Digital Finance
https://doi.org/10.1007/s42521-023-00102-z

ORIGINAL ARTICLE

StockTwits classified sentiment and stock returns

Marc-Aurèle Divernois[1] · **Damir Filipović**[1]

Received: 22 February 2023 / Accepted: 23 November 2023

Abstract
We classify the sentiment of a large sample of StockTwits messages as bullish, bearish or neutral, and create a stock-aggregate daily sentiment polarity measure. Polarity is positively associated with contemporaneous stock returns. On average, polarity is not able to predict next-day stock returns. But when we condition on specific events, defined as sudden peaks of message volume, polarity has predictive power on abnormal returns. Polarity-sorted portfolios illustrate the economic relevance of our sentiment measure.

Keywords Investor sentiment · Event study · Social media · Micro-blogs · Natural language processing

JEL Classification C55 · G14 · G17

1 Introduction

Can the stock market be predicted by analyzing social media? Recent developments in machine learning and the growing quantities of available text data from online news, social media and annual reports have triggered intensive research in finance. In their pioneering paper, Antweiler and Frank (2004) compute a bullishness measure out of 1.5 million messages posted on Yahoo! Finance and Raging Bull and find that stock messages help predict market volatility. Their results clearly reject the hypothesis that all that talk is just noise. They show that there is financially relevant information present in social media. In a similar vein, Tetlock (2007) constructs a measure of media pessimism from a Wall Street Journal column and finds that it predicts downward pressure on market prices.

✉ Damir Filipović
damir.filipovic@epfl.ch

Marc-Aurèle Divernois
divernois@gmail.com

[1] Ecole Polytechnique Fédérale de Lausanne and Swiss Finance Institute, 1015 Lausanne, Switzerland

Published online: 29 December 2023

Most of the previous financial studies of social media rely on pre-defined or manually annotated sentiment dictionaries. Such approaches are limited in various ways. How to create a sentiment classifier that understands the vocabulary of the messages posted by the investors? For instance, "bull" is an animal in everyday language but it is someone optimistic in the financial jargon. Loughran and McDonald (2012) create a word list, which helps classify tone in a financial document. However, this might not be sufficient in the context of social media because messages posted present many typos, abbreviations and slang, so one needs to have an additional layer of data preprocessing. For instance, the word "gooooooood" would not be recognized by the model if it is not corrected into "good" first. On the other hand, manually annotating and validating dictionaries is not a scalable approach to handling social media content. On another note, Cookson and Niessner (2020) employs an entropy classifier that classifies StockTwits messages into bullish versus bearish, and then uses the resulting messages to form a stock-daily disagreement measure. However, it does not include the possibility of sentiment neutral messages, which arguably exist.

Our paper overcomes these limitations. We develop a machine learning algorithm to classify the sentiment of a large sample of StockTwits messages as bullish, bearish, or neutral. The sample consists of all messages referring to US and Canadian stocks, including ETFs and other types of securities available on CRSP/Compustat, from January 2010 to March 2020. We train our machine learning classifier on the set of all user sentiment-labeled messages, which constitute about one third of the sample. We then classify the sentiment of all remaining messages. Our method scales and performs very well. It achieves an out-of-sample accuracy of 85.9%, which compares well to the anecdotal 80–85% probability that human annotators agree on the sentiment of a document, see, e.g., Wilson et al. (2005) and Chen et al. (2020).

We then construct a stock-aggregate daily sentiment polarity measure and relate it to daily stock returns. We find that polarity is positively associated with contemporaneous returns, also when controlling for lagged returns. However, unconditionally, polarity cannot predict next-day returns, which is in line with the efficient market hypothesis (EMH). We then conduct an event study. We define events as days of sudden peaks of message volume of individual tickers. We classify events as bullish, bearish, or neutral depending on the prevailing polarities. We find that bullish (bearish) events are strongly associated with large positive (negative) abnormal returns. Cumulative abnormal returns over the preceding 20 days of an event have no predictive power on the type of event. Returns normalize immediately after the jump on the event date, which again is in line with the EMH. In contrast, remarkably, we find that cumulative abnormal polarity has statistically significant predictive power on the type of event. We assess the economic relevance of our findings with the performance of cumulative abnormal polarity ranked portfolios. We find that for appropriate choices of thresholds, cumulative abnormal polarities provide valuable signals for stock market investments.

Our results and method are of broad interest for researchers that analyze social media and their interplay with stock markets. We collect and process a large dataset of messages from StockTwits posted between January 2010 and March 2020. We generate a vocabulary of one million investor sentiment-labeled terms consisting of

up to three words that frequently appear in StockTwits messages.[1] As a method, we develop a simple and efficient sentiment classifier of micro-blogs for imbalanced data. This addresses the stylized fact that bloggers post more bullish than bearish-labeled messages. In our sample, the ratio is five to one. What's more, we find that not all messages carry a substantial stock market relevant sentiment. Rather than re-sampling from the underrepresented bearish class, we thus introduce an auxiliary neutral class. We then train two independent binary classifiers. The first (second) classifies messages as bullish versus non-bullish (bearish versus non-bearish). We aggregate the two binary outcomes and classify a message as bullish (bearish) for the concordant combination bullish/non-bearish (non-bullish/bearish), and neutral otherwise. This approach is very simple and efficient, and eliminates the class imbalance bias at the same time. It can be built on any traditional binary classifier. In this study, we use logistic regression on Term Frequency-Inverse Document Frequency (TFIDF)-vectorized messages. TFIDF is a weighting scheme gauging the importance of a word in a document.

Our paper contributes to the growing literature on machine learning classification of social media and its interaction with the stock market. Most previous financial studies use Twitter as their primary source of data. Twitter has the advantage of being used by a wide range of people across the world and a few influencers can attract the attention of many investors. In 2013, following a meeting with Tim Cook (Apple CEO), Carl Icahn tweeted that he bought a large position in Apple and believed that the company is extremely undervalued. This bullish tweet caused the market capitalization of Apple to jump by $12 billion. In 2019, JPMorgan created the Volfefe Index to track Donald Trump's tweets impact on the stock market. However, it is more difficult to disentangle noise from relevant tweets in Twitter than in other more focused social media. Results from Ghoshal and Roberts (2016) show that StockTwits is significantly more informative than Twitter data. This is not surprising as StockTwits is a finance-focused platform whereas Twitter also captures irrelevant opinions on a wide range of non-finance related matters.

Our paper is the first that analyzes the predictive power of StockTwits messages on stock returns unconditionally and around specific events. Renault (2017) builds an intraday investor sentiment indicator using messages and finds that the change in investor sentiment of the first half-hour of a trading day helps forecast the last half-hour market return of that trading day. However, his classifier is based on a dictionary consisting of 8 thousand manually validated and modified terms, which limits its scalability. Renault (2020) uses larger data sets and compares various classifiers, including machine learning.

Our approach is in some parts similar to Ranco et al. (2015), who also study the relation of micro-blog sentiments with stock returns. However, they use Twitter data, whereas the finance-tailored StockTwits data we use results in higher contemporaneous correlations between stock returns and polarity. They manually annotate 100 thousand tweets, which limits the scalability of their approach. Cookson and Niessner (2020) also use logistic regression on a list of words to classify StockTwits

[1] Our collected dataset of messages and generated vocabulary are available from the authors upon request.

messages that were unclassified in the original sample as either bearish or bullish.[2] Our sample is much larger (90 million versus 1 million message in Ranco et al. (2015) and 1.5 million messages in Cookson and Niessner (2020)) and covers a longer period (10 years versus 13 months in Ranco et al. (2015) and 21 months in Cookson and Niessner (2020)).

Earlier studies of textual analysis and stock prices also include Das and Chen (2007), who provide a dictionary based approach trained on hand-classified messages extracted from the Yahoo message board in the period from July to August 2001. Boudoukh et al. (2013) use articles on selected S&P500 companies from the Dow Jones Newswire from 2000 to end of 2009. Using a proprietary textual analysis methodology available on the Dow Jones platform, they show that returns respond more to relevant news, both by type and by tone. Heston and Sinha (2017) measure sentiment of Reuters new with a proprietary Thomson Reuters neural network. They find that positive news stories are quickly incorporated into positive stock returns but negative news stories take a while to come into prices. Ke et al. (2020) extract sentiment from news articles on the Dow Jones Newswires. They train a sentiment score directly on returns. In contrast, we use user sentiment-labeled StockTwits messages as training and validation set for our sentiment classifier.

Our paper also contributes to the EMH literature by gauging how cumulative average abnormal returns and abnormal polarities behave around sudden peaks of message activity.

The remainder of the paper is structured as follows. Section 2 discusses StockTwits and stock market data. Section 3 develops our sentiment classifier based on TFIDF vectorization. Section 4 introduces the sentiment polarity measure. Section 5 relates it to stock returns. Section 6 contains the event study. Section 7 discusses the sentiment-sorted portfolio performance. Section 8 concludes. The appendix contains additional statistics and background material.

2 StockTwits and stock market data

StockTwits is a large social media platform similar to Twitter but designed for investors and traders. Users register online and can post messages about any listed stock through the prefix $ followed by the ticker of the stock. StockTwits was created in 2008 as an app built on the Twitter's API and later detached from Twitter to build a standalone social network. As of April 2019, it had over two million registered users and the number of daily posted messages has been growing exponentially, see Fig. 1.

StockTwits describes itself as "the voice of social finance and the best way to find out what is happening right now in the markets and stocks you care about". In practice, it is effectively used by finance professionals to express their opinions on individual stocks and the market as a whole. Importantly, users have the option to label their posted messages as either bullish or bearish.[3] This feature is key for

[2] As Cookson and Niessner (2020) consider a binary classifier, their maximum entropy-based method is in effect equivalent to a standard logistic regression.

[3] This optional label was effectively available as of mid-2010.

