Stübinger, Johannes; Walter, Dominik; Knoll, Julian

Working Paper

Financial market predictions with Factorization Machines: Trading the opening hour based on overnight social media data

FAU Discussion Papers in Economics, No. 19/2017

Provided in Cooperation with:
Friedrich-Alexander University Erlangen-Nuremberg, Institute for Economics

Suggested Citation: Stübinger, Johannes; Walter, Dominik; Knoll, Julian (2017) : Financial market predictions with Factorization Machines: Trading the opening hour based on overnight social media data, FAU Discussion Papers in Economics, No. 19/2017, Friedrich-Alexander-Universität Erlangen-Nürnberg, Institute for Economics, Nürnberg

This Version is available at:
https://hdl.handle.net/10419/171230

Discussion Papers in Economics

No. 19/2017

Financial market predictions with Factorization Machines: Trading the opening hour based on overnight social media data

Johannes Stübinger
University of Erlangen-Nürnberg

Dominik Walter
University of Erlangen-Nürnberg

Julian Knoll
Technische Hochschule Nürnberg Georg Simon Ohm

ISSN 1867-6707

Friedrich-Alexander-Universität Erlangen-Nürnberg
Institute for Economics
https://www.iwf.rw.fau.de/research/iwf-discussion-paper-series/

Financial market predictions with Factorization Machines: Trading the opening hour based on overnight social media data

Johannes Stübinger[a], Dominik Walter[a], Julian Knoll[b]

[a] *University of Erlangen-Nürnberg, Department of Statistics and Econometrics, Lange Gasse 20, 90403 Nürnberg, Germany*
[b] *Technische Hochschule Nürnberg Georg Simon Ohm, Keßlerplatz 12, 90403 Nürnberg, Germany*

Tuesday 24th October, 2017

Abstract

This paper develops a statistical arbitrage strategy based on overnight social media data and applies it to high-frequency data of the S&P 500 constituents from January 2014 to December 2015. The established trading framework predicts future financial markets using Factorization Machines, which represent a state-of-the-art algorithm coping with high-dimensional data in very sparse settings. Essentially, we implement and analyze the effectiveness of support vector machines (SVM), second-order Factorization Machines (SFM), third-order Factorization Machines (TFM), and adaptive-order Factorization Machines (AFM). In the back-testing study, we prove the efficiency of Factorization Machines in general and show that increasing complexity of Factorization Machines provokes higher profitability – annualized returns after transaction costs vary between 5.96 percent for SVM and 13.52 percent for AFM, compared to 5.63 percent of a naive buy-and-hold strategy of the S&P 500 index. The corresponding Sharpe ratios range between 1.00 for SVM and 2.15 for AFM. Varying profitability during the opening minutes can be explained by the effects of market efficiency and trading turmoils. Additionally, the AFM approach achieves the highest accuracy rate and generates statistically and economically remarkable returns after transaction costs without loading on any systematic risk exposure.

Keywords: Finance, social media data, Factorization Machine, overnight information, statistical arbitrage, high-frequency trading.

Email addresses: johannes.stuebinger@fau.de (Johannes Stübinger), dominik.walter@fau.de (Dominik Walter), julian.knoll@th-nuernberg.de (Julian Knoll)

1. Introduction

Within the recent past, the internet has provided an amazing amount of information reflecting real-time sentiments and perceptions about stock companies. Therefore, academic interest in online text mining for market prediction has surged over the past years. Nassirtoussi et al. (2014) gave a comprehensive review of the existing research on this topic and pointed out that the vast majority of literature uses classification algorithms. Only a small fraction applies regression analysis for describing the interactions between the media and the stock market. Following Nassirtoussi et al. (2014), this class is confined to Tetlock (2007), Schumaker et al. (2012), Hagenau et al. (2013), Jin et al. (2013), and Chatrath et al. (2014). The seminal paper of Tetlock (2007) measured the relationship between information on social media and stock markets using an ordinary least squares regression. The results evidence that news media data contain information about movements in stock market activity. Schumaker et al. (2012) and Hagenau et al. (2013) investigated the sentiments in financial news articles and their relations to the stock market by applying support vector machines. Jin et al. (2013) made forecasts by deploying a linear regression model based on news articles, historical stock indices, and currency exchange values. Chatrath et al. (2014) examined the impact of macro news on currency jumps by a stepwise multivariate regression in a Probit model. All of theses studies are not in a position to consider the effect of overnight textual data on future price changes – an obvious deficit since information in social media, news, blogs, forums, and announcements are published 24 hours a day, 7 days a week.

In contrast to these previous approaches, this paper predicts financial markets based on overnight social media data. To be more specific, we observe tweets about the S&P 500 companies during the time span in which stock markets are closed and forecast the future price changes based on the collected information. Therefore, we build a statistical arbitrage strategy based on Factorization Machines (FMs) with different complexities, namely support vector machine (SVM), second-order FM (SFM), third-order FM (TFM), adaptive-order FM (AFM). Most notable, AFM estimates automatically all hyperparameters required by the higher-order FM model. In our backtesting framework for the years 2014 and 2015, we demonstrate the efficiency of FMs in general and discover that increasing the complexity of FMs causes better performance – annualized returns after transaction costs range between 5.96 percent for SVM and 13.52 percent for AFM. Moreover, AFM achieves the highest accuracy rate with a value of 61.76 percent and possesses returns which are resistant to the impact of bid-ask spreads and do not show loadings on systematic sources of risk.

To gain more insight into this study, the rest of this paper is organized as follows. Section 2 outlines the concept of FMs. Data sample and software are described in section 3. In section 4, we outline the study design of our back-testing framework. Section 5 presents empirical results and key findings. Finally, section 6 concludes and summarizes directions for further research.

