Component 1 whose true abnormal returns are negative gives the total fraction of unskilled SMAs as 86%.\(^{22}\)

Panel B of Table 6 shows the abnormal return distribution implied by the constrained mixture model. The average abnormal return implied by the model is only 9 bps, with a standard deviation of 1.3%. Overall, the model estimates the fraction of skilled SMAs (Fraction positive) as only 14%. These estimates suggest that although aggregated SMA beliefs do appear to provide value-relevant information, as shown in Section 4.1, at the individual SMA level skill is quite scarce. To provide context, Crane and Crotty (2020), use a similar setup and estimate the fraction of skilled professional security analysts to be much larger at about 89%.

The preceding results depend on the assumption that investors in Component 0 of the mixture model have zero-ability. I test this restriction viz-a-viz unconstrained two-component and three-component mixture models using the likelihood ratio test and BIC statistic of Eq. (6) and (7). These tests suggest that the unconstrained two-component mixture model best fits the data. As a result, I re-estimate the mixture model without restrictions on \(\mu_0\) and \(\sigma_0\).

Table 7 shows the results for the unconstrained two-component mixture model. Columns (1) and (2) of Panel A capture parameter estimates corresponding to the low-skill and high-skill SMA groups respectively. The estimated fraction of low-skill SMAs is roughly 90%, with the skill distribution centered on an abnormal return of a negligible 6 bps and dispersion of 0.4%. On the other hand, the remaining 10% high-skill SMAs have an economically meaningful average abnormal return of 56 bps and dispersion of 3.8%. Panel B summarizes the overall distribution of abnormal returns implied by the unconstrained mixture model. Importantly, the

\(^{22}\)The total fraction of unskilled SMAs is computed using the estimated parameters as \(\pi_0 + \pi_1 \cdot \Phi(\frac{\mu_1 - \mu_0}{\sigma_1})\), where \(\Phi(\cdot)\) is the cumulative normal distribution function.
Table 6: SMA Belief Formation Ability: Constrained Mixture Model. The table reports result for the constrained two-component mixture model of belief formation skill. Panel A reports estimates of model parameters, where $\pi$ is the fraction of the unskilled and skilled-type group, $\mu$ is the mean of each group, $\sigma$ is the dispersion of each group’s true belief formation ability, and $\sigma_{i,j}$ is average dispersion of estimated ability of each group. The mean and dispersion of true ability for the first skill group, $\mu_0$ and $\sigma_0$ respectively, are constrained to zero. Each SMA’s average abnormal return, $ABR_i$, is computed relative to the CAPM for all publications by the SMA, where the return horizon is over trading days $t+1$ to $t+6$ with $t$ being the belief publication day. Conditional on performance group $j$, true belief formation ability is normally distributed with mean $\mu_j$ and standard deviation $\sigma_j$. Conditional on performance group $j$, estimated ability is also normally distributed with mean $\mu_j$, but standard deviation $\sigma_{i,j} = \sqrt{\sigma_j^2 + s_i^2}$, where $s_i$ is the standard error of $ABR_i$. The reported $\sigma_{i,j}$ is based on the cross-sectional average of $s_i$. Hence $\sigma_{i,j} = \sqrt{\sigma_j^2 + s^2}$. Estimates in Panel A are used to compute statistics for the cross-sectional distribution of belief formation skill reported in Panel B. SD is the standard deviation, and Fraction positive is the fraction of the distribution with positive ability. Standard errors (in parentheses) are computed as the standard deviation of the statistics from 1,000 bootstrap replications.
fraction of SMAs with truly positive average abnormal return following their belief statements, i.e. skilled SMAs, is 56% – a sizeable improvement compared to the value from the constrained two-component model.  

<table>
<thead>
<tr>
<th>Panel A: Mixture Model Parameters</th>
<th>(1) Component 0</th>
<th>(2) Component 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi$</td>
<td>0.8969</td>
<td>0.1031</td>
</tr>
<tr>
<td></td>
<td>(0.0876)</td>
<td>(0.0876)</td>
</tr>
<tr>
<td>$\mu$</td>
<td>0.0006</td>
<td>0.0056</td>
</tr>
<tr>
<td></td>
<td>(0.0002)</td>
<td>(0.0021)</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.0044</td>
<td>0.0378</td>
</tr>
<tr>
<td></td>
<td>(0.0046)</td>
<td>(0.0066)</td>
</tr>
<tr>
<td>$\sigma_{i,j}$</td>
<td>0.0147</td>
<td>0.0403</td>
</tr>
<tr>
<td></td>
<td>(0.0037)</td>
<td>(0.0060)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Mixture Return Distribution</th>
<th>Estimate</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.0012</td>
<td>(0.0002)</td>
</tr>
<tr>
<td>SD</td>
<td>0.0138</td>
<td>(0.0005)</td>
</tr>
<tr>
<td>P10</td>
<td>-0.0059</td>
<td>(0.0009)</td>
</tr>
<tr>
<td>P25</td>
<td>-0.0026</td>
<td>(0.0005)</td>
</tr>
<tr>
<td>P50</td>
<td>0.0007</td>
<td>(0.0001)</td>
</tr>
<tr>
<td>P75</td>
<td>0.0041</td>
<td>(0.0005)</td>
</tr>
<tr>
<td>P90</td>
<td>0.0076</td>
<td>(0.0008)</td>
</tr>
<tr>
<td>Fraction positive</td>
<td>0.5584</td>
<td>(0.0167)</td>
</tr>
<tr>
<td>N</td>
<td>4,145</td>
<td></td>
</tr>
</tbody>
</table>

Table 7: SMA Belief Formation Ability: Unconstrained Mixture model. The table reports result for the unconstrained two-component mixture model of belief formation skill. Panel A reports estimates of model parameters, where $\pi$ is the fraction of low and high-type SMAs, $\mu$ is the mean of each group, $\sigma$ is the dispersion of each group’s true belief formation skill, and $\sigma_{i,j}$ is average dispersion of estimated ability of each group. Each SMA’s average abnormal return ($\overline{ABR_i}$) and the standard error ($s_i$) used for estimation, as well as the estimated parameters in the table are described in Table 6. Estimates in Panel A are used to compute statistics for the cross-sectional distribution of belief formation skill reported in Panel B. P(10) — P(90) are percentiles of the implied cross-sectional distribution of SMA skill. Standard errors (in parentheses) are computed as the standard deviation of the statistics from 1,000 bootstrap replications.

