Journal of Banking and Finance 84 (2017) 25–40

Contents lists available at ScienceDirect

Journal of Banking and Finance

journal homepage: www.elsevier.com/locate/jbf

Intraday online investor sentiment and return patterns in the U.S. stock market

Thomas Renault[a,b,*]

[a] IÉSEG School of Management, Paris, France
[b] PRISM, Université Paris 1 Panthéon-Sorbonne, Paris, France

ARTICLE INFO

Article history:
Received 16 June 2016
Accepted 6 July 2017
Available online 12 July 2017

JEL classification:
G02
G12
G14

Keywords:
Asset pricing
Investor sentiment
Intraday return predictability
Textual analysis
Machine learning
Social media

ABSTRACT

We implement a novel approach to derive investor sentiment from messages posted on social media before we explore the relation between online investor sentiment and intraday stock returns. Using an extensive dataset of messages posted on the microblogging platform StockTwits, we construct a lexicon of words used by online investors when they share opinions and ideas about the bullishness or the bearishness of the stock market. We demonstrate that a transparent and replicable approach significantly outperforms standard dictionary-based methods used in the literature while remaining competitive with more complex machine learning algorithms. Aggregating individual message sentiment at half-hour intervals, we provide empirical evidence that online investor sentiment helps forecast intraday stock index returns. After controlling for past market returns, we find that the first half-hour change in investor sentiment predicts the last half-hour S&P 500 index ETF return. Examining users' self-reported investment approach, holding period and experience level, we find that the intraday sentiment effect is driven by the shift in the sentiment of novice traders. Overall, our results provide direct empirical evidence of sentiment-driven noise trading at the intraday level.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Since the pioneering work of Antweiler and Frank (2004) and Das and Chen (2007) on the predictability of stock markets using data from Internet message boards, a growing number of researchers have tried to "explore" the Web to provide forecasts for the financial markets (see, Nardo et al. (2016), for a survey of the literature). From both theoretical and empirical perspectives, two main elements can explain why messages posted by investors on the Internet could give rise to periods of departure from the efficient market hypothesis.[1]

First, given the tremendous increase in the flow of textual content published every day on the Internet, we may wonder whether value-relevant information about fundamental stock prices could be identified and exploited by traders able to process information and trade quickly. This situation would be consistent with the Grossman and Stiglitz (1980) framework of market efficiency, in which small excess returns simply represent the compensation for investors who spend time and money to continuously monitor a wide variety of information sources. Developing and maintaining infrastructures and algorithms to analyze billions of messages posted on the Internet every day has a cost, and an albeit low level of predictability can be viewed as a financial reward that helps to solve the fundamental conflict between the efficiency with which markets spread information and the incentives for acquiring information. Nonetheless, this value-relevant information should be short-lived, as fast-moving traders will compete to take advantage of any existing anomalies. Testing this hypothesis empirically would thus require combining intraday stock market data with high-granularity time-stamped textual data.

Second, studies in behavioral finance argue that stock prices may deviate temporarily from their fundamental values in the presence of sentiment-driven noise traders with erroneous stochastic beliefs (De Long et al., 1990) and limits to arbitrage (Pontiff, 1996; Shleifer and Vishny, 1997). According to Baker and Wurgler (2007), the question is no longer whether investor sentiment affects stock prices, but how to measure investor sentiment and quantify its effects. Various proxies have been used in the literature, and a significant degree of stock return predictability has been identified using investor sentiment proxies from surveys (Brown and Cliff, 2005), market data (Baker and Wurgler, 2006) or

* Correspondence to: Université Paris 1 Panthéon-Sorbonne, 17 rue de la Sorbonne, 75005 Paris, France.
E-mail addresses: thomas.renault@univ-paris1.fr, t.renault@icloud.com

[1] In the sense of Jensen (1978), "a market is efficient with respect to information set θ_t if it is impossible to make economic profits by trading on the basis of information set θ_t".

http://dx.doi.org/10.1016/j.jbankfin.2017.07.002
0378-4266/© 2017 Elsevier B.V. All rights reserved.

traditional media content (Tetlock, 2007). Recently, researchers in behavioral finance have also paid special attention to the construction of investor sentiment proxies using data from the Internet. Extracting and analyzing millions of messages published on the Web to measure investor sentiment may, at first sight, sound appealing, as it could overcome issues related to answering bias (survey-based indices), idiosyncratic non-sentiment-related components (market-based measures) or confounding causality (media-based variables). However, while encouraging results have been identified for small capitalization stocks (Sabherwal et al., 2011; Leung and Ton, 2015), until now, the empirical results for large stocks and market indices have been disappointing (Nardo et al., 2016). Computing investor sentiment using machine learning algorithms on data from Yahoo! Finance message boards, Antweiler and Frank (2004) and Das and Chen (2007) find no economically significant relation between user-generated content and stock returns. These results were confirmed recently by Kim and Kim (2014) on an extensive dataset of 32 million of messages and for a longer sample period: investor sentiment proxied by user-generated content is positively affected by previous stock performances but does not help predict future stock returns, volume or volatility.

However, today communication on social media is very different from chatter on message boards several years ago. Numerous papers report increasing use of social media by market participants, from large quantitative hedge funds to family offices and high-frequency-trading firms.[2] Little anecdotal evidence, like the integration of Twitter and StockTwits feeds into financial platforms (Bloomberg Terminal and Thomson Reuters Eikon), seems to confirm this phenomenon. Given the evolution of the regulatory framework[3] and the constantly changing nature of communication on the Internet, we believe that the "news or noise" question raised by Antweiler and Frank (2004) must be reassessed frequently. Thus, we add to the recent and expanding literature that examines new data from the Internet to forecast stock markets (see, among others, Da et al., 2015; Moat et al., 2013; Avery et al., 2016; Chen et al., 2014; Sprenger et al., 2014a) by focusing on user-generated content published on the social media platform StockTwits.

This paper contributes both to the literature on intraday return predictability and to the literature on textual analysis in finance. Analyzing ETF price dynamics, Gao et al. (2017) (GHLZ hereafter) provide empirical evidence showing that the first half-hour return predicts positively the last half-hour return. Theoretically, the market intraday momentum is consistent with an infrequent rebalancing mechanism (Bogousslavsky, 2016) and with the presence of late-informed traders in the market. Extending GHLZ model by exploring the relationship between intraday stock market returns and intraday sentiment, Sun et al. (2016) (SNS hereafter) find that the change in investor sentiment has predictive value for the intraday market returns. The signs of the estimated coefficients for the change in investor sentiment are positive on all regressions: sentiment-driven optimistic (pessimistic) traders create short-term upward (downward) price pressure, especially during the end of the trading day. One potential explanation proposed by both GHLZ and SNS is related to limits to arbitrage, as risk averse market makers might hesitate to trade against over-optimistic (over-pessimistic) noise traders during the last half-hour of the trading days to avoid exposures to overnight risks, resulting in a short-term pricing anomaly.

Regarding textual analysis in finance, one of the many challenges faced by academics and practitioners in this field concerns the methodology used to automatically convert a qualitative variable—a message, a blog post, or a tweet—into a quantitative sentiment variable. Two main methods are used for textual sentiment analysis in finance: dictionary-based approaches and machine learning techniques (see, Kearney and Liu, 2014; Das, 2014, for surveys of methods and models). Whereas dictionary-based methods that use the Harvard-IV dictionary or the Loughran and McDonald (2011) dictionary (LM hereafter) are widely used in the literature to measure sentiment in papers published in traditional media (Tetlock, 2007; Tetlock et al., 2008; Engelberg et al., 2012; Dougal et al., 2012; Garcia, 2013), textual sentiment analysis of user-generated content published on the Internet mainly relies on machine learning algorithms (Antweiler and Frank, 2004; Das and Chen, 2007; Sprenger et al., 2014b; Leung and Ton, 2015; Ranco et al., 2015). Although each method has its own advantages and limits, as we will discuss later, one simple reason that explains the predominance of machine learning techniques to quantify individual messages posted on message boards and social media is the absence of a field-specific dictionary. Messages published by online investors on the Internet are usually shorter and less formal than content published on traditional media, making the correct classification of tone difficult (Loughran and McDonald, 2016). Nonetheless, as stated by Nardo et al. (2016), "a good text classifier for a financial corpus is a good avenue for future research," as it could facilitate the comparability and enhance the replicability of previous findings.

In this paper, we first implement a novel approach to construct a lexicon of words used by investors when they share ideas and opinions about the bullishness or bearishness of the stock market on social media. Following Oliveira et al. (2016), we use a subset of 750,000 messages already tagged by online investors as bullish (positive) or bearish (negative) to automatically construct a field-specific weighted lexicon (L_1 hereafter). We also develop a field-specific non-weighted lexicon (L_2 hereafter) by examining and classifying manually all words that appear at least 75 times in the sample, adopting a methodology close to Loughran and McDonald (2011). Then, we use L_1 and L_2 to derive sentiment in a subset of 250,000 tagged messages, and we compare the out-of-sample classification accuracy with three baseline methods: a dictionary-based approach using the LM dictionary (B_1 hereafter), a dictionary-based approach using the Harvard-IV dictionary (B_2 hereafter) and a supervised machine learning algorithm using a maximum entropy classifier (M_1 hereafter). We find that L_1, L_2 and M_1 significantly outperform the standard dictionary-based approaches B_1 and B_2. Thus, the results confirm Kearney and Liu (2014) conclusion about the need to construct more authoritative and extensive field-specific dictionaries in order to enhance replicability and facilitate future work in the area.

Then, we examine the relation between online investor sentiment and intraday stock returns using an extensive dataset of nearly 60 million messages published by online investors over a five-year period, from January 2012 to December 2016. We compute five distinct intraday investor sentiment measures by aggregating the sentiment of individual messages posted on the microblogging platform StockTwits at half-hour intervals. We follow Heston et al. (2010) by dividing each trading day into 13 half-hour trading intervals, and we reassess the intraday momentum and the intraday sentiment effect documented by Gao et al. (2017) and Sun et al. (2016). We find that when investor sentiment is computed using L_1, L_2 and M_1, the first half-hour change in investor sentiment predicts positively the last half-hour S&P 500 index ETF returns. After controlling for the lagged market return and the first

[2] See, for example, http://www.wsj.com/articles/tweets-give-birds-eye-view-of-stocks-1436128047 "The Wall Street Journal - Firms Analyze Tweets to Gauge Stock Sentiment"

[3] https://www.sec.gov/rules/interp/2008/34-58288.pdf See "Commission Guidance on the Use of Company We Sites" and https://www.sec.gov/News/PressRelease/Detail/PressRelease/1365171513574 "SEC Says Social Media OK for Company Announcements if Investors Are Alerted"

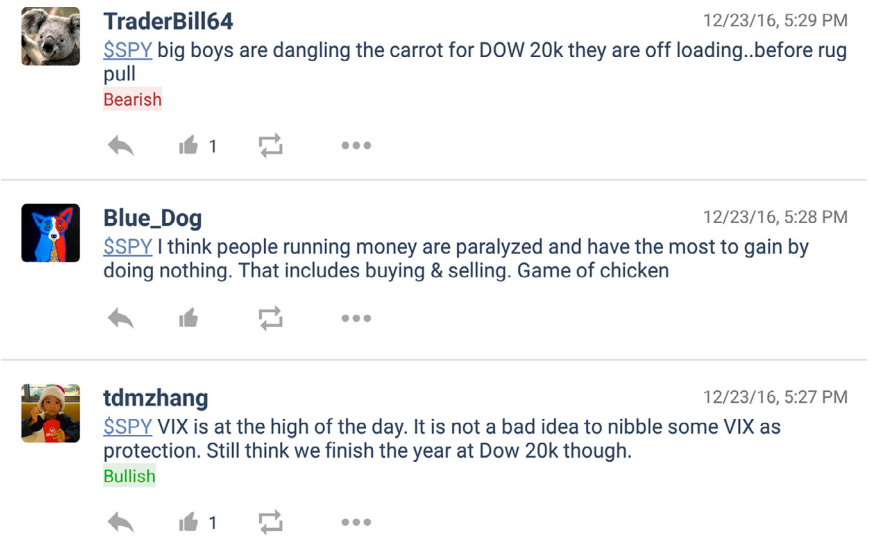

Fig. 1. StockTwits platform - Explicitly revealed sentiment. *Notes*: This figure shows a screenshot from StockTwits platform on December 23, 2016. The first message was self-classified as bearish (negative) by the investor who wrote the tweet (TraderBill64). The second message was not classified. The third was classified as bullish (positive) by the investor who wrote the tweet (tdmzhang). $SPY is the cashtag associated with the S&P 500 index ETF.

half-hour return, we find that first half-hour change in investor sentiment remains the only significant predictor of the last half-hour market return. In contrast, the predictability disappears when sentiment is computed using B_1 or B_2.