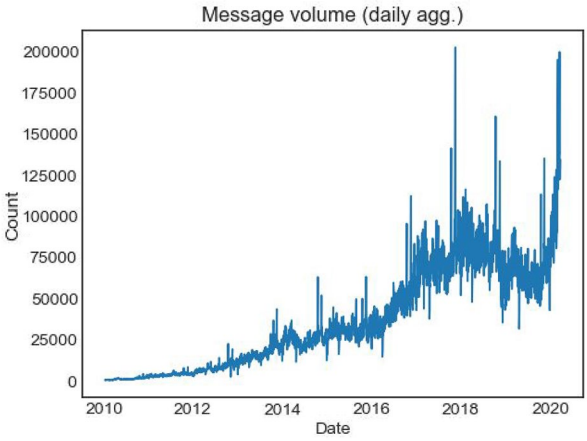


Fig. 1 Number of messages posted daily on StockTwits

our approach, as it allows for sentiment classification of all messages using machine learning trained on the user-labeled messages.[4]

The reasons for using StockTwits and not other social media data for financial studies are at least threefold. First, a major challenge in applying natural language processing is the creation of an appropriate labeled vocabulary. Loughran and McDonald (2011) show that it is essential to have a specific vocabulary to interpret finance documents (i.e., many words have a different meaning in finance than in traditional English, such as "bear trap"). In addition to that, social media slang is an additional layer of language complexity. To this extent, the functionality to self-tag bullish and bearish messages that StockTwits is extremely valuable as it allows the creation of a specific labeled vocabulary out of labeled messages. We are not aware of any other social media platform in finance offering this functionality. Second, text data from StockTwits is more reliable and less noisy than from general purpose platforms, such as Twitter, because messages focus on finance and economics matters only. Micro-bloggers have incentives to post valuable information in order to maintain or increase mentions and retweets, and thus have a greater share of voice in the forum (Sprenger et al. 2014). On the other hand, StockTwits messages might be biased and subject to malicious users that try to manipulate the market. However, market manipulations likely happen only rarely as the SEC closely monitors potential influencers to prevent any market abuse. Third, extracting data from StockTwits is easy because of its API. StockTwits' API is designed to query the database to download messages via JSON requests. We provide a short tutorial in "Tutorial for StockTwits messages extraction" of appendix.

We use stock market data from CRSP/Compustat. We extract daily closing prices, daily volume of transactions and number of shares outstanding for all US and Canadian stocks, as well as ETFs and some other types of securities, from January 2010 to March 2020. Stock prices and number of shares are adjusted to account for any distribution (i.e., dividends, stock splits) so that a comparison can be made

[4] One third of the messages in our sample have a user labeled sentiment.

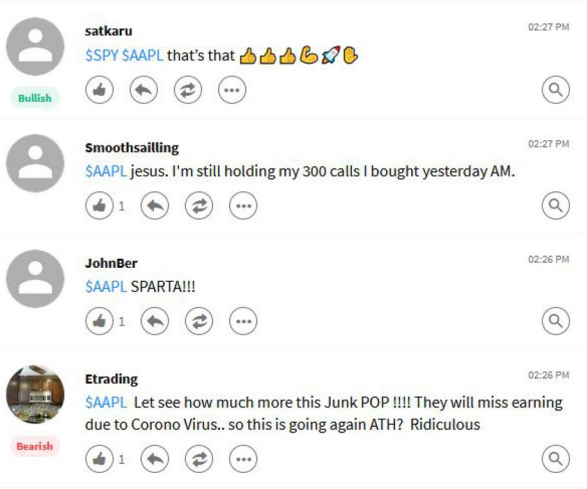


Fig. 2 Screenshot of StockTwits as of 3rd March 2020, for a query of AAPL

on an equivalent basis before and after the distribution. We use as risk-free rate the 3-month US T-bill rate, converted into daily risk-free returns. We henceforth refer to daily stock excess returns over risk-free simply as returns. Using a Python script, we then extract all messages from StockTwits for the list of tickers corresponding to the sample of US and Canadian stocks. This results in 90 million messages, which we download and store as JSON files.[5] Overall, our sample covers 8843 tickers, whereof 75% refer to ordinary common share, 15% to ETFs, and the remaining 10% to other types of securities. Henceforth, we interchangeably refer to any of these securities as either a stock or a ticker.

Every StockTwits message includes eight features: (1) the reference ticker(s), (2) a timestamp, (3) a unique message identifier, (4) the body of the message, (5) the sentiment label (bearish, bullish, or none) entered by the user, (6) a unique identifier of the user who posted the message, (7) the number of messages published by the user who posted the message, and (8) the number of followers of the user who posted the message. Our sentiment analyses builds on the first five features. The last three provide additional information on the network structure, which we briefly discuss in "User summary statistics" of appendix.

Figure 2 shows a screenshot of the StockTwits website as of 3rd March 2020, for a query of AAPL, which is the ticker for Apple. The first message is labeled as bullish by the user "satkaru", the two next are unlabeled messages that will be classified by our machine learning algorithm, and the last message is labeled as bearish by the user "Etrading".

Left plot of Fig. 3 shows the top 30 most discussed tickers on StockTwits. SPY, a large ETF that tracks the S&P500 stock market index, is the most discussed ticker, followed by Apple and other big tickers. The messages about the 15 (30) biggest tickers represent 20% (25%) of the total number of messages, which indicates that

[5] A message may refer to multiple tickers. We count any such message towards any ticker that it refers to. We give more information about this double counting in "Message count" of appendix.

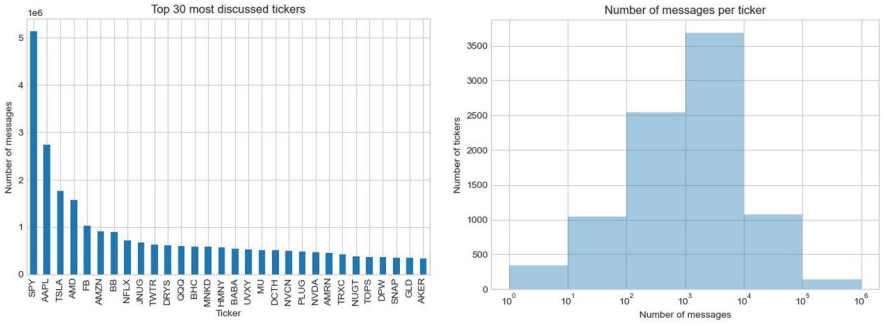


Fig. 3 Ticker summary statistics. Left graph shows the top 30 most discussed tickers on StockTwits. SPY is the ticker of a large ETF tracking the S &P 500 and AAPL is the ticker for Apple. Right graph shows the distribution of the number of messages across tickers

Table 1 Preprocessing of five sample messages

Before preprocessing	
(1)	@CassandraTwit $uvxy contango 3.5%...still long. goooooood
(2)	$FRPT Take profits while you still can
(3)	$UVXY $tvix go time boys and girls. Holding overnight again
(4)	$dnr Nice upgrade as company goes into its quiet period!
(5)	$SPY market won't reverse again towards closing. Get put options
After preprocessing	
(1)	contango still long good
(2)	take profit while you still can
(3)	go time boy and girl hold overnight again
(4)	nice upgrade as company go into its quiet period
(5)	market will not reverse again towards closing get put options

Preprocessing operations include: punctuation removal, lower casing, apostrophe handling, contraction form handling (i.e., "won't" becomes "will not"), tickers removal, users removal, URLs removal, parsing and a simple spell corrector dealing with more than two repeated characters (i.e., "goooood" becomes "good")

users talk about a wide panel of tickers and not only big firms. Right graph shows a histogram of the number of messages per ticker. The x-axis is log-scaled because due to extreme values the distribution is highly skewed.

Text messages need to be transformed into a quantitative vector to be fed into our machine learning classifier, which in turn computes a sentiment score. This transformation consists of several steps. First, we apply some preprocessing operations to the text messages: an apostrophe handler, a contraction form handler (e.g., "aren't" becomes "are not"), tickers removal, stop words removal (e.g., "a", "the", "of"),[6]

[6] We follow Renault (2020) and Saif et al. (2014) and use a restrictive list of stopwords to avoid accuracy decrease.

Fig. 4 Bullish word cloud (left), bearish word cloud (right). These correspond to the most frequent terms (up to trigrams) in user-labeled bullish (bearish) messages relative to their total appearance. The size of the terms represents their importance in the cloud

users removal, lemmatization, URLs removal and a simple spell corrector dealing with more than two repeated characters (e.g., "soooo goooood" becomes "soo good"). Table 1 shows five examples of messages before and after preprocessing.

The next step is tokenization: the slicing of a text message into smaller units called terms or tokens. In financial lingo, some words only have meaning when associated with other words (i.e., "bad apple" or "bear flag"). *N*-gram models allow accounting for words frequently occurring together with other words. The main hyperparameter in an *N*-gram model is the number N of words that form a term: a unigram is a term with only one word, a bigram is a term with two consecutive words, etc. Larger *N*-gram models increase dramatically the size of the vocabulary (i.e., the collection of all terms considered). We select $N = 1, 2, 3$ and truncate the resulting vocabulary such that it consists of the one million most frequent terms in the union of all unigrams, bigrams and trigrams.

Figure 4 represents the bullish and bearish word clouds. These represent the most frequent terms in all user-labeled bullish and bearish messages relative to their total appearance, respectively. The size of the terms represents their relative weight in the cloud. In the bullish cloud, we see terms such as "bullish divergence", "room to grow", "lot potential" which we can clearly interpret as bullish signals. In the bearish cloud, we find terms such as "recent resistance", "short setup", "bad apple" which again we can directly interpret as bearish signals. These findings are reassuring in the sense that the content of the messages on StockTwits are consistent with their labels. We checked for anomalies at random, but did not find significant issues. "Anomalies" of appendix discusses two such anomalies.

3 Sentiment classification

Left plot of Fig. 5 shows the proportions of user sentiment-labeled messages across time. In the early years of the platform, most messages were unlabeled, presumably because users were not familiar with the sentiment label yet. Albeit the proportion of unlabeled messages monotonically declines over the years, almost 60% of the more recent messages are still unlabeled. Overall, around 30 million messages are user-labeled and 60 million messages are unlabeled. We conjecture that by far not all unlabeled messages reflect market neutral opinions. Indeed, the right plot of Fig. 5

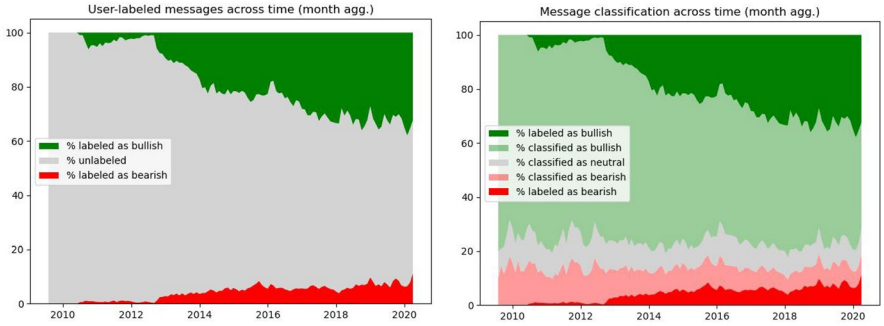


Fig. 5 Left plot shows the proportions of user-labeled messages: bullish (green), bearish (red), and unlabeled (gray) across time. Right plot shows the proportions of machine learning classified messages: bullish (light green predicted, green user-labeled), bearish (light red predicted, red user-labeled), and neutral (gray) across time. Proportions are aggregated monthly (color figure online)

reveals that a substantial part of user-unlabeled messages is machine learning classified as bullish or bearish. Hence these user-unlabeled messages contain indeed market relevant information, which we are able to capture by our algorithm.

