The Predictive Power of Social Sentiment Regarding Short-Term Stock Movements

Social Media and Business Analytics Research Project

Moritz Wilksch
wilksch@uni-potsdam.de
787297

ABSTRACT

Ever since modern-day financial markets exist, people have been trying to forecast movements in stock prices, as accurate predictions would entail economic benefits and the reduction of risks. This project examines whether social sentiment can be used to predict short-term stock movements. Using more than two years of data from Twitter, we assess the predictive power the extracted sentiment holds for 10 companies listed in the S&P500. Comparing different sentiment analysis approaches and forecasting models, we find that for three out of the ten companies, sentiment does significantly improve the forecasting performance, while a custom-built sentiment model outperforms an off-the-shelf one and tree-based models deliver better performance than linear ones. This provides evidence against the Efficient Market Hypothesis and warrants future research regarding the circumstances under which stock returns might be predictable.

Keywords

Social sentiment, Twitter, stock market, predictive power, forecasting.

INTRODUCTION

Forecasting future returns of stocks has been an active area of research ever since the advent of modern financial markets. The potential economic benefits of being able to predict security prices over a long or even short period are enormous. Most practitioners approach this problem in one of two ways: By analyzing fundamental data like balance sheets and cash flow statements to find undervalued companies to buy for a long-term investment or by conducting technical analysis to discover and exploit short-term trends and patterns in historical price movements. According to the efficient market hypothesis (EMH), neither of these approaches works to outperform the general market. The EMH in its strong form states that "security prices at any point in time fully reflect all available information" (Fama, 1970, p. 388). This implies that generating excess returns can only be done by taking on excess risks, and thus no technique or forecasting model could elevate returns above the level of the general market at the same level of risk. In recent years, however, the EMH has been challenged as its assumptions seem unrealistic and it is not able to explain certain phenomena observed in real markets. For example, the observed volatility in equity markets is higher than what would be expected under an efficient market model (Shiller, 2003).

A contending theory that aims to make more realistic assumptions about market participants is the theory of Behavioral Finance (BF). It acknowledges that market participants are subject to a wide range of cognitive biases like overconfidence and the use of heuristics (Ritter, 2003). Thus, the theory of BF allows for market inefficiencies caused by human behavior like for example overreaction or underreaction to news. This is especially interesting as we observe a rising number of retail investors and traders entering the markets (Brokernotes, 2018), and online investing communities becoming a

place of active exchange of opinions. In January 2021, a group of individual investors organized in a community on the social media platform Reddit even managed to cause a short squeeze in GameStop stock by collectively driving its price up (Burton & Parmar, 2021).
This raises the question of whether the opinion of the crowd if quantified and analyzed appropriately, holds significant predictive power regarding short-term stock movements. To examine this research question, we study ten large-cap US companies. We use sentiment analysis on data collected from the micro-blogging platform Twitter to quantify the public sentiment towards these companies and assess the predictive power it holds concerning short-term movements in the corresponding stocks. The aim is to contribute to the body of literature concerning BF and the weak and semi-strong form of EMH, depending on whether sentiment holds predictive power or all publicly available information is priced in. Practitioners might use findings of the modeling process to further improve existing models by potentially including sentiment-based features and rule out model configurations with sub-par performance.
The remainder of the paper is structured as follows: The *theoretical background* section will summarize previous research, both around financial theories in general (contrasting different forms of EMH with BF) and around the topic of public sentiment in forecasting models in specific. The *methodology* section will outline the research process, starting with data collection and sampling until model validation and performance assessment. The *results* will subsequently be presented and put into context in the *discussion*.

THEORETICAL BACKGROUND

Financial Theories

Ever since proposed by Eugene Fama in 1965, the Efficient Market Hypothesis (EMH) has been a predominant model in financial theory. The EMH is concerned with whether prices of securities at any given point in time fully reflect a particular subset of information. According to Fama (1970), a market is efficient per definition, if (1) there are no transaction costs, (2) all information is available to all market participants at no cost, and (3) all market participants agree on the implication of current information. As these assumptions seem unrealistic, Fama points out that these are only *sufficient* conditions, that is, they are not necessary and weaker forms of efficient markets exist. Specifically, he names three forms of market efficiency: weak form, semi-strong form, and strong form.
In weak-form market efficiency, the subset of information prices "fully reflect" is historical price information. Assuming that successive returns are identically distributed, this hypothesis can be expressed as

$$f\left(r_{j,t+1}\middle|\Phi_t\right) = f\left(r_{j,t+1}\right) \tag{1}$$

where $r_{j,t+1}$ denotes the return of security *j* at time *t+1*, Φ_t denotes the available information at time *t*, and *f* is the probability density of the return distribution. This model is also called the *random walk* model, as the conditional independence stated in (1) implies that security prices follow a random walk. Under these assumptions, no historical price information can be used to forecast future stock returns. This also implies that the practice of technical analysis – the study of chart patterns – cannot be used to generate excess returns. In its semi-strong form, the EMH assumes that prices reflect all *publicly available* information. While this assumption is harder to test than the weak-form EMH, Fama et al. (1969) have conducted a series of tests examining market reactions to news like stock splits, dividend- and earnings announcements. They find evidence that markets react immediately and efficiently to such events. Finally, the strong form of the EMH assumes that prices fully reflect *all information*, implying no individual can expect higher profits than the competition due to monopolistic access to information. However, there is evidence that this assumption is unrealistic as insider trading does indeed occur and generate excess returns (Lorie & Niederhoffer, 1968).

In contrast to the EMH, Behavioral Finance (BF) does not use the simplified assumption of rational agents to describe financial markets, market participants, and their interactions with one another. Rather it is the study of how psychology impacts the decision-making of individuals. Foundations for this area of research were laid in the 1970s when psychologists Daniel Kahneman and Amos Tversky started studying judgment under uncertainty. They find that humans employ certain heuristics and mental operations when assessing uncertain situations which lead to systematic and predictable errors (Tversky & Kahneman, 1974). Later, Kahneman and Tversky (1979) provide evidence that real-world decision-making does not follow the classical utility theory. Utility theory states that when presented with multiple sets of discrete, probabilistic choices X, the utility of each prospect, $U(\cdot)$, can be described as the expected utility of its outcomes:

$$U(X) = \sum_{i=1}^{n} u(x_i) \cdot \mathbb{P}(X = x_i) \tag{2}$$

They develop the "Prospect Theory" which, amongst other hypotheses, accounts for the fact that humans tend to value gains and losses in different, non-proportional ways. For example, individuals appear to be more risk-seeking when facing sure losses but more risk-averse when presented with sure gains. Subsequent research studies the effect of emotions and group behavior on decision-making processes. According to DeBondt et al. (2008), there are three classes of findings in the behavioral finance literature. First, there is a catalog of biases that human decision-makers are subject to. Second, there are speculative dynamics in asset prices, where "systematic errors of unsophisticated investors […] create profit opportunities for experts" (DeBondt et al., 2008, p. 9). This also implies that the opinions and sentiment of such investors could be used by experts to gauge the emergence of price bubbles. Finally, there are findings regarding how decision processes influence decision outcomes. This is especially applicable to corporate settings, where formal decision processes are codified. Overall, unlike in neoclassical finance theory, BF does not have a unified theoretical core (DeBondt et al., 2008). Instead, it is a collection of psychological models applied to economics and attempts to explain empirical market phenomena through the behavior of individuals (Glaser et al., 2004).

The Predictive Power of Sentiment

Publicly released news articles can have drastic effects on stock returns as markets react to their content. Consequently, there is a large body of literature examining how the sentiment in news articles affects stock returns (for example Tetlock et al., 2008; Khan et al., 2020). However, this project's objective is to examine whether the opinion of the crowd predicts stock returns. Curated news articles are therefore not of primary interest as they are published by centralized news agencies. Instead, we study the sentiment of individual investors. Yu et al. (2013) even provide evidence that social media sentiment has a stronger relationship with stock returns than sentiment extracted from traditional media. The most common starting point for social sentiment analysis is big social media platforms, as anyone can just sign up and start posting to a vast audience. Plenty of previous research has been conducted to find out whether public sentiment does indeed hold such predictive power. A happiness index calculated from Facebook posts has been shown to predict daily returns and trading volume (Karabulut, 2013; Siganos et al., 2014). Sentiment extracted from the micro-blogging platform Twitter has also been successfully used to predict short-term stock returns (Bollen et al., 2011; Mittal & Goel, 2012). Even the opinions of users in online investing forums have been shown to predict future closing prices (Li et al., 2020) or improve predictive power if combined with other sources (Nti et al., 2020). The forum "Stock Twits" seems to be of particular interest, as users can label their posts as either bullish or bearish thus providing researchers with an abundance of labeled data. In all other cases, *sentiment analysis* is the preferred technique applied on social media posts. It is the study of the extraction of opinions and attitudes from written text (Liu, 2012). Sentiment analysis techniques

employed in forecasting stock returns most often rely on a lexicon that assigns sentiment scores to single words and aggregates them (Bollen et al., 2011; Mittal & Goel, 2012; Xie, 2021). Occasionally, researchers employ self-trained machine learning models (Renault, 2019) or use pre-trained models (Audrino et al., 2020).

Overall, there is some evidence that social sentiment can be used to forecast the future return of some securities. Sentiment extracted from Twitter using the Profile of Mood States lexicon has been used as a feature in neural networks to achieve more than 75% directional accuracy for daily forecasts of Dow Jones values (Bollen et al., 2011; Mittal & Goel, 2012). Moreover, Twitter sentiment has been shown to granger-cause stock market return (Tabari et al., 2018) which can be used by machine learning models to predict future returns of the Dow Jones and NASDAQ indexes with high accuracy (Rao & Srivastava, 2014). Teti et al. (2019) show that even linear models can exploit Twitter sentiment to explain a significant amount of variance in the daily returns of 69 different technology companies. Similar results are reported by Ren et al. (2018) who not only achieve above 80% prediction accuracy but also show that adding sentiment features improves accuracy by 18%p for the Chinese SSE50 index. Even when not added to an existing financial model, sentiment indicators on their own can hold predictive power as Sun et al. (2017) illustrate for the Chinese stock market.

However, the results are not completely consistent: When applied to the Bitcoin market and the corresponding online forum bitcointalk.org, Xie (2021) finds that sentiment is mostly determined through past performance and only carries limited information for price forecasting. This is confirmed by an analysis conducted by Coqueret (2020). It suggests that the effect returns have on sentiment is much larger than vice versa, and although predictive power can be found for a small percentage of stocks, there is no clear pattern under which circumstances this is the case. Similarly, Renault (2019) states that stock returns and sentiment of five US technology stocks are highly correlated but he was unable to use this for prediction purposes. Sometimes, even if statistically significant predictive power can be found, it lacks practical significance: Acting upon predictions entails brokerage fees which often make trading strategies with minuscule upside unprofitable (Antweiler & Frank, 2004). While social sentiment has been shown to improve stock *volatility* forecasts (Audrino et al., 2020; Antweiler & Frank, 2004) and predict future trading *volume* (Oliveira et al., 2017), there is not yet a consensus as to whether or under which specific circumstances social sentiment holds predictive power regarding stock returns.