One may wonder if the results reported so far depend on the minimum number of belief statements required for computing the estimated skill ($\overline{ABR_i}$) and its standard error ($s_i$). I verify that the results are robust by requiring each SMA to have at least 10 belief statements to be included in the analysis. Results for this more stringent data requirement, reported in Tables 23.

\[ \text{Fraction positive is computed as } 1 - \left[ \pi_0 \cdot \Phi\left( \frac{\overline{ABR_i} - \mu_0}{\sigma_0} \right) + \pi_1 \cdot \Phi\left( \frac{\overline{ABR_i} - \mu_1}{\sigma_1} \right) \right]. \]

For a given percentile $P$, the corresponding quantile $q$ is computed numerically by solving $P = \pi_0 \cdot \Phi\left( \frac{q - \mu_0}{\sigma_0} \right) + \pi_1 \cdot \Phi\left( \frac{q - \mu_1}{\sigma_1} \right)$, where $\Phi(\cdot)$ is the cumulative normal distribution function.
A4 and A5 of the appendix, indicate qualitatively similar conclusions. In the unconstrained two-component model roughly 8% of SMAs are high-skill while 59% have positive abnormal returns; in the constrained model, the corresponding numbers are 18% and 10% respectively.

Overall, results from the unconstrained two-component model suggest that slightly over half of the SMAs in Seeking Alpha express beliefs that generate positive performance. However, only a small fraction (10%) are high-skill individuals whose beliefs generate economically sizeable abnormal returns. Again, compared to professional security analysts, Crane and Crotty (2020) use a similar setup to show that roughly 36% of analysts are high-skill, while 97% are skilled in generating positive abnormal returns. Qualitatively, the results from the unconstrained mixture model support earlier conclusion that most SMAs are mediocre, even though about half of them generate just positive abnormal returns. These results, combined with evidence from Section 4, suggest that investors who heed the views of SMAs are likely better off following the average belief of SMAs' rather than relying on the views of some individual SMAs, given the low prevalence of skill and the potential difficulty of identifying who the few highly skilled guys are.

5.2 SMA Characteristics and Belief Formation Ability

In view of the limited skill among SMAs, identifying SMA attributes associated with skill would help (retail) investors’ narrow their search for valuable information on social media. Moreover, research in the professional analyst literature suggest that analyst characteristics such as specialization, experience, and recommendation style are associated with analyst skill (Loh and Stulz, 2011; Barber et al., 2006; Desai et al., 2000; Mikhail et al., 1997). I, therefore, study heterogeneity in SMAs’ belief formation ability by incorporating SMA attributes observable in Seeking Alpha in modeling the probability that an SMA is high- or low-skill type. Following Crane and Crotty (2020), I parameterize the probabilities using the logistic function as follows:

\[
\pi_{i,0} = \frac{1}{1 + \exp(b_0 + b_1 x_i)}; \quad \pi_{i,1} = 1 - \pi_{i,0}
\]  

(9)
where \( x_i \) is a dummy variable that equals one if a certain SMA characteristic is above the cross-sectional median and zero otherwise. With this parametrization the density function (4) has additional parameters \( b_0 \) and \( b_1 \), and the set of parameters in the maximum likelihood problem of Eq. (5) is now \( \Theta = \{b_0, b_1, \mu_0, \mu_1, \sigma_0, \sigma_1\} \). Once these parameters are estimated the probabilities \( \pi_{i,0} \) and \( \pi_{i,1} \) for SMAs with low \( (x = 0) \) and high \( (x = 1) \) values of the characteristic can be calculated using the estimates of \( b_0 \) and \( b_1 \).

I consider the following characteristics computed over the entire sample period for each SMA. (1) Specialization, proxied by the number of unique SIC industries across which an SMA expresses beliefs. (2) Workload, proxied by the average number of belief statements published per year. (3) Experience, proxied by the number of years between an SMA’s first and last publication in Seeking Alpha. (4) Effort, proxied by the average number of words in opinion articles corresponding to each belief statement. (4) Sophistication, proxied by the average complexity (measured as the fog index) of the SMA’s writing.\(^{24}\) (5) Popularity, proxied by the average number of comments on belief statements within two days of publication. (6) Disagreement, measured as average of the absolute difference between the fraction of negative words in an SMA’s opinion article and the average fraction of negative words in the comments posted within two days of article publication. (7) “Skin in the Game” proxied by the fraction of time an SMA discloses having an investment position in a stock about which belief is being expressed.\(^{25}\)

Table 8 shows the results for SMA skill conditional on characteristics. The table reports, for each investor characteristic \( x \), the fraction of high- and low-skill type, average abnormal return and its standard deviation implied by the mixture model conditional on whether \( x \) is above or below its cross-sectional median. Supporting theoretical results on gains from specialization in information acquisition (e.g., Van Nieuwerburgh and Veldkamp, 2010), I find that SMAs that focus on fewer industries and issue fewer belief statements per year are more likely to be from the high-skill group. In particular, SMAs that specialize on a few (below median)

\(^{24}\)Fog Index is a numerical score used to measure text complexity. It estimates the years of formal education needed to understand the text on first reading. A score of 6 require Sixth grade education while a score of 17 requires College education.

\(^{25}\)To create the “skin in the game” variable, for each publication I create a dummy variable that equals 1 if the SMA discloses either a long or short position in the given stock and zero if no disclosure is made or a disclosure of no position is made. Finally, the indicator variable is averaged over the sample period for each SMA.
Table 8: SMA Characteristics and Belief Formation Ability. The table reports result for the cross-sectional distribution of SMAs’ belief formation ability from a two-component mixture model where the proportion of SMAs in each component depends on an SMA’s characteristic as shown in Eq. (9). The table reports the estimated proportion ($\pi$) of the low- and high-type SMAs for the below- ($x = 0$) and above-median ($x = 1$) SMAs for a given characteristic. Also shown in the table are the conditional Mean of SMAs’ true ability and its standard deviation (SD) implied by the mixture model. Column headers indicate the SMA characteristics. Specialization is the number of unique four-digit SIC code industries covered by an SMA. Experience is the number of years between an SMA’s first and the last belief statement. Workload is the average number of beliefs issued per year. Effort is the average number of words in opinion articles. Skin-in-Game is the proportion of times an SMA disclosed investment in a stock about which belief is being expressed. Complexity is average fog index of opinion articles. Popularity is the average number of comments on belief statements within two days of publication. Disagreement is the average of absolute value of the difference between the fraction of negative words in an SMA’s article and the fraction of negative words in comments on the article. ***, **, * indicate statistical significance of the one sided test of difference between groups at the 1%, 5% and 10% significance levels, respectively. The p-values are based on bootstrap distribution with 1,000 bootstrap replications.
industries have 35% probability of being high-skill compared to 0.4% for SMAs that cover many (above median) industries. The model-implied average abnormal return is 12 bps lower for SMAs with less industry specialization. Similarly, SMAs with fewer workload (below median publications per year) have 21% probability of being high-skill compared to 7% for SMAs with more publications per year. These results are in line with models of limited attention and cognitive capacity constraints (e.g., Hirshleifer et al., 2011; Van Nieuwerburgh and Veldkamp, 2010), which imply that SMAs who specialize less and have more workload may be less capable of effectively processing individual pieces of information to obtain precise signals. Furthermore, the results corroborate evidence in the professional analyst literature that analysts who focus on fewer industries (Desai et al., 2000) and issue fewer recommendations per year (Crane and Crotty, 2020) are more likely to outperform their peers.

