A User-Centric Analysis of Social Media for Stock Market Prediction

MOHAMED REDA BOUADJENEK*, Deakin University, Australia
SCOTT SANNER, University of Toronto, Canada
JAMIE WISE, Periscope Capital, Canada
GA WU, Twitter, Canada

Social media platforms such as Twitter or StockTwits are widely used for sharing stock market opinions between investors, traders, and entrepreneurs. Empirically, previous work has shown that the content posted on these social media platforms can be leveraged to predict various aspects of stock market performance. Nonetheless, actors on these social media platforms may not always have altruistic motivations and may instead seek to influence stock trading behavior through the (potentially misleading) information they post. While a lot of previous work has sought to analyze how social media can be used to predict the stock market, there remain many questions regarding the quality of the predictions and the behavior of active users on these platforms. To this end, this paper seeks to address a number of open research questions: Which social media platform is more predictive of stock performance? What posted content is actually predictive, and over what time horizon? How does stock market posting behavior vary among different users? Are all users trustworthy, or do some user's predictions consistently mislead about the true stock movement? To answer these questions, we have analyzed two datasets from Twitter and StockTwits covering almost 5 years of posted messages spanning 2015 to 2019. The results of this large-scale study provide a number of important insights among which: (i) StockTwits is a more predictive source of information than Twitter, leading us to focus our analysis on StockTwits; (ii) on StockTwits, users' self-labeled sentiments are correlated with the stock market but are only slightly predictive in aggregate over the short-term; (iii) there are at least three clear types of temporal predictive behavior for users over a 144 days horizon: short-, medium-, and long-term; and (iv) consistently incorrect users who are reliably wrong tend to exhibit what we conjecture to be "bot-like" post content and their removal from the data tends to improve stock market predictions from self-labeled content.

CCS Concepts: • **Information systems → Social networks**; • **Human-centered computing → Social network analysis**.

Additional Key Words and Phrases: Social Network Analysis, Stock Market Prediction, Classification.

ACM Reference Format:
Mohamed Reda Bouadjenek, Scott Sanner, Jamie Wise, and Ga Wu. 2022. A User-Centric Analysis of Social Media for Stock Market Prediction. *ACM Trans. Web* 16, 4, Article 1 (December 2022), 22 pages. https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

*This work has been primarily completed while the author was at the Periscope Capital, Toronto, Canada.

Authors' addresses: Mohamed Reda Bouadjenek, reda.bouadjenek@deakin.edu.au, Deakin University, School of Information Technology, 75 Pigdons Rd, Geelong, Victoria, Australia, 3216; Scott Sanner, ssanner@mie.utoronto.ca, University of Toronto, Department of Mechanical and Industrial Engineering, 5 King's College Rd, Toronto, Ontario, Canada, M5S 3G8; Jamie Wise, jwise@periscopecap.com, Periscope Capital, 333 Bay St. Suite 1240, Toronto, Ontario, Canada, M5H 2R2; Ga Wu, gaw@twitter.com, Twitter, 901 King St W, Toronto, Ontario, Canada, M5V 3H5.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Association for Computing Machinery.
1559-1131/2022/12-ART1 $15.00
https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

1 INTRODUCTION

The emergence of social media platforms in the mid-2000s and the expansive growth of a large active user population has enabled users to freely express their opinions [3] and thus modeling and predicting various events such as natural disasters [15, 16, 46, 62], election outcomes [19, 27], traffic flow [54], public health [36, 65], public event attendance [29], and even the stock market [9].

Long before the advent of social media, stock market price movement prediction has been an active field of research for investment purposes. The stakes certainly cannot be ignored. In 2019 alone, USD 1,452 billion of assets were traded monthly on the New York Stock Exchange only — a number that has doubled in the past 10 years[1]. The use of a variety of information sources to predict the stock market is supported by the *"efficient market hypothesis"*, which posits that investors act on all information available to move the stock price to its true valuation [6]. Broadly construed, this idea has motivated a wide variety of previous research aiming to use social media to predict the stock market [5, 9, 22, 43, 49, 50, 52, 53, 55, 58, 61].

While previous work has focused on stock market prediction itself, we argue that less work has analyzed more nuanced aspects of social media stock market discourse such as the temporal horizon of a user's predictions and associated post content, nor has much work analyzed apparent "consistently incorrect" prediction behavior of users and their characteristics. To this end, this work aims to analyze these user-centric aspects of stock market prediction from social media by asking the following progression of research questions:

- **RQ1**: Are users' self-labeled[2] stock movement sentiments sufficient for accurate stock market prediction? Here, our goal is not to develop a novel methodology but instead to perform a data science measurement analysis to understand the accuracy of users' self-labeled predictions.
- **RQ2**: Is other post content (e.g., word and emoji usage) more predictive than the self-labels? Can machine learning uncover this? Similar to RQ1, we are not concerned with building a state-of-the-art machine learning model for predicting the stock market from social media, but rather simply to understand what user content is predictive of the stock market.
- **RQ3**: Are different users better for predicting stock movements at different time horizons? If yes, is there particular content in posts that earmarks them for different predictive horizons?
- **RQ4**: Due to the massive financial stakes involved, are there consistently incorrect actors who seem to (intentionally) mislead? Can we identify "consistently correct" and "consistently incorrect" users as well as distinguishing characteristics in their posts?
- **RQ5**: Does restricting data to the subset of trustworthy users allow us to make more accurate predictions from self-labeled sentiments?

To answer these questions, we analyze social media datasets from Twitter and StockTwits covering almost 5 years of posted messages involving *cashtags* (i.e., ticker symbols preceded by a $) for stocks spanning 2015 to 2019. Then we build different classifiers to identify upward and downward stock price movements over different time horizons to predict movements up to 144 days (over 4 months) in the future.

We first empirically demonstrate that self-labeled stock sentiment features alone may be correlated with stock price movements — but only weakly, and with a short predictive time horizon. We next train machine learning methods to predict stock price movements on historical data

[1] https://finance.yahoo.com/

[2] We use the official terminology of StockTwits "self-labeled" sentiment, which refers to a feature that allows users to self-label their posted messages using a social sentiment indicator. This is achieved via a simple toggle on the StockTwits message box that allows users to communicate their view, **Bullish** or **Bearish**, on any specific asset or the market as a whole. For example, the post "*$TSLA Why would anyone buy today knowing tomorrow could be the biggest drop, in the market this year?*" (https://stocktwits.com/Stinger__/message/465939805) is labeled as **Bearish** towards *Tesla* stock by its author.

leveraging all content of a social media post. We evaluate this predictor on test data not used during training. We observe a significant boost in prediction accuracy, indicating that there is useful latent information in tweets beyond users' self-labeled stock movement sentiments. We further observe that predictors trained and evaluated on StockTwits data perform better than those for Twitter data. Based on these observations, we primarily focus on StockTwits for the remainder of our analysis.

We next analyze users in terms of the temporal horizon of their predictiveness. Here we clearly identify at least three types of temporal predictive horizon over 144 days: short-term (< 20 days), medium-term (60-100 days), and long-term (> 100 days). In general, we observe that the most reliable predictions are in the short-term and furthermore that the post content most strongly associated with different temporal prediction horizons changes with the predictive horizon.

Last but not least, we perform an analysis to identify the most "consistently correct" and alternately the most "consistently incorrect" users, identified respectively by the agreement and disagreement of their self-labeled predictions of stock movements and the actual stock market movement. We observe that the most "consistently correct" users and the most "consistently incorrect" users are distinguished by very different social media post content, which we conjecture may indicate that the most "consistently incorrect" users could be bots. By removing the most "consistently incorrect" users identified from training data, we demonstrate improvements in stock prediction performance from self-labeled content of the remaining users on held-out test data.

In summary, this article provides a *novel user-centric behavioral analysis* of stock market predictions on social media that unveils a rich and varied mixture of user predictiveness, temporal horizon accuracy, and trustworthiness. However, it is important to reiterate that our intent here is not to prescribe novel technical methodologies for predicting the stock market, but rather simply to describe user behavior and the predictiveness of user content in social media discourse on the stock market. Nonetheless, this descriptive analysis reveals a complex ecosystem of user behaviors that we suggest should be carefully considered when prescriptively leveraging social media for stock market forecasting.

The rest of this paper is organized as follows: Section 2 covers related work and puts our work in perspective. Section 3 describes our datasets, our crawling methodology, and the set of stock tickers – more precisely, exchange traded funds (ETFs) – used for this analysis. We conduct basic analysis on our datasets in Section 3. In Section 4 we describe the general methodology we use for learning stock prediction classifiers. In Section 5 we discuss the results of our user-centric analysis for stock market prediction. Finally, in Section 6 we conclude and suggest future research directions.

2 RELATED WORK

Stock market prediction has a long and rich history of research development due to the potential financial gains involved. Historical works were often based on the efficient-market hypothesis [6] that assumes investors act on all information available to move the stock price to its true valuation. In a separate vein of thinking, random walk theory [33, 34] suggests that changes in stock prices happen because of unpredictable news and other events, thus exhibiting random walk characteristics.

From a prediction perspective, many specialized techniques have been developed and are mainly divided into two categories: fundamental analysis and technical analysis (charting). The former approach studies a company's past performance as well as the credibility of its accounts [1, 37], whereas the second approach seeks to determine the future price of a stock based solely on the trends of the past price [2, 21, 66]. However, with the emergence of social media platforms and motivated by the efficient-market hypothesis, researchers have started to investigate social media's predictiveness of the stock market by proposing new technological solutions.