Analyzing users' self-reported information on their investment approach (technical, fundamental, momentum, value, growth or global macro), holding period (day trader, swing trader, position trader or long-term investor) and experience level (novice, intermediate or professional), we construct intraday investor sentiment indicators for each group of users. We find that the intraday sentiment effect is mainly driven by the shift in the sentiment of novice traders. Implementing a trading strategy using the change in novice traders' sentiment as a trading signal to buy (sell) the S&P 500 ETF during the last half-hour of the trading day before selling (buying) it at market close, we demonstrate that a sentiment-driven strategy delivers a significantly higher risk-adjusted performance compared to baseline strategies (momentum, long-only, first half-hour and random strategies).

This paper supports the role of investor sentiment in predicting intraday stock returns and adds to the existing literature for various reasons. First, our results contrast with previous findings from GHLZ by demonstrating that the intraday price momentum has disappeared during the most recent sample period. While SNS find that both the sentiment effect and lagged return variables help predict the last half-hour return on a sample period from 1998 to 2016, we find that the sentiment effect is the only predictor of the last half-hour return on a sample period from 2012 to 2016. Second, we demonstrate that the intraday sentiment-driven anomaly is very short-lived: a positive sentiment-driven price pressure on day t is followed by a price reversal on the next trading day, consistent with the noise trading hypothesis. Third, and contrary to the measure of investor sentiment based on the proprietary Thomson Reuters MarketPsych Indices (TRMI) used by SNS, our investor sentiment measure is transparent, replicable, and allows us to provide a more direct test of the noise trading hypothesis. Exploring investor base heterogeneity and focusing on users' experience level (novice, intermediate, professional), we provide to the best of our knowledge the first direct empirical evidence of intraday sentiment-driven noise trading.

The paper is structured as follows. Section 2 describes the StockTwits platform and gives details about the data. Section 3 reviews the differences between dictionary-based methods and machine-learning techniques and compares the classification accuracy of L_1 and L_2 with other baseline methods used in the literature. Section 4 explores the relation between online investor sentiment and intraday stock returns. Section 5 concludes and discusses further research.

2. Data

StockTwits is a social microblogging platform dedicated to financial markets on which individuals, investors, market professionals and public companies can publish 140-character messages to "Tap into the Pulse of the Markets". According to StockTwits.com, more than 300,000 users now use the platform to share information and ideas, producing streams that are viewed by an audience of more than 40 million across the financial web and social media platforms. In September 2012, StockTwits implemented a new feature that allows users to express their sentiment directly when they publish a message on the platform. More precisely, every time a user chooses to post a message on StockTwits, he or she can classify his or her message as "bearish" (negative) or "bullish" (positive) by simply clicking on a toggle button below his or her message. Fig. 1 shows a screenshot from the StockTwits platform, with a bearish message, an unclassified message and a bullish message.

Using the Python library *BeautifulSoup*, we extract all messages published on StockTwits between January 1, 2012, and December 31, 2016, and we store them in a MongoDB NoSQL database. For each message, we collect the following information: (1) a unique identifier, (2) the username of the user who sent the message, (3) the message content, (4) the time stamp with a one-second granularity and (5) the sentiment ("bullish", "bearish" and "unclassified") associated with the message. Table 1 shows a sample of messages from the database, with the sentiment variable associated. Our final dataset contains 59,598,856 messages from 239,996 distinct users. Overall, 9,434,321 messages are classified as bullish (15.85%) and 2,286,292 as bearish (3.84%), and the remaining are unclassified. The 4 to 1 ratio between positive and negative messages shows that online investors are, on average, optimistic about the stock markets, as already documented in the literature (see, e.g., Kim and Kim, 2014; Avery et al., 2016).

Table 2 presents descriptive statistics of StockTwits messages during the sample period. Fig. 2 represents the volume of messages per 30-min intervals during a representative week, illustrating the

Table 1
StockTwits - Sample messages.

Date	Sentiment	User	Content
2014-11-11 14:42:49	Bearish	albogolfer	$BABA Dont catch the falling knife !! Fraud company !! All Lies
2014-11-11 14:42:49	N/A	Saratoga733	@jahnman @123xyz what is time frame on the 80mil or is it only on as needed basis?
2014-11-11 14:42:49	Bearish	lbailey33	$BABA IS PURE TRASH!!! JUNK PRODUCTS!!!!
2014-11-11 14:42:51	N/A	Akcoline	$BABA would it be reason if they will be able to fill the orders of $9.6 bil
2014-11-11 14:42:54	N/A	2Struong	$BABA will dive more in the last 30 min of trading today
2014-11-11 14:42:56	N/A	Poweruhvone	@neve1064 When did SIRI IPO? What did it do for monthsyear after the IPO? Your [...].
2014-11-11 14:42:57	Bullish	Dannydyn	$TSLA BURN YOU SHORTS, BUUUUUUUUUUUUUUURRRRNNNNN!!!!
2014-11-11 14:42:58	Bullish	rosswestgate	$KERX why are we stuck? other firms would be at 35 by now lol
2014-11-11 14:42:59	N/A	DZee	@Ben_Frank_Beatdown Now that MADE ME LAUGH!!!
2014-11-11 14:43:00	N/A	seer	$PLI.CA Chen will likely be sharing Lenovo partnership talk and more tomorrow
2014-11-11 14:43:00	N/A	GrindTime	@BehavioralEdge GILD already has 100% success rate in 500+ person Japan trial. Big deal [...]
2014-11-11 14:43:00	N/A	Dawnr48	@Crackerjax if holds 28.12 good r:r maybe to 29.
2014-11-11 14:43:00	N/A	adamh9999	$CODI has good earnings and the offering is at 17.50 a share. This will go up from here.
2014-11-11 14:43:01	N/A	kim2020	this will turn tomorrow!!
2014-11-11 14:43:02	Bearish	OutOfBounds1	$SPY Oh just fall already.
2014-11-11 14:43:04	N/A	ThunderRoad	@MrMulally When that happens, this could be bigger than Apple.
2014-11-11 14:43:05	N/A	SmokeyNYY	@BiopharmaPro The funny thing is Benitec is further along with their Macular Degeneration than RXII
2014-11-11 14:43:05	N/A	FinancesStocks	$AMSWA American Software, Inc. Invites You to Join Its Second Quarter Fiscal Year [...]
2014-11-11 14:43:06	N/A	iowasooner	$APP imo probably best to sell wsmall profit & possibly revisit after next ER... .BOD aren't helping
2014-11-11 14:43:08	N/A	Gbones2692	@Daks Yes very confident with UGAZ to 24 min by mid jam
2014-11-11 14:43:08	Bullish	Beautifulla	@viktor89 Nice I got a big big smiling on my face hahaha
2014-11-11 14:43:15	N/A	Pisces_19	@LMikeH Thanks Mike! Hope this is just temporary as these providers sure need upgrades soon.
2014-11-11 14:43:16	N/A	dinovix2778	@jgp0006 it's up to you. Don't take any advice. Do your own math and sell when you feel good [...]
2014-11-11 14:43:19	N/A	kaluhiokalani71	$PCLN oh gawwd! this is tearing faces off shortz!
2014-11-11 14:43:19	N/A	stckwhisperer	@BounceBabyBounce Do you mean limit order?
2014-11-11 14:43:20	N/A	joelyboyblue	$FOSL Long some
2014-11-11 14:43:20	Bullish	Farne	@jp007 we go green hornet brotha
2014-11-11 14:43:21	N/A	Kroberg	@Trade4moneydaily Don't have a crystal ball, sorry
2014-11-11 14:43:21	N/A	regent17	@MoreCash2014 @IndianTrader I feel it, today bought 1500 shares, on may, i did the same thing at $8 !
2014-11-11 14:43:23	N/A	tubetoppuller	$ASTI float still around 10M?
2014-11-11 14:43:23	N/A	M10	$JNUG CPCharts-You got it wrong. Weak yen=strong dollar=lower gold. Wrong information [...]
2014-11-11 14:43:27	N/A	milkball	$TSLA Who is holding their calls overnight?

Notes: This table shows all messages sent on StockTwits between 14:42:49 and 14:43:27 on November 11, 2014. Messages can be classified as positive (bullish) or negative (bearish) by the investor/user who sent the tweet, or can be published without any associated sentiment (N/A).

Table 2
Descriptive statistics - StockTwits messages.

Period	Mean	Std-Dev	Min	Max	Total
30-min (All)	886.65	679.61	0	8248	59,598,856
30-min (Trading-hours)	1092.67	1918.98	0	8248	31,383,060
Daily (All)	26,649.86	32,621.16	1127	132,063	59,598,856
Daily (2012)	6,805.23	11,488.46	1127	27,831	4,204,778
Daily (2013)	10,251.82	17,786.75	2070	46,501	6,492,164
Daily (2014)	16,088.48	29,765.41	4100	59,310	10,864,373
Daily (2015)	21,766.21	42,323.00	6442	80,936	15,447,896
Daily (2016)	32,435.49	61,720.34	9153	132,063	22,589,645

Notes: This table shows descriptive statistics about the quantity of messages posted on the platform StockTwits. We present statistics at half-hour time interval and at a daily frequency for each year in our sample (2012–2016).

intraday and weekly seasonality of message posted on the social media platform. Intraday activity on StockTwits usually peaks at market opening (between 9:30 a.m. and 10:00 a.m.), decreases at lunchtime and increases again before market close (between 3:30 p.m. and 4:00 p.m.). During non-trading hours and weekends, the average number of messages per 30-min interval is approximately 10 times lower than during trading hours (over the whole sample period).

3. Textual sentiment analysis

Before assessing whether user-generated content can help predict stock returns, academics and practitioners have to implement specific procedures to convert unstructured qualitative information into structured quantitative sentiment variables. In this section, we briefly review the two distinct approaches used for textual sentiment analysis, before we detail the methodology we implement to construct field-specific lexicons and compare our results with the benchmark classifiers used in the literature.

3.1. Dictionary-based classification

In the simplest form, a dictionary-based "bag-of-words" approach consists of computing a sentiment variable by counting the number of positive words and the number of negative words in a document, using a predefined list of signed words. For example, in a simple 4-word lexicon where "good" and "love" are defined as positive and "bad" and "hate" are defined as negative, the sentence "I love Facebook $FB company" is classified as positive with a score of +1.