SMAs with above median years of experience have a lower probability (7%) of being high-skill compared to those with below median experience whose probability is 16%. This goes against the view in some papers (e.g., Seru et al., 2010) that investors get better with experience. However, this result may be driven by my measurement of experience as the number of years an SMA has participated in Seeking Alpha, which may not be highly correlated with actual investing experience. Another plausible explanation is that more skilled SMAs joined Seeking Alpha over time as the platform became more popular. Yet another possibility is that older SMAs are susceptible to experience-induced biases such as reinforcement learning, which has been shown to impair investment performance (Huang, 2019; Strahilevitz et al., 2011).

SMAs’ “skin in the game” is associated with skill. SMAs that mostly have a position in the stock about which belief is being stated have 17% probability of being high-skill compared to 5% probability for SMAs that mostly express beliefs on stocks they are not invested in. The performance of SMAs with more skin in the game is 7 bps higher. To the extent that SMAs’ truthfully disclose their investment position, this result potentially suggests that more skin in the game warrants more diligent information acquisition and processing which in turn produces superior performance. In line with this view, Campbell et al. (2019) find that having an investment position in a stock does not impair the credibility and informativeness of opinions expressed by nonprofessional analysts on social media.
Similarly, SMAs that invest more effort to write longer articles, potentially providing stronger arguments to back up their belief statement, are 9% more likely to be high-skill compared to those that write short articles. The estimated true performance of the more elaborate SMAs is 5 bps higher. Manual examination of some opinion articles show that longer articles tend to have relevant information such as financial statement analyses, projections and detailed discussion of the rationale for the stated belief. This lends further supports to the finding that elaborate SMAs are more likely to be sophisticated in producing or processing value-relevant information.

For the remaining characteristics, the results are not as pronounced. SMAs with more complex articles are only 3% more likely to be high-skill. The same holds for more popular SMAs (those that receive more comments). On the other hand, SMAs whose views other investors tend to disagree more with are less likely to be high-skill, with a probability of 7% compared to 14% for more agreeable SMAs. However, for all three characteristics the model-implied true performance for the below and above median SMAs, although statistically significant, is economically small. In untabulated results, I also find that the proportion of bearish beliefs expressed by an SMA does not matter for skill, the probability difference is indistinguishable from zero, which goes against results in the professional analyst literature (Barber et al., 2006; Crane and Crotty, 2020) that analysts who issue more sell recommendations are more likely to be high-skill.

Overall, the results show that certain SMA characteristics are associated with skill, which may help (retail) investors identify where skill lies among nonprofessional analysts. Furthermore, there are differences in how nonprofessional and professional analyst characteristics relate to skill. For instance, while professional analysts that issue more sell recommendations tend to be more skilled, it appears not to be the case for nonprofessionals.

6 How do Social Media Analysts form Beliefs

This section explores how SMAs form beliefs, in particular, whether the consensus belief influences individual SMA’s belief about stocks. Section 6.1 explores the existence of herding among SMAs and Section 6.2 examines under what conditions herding is more or less pronounced.
6.1 Evidence on Herding

Social media is a typical environment for mutual imitation, i.e., herding, due to the ease of information transmission and the ability of individuals to readily observe the actions of peers through networks of relationships. Recent developments highlight the importance of understanding herding and its nature among investors on social media. For instance, the recent frenetic trading in GameStop stock – which led to an extraordinary surge in price by 1,625% in January 2021 and the near collapse of Melvin Capital, a premier hedge fund – was in part fuelled by retail investors herding on the social media forum WallStreetBets.\(^{26}\) Understanding the broader implications of such coordinated herding on social media has now warranted ongoing congressional and regulatory monitoring in the US.\(^{27}\)

Although SMAs in Seeking Alpha are arguably more sophisticated than the WallStreetBets crowd, and their stated beliefs are backed by extended fundamental analyses and discussions which must be approved by an editorial team, it is still very plausible that SMAs’ beliefs are subject to herding in line with the general class of reputational herding and information cascade models (e.g., Banerjee, 1992; Bikhchandani et al., 1992; Scharfstein and Stein, 1990). On the other hand, since SMAs face different incentives from professional investment analysts\(^ {28}\), it is equally possible that herding is less pervasive among SMAs. For instance, given that SMAs’ financial rewards depend on their article views, an SMA may prefer to express beliefs that deviate from the consensus in order to appear knowledgeable and attract attention. Since herding might intensify mis-pricing if it is based on little or no information or promote price discovery if it is caused by fundamental information, I examine the existence and nature of herding among SMAs to shed further light on their role in financial markets.

To test for the existence of herding amongst SMAs, I adopt the herding test developed in Welch (2000) which is applicable in settings where choices are discrete (e.g., Bullish, Neutral and Bearish belief statements). The method is based on the idea that herding is an external


\(^{28}\)For instance, unlike security analysts SMAs are less likely to have conflict of interest arising from business relationships with the firms they express opinions about. They are also less likely to have personal conflicts related to job security, need to ingratiate themselves with managers and powerful investor groups. On the other hand, it is likely that SMAs care about their reputation as this could impact their compensation.
force that dynamically changes the belief transition probability matrix. To motivate the test, let \( \theta \) represent a parameter that measures whether SMA belief transition probability matrix \( P \) depends on the observed consensus \( C \). \( P \) is a 3-by-3 matrix whose elements, \( p_{i,j} \), capture the probability of an SMA moving from a previous belief statement, row \( i \), to a new belief statement, column \( j \). For example transitioning from a Bearish to Bullish belief. If \( \theta = 0 \), then \( P \) is independent of the consensus – the null hypothesis. On the other hand, \( \theta > 0 \) indicates a tendency for belief statements to follow the consensus, while \( \theta < 0 \) indicates a tendency to avoid the consensus. \( P \) is then defined as a function of \( \theta \) and a target \( C \) (the consensus) as follows:

\[
p_{i,j}(\theta, C) \equiv \frac{p_{i,j}}{D_i} \left[ 1 + (j - C)^2 \right]^{-\theta}
\]

(10

where index \( i, j \in \{1, 2, 3\} \) denote belief states: 1 represents Bearish, 2 Neutral and 3 Bullish; the denominator \( D_i = \sum_{j=1}^{3} p_{i,j} \left[ 1 + (j - C)^2 \right]^{-\theta} \) ensures that rows of the transition matrix sums to 1; \( p_{i,j} \) is the unconditional transition probability from \( i \) to \( j \) estimated from the historical SMA belief revisions; and \( C \) is the target towards which SMAs may herd, that is consensus belief about a given stock on a given day. The estimator \( \hat{\theta} \) is obtained by maximizing the log-likelihood function over the sample period, and statistical inference is based on the likelihood ratio test. Only belief revisions not older than six months are used in the estimation to avoid stale beliefs. The sample period is 2006 – 2019 due to very few belief statements in earlier periods. Appendix A.2 describes the estimation and inference procedure in more detail.29

Table 9 shows the estimated herding coefficient \( \hat{\theta} \) and the associated \( \chi^2 \) \( p \)-value for the equal-weighted consensus and characteristic-weighted consensus beliefs. The characteristics – Popularity, Effort and Specialization – for each SMA used for the weighting are defined in Section 5.2, but here are computed as of last calendar month using data over the past one year.30 The motivation for examining the characteristic-weighted consensus is to see if SMAs

29Using Monte-Carlo simulations, Welch (2000) show that the herding test is not mechanically driven by the discrete/limited number of choices or by the fact that the target (consensus) itself is the outcome of prior transitions.