Conceptual Framework

Based on previous research in the field, this project aims to examine the predictive power social sentiment holds concerning future stock returns. To capture this concept as detailed as possible, we will not only use a univariate sentiment score but also a measure of how polarized the sentiment is on any given day. Should public sentiment hold any predictive power, this would provide evidence against the weak form of the EMH as presented in equation (1) and indicate that BF might be better suited to explain modern financial markets. To make results more comparable to other research modeling future stock returns, we add basic financial indicators that are often used to predict returns (Neely et al., 2014). This results in the conceptual framework presented in figure 1.

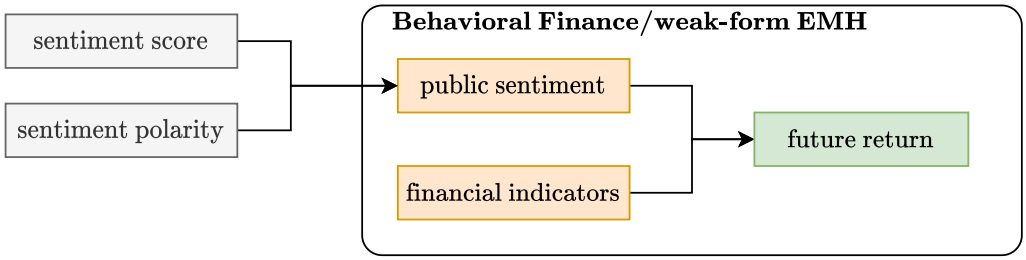


Figure 1: Conceptual framework

METHODOLOGY

Following the majority of the literature studying US markets, this project collects data from Twitter to operationalize public sentiment. Twitter is one of the largest micro-blogging platforms in the United States with over 330 million monthly active users (Twitter Inc., 2019). With a low barrier to entry, anyone can sign up and start posting on Twitter. Its impact on modern financial markets has recently become evident when Elon Musk, CEO of Tesla Inc., caused a sudden, 16% price jump in Bitcoin after tweeting Tesla would accept the cryptocurrency as payment for vehicles. Besides such high-impact messages, there are millions of posts and discussions shared by ordinary users every day. Hashtags make it easy to find topics one is interested in. Investing discussions are no exception to this rule: Many people discuss market movements, give their opinions about stocks, or share investment ideas. These discussions have become so widespread that Twitter even introduced cashtags: tags consisting of a "$" sign followed by a stock ticker to conveniently reference publicly listed companies when talking about them in a financial context. This characteristic makes cashtags a great filtering mechanism to find tweets discussing investments in specific companies.

Pre-Study

To get a reasonably good proxy of public opinion towards a stock, it must be widely discussed on Twitter. If this were not the case, very few users posting about a niche topic would dominate the "public" sentiment. As previous research suggests that the predictive power differs widely between individual companies (Mudinas et al., 2019), this study aims to compare the predictive power of social sentiment across ten different stocks instead of a single company or index. We conduct a pre-study to find the ten most talked-about companies on Twitter.

For each one of the 505 companies listed in the Standard & Poor's 500 index, we scrape all tweets containing the corresponding cashtags (Open Knowledge Foundation, 2021). Because of time constraints, the time frame for the pre-study is limited to the entire month of January 2021. The data collection is performed through *twint* (Twintproject, 2021). Twint is a python package that allows searching for tweets without adhering to the API limitations imposed by the official Twitter API, which limits search results to 3200 tweets per query. Figure 2 presents the number of tweets for each of the top 15 stocks discussed on Twitter. It is noteworthy that only a few companies make up for most of the mentions in tweets.

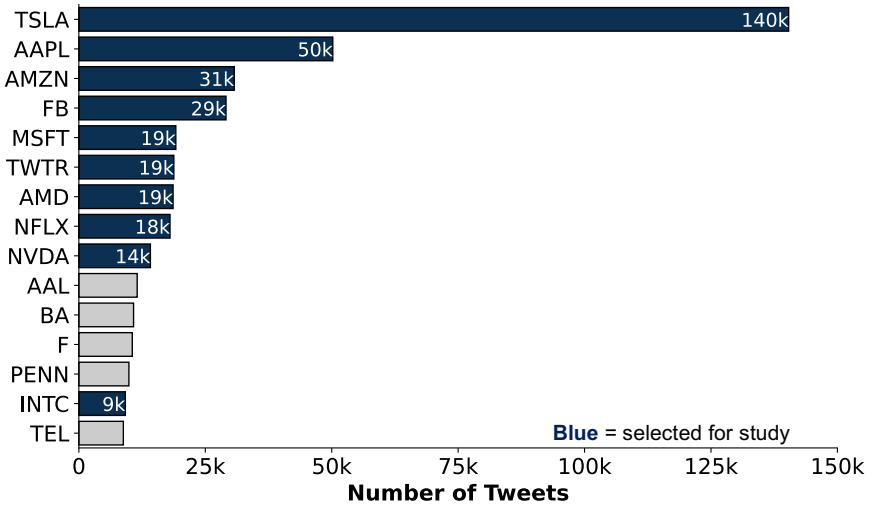


Figure 2: Pre-study results. Number of tweets per stock in January 2021

We choose the first nine stocks to be part of our study. However, as the COVID-19 pandemic coincides with parts of the study period, neither AAL, BA, F, or PENN are included as the tenth company in the sample. All these corporations had to shut down almost their entire business to comply with

government-imposed COVID-19 regulation. To create a sample that is as homogenous as possible, INTC becomes the tenth company to study. This way, the sample only consists of large-cap technology and semiconductor companies: Tesla Inc., Apple Inc., Amazon.com Inc., Facebook Inc., Microsoft Corporation, Twitter Inc., Advanced Micro Devices Inc., Netflix Inc., Nvidia Corporation, and Intel Corporation.

Data Collection

For each of the ten selected companies, the *twint* package is used to scrape the body of tweets later used for sentiment analysis. The search query remains the corresponding company cashtag (e.g. "$TSLA" for Tesla Inc.). However, to span a reasonably large time frame to build predictive models, tweets from January 1st, 2019 until April 30th, 2021 are scraped. This search is conducted in May 2021. The number of tweets obtained is presented in Table 1 as the "Initial" number of tweets.
The corresponding financial data for each company is readily available: Using the yahoo finance API, we simply query daily closing prices for this period.

Data Preprocessing

To prepare the tweets for sentiment analysis, some preprocessing is necessary. Firstly, the analysis is limited to tweets in English, as most tweets are in English anyway and most sentiment models work best when applied to English text. Secondly, a manual inspection reveals that the raw data contain large amounts of spam posts. Such tweets are either generated from bots that automatically share many tweets without real content or accounts misusing the cashtag symbol of a famous company to advertise products and services that are not related to the company as shown in figure 3.

Figure 3: Misuse of cashtags for advertising a product (Scriptstotrade, 2019)

Therefore, after filtering for language, all tweets containing more than 4 different cashtags are removed from the data set. This simple filtering mechanism removes most of the spam, leaving mostly tweets that contain text related to the mentioned cashtags. The effect on the sample size is displayed in table 1. On average, the filtering reduced the sample size per company by 49.6% (SD=8.9%p) largely due to the imposed limit to the number of cashtags, suggesting almost half of the tweets are not company-specific.
Subsequently, hyperlinks and user mentions are removed from the remaining tweets. They carry no additional information and might deteriorate the performance of sentiment analysis methods as they often contain fictional words and large numbers of symbols. Many other preprocessing techniques exist in the research area of natural language processing. The removal of stopwords and word lemmatization and stemming are most common. Stopwords are words that are needed syntactically but do not carry information, like "the", "a", "of". Lemmatization tries to return words to their dictionary form (e.g. "carried" is replaced with "carry") while stemming only removes suffixes (e.g. "carried" is replaced with "carr") (Balakrishnan, 2014). However, these preprocessing techniques should only be applied when explicitly needed by the sentiment model. Renault (2019) finds that removing stopwords imprudently can harm the accuracy of sentiment classifiers. Hence, for now, no further textual preprocessing steps are undertaken.

	Initial	Language filter	Number of cashtags filter
TSLA	1,987,635	1,688,123	1,338,920
AAPL	946,568	833,255	501,806
AMZN	669,820	597,180	331,809
FB	508,272	453,542	236,308
MSFT	359,870	319,157	162,018
TWTR	246,971	230,921	148,588
AMD	298,069	259,470	136,046
NFLX	374,735	349,502	181,948
NVDA	279,462	221,952	97,605
INTC	132,341	114,021	59,310

Table 1: Sample size (number of tweets) after each filtering step

As for the financial data, the daily closing prices are transformed into returns. The return on any given day t, r_t is the closing price on this day p_t divided by the previous closing price p_{t-1} minus 1:

$$r_t = \frac{p_t}{p_{t-1}} - 1 \tag{3}$$

This does not impact the practical data usability, as returns can easily be converted into closing prices and vice versa. However, it makes the time series stationary. A stationary time series exhibits a constant mean and variance over time. Stationarity is a major assumption for many time series models (Manuca & Savit, 1996). Moreover, it can even govern the choice of model evaluation procedure or metric (Bergmeir et al., 2014). Returns on non-trading days are backfilled, such that the correct prediction on any non-trading day is the return of the next following trading day.

Sentiment Analysis

Sentiment analysis is used to automatically (i.e. without manually annotating each data point) extract sentiments from text. Most sentiment models fall in one of two categories: Lexicon-based or machine-learning-based.

Lexicon-based approaches use a pre-assembled lexicon of emotion-laden words in which each word is assigned a sentiment score. It is assembled by human raters or researchers who also determine how the aggregate score of a sentence or paragraph is determined from individual word scores. Successfully used lexica for social media content include the Google Profile of Mood States lexicon (Bollen et al., 2011; Mittal & Goel, 2012) as well as VADER (Hutto & Gilbert, 2014) which is specifically geared towards social media content that contains typos, slang, and emoticons. Lexicon-based approaches are computationally cheap to apply to large data sets and simple to interpret. However, as they often span only a limited number of predetermined words, they tend to classify many documents in absence of any known words as “neutral”.

Alternatively, one can train machine learning models on large corpora of labeled text to learn words and word combinations that are frequently used in a positive or negative context. This makes them more flexible and expressive, as no human has to come up with words associated with specific emotions. Rather, the model itself can find words that most reliably predict a positive or negative sentiment. If trained on large enough corpora, the model can handle most words used in colloquial language. By adequately choosing the training data set, researchers can build machine learning models for domain-specific tasks or for generalizing well over a range of different tasks.