Three main procedures can be implemented to create lexicons for sentiment analysis. The first technique relies on pure experts' views, in which researchers create from scratch a list of positive and negative words, based on their knowledge and expertise. The

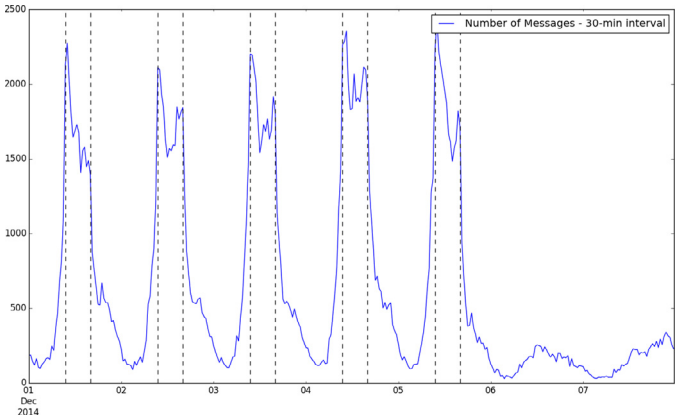

Fig. 2. StockTwits - Number of messages per 30-min interval. *Notes*: This figure shows the number of messages published on the platform StockTwits for each 30-min interval on a representative week, from Monday, December 1, to Sunday, December 7, 2014. Dashed vertical lines represent market opening hours (9:30 a.m.) and market closing hours (4 p.m.).

second technique, used, for example, to construct the LM dictionary, is a two-step process in which a vector of words is automatically generated by analyzing a list of non-classified documents. Then, each word is manually classified as positive, negative or neutral by an expert.[4] The last technique consists of creating or extracting a list of pre-classified documents and, for each word, computing statistical measures based on the term's frequency (and/or document frequency) in each class of documents. Term frequency thresholds are then used to classify each word as positive, neutral or negative.

Although a dictionary-based approach is easy to implement, and if the list of signed words is public, enables replicability, this approach has some limitations. First, it is necessary to develop field-specific dictionaries for each domain of research, as a word may not have the same meaning in two different contexts. For example, words like "liability", "capital" and "cost" are classified as negative in the Harvard-IV psychosocial dictionary but should be considered otherwise in finance (Loughran and McDonald, 2011). Furthermore, even in a given area like financial markets, formal articles written by financial journalists on traditional media are very different from user-generated content published by individual investors on the Internet. According to Loughran and McDonald (2016), the use of slang, sarcasm, emoticons and the constantly changing vocabulary on social media makes accurate classification of tone difficult. Second, except for rare exceptions (Jegadeesh and Wu, 2013), the vast majority of dictionary-based approaches uses an equal-weighting scheme, where each word in the dictionary is supposed to have the same explanatory power. Although term-weighting has the potential to increase the accuracy of textual analysis, the large number of available weighting procedures may give too many degrees of freedom to researchers in selecting the best possible empirical specification (Loughran and McDonald, 2016), creating a risk of overfitting.

3.2. Machine learning classification

The objective of a machine learning classification is to provide a prediction of Y given a set of features X. For a 2-class sentiment analysis problem, Y represents sentiment classes $Y_1 = positive$ and $Y_2 = negative$ and X is a vector of words. A supervised learning classification problem can be decomposed in three steps: (1) learn in-sample, (2) measure accuracy out-of-sample and (3) predict. First, a training dataset of n documents d pre-classified as positive or negative is used to fit the algorithm (see, Pang et al. (2002) for a description and a mathematical explanation of three of the most widely used classifiers in the literature: naive Bayes, support vector machine and maximum entropy). Then, features identified during the learning phase are used to predict the Y class on a testing dataset of n' pre-classified documents d'. Classification accuracy is computed by comparing the classifier prediction to the known value of Y for all documents in d'. When the accuracy of the prediction cannot be improved by modifying or fine-tuning the parameters and/or is in line with previous findings in the literature, then the algorithm is used to predict the outcome Y for all documents where class Y is unknown.

A machine learning technique has many advantages compared to a dictionary-based approach. Instead of relying on a (somehow subjective and limited) list of signed words, it allows the automatic construction of a very large set of features specific to the domain of interest and to the type of data. Furthermore, machine learning algorithms can provide answers to problems related to the weighting procedure or the non-independence of words in a sentence. However, this does not come without limitations. The first difficulty is to create or extract a sufficiently large list of labeled documents to construct a training dataset and a testing dataset. In most cases, documents are labeled manually by the author(s) or by financial expert(s) so there is subjectivity.[5] Second, machine learning accuracy can be very sensitive to the size and the construction of the training dataset. For example, Antweiler and Frank (2004) manually labeled only 1000 messages from Yahoo! Finance message boards (55 negative, 693 neutral and 252 positive) to train their classifier, raising concerns about the accuracy of the classification when the algorithm is fitted on such a low number of messages. Third, supervised classification accuracy can change significantly depending on the algorithm used (naive Bayes, support vector machine, maximum entropy, random forests, neural network...) and few fine-tuning arbitrary parameters. As most papers use a (private) manually labeled training dataset and a specific set of (often) unpublished rules, filters or parameters to fit the data, replicability and comparison across studies are often impossible.

3.3. Creating an investor lexicon

To create our lexicon, we follow Oliveira et al. (2016) automated procedure by focusing on messages in which sentiment is explicitly revealed by online investors. We first randomly select a list of 375,000 "bullish" messages and 375,000 "bearish" messages published on StockTwits between June 2013 and August 2014. As in Pang et al. (2002), we impose a maximum of 375 messages per user and per class (or 0.1% of the whole corpus) to avoid domination of the corpus by a small number of prolific reviewers. We implement a data cleaning process similar to Sprenger et al. (2014b), except that we choose to keep the punctuation (question marks and exclamation marks) and we do not remove the morphological endings from words. To take negation into account, we add the prefix "negtag_" to all words following "not", "no", "none", "neither", "never" or "nobody".

Although various natural language processing approaches could have been applied (lemmatization, stemming, part-of-speech tag-

[4] For example, Loughran and McDonald (2011) extract all words occurring in at least 5% of 121,217 10-K reports downloaded directly from the Security and Exchange Commission website, before manually classifying the "eligible words" as positive, negative or neutral.

[5] A system in which each message is classified by two different reviewers can be implemented to partly overcome this issue. However, as shown by Das and Chen (2007) on a sample of 438 messages posted on Yahoo! Finance message boards, the level of agreement between two human experts can be very low, with a mismatch percentage of 27.5% in their sample.

Table 3
StockTwits messages - Data pre-processing.

Message before pre-processing	@lololemon $BABA IS PURE TRASH !!
Message after pre-processing	usertag cashtag is pure trash ! !
Message before pre-processing	$FB dropping now! not good :(
Message after pre-processing	cashtag dropping now ! negtag_good emojineg
Message before pre-processing	$MSFT Short the POP
Message after pre-processing	cashtag short pop
Message before pre-processing	$GILD moves like Jagger! http://stks.co/r0nUR
Message after pre-processing	cashtag moves like jagger ! linktag

Notes: This table shows four examples of messages before and after data pre-processing (removing stopwords, adding prefix for negation, replacing users' mention by "usertag", tickers by "cashtag", links by "linktag"...).

ging), we choose to use a conservative approach by removing only three stopwords from all messages ("a", "an" and "the").[6] We also convert positive emoticons into a common word "emojipos" and negative emoticons into a common word "emojineg"[7], as in Go et al. (2009). We replace all tickers ($SPY, $AAPL, $BOA, $XOM...) with a common word "cashtag", all links by a common word "linktag", all numbers by a common word "numbertag" and all mentions of users by a common word "usertag". Table 3 shows several examples of messages before and after data pre-processing.

We use a bag-of-words approach to extract all unigrams (one word) and bigrams (two words) appearing at least 75 times in the sample of 750,000 messages. While the Harvard-IV and the LM dictionary consider only unigrams, we find that adding bigrams provides additional information and improves the accuracy of the classification.[8] For each of the 19,665 terms t identified (5786 unigrams and 13,879 bigrams), we count the number of occurrences of t in the 375,000 bullish documents ($n_{d_{pos},t}$) and the number of occurrences of t in the 375,000 bearish documents ($n_{d_{neg},t}$). We define the sentiment weight (SW) for each word as:

$$SW(t) = \frac{n_{d_{pos},t} - n_{d_{neg},t}}{n_{d_{pos},t} + n_{d_{neg},t}} \quad (1)$$

Table 4 shows a list of selected n-grams with their associated sentiment weight. For example, the word "buy" was used 20,837 times in bullish messages and 12,654 times in bearish messages, leading to a SW of 0.2443. Interestingly, we find that the bigrams "buy !" and "strong buy" convey a much more positive sentiment than the unigram "buy", with an SW equal to 0.6052 and 0.8250, respectively. The bigram "buy ?" is approximately neutral (SW equals 0.0331) whereas "negtag_buy" ("not buy", "never buy"...) conveys a negative sentiment (SW equals −0.4534).

Then, we sort all 19,665 n-grams by their SW, and we define a weighted field-specific lexicon L_1 by considering all terms in the first quintile (negative terms) and all terms in the last quintile (positive terms). Manually examining all words included in lexicon L_1 (approximately 8000 n-grams), we identify a few anomalies and misclassifications. For example, the word "further" is classified as negative, as it appears 1260 times in the 375,000 negative documents and 506 times in the 375,000 positive documents, leading to an SW of −0.4270 (in the first quintile). Analyzing the n-gram frequencies, we find that the word "further" is often used in combination with verbs like "drop," "down" and "fall" ("drop further", "down further," "fall further"), in such a way that the negativity does not come from the word "further" by itself but from the verb associated with it in the bigrams. Another anomaly is related to non-equity assets. For example, the unigram "commodity" is considered negative in L_1, because, during the sample period, commodity prices dropped, and investors were mainly commenting on past movements using bearish vocabulary. The same is true for the unigrams "Euro" and "EURUSD" as the euro currency depreciates sharply against the dollar during the sample period.

Thus, we adopt a methodology close to Loughran and McDonald (2011) to create a manually cleaned equal-weighted field-specific lexicon. More precisely, we examine all n-grams in L_1, and we manually classify each n-gram as positive (+1), negative (−1) or neutral (0). We also add typical inflections of root words defined as positive or negative to extend our lexicon. For example, we manually classify the words "bankrupt" and "bankruptcy" as negative, and we add the inflections "bankrupts", "bankrupted", "bankrupting" and "bankruptcies". We end up with a total of 543 positive terms and 768 negative terms, and we denote this lexicon L_2. L_1 and L_2 are available online.[9]

3.4. Message sentiment and classification accuracy

To assess the accuracy of L_1 and L_2, we use a time-order evaluation holdout. We randomly select a list of 125,000 bullish messages and 125,000 bearish messages published on StockTwits between September 2014 and April 2015. We use the same pre-processing techniques and the same limit of messages for a given user as for the training dataset (maximum 0.1% of the whole corpus). For each message, we compute a sentiment score by considering five classifiers:

- L_1 - Weighted field-specific lexicon: approximately 4000 negative outlook terms and 4000 positive outlook terms. $SW(t)$ as defined previously.
- L_2 - Manual field-specific lexicon: 768 negative outlook terms and 543 positive outlook terms. $SW(t)$ equals 1 for positive terms and −1 for negative terms.
- B_1 - Loughran–McDonald dictionary: 2355 negative outlook terms and 354 positive outlook terms. $SW(t)$ equals 1 for positive terms and −1 for negative terms.
- B_2 - Harvard-IV psychosocial dictionary: 2007 negative outlook terms and 1626 positive outlook terms. $SW(t)$ equals 1 for positive terms and −1 for negative terms.
- M_1 - Supervised machine learning algorithm (maximum entropy): Implemented using scikit-learn, a machine learning package in Python. Default parameters and equal prior probabilities.