30For a given SMA’s belief revision about stock \( k \), the consensus is based on belief statements of other SMAs on the same stock over the past 6-month period. The consensus is estimated if there are at least 2 belief statements.
herd more towards views of peers with characteristics identified in Section 5.2 to be associated with skill. Panel A of Table 9 shows that the estimated herding coefficient is around 0.24 and is significant, suggesting that SMAs herd towards the consensus when stating their beliefs. Furthermore, the coefficient is not very sensitive to the consensus weighting scheme, which tends to suggest that the views of more skilled SMAs do not have stronger pulling power than the average belief.

Panel A Estimated Herding Coefficient

<table>
<thead>
<tr>
<th>Concensus is</th>
<th>$\hat{\theta}$</th>
<th>$\chi^2$ p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal-weighted</td>
<td>0.237</td>
<td>0.000</td>
</tr>
<tr>
<td>Popularity-weighted</td>
<td>0.231</td>
<td>0.000</td>
</tr>
<tr>
<td>Effort-weighted</td>
<td>0.220</td>
<td>0.000</td>
</tr>
<tr>
<td>Specialization-weighted</td>
<td>0.233</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Panel B Probability of Hitting Target

<table>
<thead>
<tr>
<th>Target</th>
<th>-10</th>
<th>-1</th>
<th>0</th>
<th>0.15</th>
<th>0.25</th>
<th>0.5</th>
<th>1</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Bearish)</td>
<td>0.000</td>
<td>0.051</td>
<td>0.200</td>
<td>0.239</td>
<td>0.268</td>
<td>0.346</td>
<td>0.521</td>
<td>1.000</td>
</tr>
<tr>
<td>2 (Neutral)</td>
<td>0.000</td>
<td>0.042</td>
<td>0.081</td>
<td>0.089</td>
<td>0.094</td>
<td>0.110</td>
<td>0.149</td>
<td>0.989</td>
</tr>
<tr>
<td>3 (Bullish)</td>
<td>0.000</td>
<td>0.382</td>
<td>0.719</td>
<td>0.757</td>
<td>0.781</td>
<td>0.831</td>
<td>0.899</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 9: Herding among SMAs. The table reports results for the herding test of Eq. (10) in Panel A and the economic significance of herding in Panel B. Panel A shows the estimated herding coefficient $\hat{\theta}$ and $\chi^2$ p-value for different targets (consensus estimates). Panel B shows the probability of a belief revision hitting a hypothetical Bearish, Neutral or Bullish target for different values of $\hat{\theta}$. If $\hat{\theta} = -\infty$, the target will always be avoided. If $\hat{\theta} = 0$, the probability of hitting the target is equal to the unconditional probability of hitting the target. If $\hat{\theta} = \infty$, the target will always be hit. Values in Panel B were produced using the unconditional transition matrix. The sample period is 2006 – 2019.

Panel B of Table 9 shows the economic implication of the estimated herding coefficient. The column with $\hat{\theta} = 0$ captures the unconditional probabilities of an SMA stating a Bearish, Neutral or Bullish revision. Focusing on the column with $\hat{\theta} = 0.25$, it follows that the estimated herding coefficients reported in Panel A implies that herding increases the probability of hitting a hypothetical Bullish (Bearish) belief target by about 6 (7) percentage points. This indicates a moderate level of herding amongst SMAs, which is interesting given that SMAs could deviate from the consensus in an attempt to attract attention and readership.

by other SMAs over the 6-month period. The results are robust to estimating the consensus over a longer window of 1-year. For the characteristic-weighted consensus, I replace missing characteristics with the median value.
6.2 When is Herding more Pronounced?

An important question, then, which has implications for market efficiency, is under what conditions SMAs are more likely to herd. For instance, if herding is more when the consensus turns out to be wrong, it would suggest that SMAs’ herding is not informed by fundamental information and therefore consistent with models of herding based on little or no information (e.g., Scharfstein and Stein, 1990). To examine SMAs’ herding under different economic conditions, I modify Eq. (10) by making $\theta$ a function of some variable $y$: $\theta(y) = \theta_0 + \theta_1 y$. If $\theta_1 = 0$, herding does not depend on $y$. On the other hand, $\theta_1 > 0$ indicates that herding increases with $y$, while $\theta_1 < 0$ indicates that herding decreases with $y$.

Table 10 shows results for the estimated $\hat{\theta}_0$ and $\hat{\theta}_1$ and the associated $\chi^2 p$-values. Panel A shows results where $y$ is either the indicator variable for NBER recession (Recession), Indicator variable for market volatility computed with daily data over the past one year exceeding its sample median (Market Volatility), or Consensus Optimism (CO) measured as $CO = C - 2$, where $C$ is the equal-weighted consensus. Because beliefs are labeled as 3 (Bullish), 2 (Neutral) and 1 (Bearish), $CO > 0$ ($CO < 0$) implies an optimistic (pessimistic) consensus. Panel B shows results where $y$ is a measure of Consensus Correctness (CC) quantified as consensus optimism times future stock return: $CC = CO \times R(h)$, where $R(h)$ is the future horizon $h$ (starting $t + 1$) return of a stock. $CC$ is positive when the consensus is correct ex-post, that is an optimistic (pessimistic) consensus is followed by a positive (negative) future stock return.