In this project, the performance of both approaches of sentiment analysis for stock forecasting will be compared against each other. Multiple lexica have been developed for this setting. Al-Shabi (2020)

benchmarks the 5 most popular: VADER, SentiWordNet, SentiStrength (which is derived from LIWC), Liu and Hu Lexicon, and AFINN. He finds VADER to perform best on both data sets. While Bollen et al. (2011) and Mittal & Goel (2012) report outstanding forecasting performance using a self-developed extended lexicon based on the Profile of Mood States questionnaire, this tool is not openly available. Thus, we choose VADER for lexicon-based sentiment analysis in this project. Unlike other algorithms in natural language processing, it does not require the text data to be preprocessed. Instead, it takes advantage of different capitalizations which are often used to emphasize a point, or emoticons which consist of punctuation symbols both of which would normally be removed as part of preprocessing the text. For this project, numbers, cashtags, and hashtags are removed from the tweets nevertheless as they do not constitute natural language. Following the recommendation by Hutto and Gilbert (2014), VADER's compound score is used to measure sentiment. To also calculate the percentage of positive and negative tweets per day, they recommend the following classification rules: negative if $\text{compound} \leq -0.05$, neutral if $-0.05 < \text{compound} < 0.05$, and positive if $\text{compound} \geq 0.05$.

Although VADER is specifically geared towards application on social media content, it cannot cope well with domain-specific language. In finance, certain terms like "buying calls" or "going short" convey the authors' expectation of future stock movements, however, they can't be handled by lexicon-based approaches simply since these terms are not used in the regular everyday language the lexicon is based on. To cope with this shortcoming, we build a sentiment model specific to the collected corpus of tweets. Note that Ranco et al. (2015) argue that to reach the performance of human experts, a manually labeled training set size of around 100,000 tweets is necessary. However, Renault (2019) shows that it is possible to create a well-performing classifier with as little as 2,500 messages as the marginal performance increase diminishes. Due to resource constraints, a training set size of $n = 3000$ is chosen.

First, we sample a stratified random sample of tweets from the clean tweet data set. To have a balanced training dataset, 300 tweets per company are sampled and subsequently annotated by the author as negative, neutral, or positive. Table 2 displays the codebook according to which tweets were classified.

Negative	Neutral	Positive
• Stock is overvalued, overbought • Sold the stock (Also: locked in profit) • Not buying the stock (except for: would like to buy, but can't for personal reasons) • Shorted stock • Buying puts, selling calls • Negative news, e.g. anti-trust hearings and class-action lawsuits • Banning company's products & services (e.g. never consider buying an iPhone)	• Neutral news headlines • Unconfirmed rumors • Promotion of products unrelated to the ticker • No clear positive or negative sentiment in a tweet • Seeking opinions (e.g. "Shall I buy X or Y?", "What do you think about stock X?") • Balanced stock analysis without recommendation • Absolute price or earnings numbers without direction or interpretation (e.g. stock trades at $10) • Price target unchanged	• Acquisitions and product announcements • The stock is a bargain, undervalued, oversold • Stock reaching all-time highs • Just bought stock, planning to buy stock, want to buy stock • Buying calls, selling puts • Positive earnings releases, growing revenue, profits, or customer base (Not: Absolute numbers without judgment) • Holding the stock (i.e. keep holding, not selling)

• Mocking the company, using it as a negative example • Price target lowered	• Buying the company's products or services (e.g. buying a Tesla car) • Using the company as a positive example • Price target raised

Table 2: Codebook for classifying tweets

Second, the tweets are preprocessed by removing numbers, hashtags, and cashtags as well as converted to lower case. Additionally, each word is stemmed to reduce the number of unique words as the train set size is small. Following Renault's (2019) recommendation, stopwords and emojis are not removed as this hurts classification accuracy. Third, the text is converted to a $|D| \times |V|$ bag-of-word matrix, where D denotes the document corpus (so $|D| = n$), V denotes the set of unique words over all documents, and each entry $f_{d,t}$ is the number of times term t occurred in document d. Fourth, the term frequency – inverse document frequency (TF-IDF) weighting is applied to the matrix. TF-IDF reduces the weight of very common words that do not help differentiate between documents (e.g. "and" or "the") and assigns a higher weight to words that help separate a document from the rest. Mathematically, it is defined as:

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D) \tag{4}$$

$$\text{tf}(t, d) = \frac{f_{d,t}}{\sum_{t' \in d} f_{d,t'}} \tag{5}$$

$$\text{idf}(t, D) = \log \frac{n}{|\{d \in D : t \in d\}|} \tag{6}$$

It has been shown that the TF-IDF transformation can improve the performance of sentiment classification models (Pimpalkar & Raj, 2020). Fifth, we use scikit-learn (Pedregosa et al., 2011) to train a logistic regression model to classify tweets into three classes. The hyperparameter C (inverse L_2 regularization strength) is tuned using random search ($C \in [10^{-5}, 10]$). The accuracy of the model is calculated on the hold-out set in a 5-fold cross-validation procedure to get a realistic, unbiased estimate of its performance on unseen data.

Feature Engineering and Data Set Characteristics

As our research question aims at short-term stock movements, all data are aggregated on a daily level. The prediction target is a binary indicator of whether the following day's return is positive or negative. Thus, we aim to build a model that predicts a positive or negative return at time $t + 1$ based on information available at time t:

$$f\left(\mathbb{I}\left[r_{j,t+1} > 0\right] \middle| \Phi_t\right) \tag{7}$$

Here, $\mathbb{I}[\cdot]$ denotes the indicator function that equals 1 if the enclosed condition is fulfilled and 0 otherwise and f denotes the probability distribution over the two classes "positive" (1) and negative (0). A return of 0 is classified as negative, as it does not support a buying decision. Note the resemblance with equation (1) – the weak-form of the efficient market hypothesis. Under the weak-form EMH constructing such a model should not be possible.

While the stock returns are already daily returns, the sentiment score must be averaged over tweets for each day. As Clapham et al. (2019) suggest, the sentiment score of each tweet is weighted by its number of impressions, which we define as the sum of likes and retweets plus one to not completely

discard tweets without any impressions. As the divergence of public sentiment could also carry meaningful information, we add the percent of positive and negative tweets per day as features. Note that these do not have to sum to 1, as there are also neutral tweets. However, as Teti et al. (2019) find that neutral tweets lack any significant predictive power, and adding them would induce multicollinearity, their percentage is not included as a feature.
Furthermore, we add several financial indicators. Most financial indicators are based on moving averages or lags. Hence, an important hyperparameter is the window- or lag size they are calculated on. Findings from Shynkevich et al. (2017) suggest that the window length for technical indicators in machine learning models should be proportional to the length of the forecast horizon. They find that for one-day-ahead forecasting window lengths up to five days work best. Indicators calculated on larger windows do not hold as much predictive power. Thus, we add 2, 3, 4, and 5-day moving averages and lags for both the return and the sentiment score. Moreover, we add the trading volume and number of tweets for each day as features. We engineer an additional feature "days_in_trend" which is the number of days a stock has been trading in an up- or downtrend. The date is deconstructed into two features: "day_of_week" and "month_of_year" to indicate the weekday and month of each data point. Therefore, the final data set for forecast modeling consists of 851 data points (one for each day from January 1st, 2019, until April 30, 2021) and 25 features.

Forecasting Models

The data used for this forecasting task constitute a time series. Such data is commonly modeled with autoregressive models, that is, models that regress a response variable onto its own historical lags. These models are linear but can exploit autocorrelations that might be present in the dependent variable. As we do not only want to regress returns on their historical values but want to add sentiment indicators as predictors, we use a vector autoregression model with exogenous variables (VARX). An important hyperparameter for any VARX model is the *order* of the model which describes how many lags of the endogenous variables are used as predictors. We build the VARX model on the combined train and validation set for each order in $\{1, 2, \dots, 8\}$ and use the Akaike Information Criterion (AIC) to assess which order yields the best model (Liew, 2004). The AIC trades off the model performance (i.e., error) against model complexity such that for two equally well-performing models the one with fewer parameters is chosen. Due to numerical limitations, the VARX model only regresses the vector $Y = \langle \text{return}, \text{sentiment} \rangle$ on its own lags and on $X = \langle \text{num\_tweets}, \text{pct\_pos}, \text{pct\_neg}, \text{volume} \rangle$ as exogenous variables. Adding more features as endogenous variables was not possible as the model would not converge anymore.
Finally, we compare two types of machine learning models: a random forest (Breiman, 2001) and an implementation of gradient boosted trees called LightGBM (Ke et al., 2017). Both methods use decision trees as their base classifiers but combine their predictions in different ways. Random Forests use bootstrap aggregating (bagging). Its trees are built independently on different bootstrapped resamples of the training data set. At each split in each tree, only a subset of the features is considered for the splitting criterion, thereby decorrelating the decision trees. The final prediction is the majority vote cast by all trees. Ballings et al. (2015) find a random forest to perform best in a stock forecasting problem based on fundamental and economic data, as it can learn non-linear associations as well as feature interactions.
In contrast to this, LightGBM uses boosting: Its trees are built sequentially where each tree tries to predict the error of the cumulative predictions of the previous trees. In between steps, the data set is reweighted to assign more importance to instances that are hard to classify. We choose LightGBM over other gradient boosting implementations like XGBoost or CatBoost, as it achieves similar performance at a fraction of the time (AlShari et al., 2021). Weng et al. (2018) suggest that boosted trees work best for short-term stock price prediction based on technical indicators, with Random Forests being close runners-up.

Random Forests but especially LightGBM models have several important hyperparameters to tune. Generally, they fall into three categories: parameters controlling the complexity of individual trees, parameters for decorrelating individual trees, and parameters influencing the learning process. The complexity of individual trees needs to be controlled to avoid overfitting. Random Forests tend to use fully-grown decision trees while boosting algorithms employ simpler trees, sometimes even so-called stumps (trees with only a single split). These characteristics can be controlled with the parameters max depth, minimum samples per leaf, and cost-complexity pruning alpha (CCP alpha) for Random Forests and alpha, lambda, max depth, and the number of leaves for LightGBM. For details on the LightGBM-specific parameters alpha and lambda see Ke et al. (2017). Besides controlling their complexity, individual trees in each model should not strongly correlate with one another. If this were the case, they would make similar mistakes, and the advantages of bagging or boosting disappear. We can decorrelate trees by controlling the maximum percentage of features used per split for Random Forests or the subsample parameter for LightGBM. Finally, the learning process for LightGBM is controlled using different boosting types like gradient-boosted decision trees, gradient-based one-side sampling (Ke et al., 2017), or "Dropouts meet Additive Regression Trees" (Rashmi & Gilad-Bachrach, 2015) as well as setting the learning rate and the number of estimators. We do not tune the number of estimators for random forests and use 100 as the standard-setting, as for a large number of estimators overfitting does not occur but the marginal decrease in variance diminishes quickly and computation time at $\text{n\_estimators} = 100$ is acceptable. For LightGBM however, this is an important hyperparameter as it can induce overfitting if set too high. Table 3 gives an overview of parameters and their corresponding search spaces. They are explored using random search, as it has been shown to be more efficient than grid search or manual search (Bergstra & Bengio, 2012).