First, we note that most research has restricted predictions to a limited number of stocks or indices, the DJIA index in [10], 5 stocks in [52], 18 stocks in [53], 24 stocks in [31], and a somewhat

larger analysis of 420 stocks in [61]. In this work, we perform a large-scale analysis of data for one thousand exchange traded funds (ETFs), mostly consisting of stocks.

Second, we remark that most of the research has focused on exploring the use of generic platforms such as Twitter [10, 43, 52, 53, 55, 59], Facebook [44, 45], or Reddit [41, 63], and other works have focused on specialized platforms such as StockTwits [7, 23, 28, 39, 48, 57, 58, 61]. In this work, we explore and compare the predictiveness of both a general platform (Twitter) and a specialized platform (StockTwits) for which we have five years of data over concurrent time spans.

Third, we observe that existing approaches on social media analysis for stock forecasting have explored a specific time horizon with a large range of features including: topic features [52], sentiment and opinion features [43, 53, 58], public mood [10], textual features [22, 43, 61], and user statistical features [58]. We explore prediction leveraging *all* post content to understand what content is predictive, and for what temporal horizon with an emphasis on a user-centric analysis.

Finally, we are not aware of prior work that has explicitly attempted to analyze "consistently correct" and "consistently incorrect" user posting behavior as we explore in this work.

3 DATA DESCRIPTION

In this section, we provide a detailed description of the datasets we use for this analysis. This includes both Twitter and StockTwits as well as the historical stock price dataset.

Twitter dataset: Twitter has become an influential factor for financial markets in recent years because of its large number of active users from the financial community [64]. The Twitter dataset we used in this paper was collected by Periscope Capital[3,4], an alternative investment manager focused on North American arbitrage opportunities. The dataset covers 5 full years of posted messages spanning 2015 to 2019, for a total number of 27,100,092 English tweets posted by 1,044,704 unique users. Each tweet contains at least one *cashtag* (Ticker symbol) that starts with the $ symbol, for a total number of 26,145 unique tickers mentioned.

We provide more detailed statistics of our Twitter dataset in Figure 1. In particular, we consider the distribution of the number of users per number of posts in Figure 1a, which follows a heavy-tailed power law distribution with exponent 2.13. Also, we show in Figure 1b the distribution of the average post disparity per user, which similarly follows a power law distribution. We can observe that there are some very active users, most likely bots, who tweet more frequently than one second; the median is about 1 tweet every 11 days for a user. Finally, in Figure 1c we show the distribution of the number of users per ticker, which also follows a power-law distribution. The top three most mentioned tickers are `$AAPL` (*Apple* stock), `$TSLA` (*Tesla* stock), and `$ETH` (*Ethereum* cryptocurrency), which have been respectively mentioned by 482,252, 476,449, and 447,599 unique users. Considering the top 1,000 most discussed tickers, the median number of unique users who have mentioned a ticker is approximately 1,300.

StockTwits dataset: As we mentioned earlier, StockTwits is a microblogging platform specifically designed for investors to communicate using similar features to those of Twitter. The StockTwits dataset we used was also collected by Periscope Capital and covers the same 5 year period of messages posted. The total number of messages is 28,839,476, which were posted by 330,099 unique users. Each post contains also at least one cashtag that starts with the $ symbol, for a total number of 12,785 unique tickers mentioned. We note that for roughly the same number of posts as Twitter, we have three times fewer users. This demonstrates that the StockTwits dataset is less sparse (more posts per user) as also evidenced by the lighter-tailed power law distribution shown in

[3]https://www.periscopecap.com/

[4]All Twitter and StockTwits data collected by Periscope Capital was analyzed while the first author was employed there.

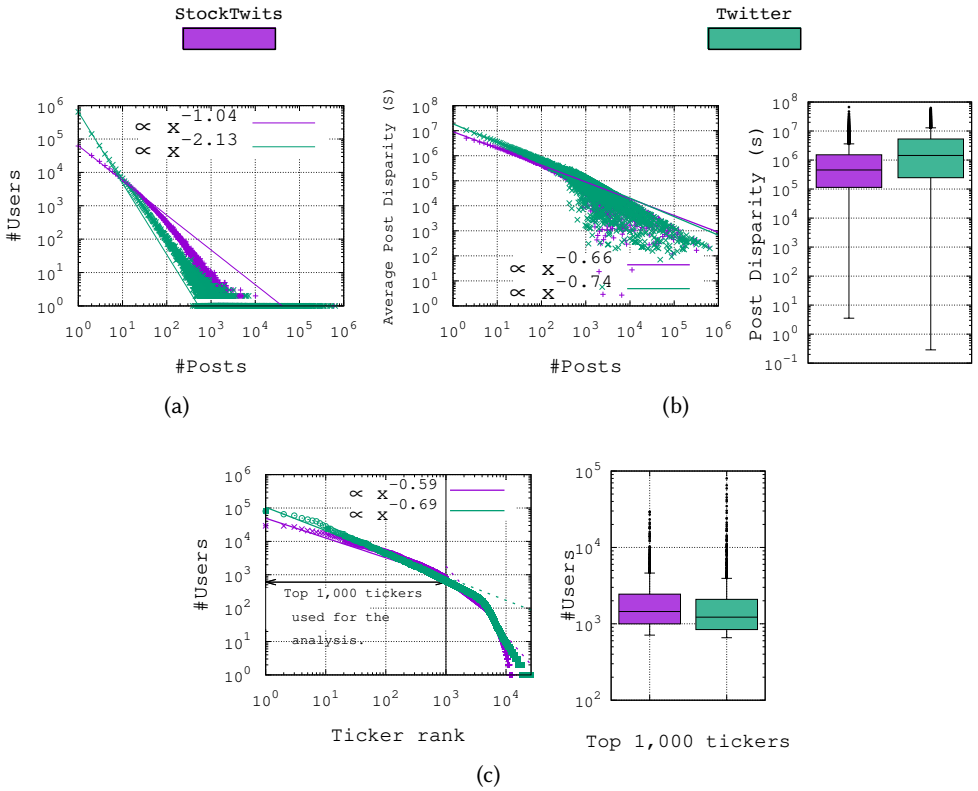


Fig. 1. Dataset statistics of **StockTwits** and **Twitter**. (a) Distribution of the number of users per number of posts. (b) Distribution of average post disparity per users posting similar number of messages. (c) Distribution of the number of users per ticker.

Figure 1a with exponent 1.04. From this data, the StockTwits community appears to be more active on a per-user basis than the general financial community on Twitter. More statistics are shown in Figures 1b and 1c. We remark that the top three most discussed tickers in our StockTwits dataset are `$SP500` (*S&P 500* stock market index), `$AMD` (*Advanced Micro Devices* stock), and `$APPL` (*Apple* stock), which have been respectively mentioned by 789,073, 546,228, and 444,405 unique users.

A distinguishing feature of StockTwits is that it allows users to self-label their posted messages using a social sentiment indicator. This is achieved via a simple toggle on the StockTwits message box that allows users to communicate their view, **Bullish** or **Bearish**, on any specific asset or the market as a whole. Figure 2 shows the growth in total posts, bearish posts, and bullish posts. There are three notable trends here: (i) the total number of messages keeps growing over time, indicating that the platform is gaining in popularity, (ii) about 20-25% of posts are sentiment-labeled, and (iii) there are in general more bullish posts than bearish posts indicating that users tend to be biased towards a positive sentiment in the market.

Stock Market Data: We used Stock Market Data provided by Yahoo! Finance[5]. The dataset contains historical price data of the 1,000 most traded stocks. These stocks correspond to the "Top 1,000 tickers" in the analysis of Figure 1c for Twitter and StockTwits. In order to illustrate the correlation

[5]https://finance.yahoo.com/

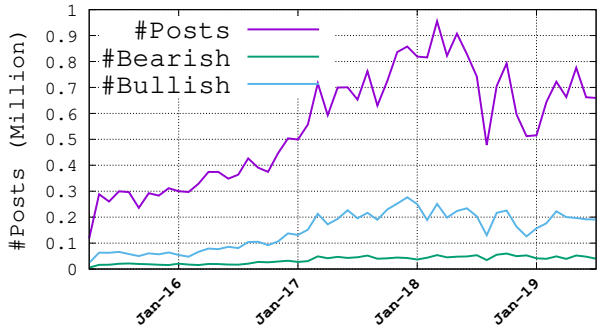
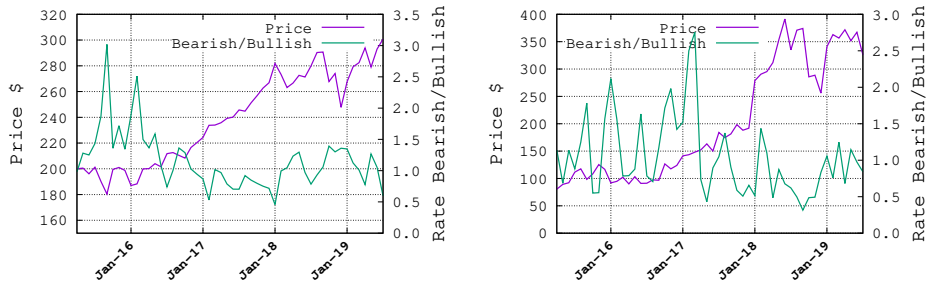
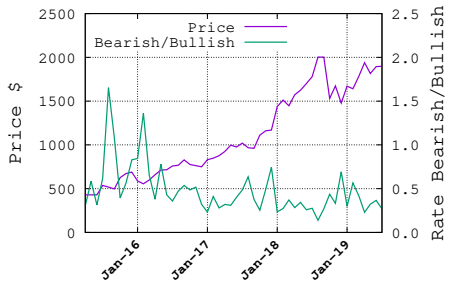


Fig. 2. Growth of the number of messages posted on StockTwits (all three lines) representing the total number of posts, bearish posts, and bullish posts. The sharp drops in the total number of posts corresponds to API issues that reduced the data that could be crawled.