The results from Panel A shows that $\hat{\theta}_1$ is negative and significant for Recession and Market Volatility, suggesting that SMAs herd less during economic downturns and episodes of high market uncertainty. The $\hat{\theta}_1$ coefficient of -0.36 for recession implies that during recessions SMAs actually go against the consensus as the implied herding coefficient is $-0.11 = 0.25 - 0.36$. This implies that in bad economic states SMAs tend to rely more on their private information. This result is in line with the theoretical literature that predicts polarization of beliefs in bad times (e.g., Cujean and Hasler, 2017), because some agents give little weight to news in bad times. It also tallies with the nature of herding amongst professional analysts who tend to herd more in good times (see, Welch, 2000). On the other hand, unlike professional analysts, I find that

\[31\text{The results are robust to using abnormal returns relative to the CAPM or the three-factor model.}\]
Table 10: SMAs’ Herding Conditional on Economic States. The table reports results for herding test conditional on the realization of some variable y, such that estimated herding coefficient \( \hat{\theta} = \hat{\theta}_0 + \hat{\theta}_1 y \). If \( \hat{\theta}_1 < 0 \), herding is decreasing in y. If \( \hat{\theta}_1 > 0 \), herding is increasing in y. Panel A shows results where y is either the indicator for NBER recession (Recession), indicator for market volatility higher than its sample median (Market Volatility), and Consensus Optimism (CO) measured as the equal-weighted consensus minus 2. Panel B reports results when y is Consensus Correctness measured as CO \( \times R(h) \), where R(h) is future stock return over horizon h \( \in \{5, 63, 126\} \) trading days. The sample period is 2006 – 2019.

SMAs herd less when the consensus is more optimistic. In terms of economic magnitude, with \( \hat{\theta}_1 \) of -0.15 and a standard deviation of 0.39 for consensus optimism, a standard deviation increase in consensus optimism reduces the herding coefficient by about 6 percentage points – a 24% decline relative to its unconditional value. This result supports the view that SMAs favor more attention-grabbing deviation from the consensus in order to attract readership.

Panel B of Table 10 shows that across different future return horizons h \( \in \{5, 63, 126\} \) trading days, the incremental herding coefficient \( \hat{\theta}_1 \) for consensus correctness is positive. This implies stronger herding when the consensus is correct, that is when the consensus is optimistic (pessimistic) and the future return turns out to be positive (negative).\(^{32}\) In terms of economic magnitude, the standard deviation of CO \( \times R(h) \) for the 5-day and 63-day horizons are 4.5% and 14.8% which implies that herding increases by about 7% of its unconditional value for a standard deviation increase in consensus correctness. That the incremental herding when the consensus is correct remains positive even for the 6-month future return horizon suggests that SMAs herd on fundamental information in the consensus rather than sentiment. The implication, therefore, is that instead of herding irrationally (Simonsohn and Ariely, 2008) or herding based on no information (Scharfstein and Stein, 1990), SMAs tend to learn fundamental information from the belief statements of their peers to improve forecasts, consistent with information-based

\(^{32}\)The statistical significance of the incremental herding coefficient decreases with horizon, turning insignificant for the 126-day and 252-day (not shown in the table) horizons.

<table>
<thead>
<tr>
<th>Panel A:</th>
<th>Recession</th>
<th>Market Volatility</th>
<th>Conc. Optimism (CO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>State of Economy</td>
<td>( \hat{\theta}_0 )</td>
<td>( \hat{\theta}_1 )</td>
<td>( \hat{\theta}_0 )</td>
</tr>
<tr>
<td>Estimate</td>
<td>0.253</td>
<td>-0.355</td>
<td>0.270</td>
</tr>
<tr>
<td>( \chi^2 ) p-value</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B:</th>
<th>CO ( \times R(h) )</th>
<th>CO ( \times R(h) )</th>
<th>CO ( \times R(h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conc. Correctness</td>
<td>( \hat{\theta}_0 )</td>
<td>( \hat{\theta}_1 )</td>
<td>( \hat{\theta}_0 )</td>
</tr>
<tr>
<td>Estimate</td>
<td>0.237</td>
<td>0.371</td>
<td>0.236</td>
</tr>
<tr>
<td>( \chi^2 ) p-value</td>
<td>0.000</td>
<td>0.006</td>
<td>0.000</td>
</tr>
</tbody>
</table>
herding models (e.g., Banerjee, 1992). An alternative explanation, which I cannot differentiate, is that SMAs independently follow the same fundamental information.\footnote{There is also a possibility that the consensus moves prices. I, however, favor the fundamental information story because as shown in Section 4.2 aggregated beliefs contain value-relevant information.} Whichever the case, the result suggests that to some extent observational learning among SMAs would likely improve price discovery as investors trade on SMAs’ belief statements.

In sum, results in this section show that SMAs herd towards the consensus. Herding is less prevalent during economic downturns, high market uncertainty and when the consensus is more optimistic. However, herding appears to be information-based, as SMAs herd more on ex-post correct consensus.

\section{Conclusion}

This paper uses natural language processing and machine learning to extract the beliefs of non-professional social media investment analysts (SMAs) about a large cross-section of stocks from opinions expressed in investment social media. The paper then examines the informativeness of SMA beliefs about individual stocks, the distribution of belief formation skill among SMAs, and the existence and nature of herding among SMAs.

The analyses show that, on average, SMA beliefs contain value-relevant information. Studying the distribution of belief formation skill, the paper shows that while about half of SMAs are able to generate just positive abnormal returns following belief statements, only a small fraction, 10%, of SMAs state beliefs that produce economically meaningful performance of 56 bps over a 5-day window. These results suggest that as a group SMAs appear to add value, although skill is considerably limited at the individual SMA level.

The paper further shows that SMA characteristics, such as specialization, effort, skin in the game, popularity, and disagreement with other investors, are associated with belief formation skill. More so, SMAs herd in their belief statements, with herding being less prevalent in bad times and when the consensus is more optimistic. On the other hand, herding is more pronounced when the consensus is correct ex-post, suggesting that herding among SMAs is likely information-based.
In sum, the paper suggests that SMAs as a group seem to improve price discovery and are not just the mad crowd recent social media events might tempt one to think. As such, (retail) investors could benefit from SMAs’ average beliefs. More broadly, some aspects of social media appear to hold good promise for financial markets.
References


A Appendix

A.1 Description of the Support Vector Classifier ML Algorithm

To get a glimpse of how linear SVC works, consider a simple binary classification task. The SVC algorithm solves the following optimization problem:

\[
\min_{w_0, w} \sum_{n=1}^{N} [1 - y_n(w_0 + w \cdot x_n)]^+ + c \cdot (w \cdot w)
\]  

(A1)

where \( y \) is the vector of binary labels consisting of entries -1 and +1 used in place of the actual class labels – for instance, -1 could denote “Bearish Belief” while +1 denotes “Bullish Belief”; \( w_0 \) and \( w \) are weights to be optimized on; \( c \) is a scalar regularization parameter; \([1 - y_n(w_0 + w \cdot x_n)]^+\) is the “hinge” loss function, with the superscript “+” indicating that only positive parts should be considered; and \( N \) is the number of observations. As shown in Hastie et al. (2009), for a given regularization parameter \( c \) the solution to (A1) is a weighted average of the regressors:

\[
\hat{w} = \sum_{n=1}^{N} \hat{\alpha}_n y_n x_n,
\]

(A2)

where only some of the \( N \) observation weights \( \hat{\alpha}_n \) are nonzero. These observations with nonzero \( \hat{\alpha}_n \), called the support vectors, are the ones relatively close to the separating hyperplane, i.e., points that are most difficult to classify. With \( \hat{w} \), SVC predicts the classes using the decision function:

\[
\hat{G}(x_n) = \text{sign}[\hat{w}_0 + \hat{w} \cdot x_n]
\]

(A3)

SVC therefore selects a relatively small number of observations for identifying the coefficients \( \hat{w} \). This in turn implies that instead of estimating \( M \) coefficients for the large feature space, only a relatively small set of parameters (selecting the support vectors and \( \hat{\alpha}_n \)) is estimated, yielding nonzero coefficients for only the important features.