Among the user-labeled messages we find five times more bullish than bearish ones. This ratio indicates that investors are on average optimistic about the market, which is consistent with findings in the literature, e.g., Renault (2017). Such an imbalance is a well-known issue in machine learning classification as it creates a bias towards the over-represented class, see Chawla et al. (2004). There are various ways to tackle class imbalance. An all-purpose standard approach in machine learning is to over-sample the minority class, which consists of randomly re-sampling from the minority class and thus artificially re-balance the class sizes in the data. We use a different approach, which is tailored for our setup. As not every message carries a substantial stock market relevant sentiment, we deviate from the bullish–bearish dichotomy. Instead, we create an auxiliary neutral sentiment class to account of messages that do not take a clear stand. See "Examples of classified messages" of appendix for examples of such neutral messages.

We first randomly select 80% of the user sentiment-labeled messages as training set (in-sample) and keep the remaining 20% as test set (out-of-sample). On the training set, we then train two independent binary classifiers. The first (second) classifies messages as bullish versus non-bullish (bearish versus non-bearish). We aggregate the two binary outcomes and classify a message as bullish (bearish) for the concordant combination bullish/non-bearish (non-bullish/bearish), and neutral otherwise. This approach is very simple and efficient, and eliminates the class imbalance bias at the same time.[7] It builds on any traditional binary classifiers.

Here we use logistic regression on Term Frequency-Inverse Document Frequency (TFIDF)-vectorized messages, as in, e.g., Yildirim et al. (2018), Qasem et al. (2015),

[7] Our approach shares similarities with conformal prediction, which predicts a set of classes that covers an instance with some probability, see, e.g., Shafer and Vovk (2008). We then interpret both prediction sets, the empty set as well as the set consisting of bullish and bearish, as neutral class.

Erdemlioglu et al. (2017). We choose TFIDF over more sophisticated algorithms such as BERT because of simplicity reasons and that it is already showing good performance. Also, TFIDF is easier to interpret and generates as a side product a dictionary of finance related terms. TFIDF is a widely used method to transform a text, in our case a message m, into a numerical vector, $TFIDF_m$. The dimension of this vector is equal to the size of the vocabulary (the collection of all terms across all messages). The components of the vector encode the importance of the corresponding terms t in the message m, as formally defined by $TFIDF_{m,t} = TF_{m,t} \cdot IDF_t$. The first factor measures how frequently term t appears in the message,

$$TF_{m,t} = \frac{\sum_{i=1}^{N_m} \mathbf{1}_{t=t_{m,i}}}{N_m}, \tag{1}$$

where N_m denotes the number of terms $t_{m,i}$ in message m. The second factor measures how important term t is to the message,

$$IDF_t = \log\left(\frac{V}{\sum_{j=1}^{V} \mathbf{1}_{t\in m_j}}\right), \tag{2}$$

where V denotes the total number of messages m_j. A term t appearing in many documents (such as "the", "is", "of") is likely to have low information content, hence a low IDF_t. Logistic regression estimates the conditional probability of a message m being bullish given $TFIDF_m$,

$$\mathbb{P}[\text{message } m \text{ is bullish} \mid TFIDM_m] = \frac{e^{\beta_0+\beta_1\cdot TFIDF_m}}{1+e^{\beta_0+\beta_1\cdot TFIDF_m}}. \tag{3}$$

The parameters (β_0, β_1) are estimated by maximum likelihood, and we use the Python scikit-learn library Pedregosa et al. (2011) for the implementation. Message m is classified as bullish (non-bullish) if the conditional probability (3) is larger (smaller) than an auxiliary chosen threshold, as discussed below. In a similar way, we perform an independent logistic regression for the bearish versus non-bearish classification.

Every message then classifies into one of the following combinations: (non-bullish, bearish), (bullish, bearish), (non-bullish, non-bearish), (bullish, non-bearish). For the first and last combinations, the two algorithms agree and the final classification is defined to be bearish (non-bullish, bearish) or bullish (bullish, non-bearish), respectively. For the two middle combinations, (bullish, bearish) and (non-bullish, non-bearish), the two algorithms disagree, so that the final classification is defined to be neutral. Formally, every message m is mapped onto either

$$m \mapsto \begin{cases} \text{(non-bullish, bearish)} & =: \text{bearish} \\ \text{(bullish, bearish)} & =: \text{neutral} \\ \text{(non-bullish, non-bearish)} & =: \text{neutral} \\ \text{(bullish, non-bearish)} & =: \text{bullish.} \end{cases} \tag{4}$$

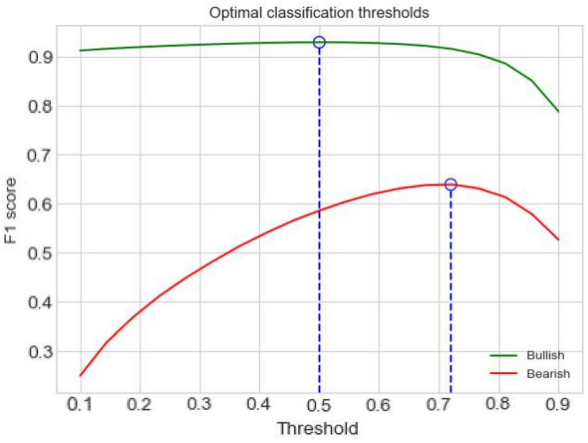


Fig. 6 Optimal classification thresholds. The green (red) line is the F1 score for the bullish versus non-bullish (bearish versus non-bearish) classifier as a function of the threshold. Circles indicate the maximal F1 scores (color figure online)

To select optimal classification thresholds, we maximize the F1 scores.[8] The F1 scores of the two binary classifiers differ because they depend on which class is defined as the positive one. We recap the definition of the F1 score in "Classifier performance" of appendix. Figure 6 shows the F1 scores as functions of the threshold. Circles indicate the maximal F1 scores, along with the corresponding optimal thresholds, 0.50 and 0.72, respectively.

If the sentiment score of a message is bigger (smaller) than 0.72 (0.50), respectively, then both classifiers agree and the sentiment of the message is classified as bullish (bearish), respectively. If the sentiment score is between 0.50 and 0.72, the classifiers disagree, (bullish, bearish), and we consider the message as neutral. Finally, we overwrite the predicted sentiment of any message by the user-labeled sentiment whenever the latter is available. Research in sentiment classification shows that human annotators tend to agree about 80–85% of the time when evaluating the sentiment of a document (see, e.g., Wilson et al. (2005) and Chen et al. (2020)). This is a benchmark for the accuracy that a sentiment classifier should meet or beat. The out-of-sample accuracy of our combined classifier is 85.9%. "Classifier performance" of appendix provides in-sample and out-of-sample confusion matrices for our combined classifier.

Right plot of Fig. 5 shows the proportions of our machine learning classifications across time. Percentages of bearish (user-labeled and classified as bearish) and bullish (user-labeled and classified as bearish) messages are stable over time, suggesting that our classification method is robust. Even if most messages were not user-labeled in the early years of the platform, as seen in the left plot of Fig. 5, we are now able to classify the sentiment of all messages in the sample, including a neutral class. Consistent with the over-representation of bullish messages observed in the user-labeled messages, there are substantially more messages classified as bullish than bearish. Examples of classified messages are given in "Examples of classified messages" of appendix.

[8] It is common practice to learn the optimal thresholds in-sample, as this is part of the training process.

4 Polarity

We next aggregate message sentiments on a daily ticker-level and across the market. Thereto, we denote by $C_{i,t,j} = 1, 0, -1$ for bullish, neutral, bearish, respectively, the sentiment of the jth message about ticker i on day t.[9] We follow Antweiler and Frank (2004) and define an average sentiment measure, which we call the sentiment polarity of ticker i on day t, as

$$P_{i,t} = \frac{\sum_{j=1}^{V_{i,t}} (\mathbf{1}_{C_{i,t,j}=1} - \mathbf{1}_{C_{i,t,j}=-1})}{\sum_{j=1}^{V_{i,t}} (\mathbf{1}_{C_{i,t,j}=1} + \mathbf{1}_{C_{i,t,j}=-1})}, \tag{5}$$

where $V_{i,t}$ denotes the number of messages about ticker i on day t.[10]

As an aggregate, we define the market polarity as a weighted average over all tickers

$$P_t^M = \frac{\sum_i V_{i,t} \cdot P_{i,t}}{V_t^M}, \tag{6}$$

where $V_t^M = \sum_i V_{i,t}$ denotes the number of messages on day t. Figure 7 shows a scatter plot of the market polarity P_t^M versus the polarity of SPY. The slope coefficient of the regression line is statistically significantly positive and the contemporaneous Pearson correlation coefficient is 0.53, suggesting that the market polarity is an accurate measure of the aggregated sentiment of the market.[11] Also, consistent with Fig. 5, SPY and market polarities are bullish-biased.

5 Relation to stock returns

For the following time-series analysis and event studies we restrict our sample. There are two reasons for doing so. First, we keep computational cost at a reasonable level. Second, and more importantly, the time series of ticker-individual polarities exhibit spikes and are too noisy if the daily message volumes $V_{i,t}$ are too small. In fact, Stocktwits is neither regulated nor moderated, so one needs to filter the information before credibly relating it to stock returns. Even if Stocktwits has

[9] We follow the close-to-close convention. First, we remove all non-business days from the sample, whereby messages posted on non-business days count towards the next business day. "Day t" then stands for the time interval from 4pm on the previous business day $t-1$ to 4 pm on business day t. This convention is consistent with the stock return data, which are close-to-close, and thus avoids any look-ahead bias of our sentiment polarity.

[10] If $V_{i,t} = 0$ then we set $P_{i,t} = 0$. Similar average sentiment measures are studied in Ranco et al. (2015) and Cookson and Niessner (2020), but without a neutral class. A related, derived measure is disagreement, which is defined as the standard deviation of the message sentiments $C_{i,t,j}$, conditional on not being neutral, $C_{i,t,j} \neq 0$, which formally is given by $\sqrt{1 - P_{i,t}^2}$. Disagreement as such is less informative than polarity as it disregards the sign of $P_{i,t}$.

[11] We do not expect P_t^M to be equal to the SPY polarity because the underlying sets of stocks differ: market polarity contains stocks that are not in the S&P500 and vice versa.