(a) S&P500.

(b) NFLX.



(c) AMZN.

Fig. 3. Price of stocks compared to the Ratio between #Bearish and #Bullish posts on StockTwits vs. time (in years). We observe clearly that there is a mirror effect between the two lines indicating that an increase in the #Bearish posts compared to #Bullish posts is followed by a drop in the stock price and vice versa.

between social media content and the stock market valuation, we refer to Figure 3. Here, we show a comparative analysis of the price of three different stocks (*S&P 500* index[6], *Netflix*, and *Amazon*)

[6]The S&P 500 is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States. It is one of the most commonly followed equity indices, and is considered to be one of the best representations of the U.S. stock market.

with respect to their respective ratio of #Bullish posts by #Bearish posts on StockTwits over the time horizon. We observe three trends here: (i) The stock market, at the index level, always tends to increase over a long period of time. (ii) The ratio of #Bullish posts by #Bearish posts tends to have a high variance, which shows there can be a quick shift of users' opinion over time, most likely in reaction to market news. (iii) Most importantly, there is a mirror effect between the price of a stock and its ratio of #bullish posts by #bearish posts over the time horizon (especially during periods of price drop or stabilization), indicating an apparent mutual anti-correlation.

4 CLASSIFICATION METHODOLOGY

In this section, we describe the formal framework we use for our analysis. We begin by describing our notation and then our classification formulation. Next, we describe the dataset preparation methodology for use by the machine learning classification algorithm. Finally, we provide a brief description of the classification training methodology employed in this work.

4.1 Notation

We present our analysis and classification methodology using the following notation:

- u, w, e, c: respectively a user u, a word w, an emoji e, and a cashtag symbol c (e.g., \$GOOG, \$MSFT, \$AMZN);
- d_c[7]: a post that mentions the cashtag symbol c. Each post has an author u, a timestamp, a set of words, may contain one or more emojis, and maybe sentiment self-labeled by its author u {1 (**Bullish**), 0 (**Bearish**)};
- $D_c = \{d_{c,1}, d_{c,2}, \dots\}$: a set of n posts that mention the same cashtag c, and in which each post $d_{c,j}$ (for $j \in \{1, \dots, n\}$) was issued during the same time window $[t_s, t_e]$. In this paper, we set the size of this time window to 7 days;
- $C = \{D_c^{(1)}, \dots, D_c^{(m)}\}$: a collection of m sets of posts all pertaining to cashtag c collected at disjoint time intervals;

Having now defined our notation, we proceed to describe our analysis methodology in detail.

4.2 The classification formulation

Our primary objective in this article is to understand user-centric behavior in stock market social media discourse and how it relates to stock market prediction. In particular, given a set of posts issued during a specific time window that mention a specific cashtag, we want to build a binary classifier to predict the upward or downward movement of that ticker in the market at a certain point in time.

Formally, given a training set $\mathbb{X}$ of labeled sets of posts $\langle D_c^{(i)}, y_k^{(i)} \rangle \in \mathbb{X}$, where $y_k^{(i)} \in \{0, 1\}$ is a binary label associated with $D_c^{(i)}$ to indicate whether the cashtag c is moving upward or downward in k days from t_s, we wish to train the function $f_k(D_c)$ to estimate the probability $p(y_k = 1|D_c)$, i.e, the probability that the cashtag c is moving upward in k days from t_s of D_c. In this work, we assume a content pooling approach for a set of posts D_c, which comprises a data pre-processing step consisting of merging all posts in D_c together and modeling them as a single document.

4.3 Dataset preparation

Given our classification definition below, our goal here is to show how we construct and prepare our dataset $\mathbb{X}$ of labeled posts $\langle D_c^{(i)}, y_k^{(i)} \rangle \in \mathbb{X}$. Therefore, we will first depict how the sets of posts are selected from the raw datasets (the Twitter or StockTwits dataset described earlier) and how

[7] d refers to a document in text classification jargon.

they are labeled using the stock market historical data. Then, we will describe the features used and the way we temporally split the dataset into train, validation and test sets. This splitting step is critical in our analysis of long-term generalization of the classifiers.

4.3.1 Extracting sets of posts: Preparing and preprocessing data is a critical step in any machine learning project before tackling the problem at hand. Here, it consists of gathering and acquiring the collection C of sets of posts from the raw datasets. In brief, for each 7 day period between "01/01/2015" and "31/12/2019", and for each cashtag c, we select all posts that mention the cashtag c. Then, we consider only sets of posts in which the cashtag is mentioned in more than 500 posts. This latter thresholding step aims to ensure that predictions for each D_c leverage a broad set of user tweets.

4.3.2 Labeling set of posts: The next step consists of labeling the collection of a set of posts C. Specifically, we aim to assign a binary label $y_k^{(i)}$ for each set of post $D_c^{(i)}$, which indicates that the cashtag c is moving upward or downward (respectively 1 or 0) in k days from t_s. We recall that t_s is the date associated to the earliest post in $D_c^{(i)}$. We use the historical stock market data described earlier for that goal.

When providing a ground truth label for our collection of sets of posts C, it is critical to note that the stock market, as summarized by an index (e.g., S&P, DJIA), always tends to go up over a long period of time as evidenced in Figure 3a. In light of this observation, we are interested in assessing user predictions relative to overall stock market performance.[8] To accomplish this, we first estimate the performance of the market as a whole using the S&P500 index from t_s to $t_s + k$ as follows:

$$\begin{cases} \text{Open Price} = \text{price of } \textbf{S\&P500} \text{ at } t_s \\ \text{Close Price} = \text{price of } \textbf{S\&P500} \text{ at } t_s + k \text{ days} \\ \text{market\_performance} = \frac{\text{Close Price - Open Price}}{\text{Open Price}} \end{cases}$$

Similarly, we then estimate the performance of the stock c on the same period from t_s to $t_s + k$ as follows:

$$\begin{cases} \text{Open Price} = \text{price of } c \text{ at } t_s \\ \text{Close Price} = \text{price of } c \text{ at } t_s + k \text{ days} \\ \text{c\_performance} = \frac{\text{Close Price - Open Price}}{\text{Open Price}} \end{cases}$$

Finally, if the performance of c is higher than the performance of the market, we assume that c is moving upward (1), otherwise, we assume that c is moving downward (0).

At this point, we have described how we create the dataset $\mathbb{X}$ of labeled posts $\langle D_c^{(i)}, y_k^{(i)} \rangle \in \mathbb{X}$ using the raw datasets. Next, we will describe what features are extracted from each set of posts.

4.3.3 Classification features: As we mentioned previously, in this work, we assume a content pooling approach for a set of posts D_c. Content pooling is widely used when dealing with microblogging data [4, 17, 42, 51], and comprises a data pre-processing step in which all posts in D_c are merged together and modeled as a single document before being fed to a classification algorithm.

The set of features that we consider for each *pooled* set of posts D_c are the following: (i) *users* (authors of the posts), (ii) *emojis*, (iii) *words*, and (iv) *VADER sentiment* features for each post in D_c.

[8] One could certainly analyze user predictions in an absolute sense, but they may be trivially correct most of the time because they simply predict an increase. In this work, we are primarily interested in understanding the behavior of prescient users who are able to "beat the market" in contrast to those who do not. In this sense, "correct" for us is "beating the market".

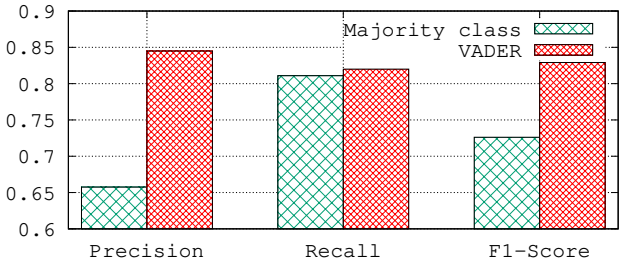


Fig. 4. Performance comparison of VADER to a baseline majority class predictor using self-labeled stock sentiment predictions from StockTwits as ground truth.

While *users*, *emojis*, and *words* correspond to respective sets of unique entities for each type, we pause for a moment to discuss our use of *VADER sentiment*.

It is important to remark that the only use of VADER sentiment in this work is to provide features for the machine learning classifier – VADER sentiment is not considered as a ground truth label for any analysis in this article and VADER sentiment is also different from the self-labeled sentiment available on StockTwits. More specifically, VADER[9] is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media [40]. We have chosen VADER to provide sentiment features for posts for four main reasons: (i) since Twitter does not provide the self-labeled sentiment of StockTwits, we desire sentiment features that can be derived for both Twitter and StockTwits, (ii) VADER is a human-validated sentiment analysis method specifically developed for Twitter and social media contexts [40], (iii) VADER is considered as a state-of-the-art sentiment classification tool [11, 20, 60] used for diverse applications (e.g., customer reviews [12], drug opinion [13], Bitcoin sentiment [56], politics [24], and banks [14]), and (iv) finally, as shown in Figure 5, a preliminary experimental evaluation we have performed on VADER indicates that its predictions are reasonable proxies for self-labeled sentiment and thus may serve as useful features for the classifier. To elaborate, since this is an imbalanced data problem, we look at precision, recall, and F-score in Figure 5 and compare to the baseline most frequent class for reference. In brief, precision, recall, and F-score are all well above 0.8 (as well as the baseline) indicating strong agreement of VADER with self-labeled stock sentiment.