For L_1, L_2, B_1 and B_2, the individual message sentiment score is defined as the average $SW(t)$ of the terms present in the message.

[6] We choose a conservative approach as we find that the words "short", "shorts", "shorted", "shorter", "shorters" and "shorties" are used by online investors to express very distinct feelings. The same is true for the words "call", "calls", "called", "calling", "caller", "callers" and for a subsequent number of words.

[7] ;) :) :-) =) :D as "emojipos". :(:-(=(as "emojineg

[8] For example, the sentence "What a bear trap!" should be not be classified as negative (i.e., "bear trap" is an expression used in technical analysis to indicate that a security should go up) even if "bear" and "trap" are individually considered negative.

[9] http://www.thomas-renault.com.

Table 4
Selected sample of n-grams and associated Sentiment Weight (SW).

n-grams	n_{total}	n_{pos}	n_{neg}	SW
awesome	1447	1077	370	0.4886
bear	5669	1506	4163	−0.4687
bear trap	393	250	143	0.2723
beast mode	182	172	10	0.8901
bottomed-out	137	127	10	0.8540
bullish	11,483	7812	3671	0.3606
bullish engulfing	121	112	9	0.8512
buy	33,491	20,837	12,654	0.2443
buy !	765	614	151	0.6052
buy ?	302	156	146	0.0331
cashtag junk	95	1	94	−0.9789
down	42,391	11,388	31,003	−0.4627
down further	145	25	120	−0.6552
emojineg	1885	401	1484	−0.5745
emojipos	15,223	10,091	5132	0.3258
great	11,952	8380	3572	0.4023
great fundamentals	126	120	6	0.9048
intraday	1334	557	777	−0.1649
investor	1493	869	624	0.1641
like	35,756	17,845	17,911	−0.0018
media	1038	557	481	0.0732
negtag_buy	1577	431	1146	−0.4534
negtag_short	781	290	491	−0.2574
optimism	185	91	94	−0.0162
poor	1467	333	1134	−0.5460
poor fundamental	136	0	136	−1.0000
price	20,730	10,393	10,337	0.0027
pump	4501	659	3842	−0.7072
scam	1540	116	1424	−0.8494
sell	23,183	6637	16,546	−0.4274
sentiment	1982	619	1363	−0.3754
short	47,856	10,022	37,834	−0.5812
stock	32,781	13,928	18,853	−0.1502
strong	8223	5966	2257	0.4511
strong buy	557	507	50	0.8205
timber	398	17	381	−0.9146
today	38,761	21,604	17,157	0.1147
trading	8383	3934	4449	−0.0614
trap	1867	426	1441	−0.5437
up	61,337	37,823	23,514	0.2333
up	786	720	66	0.8321
word	817	473	344	0.1579

Notes: This table shows the Sentiment Weight (SW) of a sample of selected words. For example, over the 750,000 messages we use to construct our lexicon, the word "buy" appears 33,491 times in the positive training dataset (375,000 messages) and 20,837 times in the negative training dataset (375,000 messages), leading to a sentiment weight *SW* of (33,491 - 20,837) / (33,491 + 20,837) = 0.2443. Red and green colors represent n-grams with a *SW* respectively in the first and last quintile (when sorting all 19,665 n-grams by their *SW*).

Given the standardized number of words in each document (maximum 140 characters), we find that using a simple relative word count weighting scheme gives slightly better results than a Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme (see Appendix A for details). This result is consistent with those of Smailović et al. (2014), who find, using data from Twitter, that the term-frequency (TF) approach is statistically significantly better than the TD-IDF based approach. For M_1, individual message sentiment score is given by the probability estimates that a message m belongs to the bullish or the bearish class. See Appendix B for a detailed description. For all messages in the testing dataset, we compare the sentiment expressed by the investor who sent the message (the real sentiment) with the sentiment score computed using the five classifiers (the estimated sentiment). We compute the percentage of correct classification excluding unclassified messages *CC* (i.e, bearish-declared messages with a sentiment score lower than 0 and bullish-declared messages with a sentiment score greater than 0), the percentage of correct classification per class (CC_{bull} and CC_{bear}, respectively), the percentage of classified messages *CM* (message with a sentiment score different from zero) and the percentage of classified messages per class (CM_{bull} and CM_{bear}). Table 5 presents the results.

Table 5
Classification accuracy - Investor social lexicons.

Classifier	*CC*(%)	CC_{bull}(%)	CC_{bear}(%)	*CM*(%)	CM_{bull}(%)	CM_{bear}(%)
L_1	74.62	73.98	**75.24**	**90.03**	**89.32**	**90.73**
L_2	**76.36**	**79.10**	73.72	61.78	60.61	62.95
B_1	63.06	57.99	67.86	27.70	26.88	28.50
B_2	58.29	63.63	53.02	58.09	57.72	58.47
M_1	75.16	75.98	74.36	**90.03**	**89.32**	**90.73**

Notes: This tables shows the out-of-sample classification accuracy for classifiers L_1, L_2, B_1, B_2 and M_1, computed on 250,000 messages from the testing dataset (125,000 positive and 125,000 negative). We report the percentage of correct classification excluding unclassified messages *CC*, the percentage of correct classification per class (respectively CC_{bull} and CC_{bear}), the percentage of classified messages *CM* (message with a sentiment score different from zero) and the percentage of classified messages per class (CM_{bull} and CM_{bear}).

We find a percentage of correct classification of 74.62% for L_1 and 76.36% for L_2. As the number of features is much greater in L_1 (approximately 8000 n-grams) than in L_2 (approximately 1300 n-grams), the percentage of classified messages *CM* is greater for L_1 (90.03%) than for L_2 (61.78%), leading to an expected arbitrage between accuracy and exhaustiveness. Interestingly, and contrary to Oliveira et al. (2016), we find that the accuracy and the percentage of the classified messages are nearly equivalent for the bullish and bearish messages for L_1.[10] However, the percentage of correct classification of benchmark dictionary-based approaches B_1 (LM) and B_2 (Harvard-IV) is significantly lower, with an accuracy of 63.06% and 58.29%, respectively. Furthermore, the percentage of classified messages in B_1 is very low (27.70%) as numerous messages published on social media do not contain any words included in the LM word lists. The LM dictionary was created by examining formal corporate 10-K reports in such a way that it is not well suited to analyze informal messages published on social media. This first result confirms Kearney and Liu (2014) discussion on the need to construct more authoritative and extensive field-specific dictionaries in order to improve textual analysis classification.

We also find that the classification accuracy of the supervised machine learning method M_1 is slightly better (75.16%) than that of L_1 (74.62%). However, as we will show later, results for the relation between investor sentiment and stock returns are qualitatively similar when intraday investor sentiment indicators are computed using L_1, L_2 or M_1. As field-specific dictionary-based approaches are more transparent than machine learning techniques, we believe that researchers should consider thoroughly implementing both methods when quantifying textual content published on the Internet. This dual approach would enhance the replicability and comparability of the findings while ensuring that the results are robust to the methodology used to convert a text into a quantitative sentiment variable. Thus, we re-affirm Loughran and McDonald (2016) conclusion by recommending that alternative complex methods (machine learning) should be considered only when they add substantive value beyond simpler and more transparent approaches (bag-of words).

[10] As we focus our analysis on financial messages published on social media with self-reported sentiment, we cannot compare directly the accuracy of our field-specific approach with previous results from the literature on textual analysis. However, out-of-sample classification accuracy between 75% and 80% is standard on user-generated content sentiment analysis (see Pang et al. (2002), Go et al. (2009) or Smailović et al. (2014), among others).

Table 6
Intraday investor sentiment indicators - Correlation matrix.

	s_{L1}	s_{L2}	s_{B1}	s_{B2}	s_{M1}
s_{L1}	1.0000				
s_{L2}	0.6250	1.0000			
s_{B1}	0.2292	0.3365	1.0000		
s_{B2}	0.2328	0.3000	0.3112	1.0000	
s_{M1}	0.9341	0.6581	0.2629	0.2361	1.0000

Notes: This tables shows the correlation matrix of our five intraday investor sentiment indicators s_x, where x={L_1, L_2, B_1, B_2, M_1}.

4. Intraday online investor sentiment and stock returns

In this section, we explore the relation between online investor sentiment and intraday stock returns. We first detail the methodology we use to derive the investor sentiment indicators by aggregating the sentiment of individual messages. Then, we reassess the intraday momentum patterns documented by GHLZ by considering an augmented sentiment-based model. Last, we analyze whether users' self-reported investment approach, holding period and experience level contain value-relevant information to understand the reason behind the intraday sentiment effect.

4.1. Intraday investor sentiment indicators

We use our five classifiers to derive a sentiment score between −1 and +1 for all 59,598,856 messages published on StockTwits between January 1, 2012, and December 31, 2016. Then, we compute five intraday investor sentiment indicators by averaging, at half-hour intervals, the sentiment score of individual messages published per 30-min period. We denote those indicators s_x where x={L_1, L_2, B_1, B_2, M_1}. To control for the increase in message volume and the seasonality of posting patterns on social media, we standardize s_x by dividing each indicator by its rolling one-week standard deviation. Table 6 shows the correlation between the five s_x indicators.

The very high correlation coefficient between s_{L1} and s_{M1} (0.9341) seems to confirm that quantifying the sentiment of individual messages using a weighted field-specific lexicon is competitive with more complex machine learning methods. However, the correlation coefficients of s_{B1} and s_{B2} with our field-specific approach are low (from 0.2292 to 0.3365) demonstrating that the methodology used to derive quantitative indicators from textual content can widely affect investor sentiment measures.

4.2. Predictive regressions

Following Heston et al. (2010), we divide each trading day into 13 half-hour intervals. We denote $r_{i,t}$ the ith half-hour return of the S&P 500 ETF on day t. As in GHLZ, $r_{1,t}$ is the first half-hour return using the closing price on day -1 and the price at 10:00 a.m. on day t. $r_{13,t}$ denotes the last half-hour return using the ETF price at 3:30 p.m. and 4:00 p.m. on day t. In a similar fashion, we denote $\Delta s_{i,t}$ the change in intraday investor sentiment in the i−th half-hour trading interval on day t. For example, $\Delta s_{1,t}$ denotes the difference between the first half-hour investor sentiment (the average sentiment of all messages sent between 9:30 a.m. and 10:00 p.m.) on day t and the last half-hour sentiment on day $t-1$ (the average sentiment of all messages sent between 3:30 p.m. and 4:00 p.m. on the previous trading day). $\Delta s_{13,t}$ denotes the difference between the last half-hour investor sentiment and the 12th half-hour investor sentiment on day t.

As in SNS, we run predictive regressions to explore the relation between changes in intraday investor sentiment and the half-hour S&P 500 index ETF return. Given GHLZ empirical evidence showing that the first half-hour return predicts the last half-hour return, we also include the first half-hour change in investor sentiment. Thus, we consider the following model:

$$r_{i,t} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 \Delta s_{i,t-1} + \epsilon_t \quad (2)$$

where i represents the ith half-hour time interval. Table 7 shows the regression results for i={11,12,13}.[11] We present the results when investor sentiment is computed using the five classifiers (L_1, L_2, B_1, B_2 and M_1). The regressions are based on 1258 observations (251 or 252 trading days per year from 2012 to 2016).

We find evidence that when investor sentiment is computed using L_1, L_2 or M_1, the first half-hour change in investor sentiment predicts the last half-hour stock market return. Coefficients are significant and positive at the 0.1% level when investor sentiment is computed with L_1 or M_1 and at the 1% level when investor sentiment is computed with L_2. The R^2 values of 1.35% (L_1) and 1.33% (M_1) are comparable to those reported by SNS on the predictability of the last half-hour return using the change in investor sentiment based on the Thomson Reuters MarketPsych Indices (1.43%). However, when investor sentiment is computed using B_1 or B_2, we do not find any predictability. This finding reinforces our conclusion that the Loughran–McDonald and the Harvard-IV psychosocial

Table 7
Predictive regressions - Investor sentiment and half-hour market return.