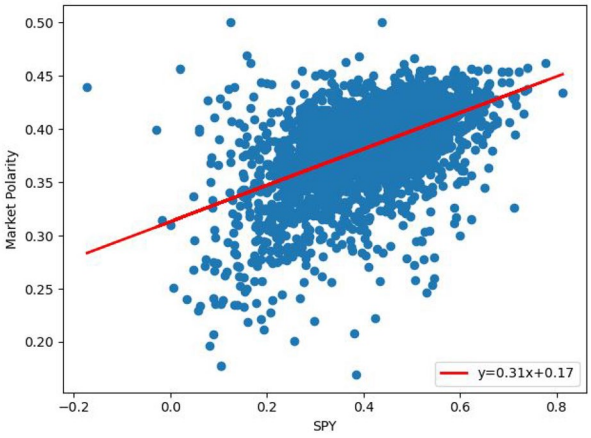


Fig. 7 Market polarity versus SPY polarity. The red line shows the linear regression line and coefficients (color figure online)

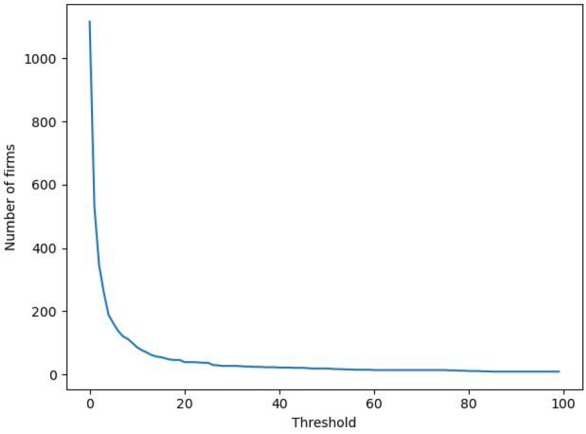


Fig. 8 Coverage as a function of the median threshold. A lower threshold increases the coverage at the expense of a bigger bias in the polarity

valuable information from respected contributors, a blog[12] describes the concerns that may rise when using Stocktwits as a financial information provider, namely self-promotion, lack of credibility and other noise. To diversify noise and better extract information, we therefore exclude tickers from our sample that are rarely discussed. To do this, we compute the median of daily message volumes $V_{i,t}$ for each ticker and exclude tickers with a median of less than 50 from our sample. Decreasing the median threshold increases the coverage at the expense of more noise in the daily polarity. Figure 8 shows the coverage as a function of the median threshold. To increase the coverage one would need to decrease the threshold a lot (e.g., decreasing the median threshold to 40 from 50 would increase the number of tickers covered to merely 22 from 19). A median threshold of 50 results in a balanced compromise between noise and coverage.

[12] https://www.warriortrading.com/stocktwits-review, last accessed on 1st of July 2022.

Table 2 Coverage after the trimming process

Ticker	Name	Market capitalization
AAPL	Apple	1287
AMD	Advanced Micro Devices	53
AMRN	Amarin	7
AMZN	Amazon	920
BABA	Alibaba	571
BAC	Bank of America	311
BB	BlackBerry	4
FB	Facebook	585
GLD	Gold ETF	59
IWM	Small-Cap ETF	55
JNUG	Direxion	0.5
MNKD	MannKind Corporation	0.2
NFLX	Netflix	142
PLUG	Plug Power	1
QQQ	Nasdaq100 ETF	134
SPY	S&P500 ETF	391
TSLA	Tesla	76
TWTR	Twitter	25
UVXY	VIX ETF	0.8

List of tickers and corresponding market capitalization as of 31st of December 2019

Table 2 shows the list of the 19 tickers above this threshold and their associated market capitalization as of 31st of December 2019. It appears that these most discussed tickers cover all sizes of stock, and hence we avoid big-firm bias. Also, it includes not only single firms but also ETFs on alternative investments. Many tickers are technology stocks, which arguably are the most discussed on social media. However, other sectors are also covered, and so even if we have a restrictive universe, it is well diversified. Even though the restricted sample contains a relatively small number of tickers, the event studies in the following section are still based on more than 1000 events.

Now, to understand how our polarity measure is related to investor sentiment, we perform linear regressions of contemporaneous and next-day returns on polarity and lagged returns as control variables:

$$R_{i,t} = \alpha^{\text{cont}} + \beta^{\text{cont}} \cdot P_{i,t} + \gamma_1^{\text{cont}} \cdot R_{i,t-1} + \gamma_2^{\text{cont}} \cdot R_{i,t-2} + \epsilon_{i,t}^{\text{cont}}, \tag{7}$$

$$R_{i,t+1} = \alpha^{\text{next}} + \beta^{\text{next}} \cdot P_{i,t} + \gamma_1^{\text{next}} \cdot R_{i,t-1} + \gamma_2^{\text{next}} \cdot R_{i,t-2} + \epsilon_{i,t}^{\text{next}}. \tag{8}$$

Table 3 Results from linear regressions of contemporaneous and next-day stock returns on polarity

	$R_{i,t}$	$R_{i,t+1}$
Constant	− 0.0049***	− 0.0001
	(0.000)	(0.000)
$P_{i,t}$	0.0094***	0.0002
	(0.001)	(0.001)
$R_{i,t-1}$	− 0.0203	0.0007
	(0.011)	(0.007)
$R_{i,t-2}$	− 0.0069	0.0033
	(0.006)	(0.006)
R^2	0.012	0.000
No. Obs	34,062	34,062

Stock returns are trimmed at the 5% percentile on both sides to remove anomalies. As robustness check, results still hold when stock returns are trimmed at lower levels (1 percentile and no trim). Standard errors are reported in parentheses. Statistical significance at the 99%, 95%, and 90% level is indicated with ***, **, *, respectively

Table 3 shows that the regression coefficient β^{cont} is positive and significant for contemporaneous returns. This indicates that polarity is a good contemporaneous proxy for the sentiment of investors. Further supporting evidence is given by the correlation between polarity and contemporaneous returns at the ticker level.

Figure 9 shows the time series during 2019 for the top 6 most discussed tickers. Correlations are always positive and range between 0.1 and 0.3. In contrast, regressing next-day returns reveals that polarity has no predictive power for next-day stock returns unconditionally. In the following section we show how polarity has predictive power around specific events.

6 Event study

Event studies constitute a statistical method widely used in financial econometrics, see, e.g., MacKinlay (1997). In general, they are used to measure the effect of events on the market value of stocks. Well-known applications of event studies include the testing of various forms of the efficient market hypothesis (EMH) [see Fama et al. (1969) and Fama (1991)].

6.1 Events

We define events as days with an unusual large number of messages for individual tickers. We conjecture that a sudden peak in StockTwits message volume indicates that an important corporate or stock market event is happening on the day of the peak. Figure 10 shows that increases (decreases) in message volumes are positively associated with increases (decreases) in contemporaneous weekly stock transaction

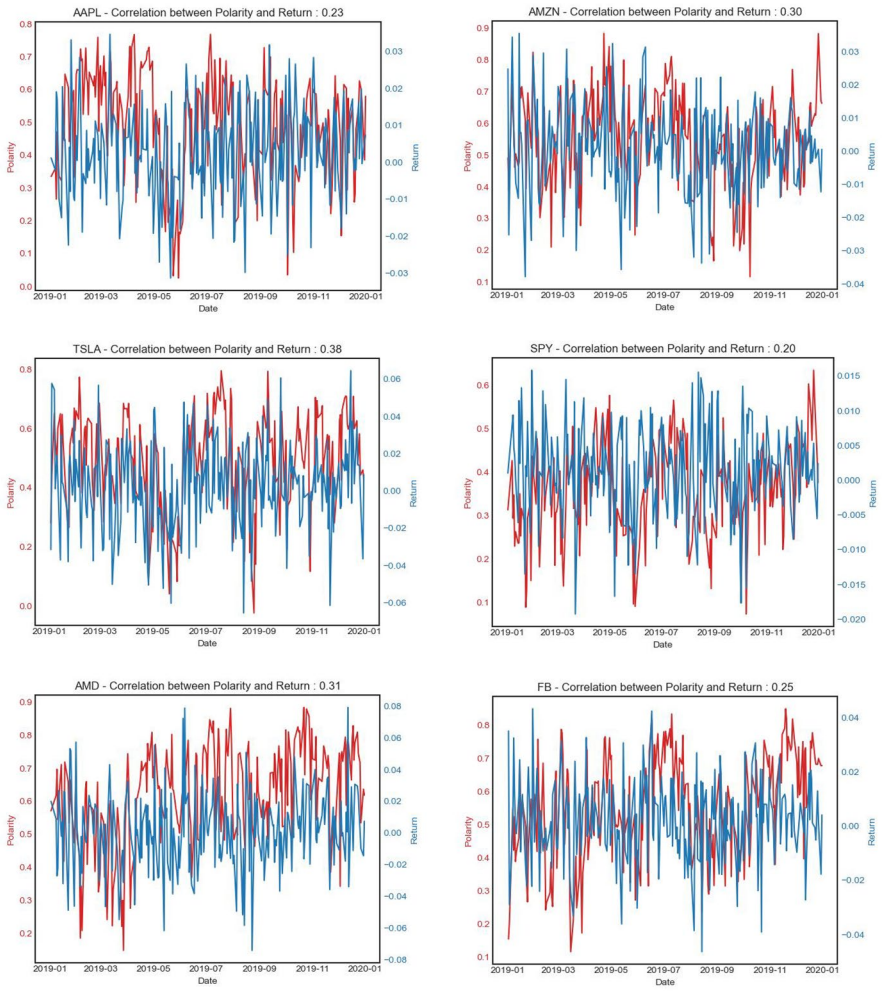


Fig. 9 Time series of daily polarity (red-left axis) and daily stock returns (blue-right axis) since 1st of January 2019 for the top 6 most discussed tickers. Pearson correlation between the two time series is shown in the title (color figure online)

volumes. These co-movements suggest that message volume peaks are a good proxy for corporate and stock market events.

To measure unusual activity peaks, we use as benchmark model a one-year rolling window regression of daily relative message volume changes of ticker i on daily relative total message volume changes.[13] Formally,

[13] Accounting for the lead time of the one-year rolling estimation window, the event study effectively applies to the shorter period from January 2011 to March 2020.

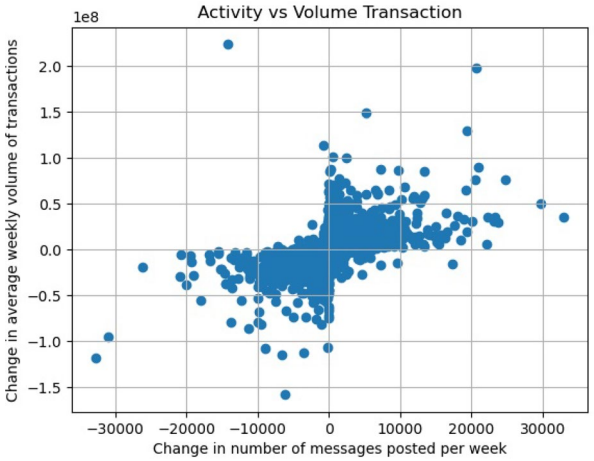


Fig. 10 Changes in weekly volume of transactions on the *y*-axis versus changes in message activity on the *x*-axis. Activity is measured in weekly messages posted per ticker

$$\frac{\Delta V_{i,t}}{V_{i,t-1}} = \alpha_i^V + \beta_i^V \cdot \frac{\Delta V_t^M}{V_{t-1}^M} + \epsilon_{i,t}, \tag{9}$$

which gives the one-year rolling estimates $\hat{\alpha}_i^V$ and $\hat{\beta}_i^V$. We then define the abnormal message volume changes for ticker i on day t as

$$AV_{i,t} = \frac{\Delta V_{i,t}}{V_{i,t-1}} - \left(\hat{\alpha}_i^V + \hat{\beta}_i^V \cdot \frac{\Delta V_t^M}{V_{t-1}^M} \right). \tag{10}$$

We define an event for ticker i as any day t where the standardized abnormal volume exceeds two,

$$\frac{AV_{i,t} - \hat{\mu}_{AV_i}}{\hat{\sigma}_{AV_i}} > 2, \tag{11}$$

where $\hat{\mu}_{AV_i}$ and $\hat{\sigma}_{AV_i}$ denote the one-year rolling empirical mean and standard deviation.