To compute aggregate features over a set of posts D_c, for each user feature we calculate the number of times that the user mentioned c, and for each emoji and/or word feature the number of time it has been associated with c. For sentiment features, for each post, VADER outputs four scoring values that we use: a positive sentiment score, a neutral sentiment score, a negative sentiment score, and a compound score. From this we then calculate several aggregate VADER statistics of the posts in D_c: the sum, standard deviation, minimum, maximum, arithmetic mean, geometric mean, harmonic mean, and coefficient of variation of the scores. These aggregate statistics then serve as our VADER sentiment features.

4.3.4 Dataset splitting: We now describe the *temporal* splitting of the dataset $\mathbb{X}$ of labeled set of posts $\langle D_c^{(i)}, y_k^{(i)} \rangle \in \mathbb{X}$ for training, validation parameter tuning, and test evaluation purposes, respectively $\mathbb{X}^{\text{train}}$, $\mathbb{X}^{\text{val}}$ and $\mathbb{X}^{\text{test}}$. To avoid data leakage in temporally overlapping train-val-test

[9]VADER: Valence Aware Dictionary and sEntiment Reasoner
https://github.com/cjhutto/vaderSentiment

Table 1. Summary of the datasets constructed for a few values of k – due to space limitation we show only a few values of k we considered. We note that test sets are intentionally balanced to ease the interpretation of the *Accuracy* metric.

Values of	k =	**4**	**14**	**29**	**44**	**59**	**74**	**89**	**104**	**119**	**134**	**149**
Twitter	#+Train	175,863	166,132	154,639	144,661	131,281	126,260	118,849	110,150	101,481	93,295	85,476
	#-Train	172,652	170,744	164,801	157,417	146,020	141,056	132,382	123,827	115,250	106,165	96,780
	#+Val	36,779	34,744	32,340	30,253	27,455	26,405	24,855	23,036	21,223	19,511	17,876
	#-Val	36,016	35,618	34,378	32,837	30,460	29,425	27,615	25,831	24,041	22,146	20,188
	#+Test	101,166	95,568	88,957	83,217	75,520	72,632	68,369	63,364	58,378	53,669	49,171
	#-Test	101,166	95,568	88,957	83,217	75,520	72,632	68,369	63,364	58,378	53,669	49,171
StockTwits	#+Train	180,629	169,829	157,082	146,945	137,377	127,986	120,208	111,390	102,627	94,552	86,530
	#-Train	180,642	179,426	174,310	166,588	158,262	149,956	141,014	131,935	122,728	113,056	103,182
	#+Val	36,856	34,652	32,051	29,983	28,030	26,114	24,527	22,728	20,940	19,292	17,655
	#-Val	36,662	36,415	35,376	33,809	32,119	30,434	28,619	26,776	24,908	22,945	20,941
	#+Test	101,713	95,632	88,454	82,746	77,358	72,070	67,690	62,724	57,790	53,243	48,726
	#-Test	101,713	95,632	88,454	82,746	77,358	72,070	67,690	62,724	57,790	53,243	48,726

splits, we define each split as follows:

$$\begin{cases} \mathbb{X}^{\text{train}} = \{\langle D_c^{(i)}, y_k^{(i)}\rangle | \forall d_s \in D_c^{(i)} : t_s \geq \text{“01/01/2015”} \wedge (t_s + k) \leq \text{“01/01/2018”}\} \\ \mathbb{X}^{\text{val}} = \{\langle D_c^{(i)}, y_k^{(i)}\rangle | \forall d_s \in D_c^{(i)} : t_s \geq \text{“01/01/2018”} \wedge (t_s + k) \leq \text{“01/06/2018”}\} \\ \mathbb{X}^{\text{test}} = \{\langle D_c^{(i)}, y_k^{(i)}\rangle | \forall d_s \in D_c^{(i)} : t_s \geq \text{“01/06/2018”} \wedge (t_s + k) \leq \text{“31/12/2019”}\} \end{cases}$$

In sum, we used 3 years of data for training, 6 months of data for validation parameter tuning, and 18 months of data for test evaluation. It is worth noting that there is absolutely no overlap between the three sets, which aims to allow a better long-term generalization of the classifiers. Finally, because the splitting depends on the value of k, the size of each subset is consequently different as shown in Table 1.

4.4 Classification algorithm

Taking as input a set of posts D_c of a given cashtag c, our goal is to combine these inputs to produce a value indicating whether the cashtag is moving *“upward”* or *“downward”* in the market at a certain point in time. To accomplish this, we mainly use the Logistic Regression (LR) classification algorithm, which is an efficient classification algorithm that is still widely-used because of its simplicity, its interpretability, and its ease of training [8]. LR has achieved state-of-the-art classification performance for diverse tasks ranging from spam filtering [30] to prediction of hospital readmission [38] and MRI data analysis [47].

Each set of posts $D_c^{(i)}$ is represented by its vector $x^{(i)}$ of n features $x^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}]$ and its associated label $y_k^{(i)} \in \{1$ **(Upward)**, 0 **(Downward)**$\}$. We used LR with L2-regularization available in the LIBLINEAR package [35]. The L2-regularization hyperparameter C was selected from the set $\{10^{-5}, 10^{-3}, \ldots, 10^{13}, 10^{15}\}$. A second hyperparameter determined the number of top Mutual Information features selected for prediction and was chosen from the set {10, 20, 30, 40, 50, 60, 70, 80, 90, 100,1000}. Both hyperparameters were optimized via a joint grid search in order to maximize accuracy on the held-out validation set. These best hyperparameters were then used to provide final results on the held-out test set.

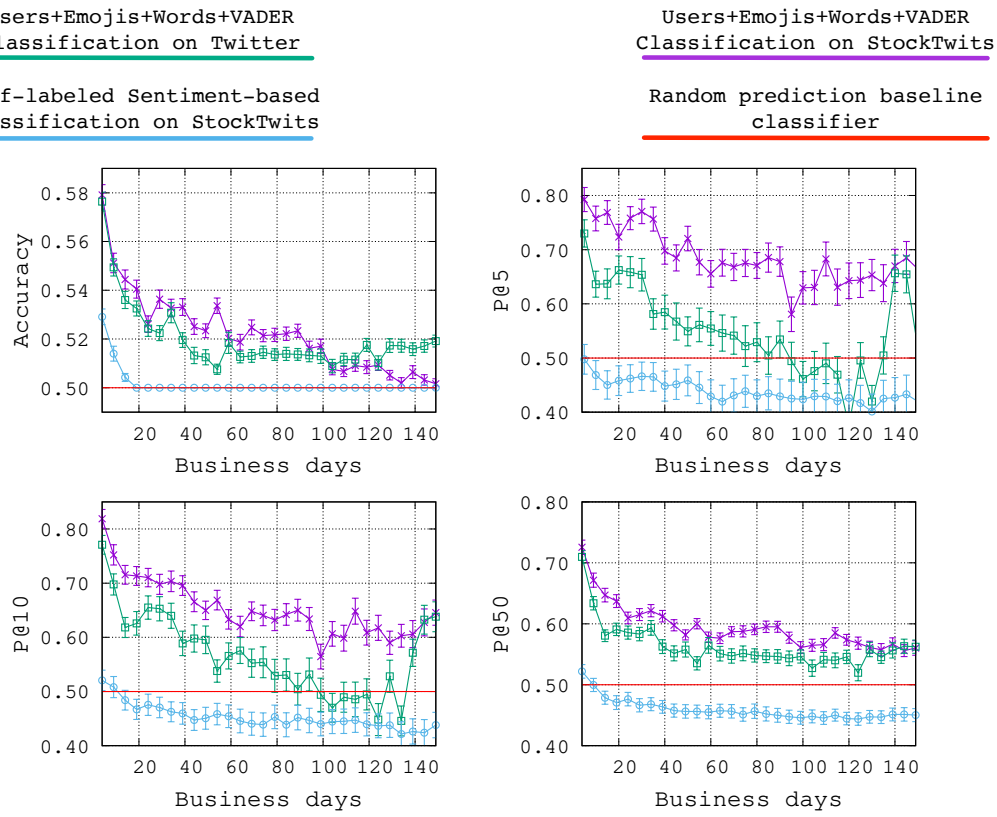


Fig. 5. Classification performance and comparison against a random prediction classifier. The results show mainly that: (i) self-labeled sentiment features on StockTwits are predictive, but only weakly and not for a long horizon, (ii) the use of all feature information in a machine learning classifier yields better predictions than self-labeled sentiment, (iii) StockTwits allows training of more accurate predictors than Twitter, and (iv) the classification performance drops significantly as the prediction time horizon increases.

5 RESULTS AND DISCUSSION

We now report and discuss the main results of the empirical evaluation, considering both the effectiveness of the classification and our interpretation of various user behaviors. Our experiments aim to address the five research questions (RQs) that we stated previously in Section 1.

5.1 Classification performance (RQ1 and RQ2)

Self-labeled social sentiment indicators on StockTwits are a valuable source of information that we may first consider for building a simple social media-based stock market predictor. Indeed, sentiment information on the stock market is explicitly provided by users through these labels. Hence, we first propose to use the ratio of `#Bullish` posts to `#Bearish` posts to predict the performance of a given ticker using the following rule: *if* $(\frac{\#Bullish}{\#Bearish} > \tau)$ *then* **Increase** *else* **Decrease** (where τ is a threshold hyperparameter tuned on the validation set). The results are shown in Figure 5 with a comparison against a random prediction baseline classifier to help establish when the prediction method is outperforming an uninformed random guess baseline. We recall that because we intentionally balanced the test sets, a random prediction baseline classifier is the best uninformed classifier with

an accuracy and a precision of 50%. Briefly, the obtained accuracy indicates that the self-labeled sentiments can be used to make limited predictions up to 20 days. Beyond that, the predictions are no better than random since the accuracy is roughly 50% and equal to the accuracy of a random prediction baseline classifier. This finding suggests that aggregated explicit self-labeled sentiments mainly relate to short-term price movements in the market.