	Parameter	**Search Space**
Random Forest	Max depth	$\text{integer} \in \{2, 5, 8, \dots, 32\}$
	Min samples per leaf	$\text{integer} \in \{1, 2, \dots, 20\}$
	Max features per split	$\text{float} \in [0, 1]$
	CCP alpha	$\text{float} \in [0, 5]$
LightGBM	Boosting Type	$\text{string} \in \{\text{gbdt}, \text{dart}, \text{goss}\}$
	Alpha	$\text{float} \in [0, 25]$
	Lambda	$\text{float} \in [0, 25]$
	Subsample	$\text{float} \in [0, 1]$
	Max depth	$\text{integer} \in \{2, 3, \dots, 32\}$
	Learning rate	$\text{float} \in [1 \times 10^{-2}, 0.5]$
	Number of leaves	$\text{integer} \in \{2, 3, \dots, 32\}$
	Number of estimators	$\text{integer} \in \{2, 3, \dots, 300\}$

Table 3: Hyperparameters and their search space for both tree-based models

Model Evaluation

In finance, forecasting the direction of future security price movements offers more value than forecasting the magnitude as most trading strategies are based on directional assumptions. Thus, we will evaluate all models on their predictive accuracy. This also applies to the VARX models. While usually, predictive regression models are evaluated on their mean squared error or similar metrics, this does not acknowledge the importance of the binary threshold which separates returns (and hence trading profits) into positive and negative. The accuracy must not be calculated on the same data set used for training the model as it would yield a heavily positively biased estimate of model performance. Especially more complex models like bagged or boosted trees would be able to memorize training examples, which yields low training error but does not generalize well. To obtain a reliable estimate, we split the data set into three parts to train, validate and tune hyperparameters, and finally obtain an

out-of-sample test performance estimate. As the data constitute a time series and are not independent and identically distributed, these splits must not be random, but sequential. Using January 2019 until October 2020 as train data, November 2020 until January 2021 as validation data and the remaining days as test data creates train, validation, and test set sizes of 670, 92, and 90 respectively. To establish a naïve baseline to compare all models against, we record prediction accuracies for simply predicting the majority class, where the majority class is calculated on the combined training and validation set. To distinctly examine whether the sentiment scores hold predictive power, we repeat the modeling process on a data set that does not contain any sentiment-based features, i.e., the sentiment score, percentage of positive and negative tweets as well as the number of tweets are dropped.

RESULTS

Sentiment Analysis

The VADER compound sentiment score is a real number in $[-1; 1]$. Figure 4 displays the distribution of this score for each of the ten companies. It is evident that the distributions between companies are almost identical (hence, we choose to not visualize companies with different colors). Moreover, most tweets have a sentiment score of zero or close to zero. This artifact either stems from the fact that not all tweets carry a clearly defined sentiment, or VADER is unable to detect some sentiment-laden tweets and incorrectly classifies them as neutral in the absence of any known words. Additionally, the greater density for positive scores in the visualization reveals that VADER identifies more positive than negative Tweets.

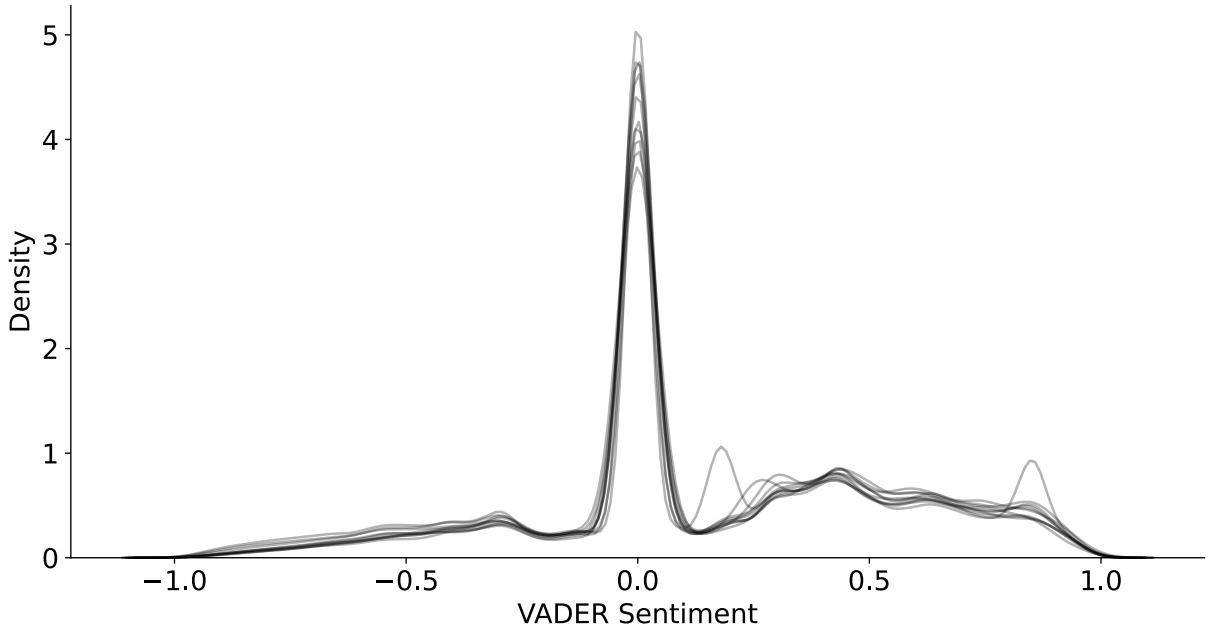


Figure 4: Distribution of VADER sentiment score per tweet for each of the ten companies

To gauge the reliability of VADER's sentiment estimates, we benchmark it on the corpus of manually labeled tweets. To obtain a class assignment from the continuous VADER scores, the thresholds -0.05 and 0.05 are used in accordance with the recommendation of Hutto and Gilbert (2014). This reveals that VADER significantly outperforms a random guessing strategy. Random guessing the sentiment class (negative, neutral, positive) based on their prior probabilities yields a mean accuracy of 36.7% (SD=0.87%p). VADER achieves a significantly higher accuracy of 43.9% (SD=0.9%p), although the magnitude of the improvement is rather small. Examining the confusion matrix presented in Table 4 reveals that VADER consistently overestimates the number of positive tweets. Therefore, the positivity skew observed in figure 4 might also be caused by a bias in VADER's estimates on this specific corpus of tweets.

	VADER		
label	**-1**	**0**	**1**
-1	**184**	143	228
0	193	**505**	578
1	151	395	**623**

Table 4: Confusion matrix of VADER sentiment estimates

We proceed with training a logistic regression model on the corpus of manually labeled tweets. Random search results calculated using a 5-fold cross-validation procedure indicate the optimal hyperparameter setting is $C = 1.192$. With this setting, the model achieves an out-of-sample accuracy of 59.8% which outperforms VADER by a large margin. The logistic regression is refit on the entire labeled data set using this hyperparameter value. A diagnosis of model coefficients reveals that it was able to learn several domain-specific terms. Table 5 displays the 20 words with the largest regression coefficients for each of the three sentiment classes. These words most strongly influence predictions in favor of each of the classes. While some words carry sentiment in the general language (e.g. avoid, weak, nice, strong) others are very finance-specific (e.g. bearish, red, call, put, bullish). Lexicon-based approaches like VADER are expected to perform well on the former but fail on the latter type of texts. However, it has to be acknowledged that although having learned very domain-specific vocabulary, the trained logistic regression model would probably not generalize well to other corpora of text due to the very small training set.

Negative	Neutral	Positive
• shit • exit • 🔻num.num • whi • lower • investing • miss • dump • cut • tech • avoid • drop • short • weak • sell • sold • put • red • bearish • down	• would • near • snap • analysi • risk • app • exampl • join • w • option • best • free • what • new • volum • chart • trader • or • tri • ticker	• well • return • higher • beat • abov • jump • up • green • upgrad • nice • strong • grow • numc • rais • long • high • bought • buy • call • bullish

Table 5: Words with largest regression coefficients for each class ("num" replaces numbers)

Finally, to obtain sentiment estimates for the rest of the tweets that have not been labeled, the trained logistic regression model is applied to the entire data set. The ratio of tweets in each sentiment class over all ten companies is displayed in figure 5. It appears that most tweets are classified as neutral or positive, while only a few ($\approx$ 10% on average) are classified as negative.

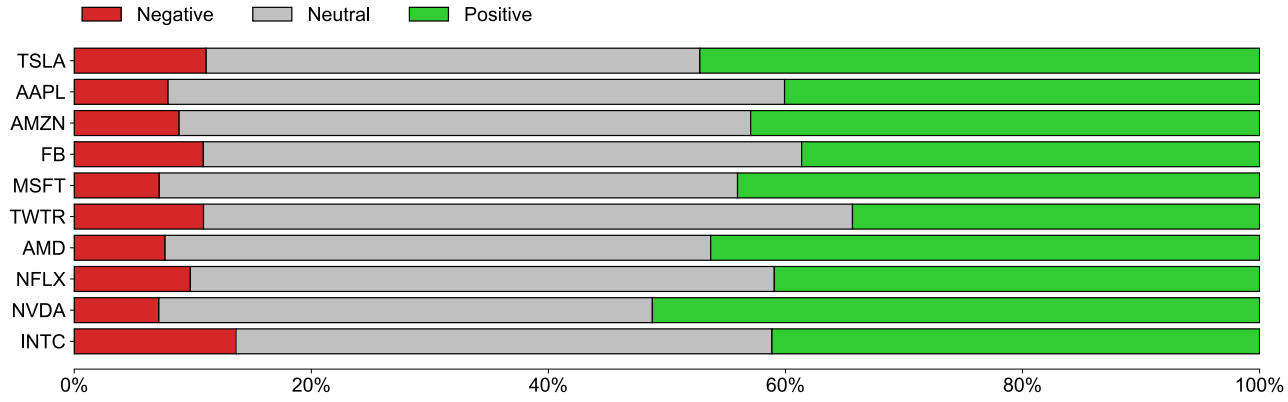


Figure 5: Sentiment class ratio over all 10 companies

Baseline Model

To gauge the predictive performance of any model, a baseline must be established first. It might be the case that the magnitude of a metric seems impressive (e.g., 99% accuracy for credit fraud) when it

turns out that a naïve approach (e.g., predicting the majority class) performs even better. As our one-day-ahead forecast is a binary classification problem, we build a baseline model which simply predicts the majority class. The majority class is calculated on the combined training and validation data. Metrics reported are calculated on the test data set. Table 6 presents the accuracy such a naïve approach achieves on the test data. It illustrates the importance of such a baseline: An accuracy above 60% for Nvidia Corp. appears to be the performance of a remarkable model, assuming market returns are equally distributed between positive and negative. However, in the case of Nvidia, the base distribution is so skewed that simply predicting "positive" all the time yields this result.