Next we define the type of the event as either bullish, neutral or bearish. We use the abnormal polarity $AP_{i,t}$ of the event date to assess how on average investors perceive the event. Figure 11 shows the distribution of abnormal polarities on event dates.

We chose to use the one-third (-0.03) and two-third percentile (0.07) of the distribution of abnormal polarities as thresholds for the type of the event. We define the type of the event for ticker i at t as

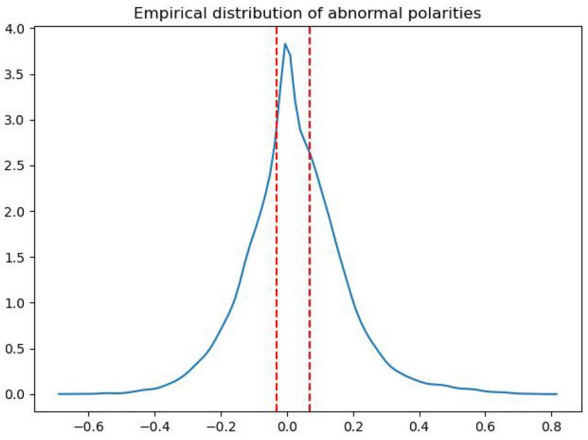


Fig. 11 Empirical distribution of abnormal polarities on event dates. Red dashed lines show the one-third and two-third percentiles. We chose to use the one-third (-0.03) and two-third percentile (0.07) of the distribution of abnormal polarities as thresholds for the type of the event

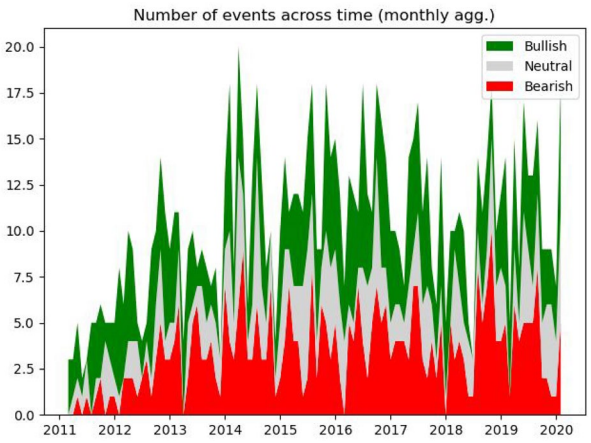


Fig. 12 Number of events of each type across time. Numbers are aggregated monthly

$$Type_{i,t} = \begin{cases} \text{Bullish} & \text{if } AP_{i,t} > 0.07, \\ \text{Neutral} & \text{if } AP_{i,t} \in [-0.03, 0.07], \\ \text{Bearish} & \text{if } AP_{i,t} < -0.03. \end{cases} \tag{12}$$

Overall, across 19 tickers, we identify 1131 events, whereof 454 bullish, 294 neutral, and 383 bearish types. This coverage is on par with previous studies (e.g., MacKinlay (1997) analyze 30 stocks and 600 events). Figure 12 shows the aggregate events and their types across time. The count of events looks stationary over time, apart from a build up phase of the platform in the early part until 2014. The distribution of event types is also balanced across time.

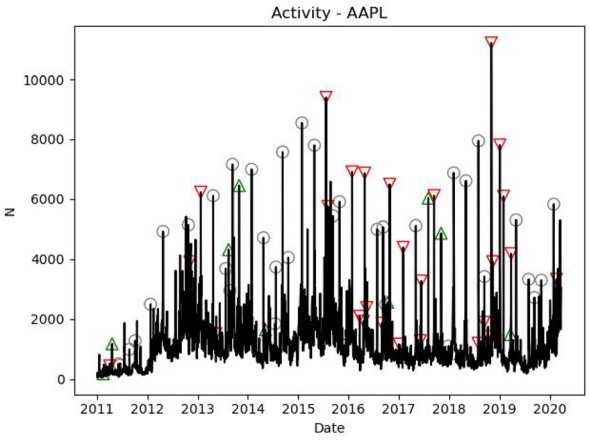


Fig. 13 Daily message volume for Apple. Events are days with an unusual high number of messages. Green upper-triangles show bullish events, gray circles are neutral events and red down-triangles represent bearish events (color figure online)

Table 4 Selected events and associated description and types for Apple

Date	Description	Type
2012-04-24	Earnings announcement	Bullish
2012-09-12	Presents iPhone 5	Neutral
2014-09-09	Presents Apple Watch	Neutral
2017-05-02	Announces drop in iPhone sales	Bearish
2017-08-31	Earnings announcement	Bullish
2017-09-12	Presents iPhone X	Bearish
2019-01-02	CEO Letter to investors	Bearish
2019-09-10	Presents iPhone 11	Neutral
2019-10-30	Earnings announcement	Bullish

This list is for illustration and non-exhaustive (9 out of 73)

As an illustration, Fig. 13 shows for Apple the time-series of message volume and the the corresponding events. Between January 2011 and March 2020, our algorithm identified 73 events for Apple. What are these events? Remarkably, we capture a variety of corporate events and disclosures. Earning announcements constitute about half of the events for Apple. Other events include Apple Keynotes (presentations that Apple gives to the press, often presenting new products), or CEO letters addressed to investors. Table 4 lists a few selected events for Apple.

6.2 Abnormal stock returns

How do stock returns behave around events? Similar to the relative message volume changes, we use as benchmark model a one-year rolling window regression of the daily returns of ticker i on the daily market returns, R_t^M, i.e., daily excess returns of the S&P500, and lagged returns as control variables,

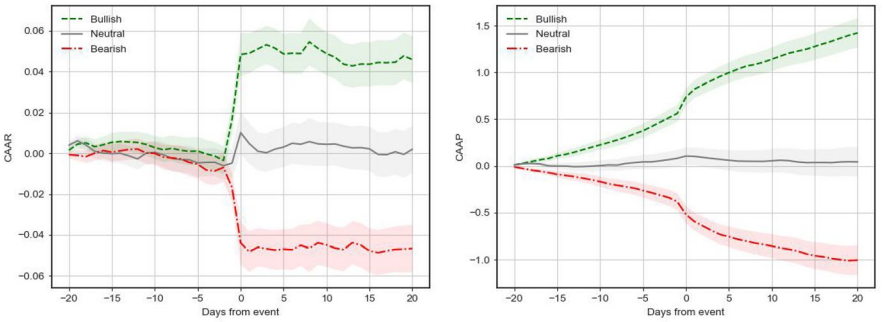


Fig. 14 Cumulative average abnormal returns (left plot) and cumulative average abnormal polarity (right plot) around identified events. CAAR and CAAP related to bearish, neutral and bullish events are displayed with the red, gray and green line, respectively. Areas around lines show confidence intervals at the 95% level

$$R_{i,t} = \alpha_i^R + \beta_i^R \cdot R_t^M + \beta_i^{(1)} \cdot R_{i,t-1} + \beta_i^{(2)} \cdot R_{i,t-2} + \beta_i^{(3)} \cdot R_{i,t-3} + \epsilon_{i,t}, \tag{13}$$

which gives the one-year rolling estimates $\hat{\alpha}_i^R$, $\hat{\beta}_i^R$ and $\hat{\beta}_i^{(j)}$. This implies the abnormal returns

$$AR_{i,t} = R_{i,t} - \left(\hat{\alpha_i^R} + \hat{\beta_i^R} \cdot R_t^M + \hat{\beta_i^{(1)}} \cdot R_{i,t-1} + \hat{\beta_i^{(2)}} \cdot R_{i,t-2} + \hat{\beta_i^{(3)}} \cdot R_{i,t-3}\right). \tag{14}$$

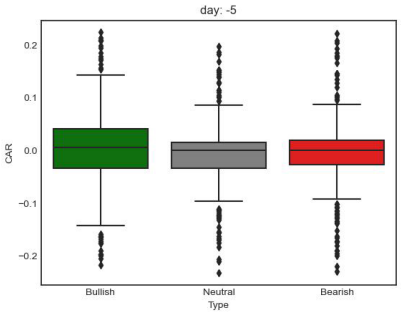
We define the cumulative abnormal returns (CAR) around a ticker i event τ as

$$CAR_i(\tau, t) = \sum_{s=-20}^{t} AR_{i,\tau+s}, \tag{15}$$

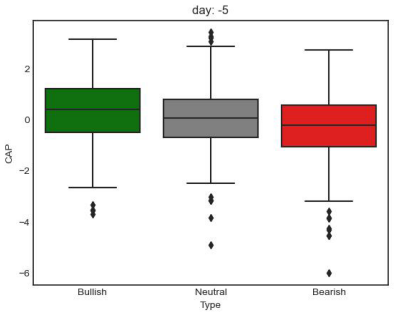
and the cumulative average abnormal returns (CAAR) across all $N = 1131$ events as

$$CAAR(t) = \frac{1}{N} \sum_{j=1}^{N} CAR_{i_j}(\tau_j, t). \tag{16}$$

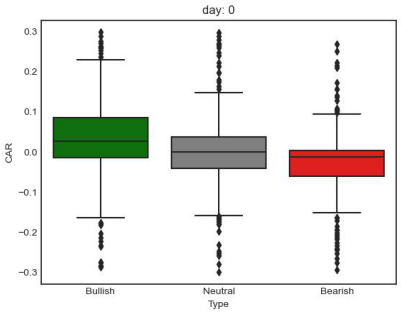
Left plot of Fig. 14 shows the CAAR around the events. This plot is consistent with MacKinlay (1997). It shows that CAAR related to bearish (bullish) events exhibits a significant downward (upward) jump at the event date, respectively. These jumps are followed by a flat CAAR during the 20 days after the event. Interestingly, there is a systematic shift in the CAAR already 1 day before the event. However, this shift is relatively small compared to the jump on the event day: one day before the event, the bullish (bearish) CAAR equals 0.020 (− 0.021). The CAAR related to the neutral events exhibits a slight upward shift around the event date but it fades away after a few days. The CAAR related to bearish events shifts already a few days before the event but this shift is not statistically significant. This is in line with Fig. 15, which shows that the CAR distributions prior to the events are not significantly different from zero, which is confirmed by the Mann–Whitney U-tests shown in Table 5. CAR has no predictive power on the type of the event: five days before an event,