We now propose to use more information beyond the self-labels, namely a machine learning classifier using the features described in Section 4: users, emojis, words, and text-based sentiment features. The results we obtained for this analysis are illustrated in Figure 5 on both Twitter and StockTwits. At first glance, we note that the performance obtained in terms of accuracy and precision is much higher than the performance of the self-labeled sentiment approach. In addition, we observe that these additional features allow us to uncover more latent information that leads our classifier to make predictions well beyond 20 days.

Finally, we note that the performance obtained using the StockTwits dataset is significantly higher than the one obtained using the Twitter dataset. We explain this by the fact that StockTwits is a social media platform that is more likely to be used by experts in the stock market as it is intended for this purpose. Consequently, in the rest of this article, we focus our analysis on the higher quality StockTwits content.

5.2 Longitudinal classification analysis (RQ3)

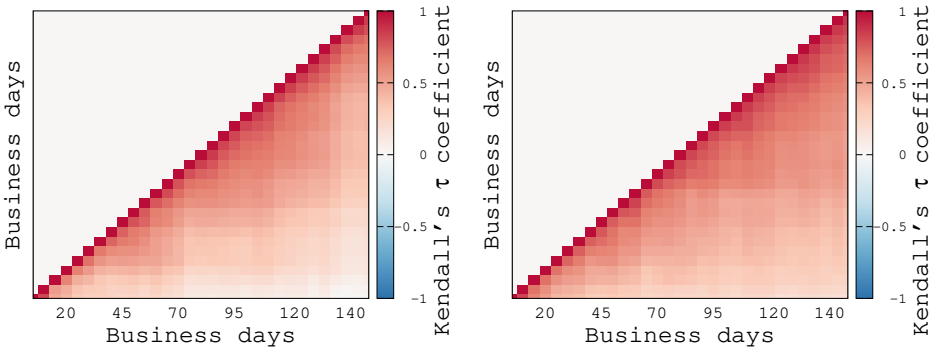
We now undertake a longitudinal study where our goal is to analyze the classification performance over time. We start by referring again to Figure 5, where we clearly observe that the classification performance drops over time as the accuracy and precision is much higher for short-term predictions than for long-term predictions. This result is intuitive given the fact that the stock market performance is likely to be affected by many external factors that increase uncertainty over long periods of time.

We now explore the relationship between post content features and the temporal horizon labels (short-, medium-, long-term) of users over time. A general method for measuring the amount of information that a feature x_j provides w.r.t. predicting a class label y_k (*"upward"* or *"downward"*) is to calculate its Pointwise Mutual Information (PMI) [25] as follows:

$$PMI(x_j, y_k) = \log \frac{P(x_j, y_k)}{P(x_j)P(y_k)} \tag{1}$$

A high PMI value indicates a more informative feature. Therefore, we rank features using their PMI values and we select the top-1000 most predictive features for each value of k. Then, we propose to use Kendall's τ correlation coefficient to compare ranked list agreement of most predictive features at two different time horizons and plot this in a heatmap in Figure 6. A high correlation is to be interpreted as a high agreement between the two ranked lists whereas a zero correlation value indicates total independence between the two lists (a negative value indicates a negative correlation). As shown in Figure 6, we can observe a strong positive diagonal (simply because Kendall $\tau = 1$ when a ranked list is compared with itself on the diagonal) and less rank correlation as the gap in the two predictive time horizons increases. In fact, we remark that there is almost no correlation between the list of ranked features in the short-, medium-, and long-term, indicating that some features are good at capturing short-term correlations, whereas others are good at capturing medium- or long-term correlations.

The above observation leads us to investigate the behavior of users to better understand their strategy in predicting the stock market. We restrict this analysis to mainly the following three types of temporal predictive behaviors of users:



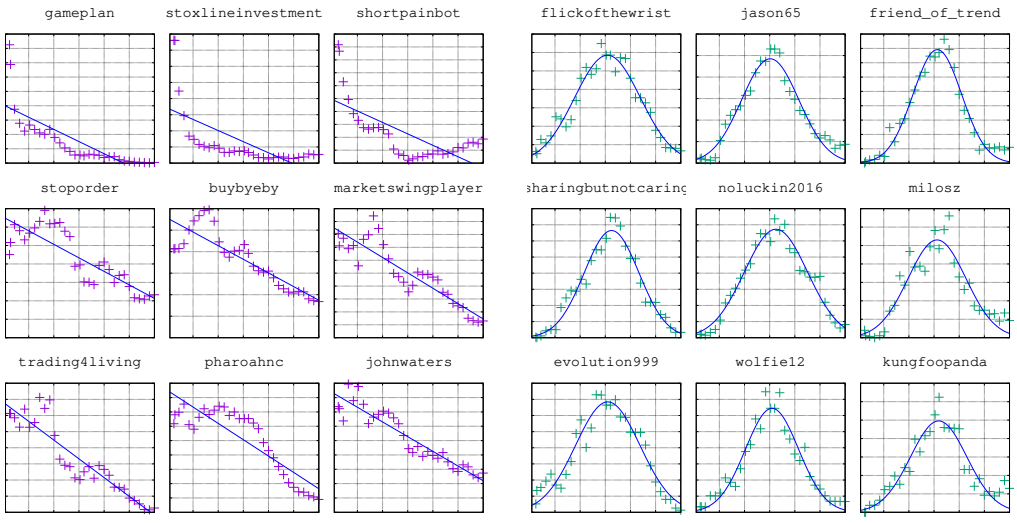
(a) $PMI(x_j, y_k = 1)$. (b) $PMI(x_j, y_k = 0)$.

Fig. 6. Heatmap showing Kendall τ correlation coefficient of feature ranking over the time horizon on StockTwits. We observe that there is a high variance in the ranking of features over the time horizon, indicating that short-, medium-, and long-term predictive features are totally different. In other words, some features are good at capturing short-term correlations, whereas others are good at capturing medium- or long-term correlations with the market.

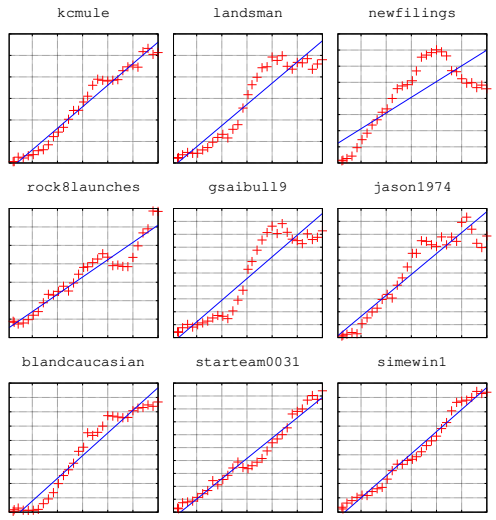
- **Short-term predictive behavior:** which we define as users who are involved in a strategy of forecasting the evolution of the stock market in a relatively short period of time – typically within a 20 day window.
- **Medium-term predictive behavior:** which we define as users who are involved in a strategy of forecasting the evolution of the stock market in a medium range of times relative to our 144 day evaluation – typically within a 60 to 100 day window.
- **Long-term predictive behavior:** which we define as users who are involved in a strategy of forecasting the evolution of the stock market in a relatively long period of time – typically more than 100 days.

We recall that each user is considered as a feature x_j for which we simply calculate the number of times that the user mentioned a cashtag symbol c. Hence, in order to identify user features x_j having one of our predictiveness patterns, we have computed for each user x_j the Mutual Information $I(x_j, y_k)$ with the stock movement label y_k for different time horizons k. Then, we proceed as follows:

- We first fit a linear regression model $y = ax+b$ to each user where y is the mutual information value and x is the time. Then we rank users using a linear combination of the Pearson correlation coefficient (which represents the ability of the linear model to fit the data) and the slope a. Taking a negative slope, this allows us to identify users with a short-term investment strategy as illustrated in Figure 7a (i.e., the informativeness of their posts decreases over the prediction time horizon). In contrast, a positive slope allows us to identify users with a long-term investment strategy as illustrated in Figure 7c (i.e., the informativeness of their posts increases over the prediction time horizon).
- To identify users with a medium-term strategy investment, we fit a Gaussian function $f(x) = a \cdot exp\left(-\frac{(x-b)^2}{2c^2}\right)$ to each user, where the parameter a is the height of the curve's peak, b is the position of the center of the peak and c (the standard deviation, sometimes called the Gaussian width) controls the width of the *bell*. Then, users are ranked by the ability of the model to fit the data and b to be around 80 days as illustrated in Figure 7b.



(a) Short-term. (b) Medium-term.



(c) Long-term.

Fig. 7. Top users exhibiting different investment strategies on StockTwits. The X-axis represents the time horizon (0 to 144 days) and the Y-axis represents the value of Mutual Information. Here we observe users with very distinct short-term, medium-term, and long-term prediction horizons as indicated by the data and best fit curves vs. time.

The results illustrated in Figure 7 show top-ranked users according to our ranking approach for each investment strategy, with their detailed statistics in Table 2. At first glance, we can clearly observe that these top users have a very good fit to the models indicating that the approach is effectively able to identify users with the corresponding investment strategies. For example, the user *"trading4living"* has a high correlation with the stock market in the short-term, which then

Table 2. Statistics of users exhibiting different temporal horizons of predictiveness on StockTwits.

Short-term.