---

Since a full treatment of the SVC algorithm is beyond the scope of this paper, I present only an illustration. See Hastie et al. (2009) for a formal exposition.
While the foregoing illustration is for binary classification, it extends to multiclass classification as in this paper. In the multiclass context, a one-vs-rest binary SVC is trained for each class to separate the class from all other classes, resulting in as many binary models as there are classes. To make a prediction, all binary classifiers are run on the test data and the class label of the classifier that has the highest score (by evaluating \( \hat{w}_0 + \hat{w} \cdot x_n \)) is returned as the prediction.

A.2 Description of Herding Test

This section provides additional details on the herding test used in the paper based on Welch (2000). Eq. (10) specifies how the transition probability depends on \( \theta \) and a target, namely the consensus \( (C) \). The estimate for \( \theta \) is obtained by maximizing the log-likelihood function given the observed transitions \( i_n \rightarrow j_n \), each with its own target \( C_n \), over all observations \( n \in [1, N] \):

\[
\hat{\theta} = \arg\max_{\theta} \sum_n \log \left[ p_{i,j}(\theta, C_n) | \{i_n, j_n\} \right]
\]  

(A4)

where \( p_{i,j}(\theta, C_n) | \{i_n, j_n\} \) is the transition probability in Eq. (10) evaluated at each realized \( n \) transition and their associated target \( C_n \) in the data. Therefore, any chosen \( \theta \) parameter implies for each observation \( n \) a probability vector \( p_i \). As Welch (2000) notes, under the assumption that draws are independent, the likelihood function is the probability of empirically observing the full sequence of transitions.

Statistical inference on the estimator \( \hat{\theta} \) is based on the likelihood ratio test. The likelihood ratio statistic is the ratio of the probability of the data for a given constant transition probability versus a transition probability that varies with \( C \) according to \( \theta \):

\[
2 \left\{ \sum_n \log \left[ p_{i,j}(\theta, C_n) | \{i_n, j_n\} \right] - \sum_n \log \left[ p_{i,j}(0) | \{i_n, j_n\} \right] \right\} \sim \chi^2_1
\]  

(A5)

Since \( p_{i,j}(0) \) is not observable, it is assumed that the \( P(0) \) matrix under the null is the empirically observed transition matrix. This assumption is conservative in that it is correct under the null and biases the test against the herding hypothesis. Finally, for the setup where \( \theta \)
depends on some variable $y$: $\theta(y) = \theta_0 + \theta_1 y$, the likelihood ratio test is performed by comparing the restricted model (with $\theta = \theta_0$) against the unrestricted model (with $\theta(y) = \theta_0 + \theta_1 y$).

For the asymptotic properties of the described maximum-likelihood estimation, see Welch (2000) who also documents excellent small sample properties of the maximum-likelihood ratio.
Microsoft (NASDAQ: MSFT) has become the company tech investors love to hate. The company’s P/E of 14.69, which is about average for an S&P 500 stock, masks the extraordinary charges in the second quarter, when the company wrote off its aQuantive ad network. Had the company even had an average quarter for June, you would have trailing years’ earnings of $2.50/share, a P/E lower than that of Ford (NYSE: F). Plus, there’s over $66 billion in cash on the balance sheet — take that out and the price is a snip. But investors aren’t buying that story anymore. The stock is back to the levels of the first of the year, before Windows 8 was rolled out. Since its peak in April it’s down 16%. And did I mention the 3.38% yield? That said, the reason investors are fleeing Microsoft makes sense. They see the ads for Windows 8, they’ve gone into the stores, and they’ve seen the Microsoft products languishing on the shelves. They’ve read the tweets and seen the reviews. They think Windows 8 does something that rhymes with trucks, and that it’s not rolling out as expected. I could argue that most software is seen in a negative light when it first comes out, that even Apple (NASDAQ: AAPL) has its share of negative reviews, and that a new user interface always takes some time to learn and get used to. But instead I’m going to talk about the cloud. When I write about cloud, I seldom mention Microsoft Azure. But Azure is a pretty good cloud, with decent price performance. And according to a recent survey by Forrester Research, which tracks corporate computing professionals, it’s getting strong reviews there. As per the survey, Azure is currently doing as well as Google (NASDAQ: GOOG) cloud services, and a full 20% expect their usage of it to grow over the next year. The reason: It’s easy to set up and easy to use. If you know Windows, you’re halfway there — you can bring all your existing tools and skills to the party. This is important stuff, because the next big step for the cloud market is the move from cloud infrastructures to cloud platforms. If Microsoft can make the jump, with anything like its current market share as a cloud platform and if its languages and other tools can be seen as the easy way to build cloud applications, that is a huge leg up in the market. And hidden within Windows 8 are all the tools you need — both you and your employees — to build and deploy cloud applications on Azure. So you’ve got a dirt cheap stock with a leg up on the cloud market of 2013. Is that worth an implied P/E of 7 when you also get $66 billion in cash? At this point, Microsoft is as discounted as it is going to get. It’s beyond cheap here.

Disclosure: I am long MSFT, AAPL, GOOG. I wrote this article myself, and it expresses my own opinions. I am not receiving compensation for it (other than from Seeking Alpha). I have no business relationship with any company whose stock is mentioned in this article.

**Figure A1: Sample SMA Article with Bullish Belief.** The figure shows an opinion article published in Seeking Alpha where the author explicitly states a bullish belief about a stock.
Summary: MMM is suitable for Enterprising Investors following the ModernGraham approach. According to the ModernGraham valuation model, the company is overvalued at the present time. The market is implying 6.96% earnings growth over the next 7-10 years, a rate of growth which is unsupported by the company’s recent results.