Ticker	**TSLA**	**AAPL**	**AMZN**	**FB**	**MSFT**	**TWTR**	**AMD**	**NFLX**	**NVDA**	**INTC**
Accuracy	0.494	0.529	0.552	0.54	0.506	0.552	0.483	0.575	0.609	0.506

Table 6: Accuracy of naïve prediction as a benchmark for each company

All real models will subsequently be compared against these baseline accuracies, and their performance will only be acknowledged when it beats the baseline. If a model fails to beat the baseline, we can assume it has not learned any meaningful, generalizable pattern.

VARX Modeling

One major assumption of the vector autoregression framework with exogenous variables is the stationarity of the endogenous and exogenous time series. Stationary time series are required to estimate accurate regression coefficients and their standard errors. To test the time series for their stationarity, we use the Augmented Dickey-Fuller test (ADF test) proposed by Dickey & Fuller (1979). It tests the null hypothesis H_0: the time series has a unit root and thus is non-stationary against the alternative hypothesis H_A: the time series does not have a unit root. Table 7 displays the p-values of the ADF test for each time series.

	TSLA	**AAPL**	**AMZN**	**FB**	**MSFT**	**TWTR**	**AMD**	**NFLX**	**NVDA**	**INTC**
vader	0.0	0.16^{NS}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
- pct_pos	0.327^{NS}	0.414^{NS}	0.003	0.001	0.015	0.0	0.017	0.001	0.0	0.0
- pct_neg	0.628^{NS}	0.067^{NS}	0.0	0.001	0.0	0.0	0.002	0.0	0.0	0.0
ml_sentiment	0.0	0.069^{NS}	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
- pct_pos	0.173^{NS}	0.391^{NS}	0.003	0.262^{NS}	0.0	0.023	0.076^{NS}	0.0	0.002	0.0
- pct_neg	0.073^{NS}	0.001	0.01	0.062^{NS}	0.071^{NS}	0.0	0.0	0.031	0.002	0.0
volume	0.059^{NS}	0.016	0.002	0.001	0.034	0.0	0.0	0.0	0.018	0.0
num_tweets	0.176^{NS}	0.066^{NS}	0.018	0.001	0.029	0.0	0.006	0.0	0.002	0.0
return	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
label	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 7: p-values of ADF test for each time series. ^{NS} = not significant at the $\alpha = 0.05$ level

Both sentiment time series are stationary for all companies except Apple Inc. The time series of stock returns are stationary as the return is the first difference of daily closing prices, which is why we use returns in the first place. Finally, several of the exogenous variables are not stationary for some companies, especially Tesla and Apple. Despite this, we will apply a VARX model for all companies, although this might impair the estimation of the coefficients' standard errors. As this study is not explanatory but predictive in nature, it does not rely on accurate standard error estimates of regression coefficients. Instead, the predictive performance is benchmarked on a hold-out set. For a detailed discussion comparing explanatory modeling with predictive modeling and the resulting influence on

the modeling process, see Shmueli (2010). Nevertheless, we do acknowledge that this might lead to convergence errors when building the models.
To find the appropriate order of the VARX model, we plot the autocorrelation of the returns and sentiment variables (see appendix A-C). They suggest orders from one through eight should be tested, as most autocorrelation occurs at these lags. Fitting the models and testing these lags reveals that most models work best using a two- or three-day order according to the model AIC. The combination of VADER sentiment used on the Intel data is an outlier and works best with an order of 5, which on the daily data coincides with the length of one whole trading week. Table 8 displays the best-performing order for each model.

	TSLA	**AAPL**	**AMZN**	**FB**	**MSFT**
ml_sentiment	2	2	2	2	2
VADER	2	2	2	3	2
	TWTR	**AMD**	**NFLX**	**NVDA**	**INTC**
ml_sentiment	2	3	2	2	2
VADER	2	3	2	3	5

Table 8: Best-performing order p for each VARX(p) model

Subsequently, the models are evaluated on the test set. The corresponding accuracy scores are displayed in table 9. Only for five out of ten companies can a model that outperforms the baseline be found. In these cases, the linear model outperforms the baseline by several percentage points. An exception is Intel: Using VADER sentiment, the model was able to achieve an accuracy improvement of more than 11%p.

ticker	**ml_sentiment**	**VADER**	**Baseline**
TSLA	0.517	**0.54**	0.494
AAPL	0.529	0.529	0.529
AMZN	0.494	0.46	0.552
FB	0.425	0.437	0.54
MSFT	**0.517**	**0.529**	0.506
TWTR	0.414	0.425	0.552
AMD	**0.529**	**0.506**	0.483
NFLX	**0.621**	0.54	0.575
NVDA	0.529	0.448	0.609
INTC	0.54	**0.621**	0.506

Table 9: Test set accuracies for VARX models

Figure 6 visualizes the performance improvement over the baseline for each VARX model (sorted by improvement of machine learning sentiment from left to right). For the five companies for which models better than the baseline exist, the models using VADER sentiment seem to perform slightly better than the models using the machine learning sentiment: Three out of the five working models use VADER sentiment, only two use the custom sentiment scores. However, besides Intel, the overall performance improvement is small in magnitude. The VARX models for Amazon, Nvidia, Facebook, and Twitter perform worse than the corresponding baselines by a large margin, indicating that even linear VARX models can overfit the training data and not generalize well.

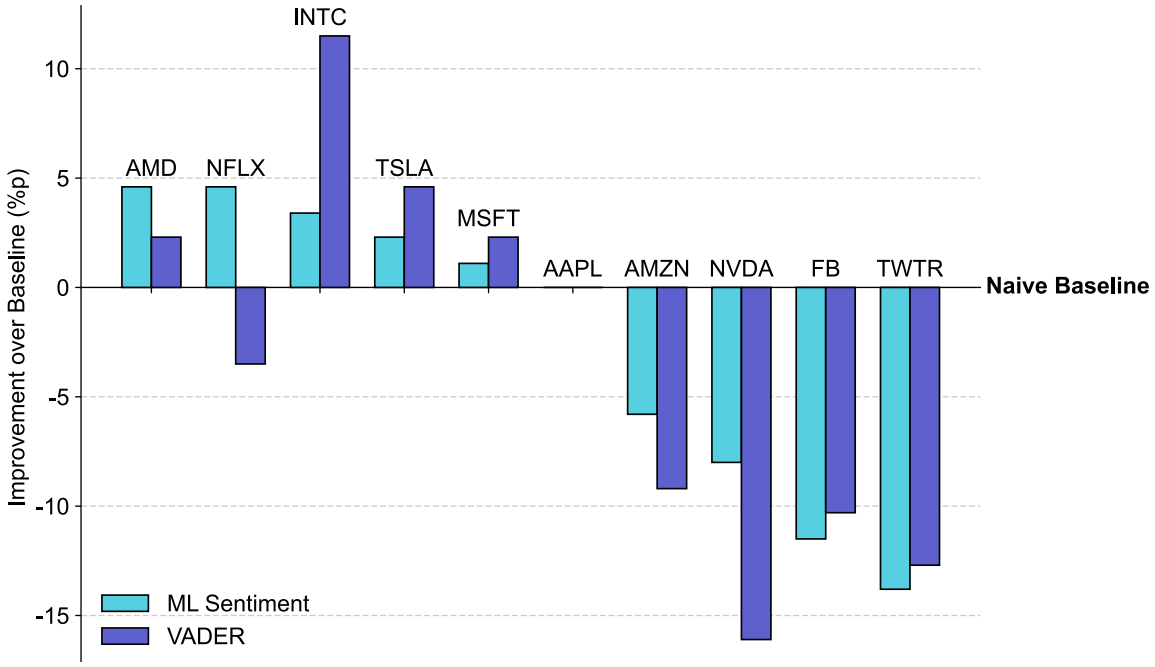


Figure 6: Accuracy improvement over baseline, colored by sentiment method

Non-linear Models

As both Random Forests, as well as LightGBMs, are ensemble models based on decision trees, they are non-parametric. Consequently, there are no assumptions that input data need to satisfy to build a valid model and we can proceed to fit the models to the data. Table 10 presents the test set accuracies for each combination of model and sentiment per ticker.

	Random Forest		LightGBM		
Sentiment:	**ml_sentiment**	**VADER**	**ml_sentiment**	**VADER**	**Baseline**
TSLA	0.478	0.478	0.444	**0.556***	0.494
AAPL	0.511	0.511	0.511	0.511	0.529
AMZN	0.511	0.50	0.467	0.511	0.552
FB	0.522	0.522	0.467	0.433	0.54
MSFT	0.489	0.489	**0.60***	0.50	0.506
TWTR	0.533	0.533	0.533	0.544	0.552
AMD	0.467	0.411	**0.544***	**0.511**	0.483
NFLX	**0.644***	0.556	**0.633**	**0.589**	0.575
NVDA	0.589	0.589	0.589	0.60	0.609
INTC	0.489	**0.589**	**0.60***	0.489	0.506

Table 10: Test set accuracy. Bold = beats baseline, * = best value for ticker

For five out of the ten companies, no model outperforms the baseline regardless of the sentiment scoring it uses. However, for the other five, such models can indeed be found. The outperformance in accuracy compared to the baseline ranges from 1.4%p for Netflix up to 9.4%p for Microsoft and Intel. When we only consider the best model for each of these four companies, each one of them outperforms the baseline by at least 6.1%p. Furthermore, the machine learning sentiment tends to produce better performing models than the VADER sentiment in all but one case. Similarly, the LightGBM model delivers better performance than the Random Forest: Only in one out of the 5 companies for which working models were found does the Random Forest outperform LightGBM.

To assess how much predictive power can be attributed to the sentiment-based variables, we repeat the modeling process, this time removing all sentiment-related features, namely the sentiment score along with its lags and moving averages, the percent of positive and negative tweets, and the number of tweets. Accordingly, the resulting forecasting models can only use the financial indicators described earlier. The performance of these models is displayed in Table 11.

Ticker	Random Forest	LightGBM
TSLA	0.478	0.456
AAPL	0.511	0.467
AMZN	0.489	0.467
FB	0.522	0.5
MSFT	0.489	0.5
TWTR	0.533	0.533
AMD	0.467	**0.556***
NFLX	0.556	0.511
NVDA	0.589	**0.667***
INTC	**0.611***	**0.567**

Table 11: Accuracy of models *without* sentiment. Bold = beats baseline, * = best for ticker

To compare the model performance with and without sentiment, we compare the best model using sentiment features against the best model using only financial features per company. As the absolute accuracy scores are misleading in the case of non-uniform prior class distribution, Figure 7 visualizes these scores relative to the company's baseline accuracy (sorted by "with sentiment" model performance improvement from left to right).