	α	β_1	β_2	Adj-R^2 (%)
11th half-hour return				
L_1	0.0000	0.0031	0.0005	−0.14
	(0.1671)	(0.4809)	(0.0568)	
L_2	0.0000	0.0057	0.0080	−0.01
	(0.2262)	(0.8112)	(0.9700)	
B_1	0.0000	0.0081	0.0038	−0.08
	(0.4161)	(0.8771)	(0.3940)	
B_2	0.0000	−0.0082	−0.0140	0.06
	(0.3183)	(−0.7383)	(−1.5655)	
M_1	0.0000	0.0047	−0.0001	−0.11
	(0.1493)	(0.7144)	(−0.0093)	
12th half-hour return				
L_1	0.0001	−0.0093	0.0050	0.06
	(1.1835)	(−1.3883)	(0.5527)	
L_2	0.0000	−0.0027	0.0036	−0.13
	(1.0038)	(−0.3930)	(0.4338)	
B_1	0.0000	−0.0096	−0.0010	−0.08
	(0.8201)	(−0.8781)	(−0.1119)	
B_2	0.0001	−0.0117	0.0031	−0.04
	(1.2040)	(−0.9928)	(0.2922)	
M_1	0.0001	−0.0055	0.0061	−0.05
	(1.0658)	(−0.7922)	(0.7040)	
Last half-hour return				
L_1	−0.0001	0.0274***	−0.0181	1.35
	(−0.9945)	(4.1448)	(−1.5949)	
L_2	−0.0000	0.0227**	−0.0086	0.71
	(−0.2838)	(3.1837)	(−0.8755)	
B_1	−0.0000	0.0075	−0.0097	−0.07
	(−0.2310)	(0.6176)	(−0.9079)	
B_2	−0.0000	0.0071	−0.0099	−0.08
	(−0.6261)	(0.6144)	(−0.7517)	
M_1	−0.0001	0.0273***	−0.0194	1.33
	(−0.9649)	(3.9754)	(−1.7576)	

Notes: This table reports the results of the equation $r_{i,t} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 \Delta s_{i,t-1} + \epsilon_t$ for i={11,12,13}. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016 (1,258 observations).

[11] As we do not find significant results for i={2,... ,10}, we do not present those results for readability.

Table 8
Predictive regressions - Investor sentiment and lagged market return.

	α	β_1	β_2	β_3	β_4	Adj-R^2 (%)
Last half-hour return						
L_1	−0.0001	0.0274***	0.0111	0.1086	0.0508	2.13
	(−1.1662)	(3.4025)	(0.5610)	(1.2903)	(1.1349)	
L_2	−0.0000	0.0216**	0.0142	0.1047	0.0523	1.68
	(−0.4378)	(2.6833)	(0.7337)	(1.2400)	(1.1456)	
B_1	−0.0000	0.0052	0.0248	0.1051	0.0392	1.10
	(−0.5873)	(0.4468)	(1.4088)	(1.2392)	(0.8589)	
B_2	−0.0000	0.0074	0.0251	0.1054	0.0391	1.12
	(−0.7841)	(0.6651)	(1.4145)	(1.2448)	(0.8590)	
M_1	−0.0001	0.0269**	0.0108	0.1062	0.0518	2.04
	(−1.0671)	(3.2612)	(0.5456)	(1.2626)	(1.1533)	
GHLZ [2012–2016]	−0.0000		0.0255	0.1039		1.10
	(−0.7003)		(1.4390)	(1.2263)		
GHLZ [1998–2016]	−0.0000		0.0673***	0.1246**		2.91
	(−0.7903)		(4.3443)	(2.7420)		

Notes: This table reports the results of the equation $r_{13,t} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 r_{1,t} + \beta_3 r_{12,t} + \beta_4 r_{13,t-1} + \epsilon_t$. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016 (1258 observations).

Table 9
Predictive regressions - News and no-news trading days.

	α	β_1	β_2	β_3	β_4	Adj-R^2 (%)	Obs.
NFP							
Release	0.0000	−0.0386	−0.0057	0.1353	0.2164	0.53	58
	(0.0185)	(−1.1573)	(−0.1732)	(0.6669)	(1.3349)		
No Release	−0.0001	0.0310***	0.0115	0.1074	0.0481	2.39	1200
	(−1.4401)	(3.6609)	(0.5373)	(1.2339)	(1.0551)		
MSCI							
Release	0.0001	0.0046	0.0426	−0.0840	0.2955**	8.88	116
	(0.7152)	(0.1700)	(1.6016)	(−0.5957)	(3.1112)		
No Release	−0.0001	0.0282***	0.0087	0.1173	0.0229	2.13	1142
	(−1.3211)	(3.3813)	(0.4071)	(1.3396)	(0.4919)		
FOMC Meetings							
Release	−0.0001	0.0193	0.0823*	0.0168	−0.1118	4.50	120
	(−0.6180)	(1.0068)	(2.3597)	(0.1069)	(−1.1740)		
No Release	−0.0001	0.0302***	0.0028	0.1162	0.0702	2.33	1138
	(−1.1176)	(3.4959)	(0.1286)	(1.2819)	(1.4009)		
NFP or MSCI or FOMC							
Release	0.0001	0.0127	0.0234	0.0019	0.1092	0.98	238
	(0.5122)	(0.8540)	(0.9985)	(0.0157)	(1.4222)		
No Release	−0.0001	0.0334***	0.0028	0.1260	0.0355	2.53	993
	(−1.5408)	(3.5410)	(0.1107)	(1.2988)	(0.6672)		

Notes: This table reports the results of the equation $r_{13,t} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 r_{1,t} + \beta_3 r_{12,t} + \beta_4 r_{13,t-1} + \epsilon_t$ for days with (release) or without (no release) macroeconomic news announcements. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016.

dictionaries are inappropriate for deriving the sentiment of short informal messages published on social media.

We then control for lagged market returns to assess if the predictability of stock index return using past change in investor sentiment is not caused by a contemporaneous correlation between sentiment and return (as documented, among others, by Kim and Kim, 2014). Based on the results in Table 7, we focus on $i = 13$ and on the first half-hour change in investor sentiment. More precisely, we consider the following model:

$$r_{13,t} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 r_{1,t} + \beta_3 r_{12,t} + \beta_4 r_{13,t-1} + \epsilon_t. \quad (3)$$

The inclusion of $r_{1,t}$ is motivated by GHLZ who find that the first half-hour return predicts the last half-hour return for a wide range of ETFs. The inclusion of $r_{13,t-1}$ is motivated by Heston et al. (2010) who identify return continuation at half-hour intervals that are exact multiples of a trading day. Table 8 presents the results.

Even after controlling for lagged market returns, the first half-hour change in investor sentiment remains the only significant predictor of the last half-hour market return. Sentiment-driven optimistic (pessimistic) traders create short-term upward (downward) price pressure at the end of the trading day. This finding provides evidence that the intraday sentiment effect is distinct from the intraday momentum effect. Interestingly, we also demonstrate that the intraday momentum effect documented by GHLZ do not hold during the most recent period. Although we find evidence of intraday momentum effect when we consider a longer time period from 1998 to 2017, with R^2 values and coefficients very similar to those reported by GHLZ on a time period from 1993 to 2013, we do not find significant intraday momentum effect when we focus on recent years (2012–2017). Academic research may have destroyed stock return predictability (McLean and Pontiff, 2016), or previous results may have been caused by data-snooping, market frictions or omitted variables. We leave this question for further research.

We also examine whether the intraday sentiment effect is driven by the release of macroeconomics news before the market opens or during the trading day. For this purpose, we re-run Eq. (3) by dividing all trading days into two groups: days with news releases and days without. We focus on three major macroeconomics announcements: Non-Farm Payroll (NFP, monthly at 8.30

Table 10
Predictive regression - Other ETFs.

Panel A: US ETF	α	β_1	β_2	β_3	β_4	Adj-R^2 (%)
SPY [S&P 500]	−0.0001	0.0274***	0.0111	0.1086	0.0508	2.13
	(−1.1662)	(3.4025)	(0.5610)	(1.2903)	(1.1349)	
DIA [Dow]	−0.0001*	0.0260***	−0.0005	0.1303	0.0441	1.97
	(−1.8996)	(3.3277)	(−0.0290)	(1.4043)	(0.9877)	
QQQ [NASDAQ]	−0.0001	0.0340***	−0.0090	0.0544	0.0289	1.26
	(−0.8698)	(3.6179)	(−0.4489)	(0.7179)	(0.6330)	
XLF [Finance]	−0.0000	0.0340***	0.0110	0.0939	0.0287	2.15
	(−0.7034)	(4.0151)	(0.8614)	(1.4558)	(0.7112)	
IYR [Real Estate]	0.0002**	0.0321***	0.0233*	−0.0091	0.0534	2.04
	(2.5444)	(4.1693)	(1.8391)	(−0.1106)	(1.5668)	
IWM [Small-Cap]	0.0001	0.0236***	0.0132	−0.0009	0.0294	0.76
	(1.3709)	(2.6280)	(1.0224)	(−0.0167)	(0.9111)	
Panel B: Non-US ETF	α	β_1	β_2	β_3	β_4	Adj-R^2 (%)
EEM [Emerging]	−0.0000	0.0215***	−0.0009	0.0808	0.0342	0.95
	(−0.5131)	(2.8544)	(−0.0922)	(1.2928)	(0.8164)	
FXI [China]	−0.0001	0.0223***	−0.0101	−0.0109	0.0636*	0.92
	(−1.0609)	(2.7922)	(−1.6133)	(−0.1602)	(1.7049)	
EFA [Non-US]	0.0000	0.0127**	−0.0016	0.0418	−0.0109	0.24
	(1.0330)	(2.1457)	(−0.2057)	(0.7786)	(−0.2509)	
VWO [Emerging]	−0.0001	0.0169**	−0.0035	0.0790	0.0447	0.75
	(−1.2608)	(2.2976)	(−0.3749)	(1.2339)	(1.0145)	
Panel C: Non-Equity ETF	α	β_1	β_2	β_3	β_4	Adj-R^2 (%)
TLT [Bond Market]	0.0001	0.0020	0.0238***	0.0092	−0.1601***	3.56
	(1.3886)	(0.3879)	(3.4643)	(0.2548)	(−4.9402)	

Notes: This table reports the results of the equation $r_{13,t,x} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 r_{1,t,x} + \beta_3 r_{12,t,x} + \beta_4 r_{13,t-1,x} + \epsilon_t$, where x={SPY, QQQ, XLF, IWM, DIA, EEM, FXI, EFA, VWO, IYR, TLT}. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016 (1258 observations).

Table 11
Predictive regression - Price reversal over the next trading day.