3M Company (NYSE:MMM) has seen quite a run-up in price over the last five years, and for many investors that alone presents a reason to turn away. However, Benjamin Graham, the father of value investing, taught that looking at the price cannot be the sole factor in investment decisions as the most important aspect to consider is whether the company is trading at a discount relative to its intrinsic value. It is through a thorough fundamental analysis that the investor is able to make a determination about a potential investment’s merits. Here is a look at how 3M Company fares in the ModernGraham valuation model. The model is inspired by the teachings of Benjamin Graham and considers numerous metrics intended to help the investor reduce risk levels. The first part of the analysis is to determine whether the company is suitable for the very conservative Defensive Investor or the less conservative Enterprising Investor, who is willing to spend a greater amount of time conducting further research. In addition, Graham strongly suggested that investors avoid speculation in order to remove the subjective elements of emotion. This is best achieved by utilizing a systematic approach to analysis that will provide investors with a sense of how a specific company compares to another company. By using the ModernGraham method one can review a company’s historical accomplishments and determine an intrinsic value that can be compared across industries. 3M Company is not a great opportunity for Defensive Investors, as the company has a low current ratio and high PEmg and PB ratios. However, the less conservative Enterprising Investor has no initial concerns and is willing to proceed to the next part of the analysis, which is a determination of the company’s intrinsic value. When estimating the intrinsic value, it is critical to consider the company’s historical earnings results in combination with a review of the market’s implied estimate for further growth. In this case, the company has grown its EPSmg (normalized earnings) from $5.14 in 2010 to an estimated $6.71 for 2014. While this is a strong level of demonstrated growth, it does not quite support the market’s implied estimate for earnings growth of 6.96% over the next 7-10 years. In order to reach that growth rate, the company would need to achieve higher growth than it has in the recent past. The ModernGraham valuation model therefore returns an estimate of intrinsic value falling below the current price, indicating the company is overvalued at the present time. Be sure to check out previous ModernGraham valuations of 3M Company for more perspective!

Disclosure: I/we have no positions in any stocks mentioned, but may initiate a short position in NKE over the next 72 hours. I wrote this article myself, and it expresses my own opinions. I am not receiving compensation for it (other than from Seeking Alpha). I have no business relationship with any company whose stock is mentioned in this article.

Figure A2: Sample SMA Article with Bearish Belief. The figure shows an opinion article published in Seeking Alpha where author explicitly states a bearish belief about a stock.
Table A1: Cashflow-Relevant News Event Categories The table reports news event categories in the RavenPack database deemed relevant for firm cash-flow. When aggregating sentiment across cash-flow news for a given firm, I focus only on these event categories.
\[
\begin{array}{cccccc}
\text{CAPM} & \text{FF3} & \text{SBM} \\
\hline
AB_{1,t} & 0.0028 & 0.0026 & 0.0030 & 0.0028 & 0.0027 & 0.0025 \\
& (2.15) & (2.06) & (2.46) & (2.34) & (2.24) & (2.16) \\
ABR_{i,t} & 0.0016 & 0.0019 & 0.0017 & 0.0013 & 0.0004 & 0.0011 \\
& (1.45) & (1.78) & (1.68) & (1.30) & (0.37) \\
ABR_{i,t-1} & -0.0020 & -0.0019 & -0.0021 & 0.0004 & 0.0011 & 0.0011 \\
& (-1.60) & (-1.58) & (-1.68) & (0.37) \\
ABR_{i,t-2} & -0.0001 & -0.0000 & 0.0004 & 0.0004 & 0.0011 & 0.0011 \\
& (-0.07) & (-0.01) & (0.37) \\
ABR_{i,t-63-\rightarrow t-3} & 0.0002 & 0.0016 & -0.0011 & 0.0011 & 0.0011 & 0.0011 \\
& (0.07) & (0.76) & (-0.48) \\
Volatility_{i,t} & -0.0044 & -0.0039 & -0.0043 & -0.0043 & -0.0043 & -0.0043 \\
& (-1.70) & (-1.58) & (-1.81) \\
Upgrade_{i,t} & 0.0011 & 0.0012 & 0.0013 & 0.0013 & 0.0013 & 0.0013 \\
& (1.98) & (2.19) & (2.38) \\
Downgrade_{i,t} & -0.0005 & -0.0005 & -0.0005 & -0.0005 & -0.0005 & -0.0005 \\
& (-0.81) & (-0.88) & (-0.86) \\
ESS_{i,t} & 0.0009 & 0.0008 & 0.0008 & 0.0008 & 0.0008 & 0.0008 \\
& (1.26) & (1.19) & (1.37) \\
\hline
\text{Obs.} & 192,968 & 192,968 & 192,968 & 192,968 & 192,968 & 192,968 \\
R^2 (%) & 1.98 & 2.03 & 1.37 & 1.42 & 1.46 & 1.52 \\
\end{array}
\]

Table A2: SMA Belief and Future Three Months Abnormal Returns The table reports results from the regression of future 3-month stock abnormal returns \((ABR_{k,t+1-\rightarrow t+63})\) on stock-level aggregate SMA belief \(AB_{k,t}\). Abnormal return for stock \(k\) is computed relative to either the CAPM, Fama and French (1993) three-factor model (FF3) or a value-weighted portfolio of firms with similar size, book-to-market and momentum characteristics (SBM) as in Daniel et al. (1997). All regressions include year-month fixed effects. Additional control variables included in columns (2), (4) and (6) are past abnormal returns: \(ABR_{k,t}\); \(ABR_{k,t-1}\); \(ABR_{k,t-2}\); \(ABR_{k,t-63-\rightarrow t-3}\); Volatility\(_{k,t}\) measured as the sum of squared daily returns in the calendar month prior to day \(t\); the number of professional stock analysts upgrading \((Upgrade_{k,t})\) and downgrading \((Downgrade_{k,t})\) stock \(k\) on day \(t\); and stock \(k\)'s News Sentiment Score on day \(t\) \((ESS_{k,t})\). Values in parentheses are \(t\)-statistics based on standard errors clustered by both stock and year-month. The sample period is 2005-2019.
Table A3: SMA Belief Predictability of Returns Conditional on Firm Characteristics. The table reports results from regressing future stock abnormal returns, $ABR_{k, t+1-\rightarrow t+1+h}$, on stock-level aggregate SMA belief, $AB_{k,t}$, and the interaction of $AB_{k,t}$ with the rank on a given stock characteristic $Rank(Z_{k,t})$:

$$ABR_{k,t+1-\rightarrow t+1+h} = \beta_0 + \beta_1 AB_{k,t} + \beta_2 AB_{k,t} \times \text{Rank}(Z_{k,t}) + \beta_3 \text{Rank}(Z_{k,t}) + \mathbf{X} \mathbf{\Gamma} + \epsilon_{k,t}$$