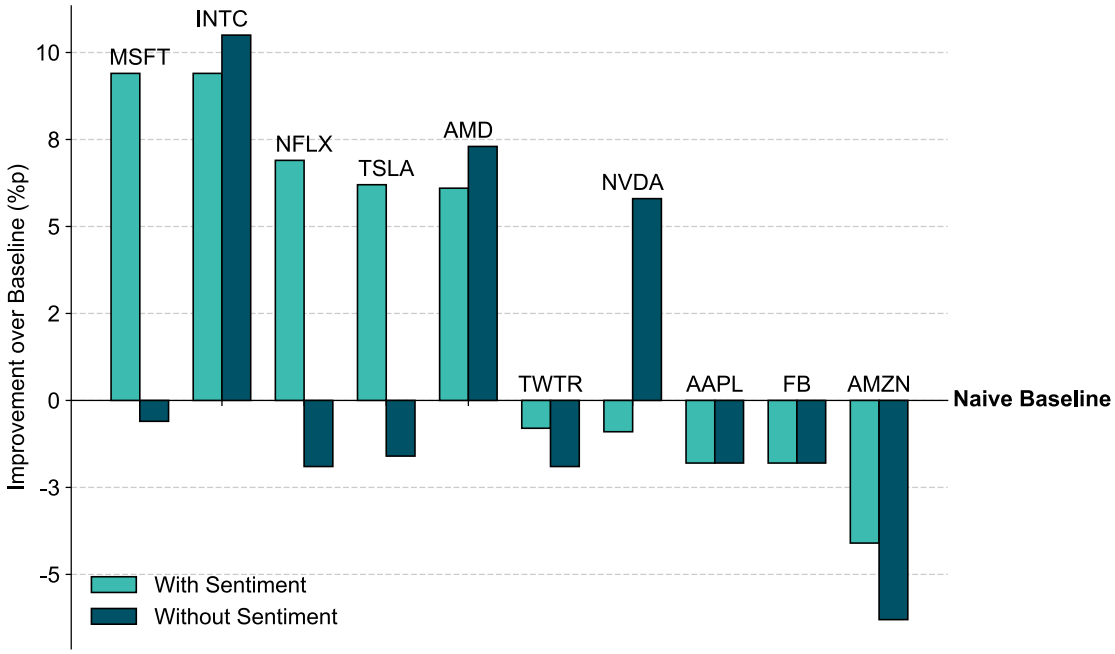


Figure 7: Accuracy improvement over baseline for the best model with and without sentiment

For six out of the ten companies, models with significant predictive power can be found, although, for Nvidia, this only holds for the model without any sentiment features. For the other four companies, comparing model performance is unnecessary, as the models have not learned any generalizable pattern that beats predicting the majority class and are thus considered inadequate. It occurs that for

Intel, the sentiment data does not seem to improve predictive power as the model without any sentiment data is more accurate. The same holds for AMD and Nvidia: while only using financial features yields a model with an accuracy of around 55.6% and 66.7% respectively, adding sentiment feature deteriorates the models' performance. In contrast to this, for Microsoft, Netflix, and Tesla, the models with sentiment features achieve significantly higher performance than their counterparts. In these cases, the sentiment data hold predictive power that can be exploited by the models.

DISCUSSION

It appears that the distribution of sentiment scores per tweet is different between VADER and the machine learning sentiment. Yet, both approaches exhibit similar phenomena: Most tweets fall into the neutral or positive category, only a minority of the tweets contain negative sentiment. As for VADER, after having benchmarked it on the labeled data set, its general positivity bias explains the large share of positive tweets. As a lexicon-based model that cannot handle finance-specific terms like "bullish" or "bearish", it is also prone to overestimating the number of neutral tweets as tweets that do not contain known words fall into this category. However, as the machine learning-based sentiment model and the labeled data themselves exhibit a larger share of positive than negative tweets, we can indeed observe a general "positivity" in the tweet data set. Similarly, in the context of economics, Soroka et al. (2018) find Twitter users to be positively biased compared to traditional news outlets, so this skew in opinion is a recognized circumstance.

When scoring both models' accuracies, VADER only outperforms a random guessing strategy by a small margin, whereas the domain-specific logistic regression model delivers significantly better performance than VADER. The problem of lexicon-based approaches handling domain-specific language is widely known and most often evaded by engineering custom models on the training corpora (e.g., Yao & Wang, 2020). Our results show why this often works: Custom models can pick up domain-specific vocabulary like "put", "bearish", "red" for the negative class, and "calls", "bullish", "buy" for the positive class. Still, the evidence suggests that among the lexicon-based sentiment analysis models VADER is the best performing one (Ahmed et al., 2017; Al-Shabi, 2020), but it is questionable as to how far generic models like VADER should be applied to very domain-specific text. Additionally, the logistic regression model could be improved even further by using a larger training corpus: Renault (2019) find that optimal results can be achieved using 10,000 to 100,000 tweets for training. Additionally, Snow et al. (2008) suggest that when outsourcing annotation to platforms like Amazon Mechanical Turk, four independent people should label all tweets for optimal reliability. Our results show that even when built on a small corpus, the custom sentiment model already outperforms off-the-shelf solutions, so investing in labeled training data pays off. It is worth mentioning that data obtained from other sources than Twitter might have to be handled in an entirely different way. The forum "StockTwits" allows users to tag their posts as bullish or bearish, thus providing large amounts of labeled data that could be used for training sentiment models. Posts on "Reddit" are often longer texts than the ones found on Twitter, sometimes containing elaborate stock analyses. Simply classifying such an analysis as one of three classes might be a too simplistic approach. Moreover, other sentiment analysis approaches that go beyond simple classification emerge in the literature. For example, Si et al. (2014) model the co-occurrence of stock tickers in tweets as a semantic network. Network-based approaches could potentially exploit relationships *between* different stocks for better predictions like association rule mining does, which is often used in recommender systems. Besides that, it is important to acknowledge that sentiment can also be expressed in the non-text form: Some Tweets contain attached images like screenshots of technical analysis charts or only consist of a single word or two. Here, all text-based sentiment models would be expected to struggle with reliably classifying such content. Other creators post their opinions as videos on YouTube potentially influencing a large audience but being hard to analyze in an automated fashion.

After developing the forecasting models, the results reinforce the notion that before building statistical forecasting models of any kind, a baseline benchmark must be established. Surprisingly, for one stock

(NVDA) this baseline accuracy is already around 61%. This can be explained by the long uptrend the stock has experienced: Predicting a positive return for every day is already a good guess. Despite the skewed class distribution for some stocks, accuracy is favorable as a metric to make results comparable to other studies. It is worth noting, however, that other metrics are sometimes used when solving the forecasting problem in a regression setting. Unfortunately, this ignores the important threshold which separates returns into positive and negative: Most strategies require a correct directional assumption to be profitable. Fitting both linear VARX models as well as non-linear tree-based models mainly puts forth three findings:

1. The social sentiment holds predictive power for several companies studied.
2. Predictions based on the custom machine learning sentiment perform better than ones based on VADER sentiment.
3. Linear models perform worse than non-linear models.

Overall, Microsoft, Netflix, and Tesla are the only three companies for which sentiment holds predictive power, that is, there is a model using sentiment features that is not only better than the benchmark but also better than its counterpart without sentiment-based features. When adding sentiment features to the model, we observe a moderate increase in predictive accuracy of around 5%p to 9%p. This effect is not quite as large as the 18%p increase Ren et al. (2018) find, but still cannot be attributed to random chance. For three other companies, Intel, AMD, and Nvidia, predictive models exist, but they do not use sentiment-based features. Finally, for the remaining four companies we were unable to find any predictive model with or without sentiment. This raises the question of whether some stocks possess inherent unpredictability, as all models failed for the same stocks. While previous research could not yet identify circumstances under which return prediction works particularly well, it has been shown that return forecasting only works for a minority of stocks: Experiments conducted by Coqueret (2020) show that a significant relationship between sentiment and return exists only for 7% of the companies in the study. This also confirms statements by Mudinas et al. (2019) who suggest that the value of sentiment for return forecasting needs to be examined on a case-by-case basis as no generalizable pattern exists. However, these results also contradict two classes of findings previously prevalent in the literature: (1) studies that do not find sentiment to be predictive at all and (2) studies that find social sentiment to be predictive with an accuracy close to 100%. As for (1), there are many factors in study design that influence whether sentiment can be considered to hold predictive power. However, the results obtained in this project suggest that (a) one should always consider multiple different companies or indices as predictive performance can vastly differ between them, and (b) the method of sentiment analysis can have a severe impact on whether the obtained sentiment scores can be used for prediction, where custom models often work better than generic off-the-shelf models. As for (2), it is unlikely that any forecasting model can achieve exceptional forecasting accuracy over extended periods of time. Stock markets still possess a large amount of inherent randomness, and well-performing predictive models can often be made obsolete by their application, as it makes the patterns they use for prediction disappear. Moreover, some studies assess model performance on the training data or very short periods of testing data which yields unreliable results. Another research approach that has been reported to deliver promising results is conducting studies on hourly rather than daily data sets (Deng et al., 2018). As sentiment and financial indicators are expected to predict short-term stock movements better than long-term returns (which are more driven by fundamentals) anyway, this provides researchers with more data points. This is especially helpful considering the advent of new machine learning and deep learning approaches that need large amounts of data to generalize well. The increased number of data points would also make the performance evaluation more reliable.
Regarding the sentiment analysis technique, for most of the companies (except for Tesla), the best model is produced by using the custom machine learning sentiment, not VADER. While the VARX

models showed a contrary pattern where VADER seems to work slightly better than the machine learning sentiment, most of the best-performing models use the custom sentiment scores. This indicates that better sentiment assessment can lead to better forecasting performance. Therefore, especially considering the bad performance of VADER on domain-specific texts, researchers should devote more resources towards the process of sentiment analysis and should carefully consider the use of off-the-shelf models trained on generic texts.

Examining the final performance of the VARX models on the test set shows that the magnitude of the performance improvement is small with only several percentage points for all companies but Intel. While fitting the VARX model it is evident that for most companies, an order (number of lags included in the model) around two is optimal. This confirms findings from Shynkevich et al. (2017) that for day-ahead forecasting only a few days of lag are needed.

Evidence from previous studies suggests that non-linear models work best for complex forecasting tasks like stock return prediction (Ballings et al., 2015; Weng et al., 2018). The modeling results from the Random Forest and LightGBM confirm these statements. Predictive models better than baseline can be found for the exact same five companies for which the VARX models also worked. However, the effect size is much larger: The non-linear models beat the baseline by more than 6%p.