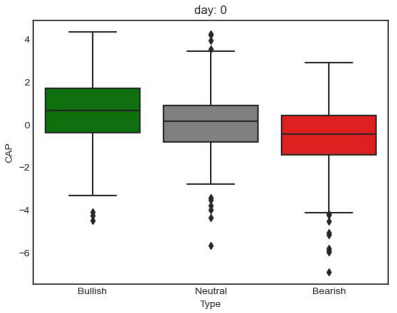
CAR 5 days before the event



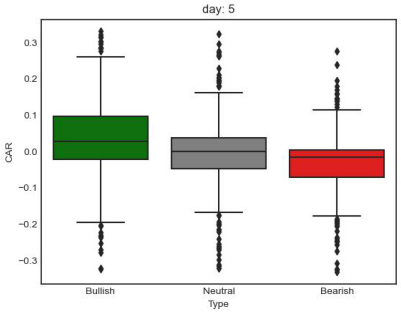
CAP 5 days before the event



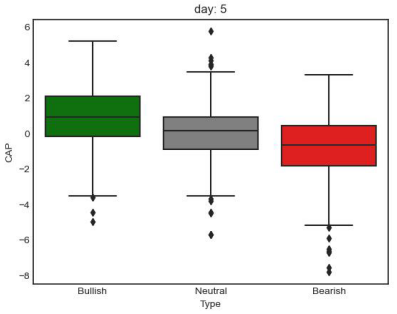
CAR on the event date



CAP on the event date



CAR 5 days after the event



CAP 5 days after the event

Fig. 15 Distributions of CAP and CAR 5 days before an event, on event date and 5 days after an event. The line inside a box shows the median while the edges of each box represent the 25% and 75% quantile of the distribution. From above the edges of a box, a distance of 1.5 times the interquartile range is measured and a whisker is drawn up to the largest and lowest observed point from the data that falls within this distance. Interquartile range is equal to the third quartile minus the first quartile

the median of the CAR distribution of the bullish events is not statistically different from the median of the neutral events. Same holds for the bearish events.

Table 5 Mann–Whitney *U*-test statistics for pairwise significant differences between distribution medians

	Alternative hypothesis	U	Z	n_1	n_2
CAR					
$\tau - 5$	H_1: $\theta_{\text{bullish}} > \theta_{\text{neutral}}$	62,628	− 1.17	452	292
	$H_1 : \theta_{\text{neutral}} > \theta_{\text{bearish}}$	54,140	− 1.73	292	380
τ	$H_1 : \theta_{\text{bullish}} > \theta_{\text{neutral}}$	71,949	2.18**	452	292
	$H_1 : \theta_{\text{neutral}} > \theta_{\text{bearish}}$	62,111	2.65***	292	380
$\tau + 5$	$H_1 : \theta_{\text{bullish}} > \theta_{\text{neutral}}$	69,631	2.13**	452	292
	$H_1 : \theta_{\text{neutral}} > \theta_{\text{bearish}}$	62,411	2.77***	292	380
CAP					
$\tau - 5$	$H_1 : \theta_{\text{bullish}} > \theta_{\text{neutral}}$	55,408	3.70***	452	292
	$H_1 : \theta_{\text{neutral}} > \theta_{\text{bearish}}$	47,998	3.00***	292	380
τ	$H_1 : \theta_{\text{bullish}} > \theta_{\text{neutral}}$	49,385	8.98***	452	292
	$H_1 : \theta_{\text{neutral}} > \theta_{\text{bearish}}$	42,101	5.36***	292	380
$\tau + 5$	$H_1 : \theta_{\text{bullish}} > \theta_{\text{neutral}}$	44,364	7.55***	452	292
	$H_1 : \theta_{\text{neutral}} > \theta_{\text{bearish}}$	40,515	6.00***	292	380

Under the null hypothesis, the two samples represent two distributions with equal median values. Statistical significance at the 99%, 95%, and 90% level is indicated with ***, **, *, respectively

6.3 Abnormal polarity

How does sentiment polarity behave around events? Similar to the above, we use as benchmark model a one-year rolling window regression of the the daily polarity of ticker i on the daily market polarity defined in (6),

$$P_{i,t} = \alpha_i^P + \beta_i^P \cdot P_t^M + \epsilon_{i,t}, \tag{17}$$

which gives the one-year rolling estimates $\hat{\alpha}_i^P$ and $\hat{\beta}_i^P$. This implies the abnormal polarity

$$AP_{i,t} = P_{i,t} - \left(\hat{\alpha}_i{}^P + \hat{\beta}_i{}^P \cdot P_t^M\right). \tag{18}$$

We define the cumulative abnormal polarity (CAP) around a ticker i event τ as

$$CAP_i(\tau, t) = \sum_{s=-20}^{t} AP_{i,\tau+s}, \tag{19}$$

and the cumulative average abnormal polarities (CAAP) across all $N = 1131$ events as

$$CAAP(t) = \frac{1}{N} \sum_{j=1}^{N} CAP_{i_j}(\tau_j, t). \tag{20}$$

Right plot of Fig. 14 shows the CAAP around the events. There are two main findings. First, in contrast to CAAR, the CAAP for bullish and bearish events is not constant after the event date, suggesting that users' sentiments about stocks tend to be biased towards recent past events. A possible explanation is that users might still post bullish (bearish) messages about a bullish (bearish) event during several days after the event. This is in contrast to the returns that immediately normalize after the event. Second, and more interestingly, the CAAP for bullish and bearish events shifts several days earlier than the CAAR. This indicates that investors are on average able to anticipate the type of an event in the near future. However, this sentiment only manifests through the social media, but not through abnormal returns.

Figure 15 illustrates this striking finding with box plots [see Dekking et al. (2005) and Tukey (1977)] showing the distributions of the CAR and CAP, for all three event types, 5 days before the event, at the event date, and 5 days after the event, respectively.

To check statistical significance, we use the Mann–Whitney *U*-test [see Mann and Whitney (1947) and Sheskin (1998)] to test whether the three samples (bullish, neutral and bearish) represent populations with different median values.[14] Table 5 shows U-test statistics for pairwise comparisons. The null is rejected in every case except for CAR at $\tau - 5$. That is, 5 days before the event, CAR has no predictive power on the type of event. This is consistent with the EMH. In contrast, 5 days before the event, CAP can predict the type of event. At the event date, the medians of the CAR shift as the abnormal returns jump for both bullish and bearish events. Also this is consistent with the EMH. Finally, 5 days after the event, the distributions of the CAR are very similar to the ones at the event date. Again, this is consistent with the EMH, as all new information is instantaneously embedded into the prices and the returns normalize after the event, immediately. The medians of the CAP 5 days after the event exhibit an extended shift compared to the ones at the event date, as investors continue to post about recent past events.

7 Sentiment-sorted portfolios

We assess the economic relevance of the sentiment polarity and construct sorted portfolios. Thereto, we define for every ticker i and day t

$$CAP_{i,t} = \sum_{s=t-14}^{t} AP_{i,s}, \tag{21}$$

[14] This interpretation only holds under stringent assumptions on the populations, namely that the two population distributions are equal up to a shift. Under the null hypothesis, the three samples represent distributions with equal medians. Let θ_i be the median of the distribution i. Formally, we test $H_0 : \theta_{\text{bullish}} = \theta_{\text{neutral}}$ against $H_1 : \theta_{\text{bullish}} > \theta_{\text{neutral}}$ and $H_0 : \theta_{\text{neutral}} = \theta_{\text{bearish}}$ against $H_1 : \theta_{\text{neutral}} > \theta_{\text{bearish}}$ 5 days before an event, on event date and 5 days after an event. We define U as the Mann–Whitney test statistic, Z as the normal approximation of the Mann–Whitney test statistic for large sample sizes, n_1 and n_2 as the sample sizes. We refer to Sheskin (1998) for the test statistic computation.

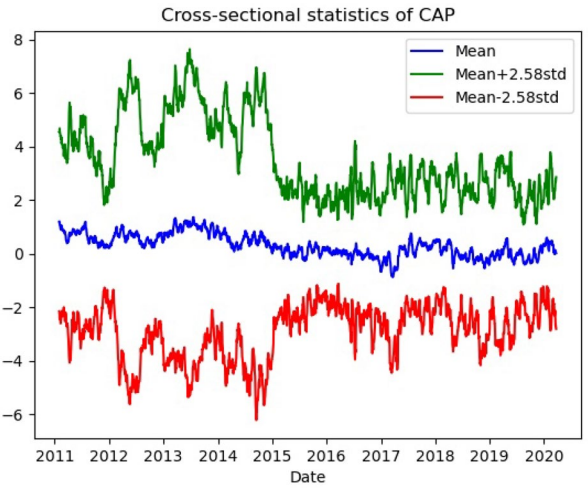


Fig. 16 Cross-sectional statistics of $CAP^{(R)}_{i,t}$

which is the running CAP over the last 14 days plus the current day t (we rebalance the portfolio at the close on day t).[15] Note the difference to (19). While we cannot predict the arrival of an event, we assume that the more $CAP_{i,t}$ deviates from zero the more likely there will be a an event on the next day. We will thus use $CAP_{i,t}$ as a baseline signal for market timing. However, as we have seen above, CAP continues to shift after an event. To avoid exposures to short-term reversals, we thus reset the running CAP after every event. Formally, let $\tau_{i,t} \leq t$ denote the most recent past event date by t of ticker i. Then we define the reset CAP

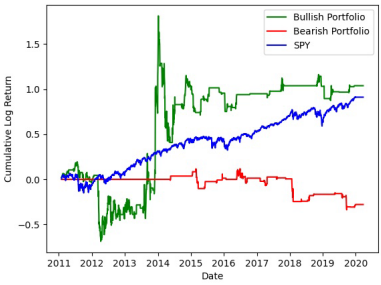
$$CAP^{(R)}_{i,t} = \sum_{s=\max\{t-14,\tau_{i,t}+1\}}^{t} AP_{i,s} = \begin{cases} CAP_{i,t}, & \text{if } \tau_{i,t} < t-14, \\ \sum_{s=\tau_{i,t}+1}^{t} AP_{i,s}, & \text{if } t-14 \leq \tau_{i,t} < t, \\ 0, & \text{if } \tau_{i,t} = t, \end{cases} \tag{22}$$

where we used the convention that $\sum_{s=t+1}^{t} \cdot = 0$.