Abnormal return for stock $k$ is computed relative to CAPM. Stock aggregate belief, $AB_{k,t}$, is normalized to unit variance. The rank for each stock characteristic $Rank(Z_{k,t})$ ranges from 1 to 10 and is obtained by sorting firms into deciles at the end of last calendar month. Firm characteristics ($Z_{k,t}$) are indicated in the column headers. Turnover is the number of shares traded divided by the number of shares outstanding per day average over the past 12 months; Price is the stock price per share at the end of a given month; Size is market capitalization as of last calendar year; BM is book-to-market ratio as of the most recent fiscal year end; Volatility is return volatility using 12-month of daily data; and Idio Vol. is idiosyncratic volatility relative to CAPM computed over the same window as Volatility. $\mathbf{X}$ captures additional controls, namely past abnormal returns ($ABR_{k,t-1}$, $ABR_{k,t-2}$; $ABR_{k,t-63\rightarrow t-3}$); number of professional stock analysts upgrading ($Upgrade_{k,t}$) and downgrading ($Downgrade_{k,t}$) stock $k$ on day $t$; Stock $k$’s News Sentiment Score on day $t$ ($ESS_{k,t}$); and year-month fixed effects. Values in parentheses are $t$-statistics based on standard errors clustered by both firm and year-month. The sample period is 2005-2019.
Panel A: Mixture Model Parameters

<table>
<thead>
<tr>
<th>Component 0</th>
<th>Component 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\pi)</td>
<td>0.8228</td>
</tr>
<tr>
<td></td>
<td>(0.0412)</td>
</tr>
<tr>
<td>(\mu)</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>(NA)</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>(NA)</td>
</tr>
<tr>
<td>(\sigma_{i,j})</td>
<td>0.0110</td>
</tr>
<tr>
<td></td>
<td>(0.0002)</td>
</tr>
</tbody>
</table>

Panel B: Mixture Return Distribution

<table>
<thead>
<tr>
<th>Estimate</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.0007</td>
</tr>
<tr>
<td>SD</td>
<td>0.0098</td>
</tr>
<tr>
<td>Fraction positive</td>
<td>0.1038</td>
</tr>
<tr>
<td>N</td>
<td>2,610</td>
</tr>
</tbody>
</table>

**Table A4: SMA Belief Formation Formation Ability: Constrained Mixture Model** The table reports result for the constrained two-component mixture model of belief formation skill using data for only SMAs with at least 10 belief statements. Panel A reports estimates of model parameters, where \(\pi\) is the fraction of the unskilled and skilled-type group, \(\mu\) is the mean of each group, \(\sigma\) is the dispersion of each group’s true belief formation ability, and \(\sigma_{i,j}\) is average dispersion of estimated ability of each group. The mean and dispersion of true ability for the first skill group, \(\mu_0\) and \(\sigma_0\) respectively, are constrained to zero. Each SMA’s average abnormal return, \(\bar{ABR}_i\), is computed relative to the CAPM for all publications by the SMA, where the return horizon is over trading days \(t+1\) to \(t+6\) with \(t\) being the belief publication day. Conditional on performance group \(j\), true belief formation ability is normally distributed with mean \(\mu_j\) and standard deviation \(\sigma_j\). Conditional on performance group \(j\), estimated ability is also normally distributed with mean \(\mu_{ij}\), but standard deviation \(\sigma_{i,j} = \sqrt{\sigma_j^2 + s_i^2}\), where \(s_i\) is the standard error of \(\bar{ABR}_i\). The reported \(\sigma_{i,j}\) is based on the cross-sectional average of \(s_i\). Hence \(\sigma_{i,j} = \sqrt{\sigma_j^2 + \bar{s}^2}\). Estimates in Panel A are used to compute statistics for the cross-sectional distribution of belief formation skill reported in Panel B. SD is the standard deviation, and Fraction positive is the fraction of the distribution with positive ability. Standard errors (in parentheses) are computed as the standard deviation of the statistics from 1,000 bootstrap replications.
Table A5: SMA Belief Formation Ability: Unconstrained Mixture model. The table reports result for the unconstrained two-component mixture model of belief formation skill using data for only SMAs with at least 10 belief statements. Panel A reports estimates of model parameters, where \( \pi \) is the fraction of low, medium and high-type group, \( \mu \) is the mean of each group, \( \sigma \) is the dispersion of each group’s true belief formation ability, and \( \sigma_{i,j} \) is average dispersion of estimated ability of each group. Each SMA’s estimated abnormal return, \( \overline{ABR}_i \), is computed relative to the CAPM for all publications by the SMA, where the return horizon is over trading days \( t+1 \) to \( t+6 \) with \( t \) being the belief publication day. Conditional on performance group \( j \), true belief formation ability is normally distributed with mean \( \mu_j \) and standard deviation \( \sigma_j \). Conditional on performance group \( j \), estimated ability is also normally distributed with mean \( \mu_j \), but standard deviation \( \sigma_{i,j} = \sqrt{\sigma_j^2 + s_i^2} \), where \( s_i \) is the standard error of \( \overline{ABR}_i \). The reported \( \sigma_{i,j} \) is based on the cross-sectional average of \( s_i \). Hence \( \sigma_{i,j} = \sqrt{\sigma_j^2 + \bar{s}^2} \). Estimates in Panel A are used to compute statistics for the cross-sectional distribution of belief formation skill reported in Panel B. SD is the standard deviation, P(10) — P(90) are the percentiles, Fraction positive is the fraction of the mixture distribution with positive ability. Standard errors (in parentheses) are computed as the standard deviation of the statistics from 1,000 bootstrap replications.

<table>
<thead>
<tr>
<th>Panel A: Mixture Model Parameters</th>
<th>(1) Component 0</th>
<th>(2) Component 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi )</td>
<td>0.9223</td>
<td>0.0777</td>
</tr>
<tr>
<td></td>
<td>(0.0268)</td>
<td>(0.0268)</td>
</tr>
<tr>
<td>( \mu )</td>
<td>0.0006</td>
<td>0.0049</td>
</tr>
<tr>
<td></td>
<td>(0.0001)</td>
<td>(0.0027)</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>0.0026</td>
<td>0.0259</td>
</tr>
<tr>
<td></td>
<td>(0.0004)</td>
<td>(0.0058)</td>
</tr>
<tr>
<td>( \sigma_{i,j} )</td>
<td>0.0113</td>
<td>0.0281</td>
</tr>
<tr>
<td></td>
<td>(0.0003)</td>
<td>(0.0053)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Mixture Return Distribution</th>
<th>Estimate</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.0010</td>
<td>(0.0002)</td>
</tr>
<tr>
<td>SD</td>
<td>0.0107</td>
<td>(0.0004)</td>
</tr>
<tr>
<td>P10</td>
<td>-0.0031</td>
<td>(0.0004)</td>
</tr>
<tr>
<td>P25</td>
<td>-0.0012</td>
<td>(0.0002)</td>
</tr>
<tr>
<td>P50</td>
<td>0.0007</td>
<td>(0.0001)</td>
</tr>
<tr>
<td>P75</td>
<td>0.0026</td>
<td>(0.0003)</td>
</tr>
<tr>
<td>P90</td>
<td>0.0046</td>
<td>(0.0005)</td>
</tr>
<tr>
<td>Fraction positive</td>
<td>0.5942</td>
<td>(0.0200)</td>
</tr>
<tr>
<td>N</td>
<td>2,610</td>
<td></td>
</tr>
</tbody>
</table>