Despite not working for all companies, the fact that multiple predictive models *without* sentiment features could be developed provides evidence against the Efficient Market Hypothesis. According to the weak-form EMH expressed in equation (1), no historic price information available at time t can be used to forecast future returns. The results show that for Intel, AMD and Nvidia however, such predictability can be found. Moreover, discovering another three models that successfully exploit sentiment features to predict future return delivers further evidence against the EMH in its semi-strong form: According to the semi-strong form EMH, publicly available information is priced into security prices and thus should not be able to predict future returns. While Fama (1970) obviously did not consider social media as a channel of information, opinions and emotions shared on such platforms should not be able to predict or influence stock returns, as they do not constitute "information generating event[s]" (Fama, 1970, p. 404) like stock splits or earnings announcements. These findings add to the body of research that suggests the EMH might not be entirely applicable to modern financial markets. Shiller (2003) describes the phenomenon of excess volatility: Stock markets are more volatile than one would expect under the EMH, indicating that other factors than newly published information drive short-term movements. Another well-known behavior pattern in financial markets is "herding": Investors tend to act similarly to one another, both intentionally and unintentionally (Spyrou, 2013). This can be amplified by social media platforms on which everyone can read about decisions other retail investors or influencers make with little time delay. Bizzi and Labban (2019) show that heavy social media use is associated with a significantly larger susceptibility to blindly following the trading choices of others compared to light social media use. This issue can also be exploited with bad intent: Sabherwal et al. (2011) suggest that online message boards can be used to use investors' herding behavior for pump-and-dump schemes. In such cases, social sentiment could not only be used for the forecasting of returns, but also for monitoring fraudulent attempts to manipulate markets. For all these empirically observed phenomena the EMH fails to acknowledge, Behavioral Finance offers better explanations as it recognizes the biases humans underly in the process of making decisions. Especially with the growing number of retail investors who lack formal education about financial markets and might not act rationally, the theories of Behavioral Finance should be considered by practitioners and regulators.

As with any research, this study is subject to several limitations. It assumes that Twitter is a good proxy for public sentiment, whereas there are other social networks and platforms on which users can discuss stock markets and share their opinions. As Twitter is a predominantly English-speaking platform, the sample of companies studied only includes large-cap companies from the United States. Results thus need not generalize to other markets, although similar studies have been conducted for Chinese markets (Ren et al., 2018; Li et al., 2020) and cryptocurrency markets (Xie, 2021). Moreover,

social sentiment can be operationalized in many ways. Here, we classify tweets into one of three classes and aggregate metrics on a daily level. Other approaches include measuring different types of emotions (Bollen et al., 2011) or using network-based approaches. Additionally, the corpus-specific sentiment analysis model – while outperforming VADER – was still built on a rather small corpus. Using more training data and multiple raters for labeling data could potentially make it more accurate. Finally, the results indicate that complex, non-linear forecasting models work better than simple ones. The application of Deep Learning to this data might thus yield further performance improvements, although more data points might be needed. Therefore, future research could experiment with different types of sentiment analysis techniques and forecasting models, ideally using samples of multiple different companies or indices. This could contribute to finding a pattern in which companies' stock movements are more predictable than others. Besides that, before using any findings in a real-world trading strategy, a backtest for return on investment (including fees) should be conducted to compare any complex strategies performance against simple buy and hold approaches.

CONCLUSION

In this project, we examined whether social sentiment holds predictive power regarding short-term stock returns. To conduct the study, we scraped 28 months' worth of Tweets for the 10 most talked-about companies on Twitter by their cashtag. Two different sentiment analysis methods were applied to classify tweets as positive, negative, or neutral: VADER, a lexicon-based prebuilt model, and a custom-built sentiment analysis classifier built on a corpus of 3,000 manually labeled tweets. The sentiment scores were aggregated on a daily level. Subsequently, linear VARX and non-linear tree-based models were used to predict the following days' binarized return (positive or negative) from daily sentiment scores and financial indicators. We find that for three out of ten companies, sentiment does hold predictive power in that the forecasting model with sentiment outperformed both its counterpart without sentiment and a naïve baseline. Mostly, the sentiment obtained from the custom-trained model performs better for return forecasting, as can classify sentiment significantly more accurately than VADER. The large accuracy improvements over a naïve baseline provide evidence against the Efficient Market Hypothesis under which no forecasting model should outperform random guessing. It shows that the emotions and opinions of the crowd of retail investors can provide valuable information and suggests that Behavioral Finance, which accounts for human traits in decision-making processes, might be a more applicable theory than the EMH. Furthermore, the observed accuracy improvement adds to the body of research regarding the prediction of stock returns: It confirms that sentiment can be used as a feature that improves predictive power, but not for all stocks. Rather, it only works for a minority of stocks, which both confirms prior research in the field and might explain why some studies which only analyze a single or very few companies or indices conclude that sentiment cannot be used for predictive modeling. Our results warrant future research in the domains of sentiment analysis and stock return forecasting, implying larger samples need to be used to further examine circumstances that are conducive to predictable returns.

REFERENCES

1. Ahmed, T., Bosu, A., Iqbal, A., & Rahimi, S. (2017, October). SentiCR: a customized sentiment analysis tool for code review interactions. In *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)* (pp. 106-111). IEEE.
2. Al-Shabi, M. A. (2020). Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. *IJCSNS*, *20*(1), 1.
3. Alshari, H. H., Saleh, A. Y., & Odabas, A. (2021). Comparison of Gradient Boosting Decision Tree Algorithms for CPU Performance. *Journal of Institue Of Science and Technology*, *37*(1).
4. 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.
5. Audrino, F., Sigrist, F., & Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, *36*(2), 334-357.
6. Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances.
7. Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert systems with Applications*, *42*(20), 7046-7056.
8. Bergmeir, C., Costantini, M., & Benítez, J. M. (2014). On the usefulness of cross-validation for directional forecast evaluation. *Computational Statistics & Data Analysis*, *76*, 132-143.
9. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).
10. Bizzi, L., & Labban, A. (2019). The double-edged impact of social media on online trading: Opportunities, threats, and recommendations for organizations. *Business Horizons*, *62*(4), 509-519.
11. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1-8.
12. Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.
13. Brokernotes (2018). The modern trader. Retrieved from https://brokernotes.co/wp-content/uploads/2017/08/BN-research-report_2018-FINAL.pdf (access: May 25th, 2021).
14. Burton, K., Parmar, H. (January 27th, 2021). Reddit Crowd Bludgeons Melvin Capital in Warning to Industry. *Bloomberg*. Retrieved from https://www.bloomberg.com/news/articles/2021-01-27/bros-on-reddit-bludgeon-melvin-capital-in-warning-to-wall-street (access: June 12th, 2021)
15. Clapham, B., Siering, M., & Gomber, P. (2019). Popular news are relevant news! How investor attention affects algorithmic decision-making and decision support in financial markets. *Information Systems Frontiers*, 1-18.
16. Coqueret, G. (2020). Stock-specific sentiment and return predictability. *Quantitative Finance*, *20*(9), 1531-1551.
17. De Bondt, W. F., Muradoglu, Y. G., Shefrin, H., & Staikouras, S. K. (2008). Behavioral finance: Quo Vadis?. *Journal of Applied Finance (Formerly Financial Practice and Education)*, *18*(2).
18. Deng, S., Huang, Z. J., Sinha, A. P., & Zhao, H. (2018). The interaction between microblog sentiment and stock return: An empirical examination. *MIS Quarterly*, *42*(3), 895-918.
19. Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, *74*(366a), 427-431.
20. Fama, E. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance, 25*(2), 383-417. doi:10.2307/2325486

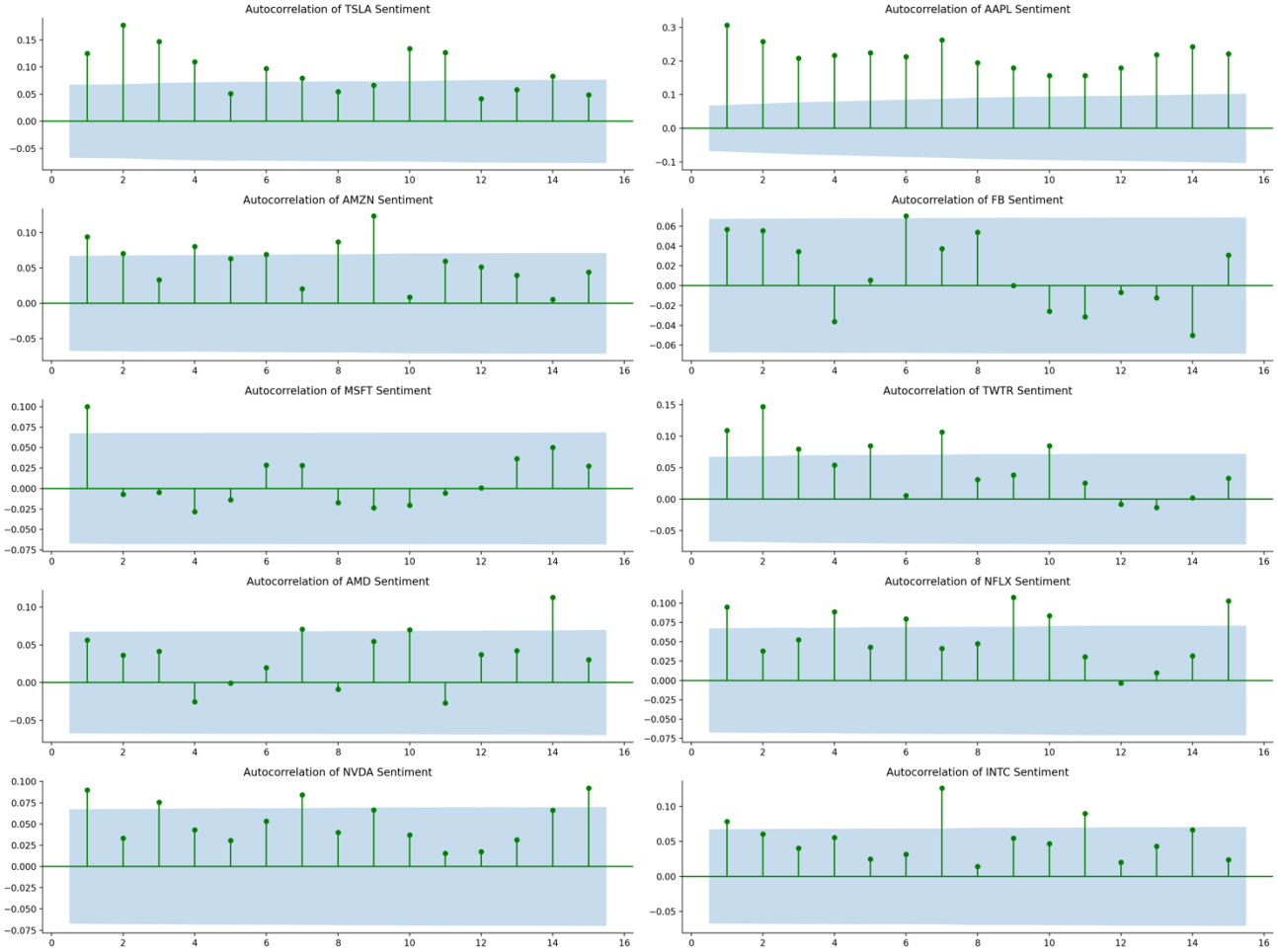
21. Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The adjustment of stock prices to new information. *International economic review*, *10*(1), 1-21.
22. Glaser, M., Nöth, M., & Weber, M. (2004). Behavioral finance. *Blackwell handbook of judgment and decision making*, 527-546.
23. Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 8, No. 1).
24. Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263-292.
25. Karabulut, Y. (2013, October). Can Facebook predict stock market activity?. In *AFA 2013 San Diego Meetings Paper*.
26. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*, 3146-3154.
27. Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. (2020). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, 1-24.
28. Li, Y., Bu, H., Li, J., & Wu, J. (2020). The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning. International Journal of Forecasting, 36(4), 1541-1562.
29. Liew, V. (2004). Which Lag Selection Criteria Should We Employ?. Economics Bulletin. 3. 1-9.
30. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.
31. Lorie, J. H., & Niederhoffer, V. (1968). Predictive and statistical properties of insider trading. *The Journal of Law and Economics*, *11*(1), 35-53.
32. Manuca, R., & Savit, R. (1996). Stationarity and nonstationarity in time series analysis. *Physica D: Nonlinear Phenomena*, *99*(2-3), 134-161.
33. Mittal, A., & Goel, A. (2012). Stock prediction using Twitter sentiment analysis. Standford University, CS229 (2011 http://cs229. Stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf), 15.
34. Mudinas, A., Zhang, D., & Levene, M. (2019). Market trend prediction using sentiment analysis: lessons learned and paths forward. *arXiv preprint arXiv:1903.05440*.
35. Neely, C. J., Rapach, D. E., Tu, J., & Zhou, G. (2014). Forecasting the equity risk premium: the role of technical indicators. *Management Science*, *60*(7), 1772-1791.
36. Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). Predicting stock market price movement using sentiment analysis: Evidence from Ghana. Applied Computer Systems, 25(1), 33-42.
37. Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume, and survey sentiment indices. *Expert Systems with Applications*, *73*, 125-144.
38. Open Knowledge Foundation (2021). S and P 500 Companies. Retrieved from https://github.com/datasets/s-and-p-500-companies.git (access: May 6th, 2021).
39. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of Machine Learning Research*, *12*, 2825-2830.