Rank	Username	#tweets	#tickers	Activity period	Disparity
1	gameplan	20,264	1,921	4-2016 to 9-2017	37mn
2	stoxlineinvestment	5,052	2,192	4-2015 to 7-2019	440mn
3	shortpainbot	55,067	4,172	10-2017 to 11-2018	9mn
4	stoporder	536	24	8-2015 to 5-2018	44h
5	buybyeby	970	29	4-2015 to 6-2019	37h
6	marketswingplayer	2,905	46	4-2015 to 1-2019	671mn
7	trading4living	53,211	174	9-2015 to 8-2019	38mn
8	pharoahnc	1,467	23	5-2015 to 8-2019	25h
9	johnwaters	5,564	150	6-2016to 8-2019	297mn

Medium-term.

Rank	Username	#tweets	#tickers	Activity period	Disparity
1	flickofthewrist	568	87	10-2016 to 7-2018	27h
2	jason65	192	10	2-2017 to 8-2017	23h
3	friend_of_trend	96	4	11-2015 to 4-2019	310h
4	sharingbutnotcaring	212	17	9-2016 to 8-2019	120h
5	noluckin2016	184	18	11-2015 to 7-2016	29h
6	milosz	42	1	11-2016 to 8-2017	168h
7	evolution999	371	17	12-2016 to 8-2019	61h
8	wolfie12	78	9	7-2017 to 1-2018	53h
9	kungfoopanda	54	2	11-2017 to 3-2018	59h

Long-term.

Rank	Username	#tweets	#tickers	Activity period	Disparity
1	kcmule	7,170	338	4-2015 to 8-2019	314mn
2	landsman	743	24	10-2015 to 8-2019	44h
3	newfilings	164,245	5,304	4-2015 to 6-2016	3mn
4	rock8launches	2,376	25	12-2016 to 12-2018	425mn
5	gsaibull9	584	4	2-2016 to 2-2018	29h
6	jason1974	1,320	75	11-2016 to 8-2019	17h
7	blandcaucasian	1,250	29	4-2015 to 7-2019	29h
8	starteam0031	906	1	4-2015 to 2-2019	37h
9	simewin1	1,768	15	10-2015 to 7-2019	18h

drops over time. In contrast, the user *"kcmule"* has a very low predictiveness of the stock market in the short-term, which then increases over time. As for the user *"jason65"*, they have only a high correlation with the market in the medium-term exhibiting a sort of bell shape with almost no predictiveness in short and long-term horizons. A key insight here is that an investor or an analyst with an investment strategy for a specific time horizon might select the corresponding subset of users who are the most predictive for that time horizon.

As a final remark, we observe that this analysis was not intended to be exhaustive of all user types in terms of their temporal predictiveness. Rather, we simple chose to fit the short-, medium-, and long-term models as three temporal predictiveness patterns that we hypothesized may be in the data and – as our results show – were in the data. This analysis does not preclude that there may be users with different temporal predictiveness patterns beyond those we have analyzed here.

5.3 "Consistently incorrect" user analysis (RQ4)

Since there are massive financial stakes involved in the stock market, there is a serious risk of having "consistently incorrect" users who may seek to mislead or influence others for self-gain or other unknown motives. It is important to remind the reader that we consider "correct" to be "beating the market" as defined previously. We conjecture that identifying these "consistently incorrect" users is critical for building a robust stock movement predictor.

To identify both "consistently correct" and "consistently incorrect" users, we can consider how the self-labeled stock sentiment aligns with the actual observed stock performance for each user. The idea is that, in general, when a user tends to express a bullish stock sentiment and the stock goes up, or when they tend to express a bearish stock sentiment and the stock goes down, then this user appears to be a "consistently correct" user. In contrast, when a user tends to express a bullish stock sentiment but the stock goes down, and when they tend to express a bearish stock sentiment but the stock goes up, then this user appears to be a "consistently incorrect" user. Formally, given a self-labeled sentiment binary feature vector $\mathbf{x}_j$ (a vector over the tweets of that user) that indicates whether a user j expresses a bullish or bearish sentiment on the cashtag in the tweet (respectively $x_j = 1$ and $x_j = 0$), and the true binary label vector $\mathbf{y}_k$ (the ground truth increase or decrease for the cashtag in each tweet at prediction horizon k), we estimate the (in)correctness score of user j at horizon k using a modified two-outcome Pointwise Mutual Information that captures the two "consistently correct" and the two "consistently incorrect" scenarios discussed above:

$$Score^{Correct}(j,k) = PMI(\mathbf{x}_j = 0, \mathbf{y}_k = 0) + PMI(\mathbf{x}_j = 1, \mathbf{y}_k = 1)$$
$$Score^{Incorrect}(j,k) = PMI(\mathbf{x}_j = 0, \mathbf{y}_k = 1) + PMI(\mathbf{x}_j = 1, \mathbf{y}_k = 0)$$

In this analysis, we only consider identifying "consistently (in)correct" users for the short-term investment strategy, thus, fixing $k \leq 40$. The ranked lists of users on this "consistently (in)correct" scale that we obtain for the different values of k are merged into two lists (one list for "consistently incorrect" users and one list for "consistently correct" users) with users ranked according to their average rank in the lists for different values of temporal predictive horizon $k \leq 40$.

Table 3 shows the top 10 "consistently correct" and "consistently incorrect" users according to this analysis. At first glance, we can notice that the behavior of "consistently correct" and "consistently incorrect" users is roughly the same — a comparable number of posts, a comparable number of mentioned stocks, and a comparable length for activity period. These similar behaviors indicate that "consistently incorrect" users tend to be as active as regular and "consistently correct" users with similar posting patterns, which may make it challenging to identify them based on activity patterns alone. Hence, next we instead seek to identify whether posting content can help distinguish "consistently correct" from "consistently incorrect" users.

Next, we have taken the Top 200 "consistently correct" and "consistently incorrect" users, then we have created a dataset $\mathbb{X}$ of labeled users $\langle x^{(i)}, y^{(i)} \rangle \in \mathbb{X}$, where $x^{(i)}$ represents a user and $y^{(i)}$ is a binary label associated with $x^{(i)}$ to indicate whether the i^{th} user is a "consistently incorrect" or a "consistently correct" user (respectively 0 and 1). Each user i is represented by its vector $x^{(i)}$ of n features representing: (i) words, (ii) emojis, (iii) mentions, (iv) hashtags, and (v) cashtags. In Figure 8 we show top words and emojis used by "consistently incorrect" and "consistently correct" users that we obtained using Pointwise Mutual Information. We can clearly observe that the language used tends to be different. In particular, "consistently correct" users tend to use more common conversational words (including profanity) and emojis, whereas "consistently incorrect" users tend to produce less readable content that mentions company and stock abbreviations and business-related emojis (e.g., the trademark sign). The sharp contrast in language and the lack of conversational content among the "consistently incorrect" users lead us to hypothesize that many "consistently incorrect" users *may* be automated bots although we have no ground truth method for determining whether a user account is actually a bot.

We also remark that bot detection may become even more difficult over time. Given recent advances in natural language processing, automated text generation is now becoming indistinguishable from human generation. In particular, the GPT-3 model [18] released by OpenAI in 2020 has been found to be so proficient at generating text that human evaluators are now failing to

Table 3. "Consistently incorrect" user analysis for short-term predictions on StockTwits. We observe the behavior of "consistently correct" and "consistently incorrect" users is roughly the same — a comparable number of posts, a comparable number of mentioned stocks, and a comparable length for activity period. However, we will observe in Figure 9 that post content of the two user classes is very different. Disparity is the average time between user posts.

Top 10 "consistently incorrect" users

Rank	Username	#tweets	#tickers	Activity period	Disparity
1	edthedaddy	1,592	33	4-2016 to 5-2018	683.958mn
2	z_md	556	153	9-2016 to 12-2018	2151.44mn
3	johnny310x	2,009	219	4-2015 to 6-2017	568.688mn
4	inverseone	6,973	560	4-2015 to 12-2017	197.527mn
5	stockflareus	403	107	8-2016 to 8-2016	21.9118mn
6	franking1969	1,700	33	4-2015 to 5-2018	958.894mn
7	operaghost88	2,437	35	4-2015 to 12-2017	571.685mn
8	dblp214	2,395	98	10-2015 to 7-2019	840.679mn
9	guest617	1,402	80	4-2015 to 2-2018	1073.34mn
10	stevenlarrykaye	2,365	28	4-2015 to 7-2019	948.709mn

Top 10 "consistently correct" users

Rank	Username	#tweets	#tickers	Activity period	Disparity
1	bullboard	926	627	9-2016 to 6-2018	982.615mn
2	mrnoyes	354	178	10-2016 to 2-2019	3565.49mn
3	panamaorange	1,175	43	4-2015 to 8-2019	1918.5mn
4	jamtrades	1,917	207	4-2015 to 8-2019	1176.19mn
5	lexcorp331	4,758	224	5-2017 to 8-2019	248.576mn
6	chris_e	1,253	184	4-2015 to 8-2019	1800.16mn
7	estockpicks	7,262	1,310	4-2015 to 8-2019	310.377mn
8	usacoder	2,616	296	7-2016 to 7-2019	604.291mn
9	doug5007	1,022	73	4-2015 to 7-2019	2195.93mn
10	marknewtoncmt	21,01	511	4-2015 to 8-2019	1073.44mn

distinguish between GPT-3 and human-authored text [26]. Therefore, we expect that in the future, the linguistic contrast between AI-based text and human-authored text may become less apparent, making it more difficult to recognize online bots if their authors wish to disguise them.

To predict "consistently incorrect" users, we have built a classifier using Logistic Regression on the text and emoji features previously described. We trained, tuned, and tested this classifier using nested 10-fold cross-validation and evaluated according to six different classification metrics (accuracy, precision, recall, F1-score, area under curve – AUC, and average precision for a ranking perspective – Avg-P). These classification results are shown in Figure 9, where we are able to achieve a strong accuracy of roughly 78% with a high precision of 88% and solid AUC of 80%.