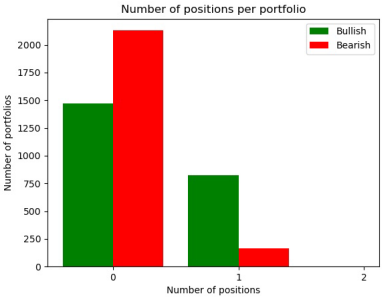
We also define time-varying thresholds on the reset CAP for market timing. For every day t, we estimate the mean μ_t and standard deviation σ_t of $CAP^{(R)}_{i,t}$ across the 19 tickers i. For a fixed multiplier x, we define $U_t(x) = \mu_t + x \cdot \sigma_t$ the upper threshold, and $L_t(x) = \mu_t - x \cdot \sigma_t$ the lower threshold. Figure 16 shows the time series of the cross-sectional mean μ_t and the 99% confidence interval, $L_t(x)$ and $U_t(x)$ for $x = 2.58$. As a robustness check of our approach, we observe that the mean is well centered at zero. We also see a regime change in the early 2015. In the first regime the standard deviation is much larger (and more volatile) than in the second regime.[16] "Sentiment-sorted portfolios for various thresholds" of appendix contains the results for the 95% ($x = 1.96$) and 99.5% ($x = 2.81$) confidence intervals.

[15] As above, we work here with the the restricted sample of 19 tickers and dates t ranging through all business days of the sample period, excluding the first 14 days (for the CAP) and the last day (for the last portfolio holding period).

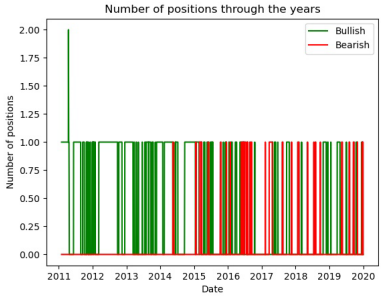
[16] We could not find an exogenous cause for this regime change.



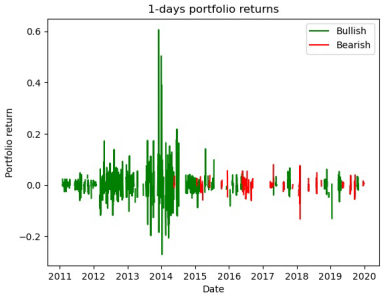
Cumulative log returns



Distribution of number of positions



Number of positions over time



Daily returns

Fig. 17 Bullish and bearish portfolios for $x = 2.58$

Based on these signals, we now construct reset CAP-sorted portfolios. Formally, we define the ticker sets $I_t^{\text{bull}} = \{i \mid CAP_{i,t}^{(R)} > U_t(x)\}$ and $I_t^{\text{bear}} = \{i \mid CAP_{i,t}^{(R)} < L_t(x)\}$. At the close of any day t, we form the equally weighted bullish (bearish) portfolio consisting of tickers in I_t^{bull} (in I_t^{bear}), and realize the 1-day returns. If any of the index sets is empty, we set the corresponding return to zero. This strategy is out-of-sample and easily implementable in practice. At the end of day t and for every stock, we extract and classify all messages to compute the polarity using (5), the abnormal polarity using (18) and the running reset CAP using (22). Next, we compute the lower $L_t(x)$ and upper $U_t(x)$ thresholds by estimating the cross-sectional mean and standard deviation of the reset CAP. Finally, we short sell all tickers where the reset CAP is smaller than the lower threshold, and buy long all tickers where the reset CAP is larger than the upper threshold.

Top-left plot of Fig. 17 shows the cumulative log returns of bullish and bearish portfolios as well as the S&P500. Overall, the portfolio performance is consistent with our approach: the bullish (bearish) portfolio outperforms (under-performs) the market. Remarkably, the upward (downward) steps suggests that our portfolio strategy succeeds to take the right positions just before an event. The remaining plots of Fig. 17 show the number of positions across time of our portfolios. Most of the

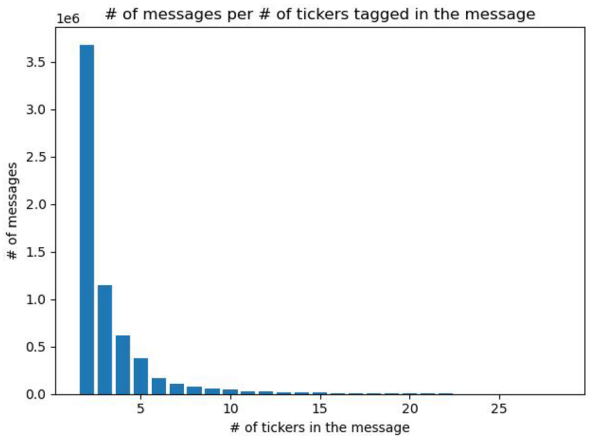


Fig. 18 Histogram of the number of tickers per message, across all messages referring to more than one ticker

returns are earned with portfolios consisting of very few tickers. This is a result of our market timing and stock picking strategy: we only invest in the top/bottom percentiles of CAP, whenever our signal is strong enough.

The portfolio performance arguably depends on the choice of the multiplier x. In particular, the smaller x the more likely the bearish portfolio exhibits positive returns. On the other hand, the larger x the more likely the bullish portfolio misses the opportunities of positive returns. A careful gauging of x, possibly asymmetric in bullish and bearish, is therefore required for a real-world implementation of these strategies. Results could also improve for a larger cross-section of stocks than the 19 of our reduced sample, which is left for future research.

8 Conclusion

We extract 90 million messages from StockTwits from January 2010 to March 2020, covering a large sample of US and Canadian stocks. Messages are either user-labeled as bullish or bearish or left unlabeled. Using the user-labeled messages as training set, we perform logistic regressions on TFIDF vectorized messages to classify all unlabeled messages as either bullish, neutral or bearish. We observe a 5-to-1 bullish-to-bearish ratio, indicating that investors are on average optimistic. We build time-series of daily sentiment polarity for individual tickers and the aggregate market. We show that daily polarity is positively associated to contemporaneous stock returns, but this result loses its significance against next-day returns. We then define events as days with sudden peak of message volume and relate them to corporate and stock market events. We show that cumulative abnormal polarity has significant predictive power on the type of event, in contract to cumulative abnormal returns. We also note that investor sentiment about a ticker tends to be biased towards recent past events. As robustness check, we show that our event study on cumulative abnormal returns is consistent with previous literature on the efficient market hypothesis.

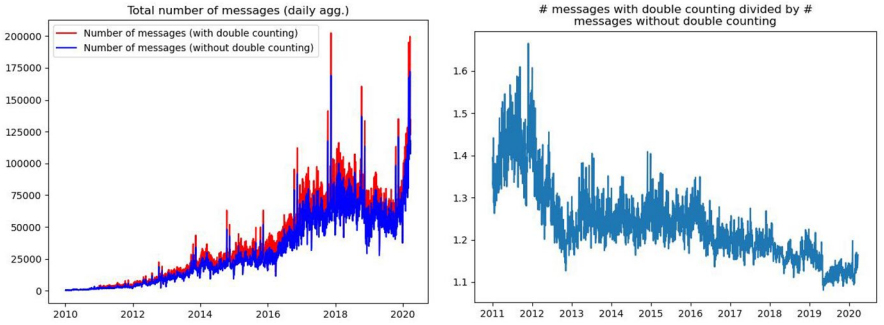


Fig. 19 Left plot shows the total number of messages with double counting (red) compared to the total number of messages without double counting (blue). Right plot shows the ratio between the number of messages with double counting and the number of messages without double counting. Numbers are aggregated daily (color figure online)

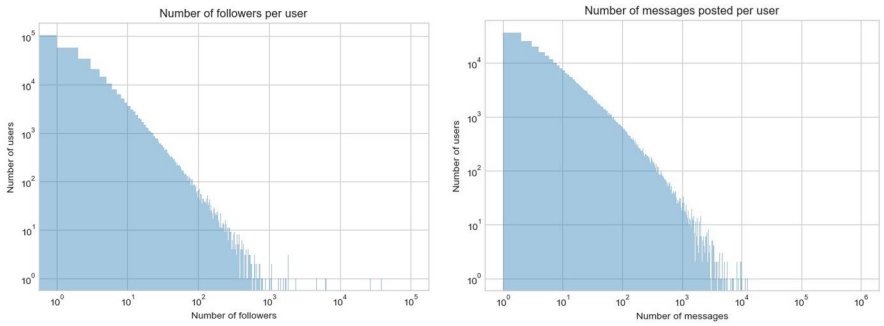


Fig. 20 User summary statistics. Left graph is a log–log histogram of the number of followers per user and the right graph shows the log–log histogram of the number of messages posted by users

The performance of sentiment-sorted portfolios illustrates the economic relevance of our sentiment measure.

Appendix

The appendix contains additional summary statistics, robustness checks, and auxiliary results.

Tutorial for StockTwits messages extraction

We use stock price data from CRSP/Compustat of all US and Canadian listed stocks from January 2010 to March 2020. From this dataset, we create the list of unique tickers for which we will extract messages. We will later be able to merge the two datasets using the date and ticker for every observation. We use the StockTwits

Table 6 Confusion matrix for the combined classifier out-of-sample

	True	
	Bullish	Bearish
Predicted		
Bullish	3,555,896	155,280
Neutral	663,541	160,423
Bearish	550,928	746,261

Application Programming Interface (API) to download messages from StockTwits. One query on StockTwits API is called a JavaScript Object Notation (JSON) request. Every message on StockTwits has a unique identifier ("msg_id") posted by a user with a unique identifier ("user_id"). JSON requests allow to query the database by ticker (called "symbol method") or by user (called "user method"). We use the query by ticker. One query only outputs the latest 30 messages concerning that ticker. However, it is possible to set a parameter ("max") to output the latest 30 messages up to this particular message identifier. This parameter allows us to crawl the message history of a ticker by recursively changing the "max" parameter to the oldest message identifier in the query. To perform a JSON request for Apple (AAPL) up to the message identifier 30,000,000, simply enter the following URL in a browser: https://api.stocktwits.com/api/2/streams/symbol/AAPL.json? &max=30000000. The page we get looks unreadable but it has always the same structure : several pairs of keys and values. The structure of JSON can easily be interpreted by modern programming languages. We create a Python script to query the API and extract the message history of every ticker in the ticker list. We store the output of every JSON request in .txt files in dedicated ticker folders.

Message count

A StockTwits message can refer to multiple tickers. Figure 18 shows the histogram of the number of tickers tagged per message. As the vast majority of message includes only one ticker, we only show on this plot messages referring to more than one ticker. The maximum number of tickers per message amounts to 28 and corresponds to 11 messages in the sample. Many messages refer to several tickers and this creates duplicates in the database because we consider the same message for all tickers tagged in the message.

Left plot of Fig. 19 shows the number of messages with and without double counting. In our sample, the number of messages without double counting is 76 million, as opposed to 90 million messages with double counting. Right plot of Fig. 19 shows the ratio between the number of messages with double counting and the number of messages without double counting. Throughout this paper, we only refer to the number of messages with double counting.