40. Pimpalkar, A. P., & Raj, R. J. R. (2020). Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, *9*(2), 49-68.
41. 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.
42. Rao, T., & Srivastava, S. (2014). Twitter sentiment analysis: How to hedge your bets in the stock markets. In State of the art applications of social network analysis (pp. 227-247). Springer, Cham.
43. Rashmi, K. V., & Gilad-Bachrach, R. (2015, February). Dart: Dropouts meet multiple additive regression trees. In *Artificial Intelligence and Statistics* (pp. 489-497). PMLR.
44. Ren, R., Wu, D. D., & Liu, T. (2018). Forecasting stock market movement direction using sentiment analysis and support vector machine. *IEEE Systems Journal*, *13*(1), 760-770.
45. Renault, T. (2019). Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. Digital Finance, 1-13.
46. Ritter, J. R. (2003). Behavioral finance. *Pacific-Basin finance journal*, *11*(4), 429-437.
47. Sabherwal, S., Sarkar, S. K., & Zhang, Y. (2011). Do internet stock message boards influence trading? Evidence from heavily discussed stocks with no fundamental news. *Journal of Business Finance & Accounting*, *38*(9-10), 1209-1237.
48. Scriptstotrade (December 25th, 2019). Tweet. Retrieved from https://twitter.com/scriptstotrade/status/1209874332069052416 (access: June 15th, 2021)
49. Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of economic perspectives*, *17*(1), 83-104.
50. Shmueli, G. (2010). To explain or to predict?. *Statistical Science*, *25*(3), 289-310.
51. Shynkevich, Y., McGinnity, T. M., Coleman, S. A., Belatreche, A., & Li, Y. (2017). Forecasting price movements using technical indicators: Investigating the impact of varying input window length. Neurocomputing, 264, 71-88.
52. Si, J., Mukherjee, A., Liu, B., Pan, S. J., Li, Q., & Li, H. (2014, October). Exploiting social relations and sentiment for stock prediction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1139-1145).
53. Siganos, A., Vagenas-Nanos, E., & Verwijmeren, P. (2014). Facebook's daily sentiment and international stock markets. *Journal of Economic Behavior & Organization*, *107*, 730-743.
54. Snow, R., O'connor, B., Jurafsky, D., & Ng, A. Y. (2008, October). Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 254-263).
55. Soroka, S., Daku, M., Hiaeshutter-Rice, D., Guggenheim, L., & Pasek, J. (2018). Negativity and positivity biases in economic news coverage: Traditional versus social media. *Communication Research*, *45*(7), 1078-1098.
56. Spyrou, S. (2013). Herding in financial markets: a review of the literature. *Review of Behavioral Finance*.
57. Tabari, N., Biswas, P., Praneeth, B., Seyeditabari, A., Hadzikadic, M., & Zadrozny, W. (2018, July). Causality analysis of Twitter sentiments and stock market returns. In *Proceedings of the first workshop on economics and natural language processing* (pp. 11-19).
58. Teti, E., Dallocchio, M., & Aniasi, A. (2019). The relationship between Twitter and stock prices. Evidence from the US technology industry. Technological Forecasting and Social Change, 149, 119747.

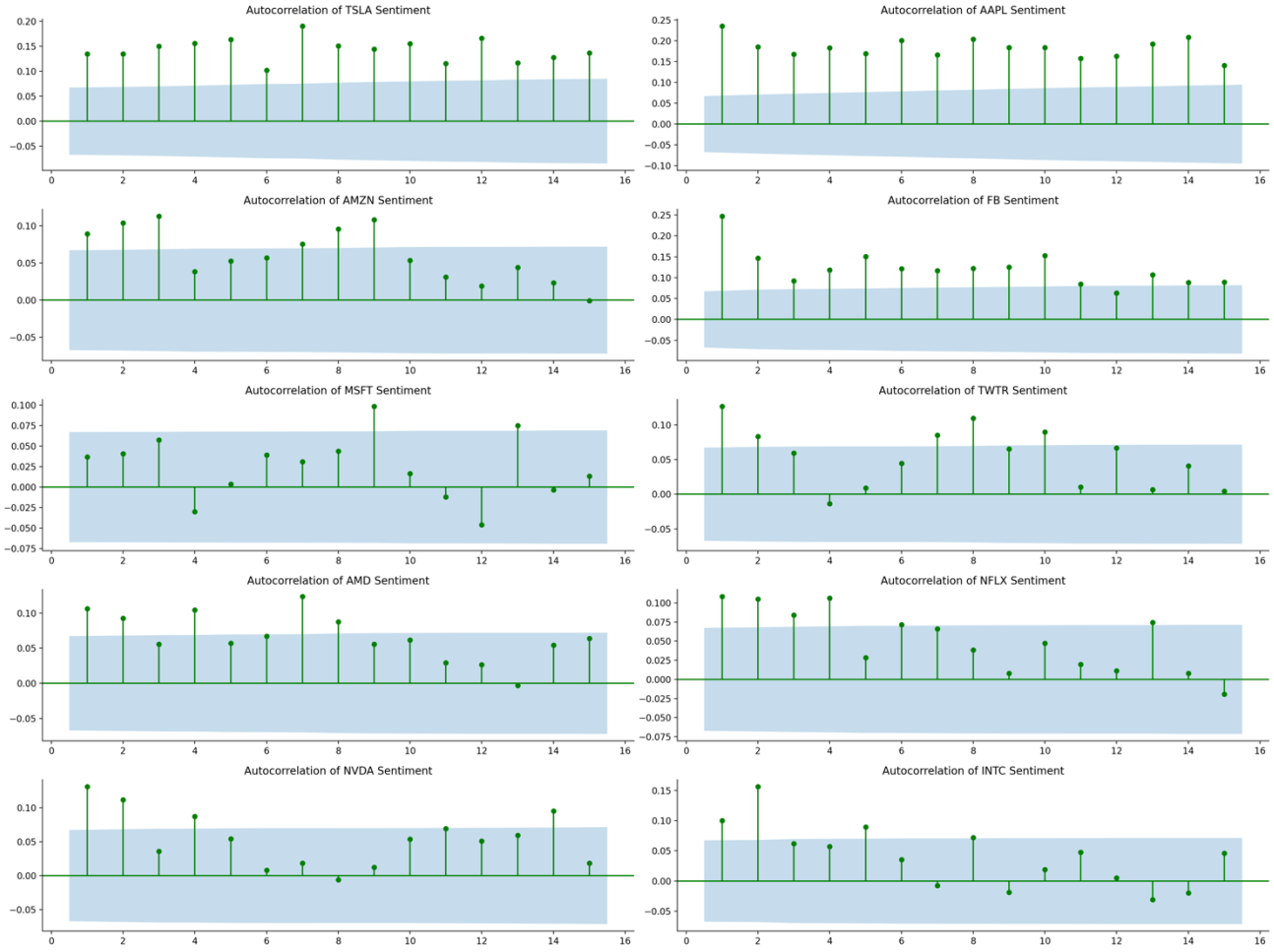
59. Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, *63*(3), 1437-1467.
60. Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124-1131.
61. Twintproject (2021). TWINT – Twitter Intelligence Tool. Retrieved from https://github.com/twintproject/twint.git (access: May 6th, 2021).
62. Twitter Inc. (2019). Selected Company Financials and Metrics. Retrieved from https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Selected-Company-Metrics-and-Financials.pdf (access: June 14th, 2021).
63. Weng, B., Lu, L., Wang, X., Megahed, F. M., & Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. *Expert Systems with Applications*, *112*, 258-273.
64. Xie, P. (2021). The Interplay Between Investor Activity on Virtual Investment Community and the Trading Dynamics: Evidence From the Bitcoin Market. *Information Systems Frontiers*, 1-17.
65. Yao, F., & Wang, Y. (2020). Domain-specific sentiment analysis for tweets during hurricanes (DSSA-H): A domain-adversarial neural-network-based approach. *Computers, Environment and Urban Systems*, *83*, 101522.
66. Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision support systems*, *55*(4), 919-926.

APPENDIX

(A) Autocorrelation of VADER Sentiment



(B) Autocorrelation of Machine Learning Sentiment



(C) Autocorrelation of Daily Returns