5.4 Improving predictions with "consistently correct" users (RQ5)

We now determine if we can improve the predictions of stock market movements using simple self-labeled predictions by removing users that we predict to be "consistently incorrect"[10] — this provides us with "Trusted user sentiment-based classification on StockTwits" that we compare to the original "Self-labeled Sentiment-based classification on StockTwits". The obtained results are presented in Figure 10. Briefly, we can observe that removing users predicted to be "consistently incorrect" slightly improves accuracy and extends self-labeled predictiveness from 20 to almost 40 days. These encouraging results indicate that classifying and removing "consistently incorrect" users can be potentially useful as a preprocessing step before building stock market prediction models from social media.

[10]This predictor outlined in the previous section was trained on different data than we test with here.

(a) “Consistently correct” users.

(b) “Consistently incorrect” users.

Fig. 8. Cloud of top words and emojis used by both “consistently correct” and “consistently incorrect” users. Profanity has been partially censored for “consistently correct” users; “consistently incorrect” users did not use profanity. We note that these tweet language characteristics suggest that the “consistently correct” users tend to use more conversational words (profanity-laced or not), while the “consistently incorrect” users tend to heavily use less conversational complex abbreviations and thus may be more likely to be bots.

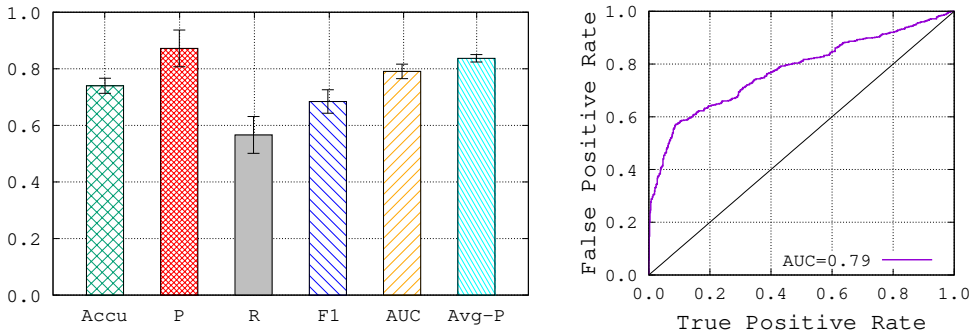


Fig. 9. Classification performance (Accuracy, Precision, Recall, F1-Score, AUC, and Average Precision) for “consistently incorrect” user identification. Nested cross validation is used with 10 folds at the second level for hyperparameter tuning and 5 folds at the top level for determining confidence intervals on the test performance. A 95% confidence interval is shown. Overall, “consistently incorrect” users can be classified correctly with relatively high Accuracy and Precision, among other metrics.

6 CONCLUSION

This work provides a user-centric behavioral analysis of stock market predictions on social media. Our results suggest that using social media for stock market prediction generalizes well over a long time horizon with higher accuracy for short-term predictions than long-term predictions. We also demonstrate that the information content of a tweet is more useful for prediction than the user’s own self-label. Furthermore, an extensive longitudinal analysis of user predictions indicate

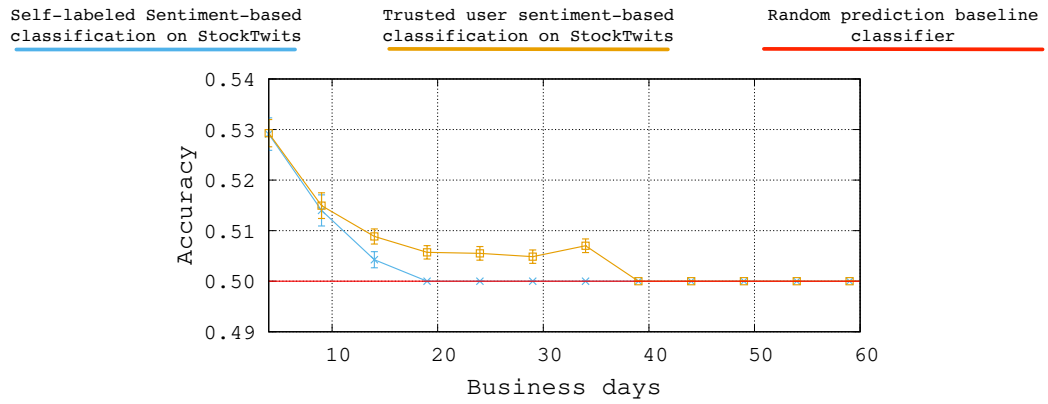


Fig. 10. Self-labeled sentiment classification performance with and without "consistently incorrect" users. The results show mainly that removing "consistently incorrect" users allows to improve the prediction accuracy and time horizon for the self-labeled sentiment classifier.

that different users are predictive for different temporal horizons. We also show that some users seem to be "consistently incorrect" (consistently making predictions that are opposite to the actual market performance) and an analysis of their tweet language characteristics suggests some of these users could be bots. We showed that we can accurately predict "consistently incorrect" users from their tweet content alone and removing them does slightly improve performance of simple stock market prediction from self-labeled tweets.

Overall, we believe the novel user-centric analysis described in this paper reveals a complex ecosystem of behaviors that must be carefully considered when leveraging social media for stock market forecasting. Among many interesting directions, future work might consider leveraging (contextualized) word embedding methods and classifiers (e.g., fine-tuned BERT [32]) with the goal of improving predictions and generalization from the data. Future work may also consider developing and evaluating an automatic portfolio optimization method based on automatically identifying and leveraging temporally targeted and "consistently correct" user data building on the user-centric analysis in this work.

ACKNOWLEDGEMENT

The project received funding from the Ontario Centres of Excellence through a Voucher for Innovation and Productivity grant, 31104.

REFERENCES

[1] Jeffrey S. Abarbanell and Brian J. Bushee. 1997. Fundamental Analysis, Future Earnings, and Stock Prices. *Journal of Accounting Research* 35, 1 (1997), 1–24. http://www.jstor.org/stable/2491464

[2] Amine Mohamed Aboussalah and Chi-Guhn Lee. 2020. Continuous control with Stacked Deep Dynamic Recurrent Reinforcement Learning for portfolio optimization. *Expert Systems with Applications* 140 (2020), 112891. https://doi.org/10.1016/j.eswa.2019.112891

[3] Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management* 58, 4 (2021), 102597. https://doi.org/10.1016/j.ipm.2021.102597

[4] David Alvarez-Melis and Martin Saveski. 2016. Topic modeling in twitter: Aggregating tweets by conversations. In *Tenth International AAAI Conference on Web and Social Media*.

[5] G. V. Attigeri, Manohara Pai M M, R. M. Pai, and A. Nayak. 2015. Stock market prediction: A big data approach. In *TENCON 2015 - 2015 IEEE Region 10 Conference*. 1–5.

[6] Louis Bachelier. 1900. Théorie de la spéculation. *Annales scientifiques de l'École Normale Supérieure* 3e série, 17 (1900), 21–86. https://doi.org/10.24033/asens.476

[7] Roy Bar-Haim, Elad Dinur, Ronen Feldman, Moshe Fresko, and Guy Goldstein. 2011. Identifying and Following Expert Investors in Stock Microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., 1310–1319. https://aclanthology.org/D11-1121

[8] Christopher M Bishop. 2006. *Pattern recognition and machine learning*. springer.

[9] Johan Bollen and Huina Mao. 2011. Twitter Mood as a Stock Market Predictor. *Computer* 44, 10 (oct 2011), 91–94. https://doi.org/10.1109/MC.2011.323

[10] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1 – 8. https://doi.org/10.1016/j.jocs.2010.12.007

[11] Venkateswarlu Bonta and Nandhini Kumaresh2and N Janardhan. 2019. A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology* 8, S2 (2019), 1–6.

[12] Anton Borg and Martin Boldt. 2020. Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications* 162 (2020), 113746. https://doi.org/10.1016/j.eswa.2020.113746

[13] Dr Bose, PS Aithal, Sandip Roy, et al. 2021. Survey of Twitter Viewpoint on Application of Drugs by VADER Sentiment Analysis among Distinct Countries. *International Journal of Management, Technology, and Social Sciences (IJMTS)* 6, 1 (2021), 110–127.

[14] Raphael Kwaku Botchway, Abdul Bashiru Jibril, Michael Adu Kwarteng, Miloslava Chovancova, and Zuzana Komínková Oplatková. 2019. A Review of Social Media Posts from UniCredit Bank in Europe: A Sentiment Analysis Approach. In *Proceedings of the 3rd International Conference on Business and Information Management (ICBIM '19)*. Association for Computing Machinery, New York, NY, USA, 74–79. https://doi.org/10.1145/3361785.3361814

[15] Mohamed Reda Bouadjenek and Scott Sanner. 2019. Relevance-Driven Clustering for Visual Information Retrieval on Twitter. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19)*. Association for Computing Machinery, New York, NY, USA, 349–353. https://doi.org/10.1145/3295750.3298914

[16] Mohamed Reda Bouadjenek, Scott Sanner, and Yihao Du. 2020. Relevance- and interface-driven clustering for visual information retrieval. *Information Systems* 94 (2020), 101592. https://doi.org/10.1016/j.is.2020.101592

[17] Mohamed Reda Bouadjenek, Scott Sanner, Zahra Iman, Lexing Xie, and Daniel Xiaoliang Shi. 2022. A longitudinal study of topic classification on Twitter. *PeerJ Computer Science* 8 (2022), e991.