Table 7 Confusion matrix for the combined classifier in-sample

	True	
	Bullish	Bearish
Predicted		
Bullish	14,433,375	532,509
Neutral	2,656,730	642,329
Bearish	1,992,535	3,071,837

User summary statistics

Left plot of Fig. 20 shows a log–log histogram of the number of followers per user and a right plot shows a log–log histogram of the number of messages posted by users. There are a few users with many followers (they can be seen as "influencers"), and many users with a few followers. In addition, most users seem to post on average between 10 and 400 messages and a few post a lot more. Overall, this appears to be a well balanced network structure. A more detailed study of the network effects on market sentiment is beyond the scope of this paper.

Anomalies

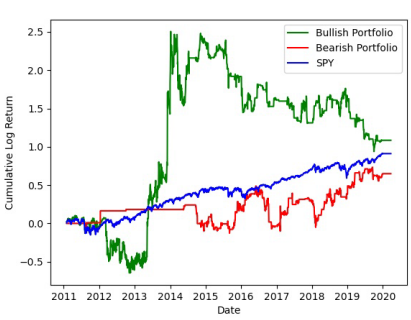
We discuss here two anomalies that appear in the word clouds in Fig. 4.

The term "aldox" in the bullish cloud caught our attention. After some research, it is an abbreviation for Aldoxorubicin, a drug against tumors and is associated with pharmaceutical messages where investors were very enthusiastic about it. An example of a related message is "aldox is on the slide. have great faith this is truly world change". That is why the term is appearing almost exclusively in bullish messages, hence in the bullish cloud.

The bearish cloud contains the term "long position open", which seems like a bullish signal. Closer inspection shows that this term frequently appears in bearish user-labeled messages of intraday alerts such as "sell $labd close labd long position. open labd short position. time: 14:53 ny price: $13.64 zquant intraday alerts". However, this anomaly is not an issue. We tested what happened when "long position open" is fed as a message into our sentiment classifier. As a message, it consists of the trigram "long position open", the two bigrams "long position" and "position open", and the single words as unigrams. This results in a bullish score of 0.91 and the message is—correctly—classified as bullish.

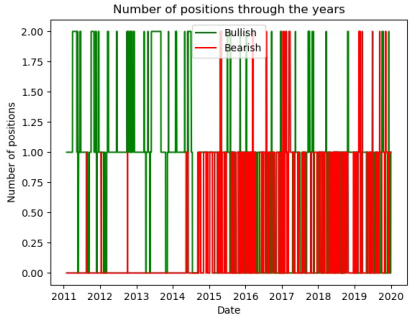
Classifier performance

We first recap the definition of the basic performance measures for a binary classifier. First, one has to choose one class as the positive class. Instances (messages) are then divided according to their predicted and actual labels into true positives TP (predicted positive, actual positive), false positives FP (predicted positive, actual

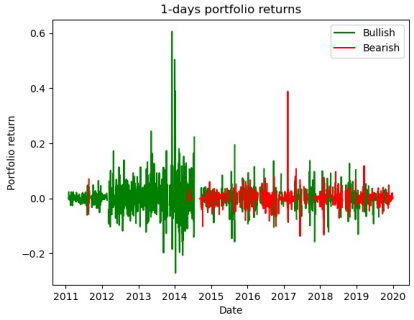


Cumulative log returns

Distribution of number of positions



Number of positions over time



Daily returns

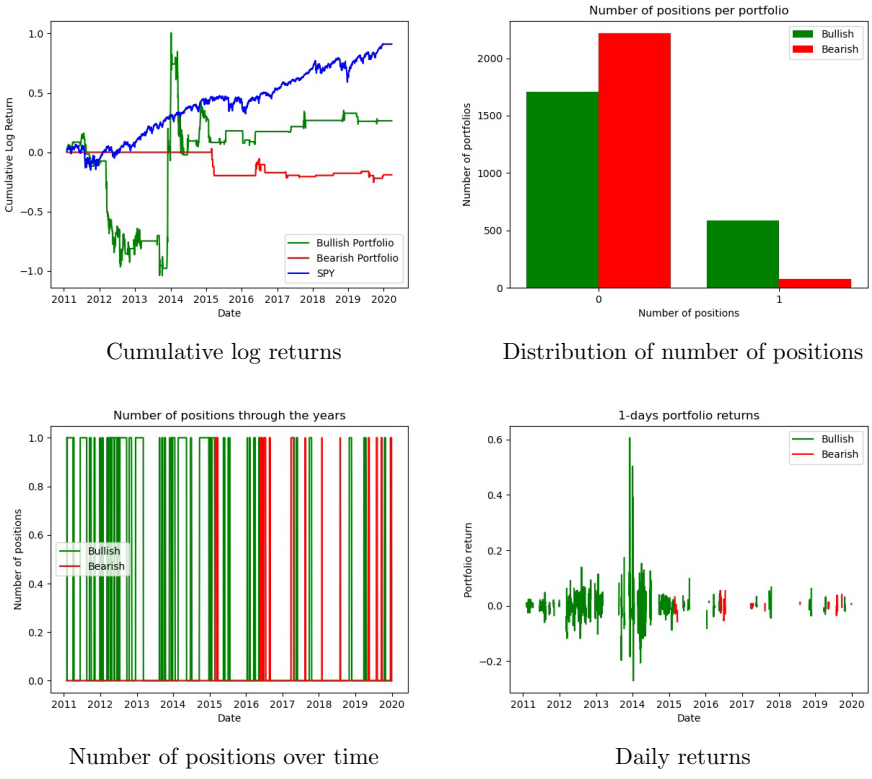
Fig. 21 Bullish and bearish portfolios for $x = 1.96$

negative), true negatives TN (predicted negative, actual negative), and false negatives FN (predicted negative, actual positive). Precision $PRE = \frac{TP}{TP + FP}$ is the proportion of true positives among the predicted positives. Recall $REC = \frac{TP}{TP + FN}$ is the proportion of true positives among the actual positives. The precision-recall trade-off is captured by the F1 score, $\frac{2 \cdot PRE \cdot REC}{PRE + REC}$, the harmonic mean of precision and recall.

Tables 6 and 7 show the confusion matrices of our combined classifier out-of-sample and in-sample, respectively. We define accuracy as the fraction of correct predictions, omitting the messages with a predicted neutral sentiment. We thus obtain an out-of-sample accuracy of 85.9%. The in-sample accuracy is 87.4%.

Examples of classified messages

Here are some representative examples of classifications. Typical messages classified as bullish contain terms such as "buy buy" or "hope the pump come soon". Whereas typical bearish messages contain terms such as "sell everything" or "start



Cumulative log returns | Distribution of number of positions

Number of positions over time | Daily returns

Fig. 22 Bullish and bearish portfolios for $x = 2.81$

short position here". Neutral messages are either empty, or irrelevant to finance (e.g., "political posturing friend"[17]), or ambiguous (e.g., "lol wow").

Sentiment-sorted portfolios for various thresholds

As the thresholds $U_t(x)$ and $L_t(x)$ are functions of the hyperparameter x, we provide for robustness check the results of our CAP-sorted portfolios for the values $x = 1.96$ (95% confidence band) in Fig. 21, and for $x = 2.81$ (99.5% confidence band) in Fig. 22.

Acknowledgements We thank Pierre Collin-Dufresne, Damien Challet, François Degeorge, Rüdiger Fahlenbrach, Thomas Renault, Michael Rockinger, and seminar and conference participants at the Applied Machine Learning Days, SIAM Financial Engineering Conference, International FinTech, InsurTech & Blockchain Forum, Swissquote Conference, SFI Research Days and SSES Annual Congress 2022, and an anonymous referee for helpful comments. No funds, grants, or other support was received.

[17] This is a reply to the message "honestly, how dumb can you be to believe that china was going to buy significant amount of agricultural products after the breakdown in trade talks. Even if they buy it will be just a little bit and not significant"

Funding Open access funding provided by EPFL Lausanne.

Data availability Data, codes and Python scripts needed to reproduce the results for this paper are available on https://github.com/marcaureledivernois/tweet_sklearn.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

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

Boudoukh, J., Feldman, R., Kogan, S., & Richardson, M. (2013). Which news moves stock prices? A textual analysis. Working Paper 18725, National Bureau of Economic Research, January 2013. http://www.nber.org/papers/w18725

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter, 6*(1), 1–6.

Chen, E., Lu, Z., Xu, H., Cao, L., Zhang, Y., & Fan, J. (2020). A large scale speech sentiment corpus. In *Proceedings of the 12th language resources and evaluation conference* (pp. 6549–6555).

Cookson, J. A., & Niessner, M. (2020). Why don't we agree? Evidence from a social network of investors. *The Journal of Finance, 75*(1), 173–228.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science, 53*(9), 1375–1388. http://www.jstor.org/stable/20122297.

Dekking, F., Kraaikamp, C., Lopuhaa, H., & Meester, L. (2005). *A modern introduction to probability and statistics*. Springer.

Erdemlioglu, D., Gillet, R. L., & Renault, T. (2017). *Market reaction to news and investor attention in real time*. Available at SSRN 3010847.

Fama, E. F. (1991). Efficient capital markets. *The Journal of Finance, 46*(5), 1575–1617.

Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The adjustment of stock prices to new information. *International Economic Review, 10*(1), 1–21.

Ghoshal, S., & Roberts, S. (2016). Extracting predictive information from heterogeneous data streams using Gaussian processes. *Algorithmic Finance, 5*(1–2), 21–30.

Heston, S. L., & Sinha, N. R. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal, 73*(3), 67–83. https://doi.org/10.2469/faj.v73.n3.3

Ke, Z., Kelly, B. T., & Xiu, D. (2020). *Predicting returns with text data*. Available at SSRN: https://ssrn.com/abstract=3389884

Loughran, T., & McDonald, B. (2011). Barron's red flags: Do they actually work? *Journal of Behavioral Finance, 12*(2), 90–97.

Loughran, T., & McDonald, B. (2012). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35–65.

MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of Economic Literature, 35*(1), 13–39.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics, 18*(1), 50–60.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Qasem, M., Thulasiram, R., & Thulasiram, P. (2015). Twitter sentiment classification using machine learning techniques for stock markets. In *International conference on advances in computing, communications and informatics*.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. *PLoS ONE, 10*(9), e0138441.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the US stock market. *Journal of Banking and Finance, 84*, 25–40.

Renault, T. (2020). Sentiment analysis and machine learning in finance: A comparison of methods and models on one million messages. *Digital Finance, 2*(1), 1–13.

Saif, H., Fernández, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter. In *9th international conference on language resources and evaluation* (pp. 810–817).

Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research, 9*, 371–421.

Sheskin, D. J. (1998). *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management, 20*(5), 926–957.

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

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Addison-Wesley.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Conference on empirical methods in natural language processing* (pp. 347–354).

Yildirim, S., Jothimani, D., Kavaklioglu, C., & Basar, A. (2018). Classification of hot news for financial forecast using NLP techniques. In *IEEE international conference on big data*.