Period	α	β_1	β_2	β_3	Adj-R^2 (%)
1st (First) Half-Hour	0.0003	0.0125	−0.0566	−0.0798	0.13
	(1.7010)	(0.5416)	(−1.0475)	(−0.5147)	
2nd Half-Hour	0.0000	−0.0039	0.0366	0.0416	0.15
	(0.2412)	(−0.3034)	(1.4359)	(0.5363)	
3rd Half-Hour	0.0000	0.0002	−0.0143	−0.0210	−0.04
	(0.1393)	(0.0232)	(−1.1623)	(−0.5346)	
4th Half-Hour	0.0001	0.0071	−0.0013	0.0510	0.14
	(1.4095)	(1.1097)	(−0.1008)	(1.4513)	
5th Half-Hour	0.0000	0.0001	0.0116	0.0162	−0.06
	(0.6032)	(0.0183)	(0.9775)	(0.3548)	
6th Half-Hour	0.0000	−0.0005	−0.0005	−0.0304	−0.10
	(0.2331)	(−0.0883)	(−0.0572)	(−0.8702)	
7th Half-Hour	0.0001*	0.0023	0.0076	0.0181	−0.02
	(2.2981)	(0.4475)	(0.8243)	(0.5778)	
8th Half-Hour	−0.0000	−0.0132*	0.0310*	−0.0181	1.21
	(−0.2736)	(−2.1961)	(2.3822)	(−0.5229)	
9th Half-Hour	−0.0000	−0.0050	−0.0017	0.0045	−0.13
	(−0.1111)	(−0.9159)	(−0.1612)	(0.1302)	
10th Half-Hour	0.0001	−0.0019	−0.0019	−0.0317	−0.12
	(1.5648)	(−0.2860)	(−0.1623)	(−0.7760)	
11th Half-Hour	0.0000	−0.0163*	0.0175	−0.0032	0.39
	(0.5419)	(−2.5282)	(1.3305)	(−0.0641)	
12 Half-Hour	0.0001	−0.0163*	0.0285	0.0161	0.71
	(1.2226)	(−2.1070)	(1.5629)	(0.3186)	
13th (Last) Half-Hour	−0.0000	0.0074	−0.0009	−0.0718	0.24
	(−0.5232)	(0.8749)	(−0.0330)	(−1.3239)	

Notes: This table reports the results of the equation $r_{i,t+1} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 r_{1,t} + \beta_3 r_{13,t} + \epsilon_t$ for i={1,12,... ,13}. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016 (1258 observations).

a.m.), the Michigan Consumer Sentiment Index (MSCI, preliminary and final releases, monthly at 10:00 a.m.) and the Federal Open Market Committee meeting (FOMC, every six weeks at 2:00 p.m.). To account for FOMC pre-meeting or post-meeting announcement drift, we include one day before and one day after the meetings. Table 9 reports the results. For readability, we present the results only when field-specific lexicon L_1 is used to derive investor sentiment, but we find similar results for L_2 and M_1, and no significant results for B_1 and B_2, as previously.

We find that the intraday sentiment effect is concentrated on days without macroeconomic news announcements. The first half-hour shift in investor sentiment is not significant on NFP days,

Table 12
Distribution of users' self-reported investment approach, holding period and experience level.

	Users		Messages	
	Number	Percentage (%)	Number	Percentage (%)
Investment approach				
Technical	29,104	12.12	13,177,530	22.11
Fundamental	9541	3.97	3,936,066	6.60
Global Macro	2425	1.01	872,404	1.46
Momentum	13,533	5.64	6,003,008	10.07
Growth	13,111	5.46	4,590,279	7.70
Value	7295	3.04	3,346,318	5.61
Holding period				
Day trader	16,462	6.86	6,046,038	10.14
Swing trader	29,956	12.48	13,223,008	22.18
Position trader	15,514	6.46	6,003,489	10.07
Long-term investor	15,026	6.26	6,344,566	10.64
Experience level				
Novice	25,686	10.70	5,260,787	8.83
Intermediate	36,082	15.03	14,499,167	24.32
Professional	14,619	6.09	11,779,219	19.76

Notes: This table reports the distribution of users' self-reported investment approach, holding period and experience level. Percentage is calculated as the number of users (or messages) who self-reported a given trading strategy in their profile divided by the total number of users (or messages) in the sample.

Table 13
Predictive regression - Investor sentiment by investment approach, holding period and experience level.

Panel A: Investment approach	[1]	[2]	[3]	[4]	[5]
$r_{1,\,t}$	0.0156	0.0248	0.0226	0.0210	0.0239
	(0.7946)	(1.3942)	(1.2225)	(1.1514)	(1.3368)
$r_{12,\,t}$	0.1065	0.1039	0.1051	0.1030	0.1032
	(1.2613)	(1.2259)	(1.2462)	(1.2275)	(1.2317)
$\Delta s_{1,\,t,\,technical}$	0.0217*				
	(2.5564)				
$\Delta s_{1,\,t,\,fundamental}$		0.0037			
		(0.4132)			
$\Delta s_{1,\,t,\,momentum}$			0.0163		
			(1.3456)		
$\Delta s_{1,\,t,\,growth}$				0.0212*	
				(2.1436)	
$\Delta s_{1,\,t,\,value}$					0.0210*
					(2.1051)
Adj-R^2(%)	1.65	1.03	1.19	1.38	1.44
Panel B: Holding period	[1]	[2]	[3]	[4]	
$r_{1,\,t}$	0.0233	0.0195	0.0208	0.0240	
	(1.2949)	(1.0120)	(1.1219)	(1.3328)	
$r_{12,\,t}$	0.1034	0.1055	0.1012	0.1037	
	(1.2256)	(1.2486)	(1.2031)	(1.2277)	
$\Delta s_{1,\,t,\,day}$	0.0154				
	(1.2547)				
$\Delta s_{1,\,t,\,swing}$		0.0178			
		(1.7557)			
$\Delta s_{1,\,t,\,position}$			0.0206*		
			(2.0494)		
$\Delta s_{1,\,t,\,long}$				0.0097	
				(1.1156)	
Adj-R^2 (%)	1.17	1.31	1.36	1.10	
Panel C: Experience level	[1]	[2]	[3]		
$r_{1,\,t}$	0.0194	0.0186	0.0194		
	(1.0796)	(0.9882)	(0.9950)		
$r_{12,\,t}$	0.1054	0.1051	0.1050		
	(1.2551)	(1.2504)	(1.2410)		
$\Delta s_{1,\,t,\,novice}$	0.0306**				
	(3.2360)				
$\Delta s_{1,\,t,\,intermediate}$		0.0243*			
		(2.2976)			
$\Delta s_{1,\,t,\,professional}$			0.0154		
			(1.7427)		
Adj-R^2 (%)	1.77	1.51	1.33		

Notes: This table reports the results of the equation $r_{13,t} = \alpha + \beta_1 \Delta s_{1,t,x} + \beta_2 r_{1,t} + \beta_3 r_{12,t} + \epsilon_t$. As the constant α is not significant in any regression, we do not report results for α for readability. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016 (1258 observations).

Table 14
Trading strategy performance.

Strategy	Mean (%)	Std Dev (%)	Sharpe ratio
Sentiment-driven strategy	4.55	3.042	1.496***
Always long strategy	−0.632	3.055	−0.207
First half-hour strategy	1.66	3.054	0.544
12th half-hour strategy	0.566	3.055	0.185

Notes: This table reports the annualized mean returns, standard deviations and Sharpe ratios of trading strategies relying on different signals to buy (sell) S&P 500 ETF index at 3:30 p.m. on day t and sell (buy) it at market close on the same trading day. Superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively, using a simulation-based p-value for the Sharpe ratio significance level.

MSCI days, and [−1:+1] days around FOMC meetings. Investor sentiment, thus, is not a mere reflection of macroeconomics news announcements. This result is consistent with the fact that on days with macroeconomic news announcements, the last half-hour return is mainly driven by the news announcements in such a way that sentiment-driven traders do not affect prices. However, on days with no news, investor sentiment affects stock prices.

As in GHLZ and SNS, we then analyze whether the sentiment effect is significant for other domestic ETFs, sector indices, international ETFs and bond ETFs. Table 10 reports the results. As above, we report only the results when we use L_1 to measure investor sentiment, but the results are similar for L_2 and M_1. We confirm that the first half-hour change in investor sentiment predicts the last half-hour return for a diverse set of ETFs. We also find that the associated R^2 decreases for international equity indices and small capitalization ETFs (Russell 2000) and is not significant for bond market ETFs. This result is consistent with the fact that users on StockTwits mainly discuss the development of the U.S. stock market indices and the cross-section of large and medium capitalization stock returns. These complementary results provide evidence that analyzing data from StockTwits allows researchers to construct a value-relevant intraday measure of U.S. investor sentiment.

Last but not least, we investigate whether the predictability identified previously is driven by fundamental end of day demand or by noise trading. To do so, we consider the following model:

$$r_{i,t+1} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 r_{1,t} + \beta_3 r_{13,t} + \epsilon_t. \quad (4)$$

If the predictability is driven by noise trading, we should identify a price reversal over the next trading day (i.e, a negative coefficient on β_1). This situation would be consistent with the presence of uninformed sentiment-driven traders causing a price run up on day t (as shown previously) followed by a price reversal afterwards on day $t+1$ when arbitrageurs step in to correct the anomaly. Table 11 shows the regression results for i={1,2,... ,13}.[12] We identify a significant price reversal on day $t+1$ during the 8th, the 11th and the 12th half-hour of the trading days, favoring the noise trading hypothesis over the fundamental end of day demand hypothesis. This result is consistent with the evidence of sentiment-driven short-term price pressures followed by price reversals documented in the literature (Tetlock, 2007; Garcia, 2013). However, to the best of our knowledge, we provide the first evidence of sentiment-driven price pressure followed by a price reversal at the intraday level, consistent with limits to arbitrage during the last half-hour of the trading day.[13]

[12] For readability, we only present our results when L_1 is used to compute investor sentiment.

[13] Sun et al. (2016) find that "there appears to be some evidence of reversal at longer horizons", but the coefficient estimates for β_1 are not significant in most of their regressions, with the exception of the 11th half-hour.

4.3. Exploring investor base heterogeneity

Contrary to the Thomson Reuters MarketPsych Index (TRMI) used by SNS as a proxy for intraday investor sentiment (a "black box" aggregate indicator), focusing on data from StockTwits allows researchers to test directly whether the predictability is driven (or not) by noise trader sentiment. StockTwits provides unique information about users' self-reported investment approach (technical, fundamental, global macro, momentum, growth, or value), holding period (day trader, swing trader, position trader, or long-term investor), and experience level (novice, intermediate, or professional). For example, using data from StockTwits and exploiting investor base heterogeneity, Cookson and Niessner (2016) find that investor disagreement robustly forecasts abnormal trading volume at a daily frequency. In a similar fashion, we assess in this subsection whether a specific type of trader or a specific trading strategy drives the sentiment effect identified previously. Although reporting the investment approach, the holding period and the experience level is not required to register to StockTwits, we still observe a self-reported trading strategy for a large number of users (84,891 users) and messages (35,436,607 messages). Table 12 presents the distribution of users by the investment approach, holding period and experience level.

As in the previous subsection, we construct intraday investor sentiment indicators at half-hour time intervals. However, instead of considering all messages, we create intraday investor sentiment indicators for each investment approach, each holding period and each experience level by considering only the messages of users who self-reported the given information in their profile. We find qualitatively similar results when we use L_1, L_2 or M_1 but no significant results when we use B_1 and B_2, confirming previous findings. For readability, we present the results only when field-specific lexicon L_1 is used to quantify individual message sentiment. As only 1.01% of users self-declared themselves as following a "Global Macro" trading approach, we remove this strategy as in Cookson and Niessner (2016). The correlation coefficient between the 12 investor sentiment indicators at half-hour time intervals range from 0.0780 (between "fundamental traders" and technical traders") to 0.6216 (between "technical traders" and "swing traders"). See Appendix C for details. We denote with $\Delta s_{1,t,x}$ the first half-hour change in investor sentiment on day t for users' self-reported characteristic x. Then, we estimate the following predictive regression:

$$r_{13,t} = \alpha + \beta_1 \Delta s_{1,t,x} + \beta_2 r_{1,t} + \beta_3 r_{12,t} + \epsilon_t. \quad (5)$$

where $r_{13,t}$ is the last half-hour return, $r_{1,t}$ is the first half-hour return, $r_{12,t}$ the 12th half-hour return and $\Delta s_{1,t,x}$ represents the change in sentiment the first half-hour of day t for each investor type $x = \{x_1, x_2, x_3\}$. We consider each investor depending on his or her trading approach (x_1 = {*technical, fundamental, momentum, growth, value*}), his or her holding period (x_2 = {*day, swing, position, long-term*}) and his or her experience (x_3 = {*novice, intermediate, professional*}). Table 13 presents the results by investment approach, holding period and experience level.