[18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. (2020). arXiv:cs.CL/2005.14165

[19] Michael P. Cameron, Patrick Barrett, and Bob Stewardson. 2016. Can Social Media Predict Election Results? Evidence From New Zealand. *Journal of Political Marketing* 15, 4 (2016), 416–432. https://doi.org/10.1080/15377857.2014.959690

[20] Jonnathan Carvalho and Alexandre Plastino. 2021. On the evaluation and combination of state-of-the-art features in twitter sentiment analysis. *Artificial Intelligence Review* 54, 3 (2021), 1887–1936.

[21] Roberto Cervello-Royo, Francisco Guijarro, and Karolina Michniuk. 2015. Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA index with intraday data. *Expert Systems with Applications* 42, 14 (2015), 5963 – 5975. https://doi.org/10.1016/j.eswa.2015.03.017

[22] Chen Chen, Wu Dongxing, Hou Chunyan, and Yuan Xiaojie. 2014. Exploiting Social Media for Stock Market Prediction with Factorization Machine. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Vol. 2. 142–149.

[23] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Issues and Perspectives from 10,000 Annotated Financial Social Media Data. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 6106–6110. https://aclanthology.org/2020.lrec-1.749

[24] Yuxiao Chen, Jianbo Yuan, Quanzeng You, and Jiebo Luo. 2018. Twitter Sentiment Analysis via Bi-Sense Emoji Embedding and Attention-Based LSTM. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*. Association for Computing Machinery, New York, NY, USA, 117–125. https://doi.org/10.1145/3240508.3240533

[25] Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16, 1 (1990), 22–29. https://www.aclweb.org/anthology/J90-1003

[26] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

[27] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication* 64, 2 (2014), 317–332. https://doi.org/10.1111/jcom.12084 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcom.12084

[28] Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 519–535. https://doi.org/10.18653/v1/S17-2089

[29] Vinicius Monteiro de Lira, Craig Macdonald, Iadh Ounis, Raffaele Perego, Chiara Renso, and Valeria Cesario Times. 2019. Event attendance classification in social media. *Information Processing & Management* 56, 3 (2019), 687–703. https://doi.org/10.1016/j.ipm.2018.11.001

[30] Bilge Kagan Dedeturk and Bahriye Akay. 2020. Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. *Applied Soft Computing* 91 (2020), 106229. https://doi.org/10.1016/j.asoc.2020.106229

[31] Ali Derakhshan and Hamid Beigy. 2019. Sentiment analysis on stock social media for stock price movement prediction. *Engineering Applications of Artificial Intelligence* 85 (2019), 569 – 578. https://doi.org/10.1016/j.engappai.2019.07.002

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. [n.d.]. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[33] Eugene F. Fama. 1991. Efficient Capital Markets: II. *The Journal of Finance* 46, 5 (1991), 1575–1617. https://doi.org/10.1111/j.1540-6261.1991.tb04636.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.1991.tb04636.x

[34] Eugene F. Fama, Lawrence Fisher, Michael C. Jensen, and Richard Roll. 1969. The Adjustment of Stock Prices to New Information. *International Economic Review* 10, 1 (1969), 1–21. http://www.jstor.org/stable/2525569

[35] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of machine learning research* 9, Aug (2008), 1871–1874.

[36] Ronen Feldman, Oded Netzer, Aviv Peretz, and Binyamin Rosenfeld. 2015. Utilizing Text Mining on Online Medical Forums to Predict Label Change Due to Adverse Drug Reactions. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. Association for Computing Machinery, New York, NY, USA, 1779–1788. https://doi.org/10.1145/2783258.2788608

[37] Anthony C Greig. 1992. Fundamental analysis and subsequent stock returns. *Journal of Accounting and Economics* 15, 2 (1992), 413 – 442. https://doi.org/10.1016/0165-4101(92)90026-X

[38] Shagun Gupta, Dennis T. Ko, Paymon Azizi, Mohamed Reda Bouadjenek, Maria Koh, Alice Chong, Peter C. Austin, and Scott Sanner. 2020. Evaluation of Machine Learning Algorithms for Predicting Readmission After Acute Myocardial Infarction Using Routinely Collected Clinical Data. *Canadian Journal of Cardiology* 36, 6 (2020), 878–885. https://doi.org/10.1016/j.cjca.2019.10.023

[39] Patrick Houlihan and Germán G Creamer. 2019. Leveraging social media to predict continuation and reversal in asset prices. *Computational Economics* (2019), 1–21.

[40] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the 8th International AAAI Conference on Web and Social Media a (ICWSM-14)*.

[41] Dennis Huynh, Garrett Audet, Nikolay Alabi, and Yuan Tian. 2021. Stock Price Prediction Leveraging Reddit: The Role of Trust Filter and Sliding Window. In *2021 IEEE International Conference on Big Data (Big Data)*. 1054–1060. https://doi.org/10.1109/BigData52589.2021.9671412

[42] Zahra Iman, Scott Sanner, Mohamed Reda Bouadjenek, and Lexing Xie. 2017. A Longitudinal Study of Topic Classification on Twitter.. In *Proceedings of the 11th International AAAI Conference on Web and Social Media a (ICWSM-17)*. 552–555.

[43] Fang Jin, Wei Wang, Prithwish Chakraborty, Nathan Self, Feng Chen, and Naren Ramakrishnan. 2017. Tracking Multiple Social Media for Stock Market Event Prediction. In *Advances in Data Mining. Applications and Theoretical Aspects*, Petra Perner (Ed.). Springer International Publishing, Cham, 16–30.

[44] Yigitcan Karabulut. 2013. Can facebook predict stock market activity?. In *AFA 2013 San Diego Meetings Paper*.

[45] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790.

[46] Shamanth Kumar, Geoffrey Barbier, Mohammad Ali Abbasi, and Huan Liu. 2011. TweetTracker: An analysis tool for humanitarian and disaster relief.. In *Proceedings of the 5 International AAAI Conference on Web and Social Media a (ICWSM-11)*. 78–82.

[47] Kyoungjae Lee and Xuan Cao. 2021. Bayesian group selection in logistic regression with application to MRI data analysis. *Biometrics* 77, 2 (2021), 391–400.

[48] Quanzhi Li and Sameena Shah. 2017. Learning Stock Market Sentiment Lexicon and Sentiment-Oriented Word Vector from StockTwits. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*.

Association for Computational Linguistics, Vancouver, Canada, 301–310. https://doi.org/10.18653/v1/K17-1031

[49] Xueming Luo, Jie Zhang, and Wenjing Duan. 2013. Social media and firm equity value. *Information Systems Research* 24, 1 (2013), 146–163.

[50] Nader Mahmoudi, Paul Docherty, and Pablo Moscato. 2018. Deep neural networks understand investors better. *Decision Support Systems* 112 (2018), 23 – 34. https://doi.org/10.1016/j.dss.2018.06.002

[51] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling. In *Proceedings of the ACM SIGIR Conference (SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 889–892. https://doi.org/10.1145/2484028.2484166

[52] Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction. In *Proceedings of the ACL-IJCNLP (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, 1354–1364. https://doi.org/10.3115/v1/P15-1131

[53] Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin. 2015. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications* 42, 24 (2015), 9603 – 9611. https://doi.org/10.1016/j.eswa.2015.07.052

[54] Ming Ni, Qing He, and Jing Gao. 2016. Forecasting the subway passenger flow under event occurrences with social media. *IEEE Transactions on Intelligent Transportation Systems* 18, 6 (2016), 1623–1632.

[55] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi. 2016. Sentiment analysis of Twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*. 1345–1350.

[56] Toni Pano and Rasha Kashef. 2020. A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19. *Big Data and Cognitive Computing* 4, 4 (2020). https://doi.org/10.3390/bdcc4040033

[57] Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Is Domain Adaptation Worth Your Investment? Comparing BERT and FinBERT on Financial Tasks. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 37–44. https://doi.org/10.18653/v1/2021.econlp-1.5

[58] Juan Pineiro-Chousa, Marcos Vizcaíno-González, and Ada María Pérez-Pico. 2017. Influence of Social Media over the Stock Market. *Psychology & Marketing* 34, 1 (2017), 101–108. https://doi.org/10.1002/mar.20976 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.20976

[59] Tushar Rao and Saket Srivastava. 2012. *Using twitter sentiments and search volumes index to predict oil, gold, forex and markets indices.* Technical Report. Institute of Technology, Delhi, India.

[60] Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5, 1 (2016), 1–29.

[61] Andrew Sun, Michael Lachanski, and Frank J. Fabozzi. 2016. Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis* 48 (2016), 272 – 281. https://doi.org/10.1016/j.irfa.2016.10.009

[62] Charles Thomas, Richard McCreadie, and Iadh Ounis. 2019. Event Tracker: A Text Analytics Platform for Use During Disasters. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1341–1344. https://doi.org/10.1145/3331184.3331406

[63] Charlie Wang and Ben Luo. 2021. Predicting $ GME Stock Price Movement Using Sentiment from Reddit r/wallstreetbets. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*. 22–30.

[64] Steve Y. Yang, Sheung Yin Kevin Mo, and Anqi Liu. 2015. Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance* 15, 10 (2015), 1637–1656. https://doi.org/10.1080/14697688.2015.1071078 arXiv:https://doi.org/10.1080/14697688.2015.1071078

[65] Andrew Yates, Nazli Goharian, and Ophir Frieder. 2015. Extracting adverse drug reactions from social media. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[66] Konstantinos Saitas Zarkias, Nikolaos Passalis, Avraam Tsantekidis, and Anastasios Tefas. 2019. Deep Reinforcement Learning for Financial Trading Using Price Trailing. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3067–3071.