Analyzing each investment approach separately, and controlling for lagged market returns, we find significant results for traders with technical, growth and value investing strategies and for position traders (i.e., holding periods from a few days to a few weeks). We also find that the significance of the results decreases with traders' self-reported experience. The first half-hour change in novice investor sentiment is significant at the 1% level (Adj-R^2 equal to 1.77%) whereas the first half-hour change in intermediate investor sentiment is significant only at the 5% level (Adj-R^2 equal to 1.51%), and the first half-hour change in professional investor sentiment is not significant (Adj-R^2 equal to 1.33%). We also consider all possible approach and experience, approach and pe-

riod, and period and experience doublets (60 combinations). We find that the last half-hour return is robustly forecasted by the first half-hour change in novice investor sentiment. Looking at 10 doublets with the highest Adj-R^2, we find that all the best combinations, except one, include the change in novice investor sentiment (with R^2 ranging from 1.69 to 2.05). The only other characteristic that adds value when combined with the "novice experience" is the trading approach "technical analysis" (significant at the 10% level).

Last, we simulate a trading strategy buying (selling) the S&P 500 ETF at 3.30 p.m. on days with an increase in novice investor sentiment during the first half-hour of that day, and selling (buying) at 4:00 p.m. We present the results when the performance of the trading strategies is evaluated using the Sharpe ratio, but the results are robust to the performance evaluation metrics as all trading strategies exhibit very similar volatility. We compare the performance of a "sentiment-driven" strategy with an *Always Long Strategy* buying the ETF at the beginning of the last half-hour and selling it at market close. We also consider a *First Half-Hour Return Strategy* buying (selling) the ETF on days with a positive (negative) first half-hour return and selling (buying) it at market close, and a *12th Half-Hour Return Strategy* buying (selling) the ETF on days with a positive (negative) 12th half-hour return and selling (buying) it at market close. As in Roger (2014), we compare the Sharpe ratio of each strategy to the simulated Sharpe ratio distribution by generating 10,000 strategies randomly buying (selling) the S&P 500 ETF. Table 14 reports the results.

We find that the average annualized return of a strategy using half-hour change in novice investor sentiment as a trading signal is equal to 4.55%, with a Sharpe ratio of 1.496. Although the annualized return might not seem impressive at first sight, the return is remarkable as we hold a position only during 30 min per day and we do not keep any position overnight. Comparing the location of the sentiment-driven strategy Sharpe ratio in the simulated Sharpe ratio distribution, we find that only 9 random strategies out of the 10,000 simulated ones have a Sharpe ratio greater than 1.496. Thus, the observed profitability is significant at the 0.1% level. We also demonstrate that a sentiment-driven strategy significantly outperforms other benchmark strategies. Overall, the results provide empirical evidence of sentiment-driven noise trading at the intraday level.

4.4. Discussion of empirical results

According to GHLZ, there are two explanations for why the first half-hour return predicts the last half-hour return. First, strategic informed traders might time their trade for periods of high trading volume. On days with positive overnight night news, informed traders are likely to trade very actively at the market opening before reinforcing their position during the last half-hour. Second, on days with a sharp overnight and first half-hour increase in the stock market index, some traders might expect a price reversal over the following hours and short the market. As typical day traders are flat at the end of the day, they are likely to unwind their position during the last half-hour return which, in turn, will push prices up. Closer to our paper, SNS provide two reasons to explain why investor sentiment has predictive value for intraday market returns and why the sentiment effect is concentrated on the end of the trading day. First, due to risk aversion, investors trading the S&P 500 index ETF might prefer to wait a few hours before taking a position on the market. Second, risk-averse arbitrageurs may be more likely to trade against sentiment traders at the beginning of the day than later in the day due to the uncertainty introduced by overnight news.

Our findings provide direct empirical evidence for the two hypotheses proposed by SNS. First, we find that when investors are more optimistic during the first 30 min on day t than during the last 30 min of day $t-1$, the S&P 500 index ETF significantly increase during the last half-hour of the trading day. However, all other variations in investor sentiment ($\Delta s_{i,t}$ for i={2,... 12}) are not significant in predictive regressions. This finding illustrates the "timing effect" as investors seem to prefer to wait until "the dust is about to settle" before buying or selling the S&P 500 index ETF based on their initial sentiment. This finding is also consistent with the explanation based on the presence of late-informed investors provided by GHLZ.

Furthermore, analyzing users' self-reported experience, we find that the last half-hour predictability is driven by the shift in the sentiment of novice traders, and, to a lesser extent, by the shift in the sentiment of traders following technical analysis strategies. This finding is consistent with Hoffmann and Shefrin (2014) who find, using private data from a sample of discount brokerage clients, that individual investors who use technical analysis are disproportionately likely to speculate in the short-term stock market. Examining the impact of aggregate investor sentiment on trading volume and long-run price reversal, SNS document that the investor sentiment effect is driven by noise trading. In this paper, using self-reported experience level instead of making indirect inferences by analyzing market reactions, we provide, to the best of our knowledge, the first direct empirical evidence of intraday sentiment-driven noise trading.

5. Conclusion

Improving the transparency and replicability of results are of utmost importance for the big-data and finance environment. Although developing public field-specific lexicons will obviously not solve all issues related to replicability and comparability, it still constitutes an important step to facilitate further research in this area, as stated by Nardo et al. (2016) in a recent survey of the literature of financial market prediction using the Web. In the first part of this paper, we construct a lexicon of words used by online investors when they share opinions and ideas about the bullishness or bearishness of the stock market by using an extensive dataset of messages for which sentiment is explicitly revealed by investors. We demonstrate that a transparent and replicable approach significantly outperforms the benchmark dictionaries used in the literature while remaining competitive with more complex machine learning algorithms. The findings provide empirical evidence to Kearney and Liu (2014) conclusion about the need to develop a more authoritative field-specific lexicon and of Loughran and McDonald (2016) recommendations that alternative complex methods (machine learning) should be considered only when they add substantive value beyond simpler and more transparent approaches (bag-of words).

In the second part, we explore the relation between online investor sentiment and intraday S&P 500 index ETF returns. We find that the first half-hour change in investor sentiment predicts the last half-hour return, even after controlling for lagged market returns (first half-hour return and lagged half-hour return). This finding holds for a wide range of ETFs and is robust to macroeconomic news announcements. We also demonstrate that the short-term sentiment-driven price pressure is followed by a price reversal on the next trading day, consistent with the noise trading hypothesis. Then, analyzing users' self-reported investment approach, holding period and experience level, we find that the sentiment effect is mainly driven by the shift in the sentiment of novice traders. We confirm this result by showing that a strategy that uses changes in novice investors' sentiment as trading signals significantly outperforms other baseline strategies (risk-adjusted performance). Overall, the results provide direct empirical evidence of intraday sentiment-driven noise trading.

Table A.15
Classification accuracy - TD-IDF and relative word count weighting scheme.

Classifier	$CC(\%)$	$CC_{bull}(\%)$	$CC_{bear}(\%)$	$CM(\%)$	$CM(\%)$	$CM_{bear}(\%)$
L_1 (TF-IDF)	74.53	73.82	75.23	89.96	89.31	90.61
L_1 (Word Count)	74.62	73.98	75.24	90.03	89.32	90.73

Notes: This tables shows the out-of-sample classification accuracy when terms' weight are computed using a relative word count weighting scheme or a TF-IDF weighting scheme. We also present results from a simple relative word count weighting scheme (as used in the paper). We report the percentage of correct classification excluding unclassified messages *CC*, the percentage of correct classification per class (respectively CC_{bull} and CC_{bear}), the percentage of classified messages *CM* (message with a sentiment score different from zero) and the percentage of classified messages per class (CM_{bull} and CM_{bear}).

Although we focused on the predictability of aggregate market returns, we believe that the evolution of intraday investor sentiment over time and across users with different trading approaches, experiences and investment horizons can also be useful in many other situations, such as explaining the cross-section of average stock returns or forecasting stock market volatility. We encourage further research in this area by making public the field-specific weighted lexicon we developed for this paper.

Appendix A. Weighting scheme

The standard TF-IDF weighting scheme, often used in information retrieval and text mining, can be computed as:

$$tf\text{-}idf(t,d) = \frac{n_{d,t}}{n_{d,T}} * \log \frac{N_d}{N_{d,t}} \tag{A.1}$$

where t is a term (unigram or bigram), d is a collection of documents, $n_{d,\,t}$ is the number of occurrences of term t in documents d, $n_{d,\,T}$ is the total number of terms in documents d, N_d is the total number of documents d, $N_{d,\,t}$ is the total number of documents d containing term t. Then, the sentiment weight for each term t can be computed as in Oliveira et al. (2016) as:

$$SW_{tf\text{-}idf}(t) = \frac{tf\text{-}idf(t,d_{pos}) - tf\text{-}idf(t,d_{neg})}{tf\text{-}idf(t,d_{pos}) + tf\text{-}idf(t,d_{neg})}, \tag{A.2}$$

where d_{pos} is a collection of positive documents, and d_{neg} is a collection of negative documents. In the paper, we choose to adopt a very simple relative word count (wc) term-weighting, defined as:

$$SW_{wc}(t) = \frac{n_{d_{pos},t} - n_{d_{neg},t}}{n_{d_{pos},t} + n_{d_{neg},t}} \tag{A.3}$$

Given the maximum length of the messages published on social media (140 characters), $N_{d,\,t} \approx n_{d,\,T}$ (as a given word very rarely appears twice in the same tweet). Furthermore, in our empirical analysis, the number of bullish (positive) documents in the training dataset is equal to the number of bearish (negative) documents (375,000) ($n_{d_{pos},T} \approx n_{d_{neg},T}$ and $N_{d_{pos}} \approx N_{d_{neg}}$). From previous equations, it thus can be easily seen that $SW_{tf\text{-}idf}(t) \approx SW_{wc}(t)$.

Analyzing all n-grams that appear at least 75 times in our training dataset, we find an absolute difference between $SW_{tf\text{-}idf}(t)$ and $SW_{wc}(t)$ equal to 0.024. Comparing out-of-sample classification accuracy, we find qualitatively similar results when a TF-IDF scheme is used to compute the terms' weight and to identify relevant features (n-grams). Table A.15 presents the out-of-sample classification accuracy of a subset of 250,000 messages. Furthermore, the results for the predictability of intraday returns are qualitatively similar when investor sentiment is derived using a relative word-count weighting scheme or a TF-IDF scheme. Table A.16 presents the results. Overall, we find that the results are robust to the method used for term-weighting. As the term-weighing scheme lacks theoretical motivation (Loughran and McDonald, 2016), we favor the simplest approach due to the standardized (and short) size of the messages posted on social media. Recently, Smailović et al. (2014) confirmed that the TF approach is statistically significantly better than the TD-IDF-based approach to data from Twitter.

Appendix B. Message classification

We compute a sentiment score between -1 and $+1$ for all messages published on StockTwits ($SS(m)$) by adopting dictionary-based approaches and a machine learning method.

Dictionary-based approaches

For dictionary-based approach L_1, we use a methodology similar to Oliveira et al. (2016). Message sentiment is equal to the average $SW(t)$ of the terms present in the message and included in lexicon L_1. When a bigram is present in the text, we do not take into account the score of the individual unigram included in the bigram to avoid double counting. For example, considering the following message:

Table A.16
Predictive regressions - Investor sentiment and half-hour market return.

	α	β_1	β_2	$AdjR^2$ (%)
L_1 (TF-IDF)	−0.0001 (−1.3099)	0.0316*** (3.9785)	−0.0083 (−0.6618)	1.36
L_1 (Word Count)	−0.0001 (−1.4169)	0.0312*** (4.1339)	−0.0087 (−0.6879)	1.44

Notes: This table reports the results of the equation $r_{13,t} = \alpha + \beta_1 \Delta s_{1,t} + \beta_2 \Delta s_{12,t} + \epsilon_t$ when the change in investor sentiment is computed using a relative word count weighting scheme or a TF-IDF weighting scheme. Robust t-statistics are reported in parenthesis and superscripts ***, **, and * indicate statistical significance at the 0.1%, 1% and 5% level, respectively. The sample period is from January 1, 2012 to December 31, 2016 (1258 observations).

TimCGriffith
Tim Griffith

Nov. 5 2015 at 6:38 PM

$SPY Want to see a bloodbath, take a look at the short attack on **$STRP**! A scam company like **$VRX** called on their BS! Bearish

http://stocktwits.com/message/45003236

Using the field-specific lexicon L_1, we find that the following terms are present in the message above (within the brackets the *SW* computed as in Eq. 1):

- cashtag ! [$SW = 0.3069$]
- cashtag called [$SW = -0.3033$]
- bloodbath [$SW = -0.6600$]
- short [$SW = -0.5811$]
- scam [$SW = -0.8493$]

Taking the average $SW(t)$, we find a sentiment score equals −0.4069. In this example, the classification is correct as the message was classified as "Bearish" by the user who sent the tweet, and we obtain a sentiment score lower than 0. We use a similar methodology to compute $SS(m)$ for the other dictionary-based approaches L_2, B_1 and B_2, except that we consider an equal-weighting scheme by giving all words in the positive lists a weight of +1 and all words in the negative lists a weight of +1. Using the previous example, we identify the following terms:

- L_2 : bloodbath [−1], short [−1], scam [−1]
- B_1 : None of the words are present in the LM dictionary
- B_2 : short [−1], attack [−1], company [+1], like [+1]

We end up with a sentiment score for the message equal to −1 for L_2, 0 for B_1 (no term identified) and 0 for B_2 (two positive terms and two negative terms).

Machine learning methods

We experiment three machine algorithms as in Pang et al. (2002) and Go et al. (2009): naive Bayes (NB), maximum entropy (MaxEnt) and support vector machines (SVM). We report results only for MaxEnt, as we find that MaxEnt provides better results than NB (we conjecture due to the overlapping in NB) and similar (but with a lower computational complexity) than SVM. For MaxEnt, the probability that document d belongs to class c given a weight vector δ is equal to:

$$P(c|d,\delta) = \frac{exp[\sum_i \delta i f_i(c,d)]}{\sum_c exp[\sum_i \delta i f_i(c,d)]} \quad \text{(B.1)}$$

where $f_i = \{f_1, f_2, ..., f_m\}$ is a predefined set of m features (unigram or bigram) that can appear in a document. The weight vector is found by numerical optimization of the lambdas to maximize the conditional probability. We use the "liblinear" package for this purpose. Considering the previous message ($SPY Want to see a bloodbath, take a look at the short attack on $SRTP! A scam company like $VRX called on their BS!), we find using MaxEnt: $P(c_{pos}) = 0.12$ and $P(c_{neg}) = 0.88$. To obtain an $SS(m)$ between −1 and +1, we define:

$$SS(m)_{MaxEnt} = (P(c_{pos}|m,\delta) - 0.5) * 2. \quad \text{(B.2)}$$

In the previous example, we find $SS_{MaxEnt} = -0.76$. We then consider all messages with an $SS_{MaxEnt} < 0$ (equivalent to a $P(c_{pos}) < 0.5$) as negative, and all messages with an $SS_{MaxEnt} > 0$ as positive. When a message does not contain any features included in $\{f_1, f_2, ..., f_m\}$, then $SS_{MaxEnt} = 0$, and we consider the message as unclassified.

Appendix C. Trading strategy correlation

Intraday investor sentiment - Self-reported trading strategy correlation

	Technical	Fundamental	Momentum	Growth	Value	Day	Swing	Position	Long-Term	Novice	Intermediate	Professional
Technical	1.000											
Fundamental	0.1037	1.000										
Momentum	0.1664	0.0844	1.000									
Growth	0.1154	0.1202	0.1170	1.000								
Value	0.1126	0.0780	0.0792	0.0984	1.000							
Day	0.4816	0.1103	0.2429	0.0950	0.0889	1.000						
Swing	0.6216	0.1978	0.3520	0.2193	0.1464	0.1806	1.000					
Position	0.3146	0.2421	0.2412	0.2295	0.2240	0.1224	0.1880	1.000				
Long	0.1659	0.3569	0.1374	0.3829	0.4118	0.0878	0.1597	0.1585	1.000			
Novice	0.2309	0.1867	0.2425	0.3285	0.1534	0.1753	0.3131	0.2035	0.3535	1.000		
Intermediate	0.4778	0.2716	0.3401	0.2846	0.1905	0.3161	0.4873	0.4588	0.2837	0.1773	1.000	
Professional	0.4778	0.2411	0.2261	0.1687	0.3019	0.3804	0.4224	0.3631	0.2986	0.1386	0.2050	1.000

Notes: This tables shows the correlation matrix of intraday investor sentiment indicators for each investment approach, each holding period and each experience level. Results are presented when investor sentiment indicators are computed from individual message quantification using L_1.

References

Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? the information content of internet stock message boards. J. Finance 59 (3), 1259–1294.

Avery, C.N., Chevalier, J.A., Zeckhauser, R.J., 2016. The "CAPS" prediction system and stock market returns. Rev. Finance 20 (4), 1363–1381.

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. J. Finance 61 (4), 1645–1680.

Baker, M., Wurgler, J., 2007. Investor sentiment in the stock market. J. Econ. Perspect. 21 (2), 129–152.

Bogousslavsky, V., 2016. Infrequent rebalancing, return autocorrelation, and seasonality. J. Finance 71 (6), 2967–3006.

Brown, G.W., Cliff, M.T., 2005. Investor sentiment and asset valuation. J. Bus. 78 (2), 405–440.

Chen, H., De, P., Hu, Y.J., Hwang, B.-H., 2014. Wisdom of crowds: the value of stock opinions transmitted through social media. Rev. Financ. Stud. 27 (5), 1367–1403.

Cookson, J. A., Niessner, M., 2016. Why don't we agree? evidence from a social network of investors. Working Paper, Colorado University.

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

Das, S.R., 2014. Text and context: language analytics in finance. Found. Trends Finance 8 (3), 145–261.

Das, S.R., Chen, M.Y., 2007. Yahoo! for amazon: sentiment extraction from small talk on the web. Manag. Sci. 53 (9), 1375–1388.

De Long, J.B., Shleifer, A., Summers, L.H., Waldmann, R.J., 1990. Noise trader risk in financial markets. J. Polit. Econ. 98 (4), 703–738.

Dougal, C., Engelberg, J., Garcia, D., Parsons, C.A., 2012. Journalists and the stock market. Rev. Financ. Stud. 25 (3), 639–679.

Engelberg, J.E., Reed, A.V., Ringgenberg, M.C., 2012. How are shorts informed? short sellers, news, and information processing. J. Financ. Econ. 105 (2), 260–278.

Gao, L., Han, Y., Li, S. Z., Zhou, G., 2017. Market intraday momentum. Working Paper, Washington University in St. Louis.

Garcia, D., 2013. Sentiment during recessions. J. Finance 68 (3), 1267–1300.

Go, A., Bhayani, R., Huang, L., 2009. Twitter sentiment classification using distant supervision. Working paper. Stanford University.

Grossman, S.J., Stiglitz, J.E., 1980. On the impossibility of informationally efficient markets. Am. Econ. Rev. 70 (3), 393–408.

Heston, S.L., Korajczyk, R.A., Sadka, R., 2010. Intraday patterns in the cross-section of stock returns. J. Finance 65 (4), 1369–1407.

Hoffmann, A.O., Shefrin, H., 2014. Technical analysis and individual investors. J. Econ. Behav. Organ. 107, 487–511.

Jegadeesh, N., Wu, D., 2013. Word power: a new approach for content analysis. J. Financ. Econ. 110 (3), 712–729.

Jensen, M.C., 1978. Some anomalous evidence regarding market efficiency. J. Financ. Econ. 6 (2), 95–101.

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

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

Leung, H., Ton, T., 2015. The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks. J. Bank. Finance 55, 37–55.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. J. Finance 66 (1), 35–65.

Loughran, T., McDonald, B., 2016. Textual analysis in accounting and finance: a survey. J. Account. Res. 54 (4), 1187–1230.

McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? J. Finance 71 (1), 5–32.

Moat, H.S., Curme, C., Avakian, A., Kenett, D.Y., Stanley, H.E., Preis, T., 2013. Quantifying wikipedia usage patterns before stock market moves. Sci. Rep. 3.

Nardo, M., Petracco, M., Naltsidis, M., 2016. Walking down wall street with a tablet: a survey of stock market predictions using the web. J. Econ. Surv. 30 (2), 356–369.

Oliveira, N., Cortez, P., Areal, N., 2016. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. Decis. Support Syst. 85, 62–73. http://www.sciencedirect.com/science/article/pii/S0167923616300240.

Pang, B., Lee, L., Vaithyanathan, S., 2002. Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Vol. 10. Association for Computational Linguistics, pp. 79–86.

Pontiff, J., 1996. Costly arbitrage: evidence from closed-end funds. Q. J. Econ. 111 (4), 1135–1151.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., Mozetič, I., 2015. The effects of twitter sentiment on stock price returns. PLoS One 10 (9).

Roger, P., 2014. The 99% market sentiment index. Finance 35 (3), 53–96.

Sabherwal, S., Sarkar, S.K., Zhang, Y., 2011. Do internet stock message boards influence trading? evidence from heavily discussed stocks with no fundamental news. J. Bus. Finance Account. 38 (9–10), 1209–1237.

Shleifer, A., Vishny, R.W., 1997. The limits of arbitrage. J. Finance 52 (1), 35–55.

Smailović, J., Grčar, M., Lavrač, N., Žnidaršič, M., 2014. Stream-based active learning for sentiment analysis in the financial domain. Inf. Sci. (Ny) 285, 181–203.

Sprenger, T.O., Sandner, P.G., Tumasjan, A., Welpe, I.M., 2014. News or noise? using twitter to identify and understand company-specific news flow. J. Bus. Finance Account. 41 (7–8), 791–830.

Sprenger, T.O., Tumasjan, A., Sandner, P.G., Welpe, I.M., 2014. Tweets and trades: the information content of stock microblogs. Eur. Financ. Manag. 20 (5), 926–957.

Sun, L., Najand, M., Shen, J., 2016. Stock return predictability and investor sentiment: a high-frequency perspective. J. Bank. Finance 73, 147–164. http://www.sciencedirect.com/science/article/pii/S0378426616301595.

Tetlock, P.C., 2007. Giving content to investor sentiment: the role of media in the stock market. J. Finance 62 (3), 1139–1168.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: quantifying language to measure firms' fundamentals. J. Finance 63 (3), 1437–1467.