2. Factorization Machines

FMs are general predictors in machine learning introduced by Rendle (2010). They aim to extract structure from training examples in the form of a statistical model. The application of the derived model to new cases with the same structure as the training examples allows for a classification or prediction given the new situation. In this context, each training example is represented by a feature vector $\boldsymbol{x}$ containing information about the specific case and a target value y, which reflects the value that is predicted with the statistical model. All the feature vectors of the training examples can be collected in a feature matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, while the target values form the corresponding target vector $\boldsymbol{y} \in \mathbb{R}^n$.

In the following, we describe the well-known second-order FM model (Rendle, 2010) as well as the higher-order FM approach (Knoll, 2016b), which is a more complex generalization of the second-order FM model. Furthermore, we explain how to find the necessary hyperparameters employing the adaptive-order algorithm and give an insight in the typical learning methods used to optimize FM models.

2.1. Second-order Factorization Machines

The second-order FM model equation based on the feature vector $\boldsymbol{x} \in \mathbb{R}^p$, containing data about p features, can be expressed as follows (Rendle, 2010):

$$\hat{y}(\boldsymbol{x}) := v^{(0)} + \sum_{j_1=1}^{p} x_j v_{j_1,1}^{(1)} + \sum_{j_1=1}^{p} \sum_{j_2=j_1+1}^{p} x_{j_1} x_{j_2} \sum_{f=1}^{k_2} v_{j_1,f}^{(2)} v_{j_2,f}^{(2)}, \tag{1}$$

with the model parameters $v^{(0)} \in \mathbb{R}$, $\boldsymbol{V}^{(1)} \in \mathbb{R}^{p \times 1}$, and $\boldsymbol{V}^{(2)} \in \mathbb{R}^{p \times k_2}$. Hence, the second-order FM model is able to factorize all pairwise interactions based on the dot product of two vectors of the matrix $\boldsymbol{V}^{(2)}$. This means that the parameters have different functions in this model: $v^{(0)}$ captures a global intercept, $v_{1,1}^{(1)}, \ldots, v_{p,1}^{(1)}$ capture a linear weight for each feature, and $v_{1,1}^{(2)}, \ldots, v_{p,k_2}^{(2)}$ capture the second-order interactions between two features, respectively. By adjusting the hyperparameter $k_2 \in \mathbb{N}_0$ (the number of second-order factors), the amount of model parameters factorizing a second-order interaction can be determined. Consequently, the higher the k_2 the more precisely the model

fits to the training data. Contrary, a rather low k_2 abstracts more from the data in order to produce a generalized prediction.

2.2. Higher-order Factorization Machines

The higher-order FM model factorizes not only second-order interactions but all interactions up to the dth order (Rendle, 2010):

$$\hat{y}(\boldsymbol{x}) := v^{(0)} + \sum_{l=1}^{d} \sum_{j_1=1}^{p} \dots \sum_{j_l=j_{l-1}+1}^{p} \Big(\prod_{m=1}^{l} x_{j_m} \Big) \Big(\sum_{f=1}^{k_l} \prod_{m=1}^{l} v_{j_m,f}^{(l)} \Big), \tag{2}$$

with $d \in \mathbb{N}$, $v^{(0)} \in \mathbb{R}$, $\boldsymbol{V}^{(l)} \in \mathbb{R}^{p \times k_l}$, and $k_l \in \mathbb{N}_0$. The linear weights for each feature can be included (excluded) in the higher-order terms by setting $k_1 = 1$ ($k_1 = 0$). In general, the term $\sum_{f=1}^{k_l} \prod_{m=1}^{l} v_{j_m,f}^{(l)}$ factorizes interactions of order l, from $l = 1$ for linear weights up to $l = d$ for the highest included order.

Following this approach, a third-order FM is based on the following model:

$$\begin{aligned} \hat{y}(\boldsymbol{x}) := v^{(0)} &+ \sum_{j_1=1}^{p} x_j v_{j_1,1}^{(1)} + \sum_{j_1=1}^{p} \sum_{j_2=j_1+1}^{p} x_{j_1} x_{j_2} \sum_{f=1}^{k_2} v_{j_1,f}^{(2)} v_{j_2,f}^{(2)} \\ &+ \sum_{j_1=1}^{p} \sum_{j_2=j_1+1}^{p} \sum_{j_3=j_2+1}^{p} x_{j_1} x_{j_2} x_{j_3} \sum_{f=1}^{k_3} v_{j_1,f}^{(3)} v_{j_2,f}^{(3)} v_{j_3,f}^{(3)}, \end{aligned} \tag{3}$$

with $k_3 \in \mathbb{N}_0$ and the model parameters $\boldsymbol{V}^{(3)} \in \mathbb{R}^{p \times k_3}$ factorizing the third-order interactions.

2.3. Adaptive-order Factorization Machines

One crucial task when using higher-order FMs is to determine values for the hyperparameters d and k_2 up to k_d. One could conduct a preliminary cross-validation grid search varying each of these hyperparameters. The disadvantage of this method is that it is very time-consuming due to the growing number of dimensions with each additional order of the FM approach. Thus, Knoll et al. (2017) proposed the adaptive-order algorithm to find faster a reasonable hyperparameter constellation. The algorithm is divided in three steps:

- In Step 1, the highest considered order d and a score r_l ($l \in \{2, \dots, d\}$) reflecting the importance of each of the considered orders are determined. In this context, r_l is defined as the median root mean squared error (RMSE) of a 10-fold cross-validation procedure conducted with a FM model containing linear weights and one factor at the lth order. Starting with the third order, the highest included order d is found when r_d is lower than r_{d+1}.

- Step 2 determines the most favorable setting among FMs of the dth order which contain the proportion of the number of factors for each order given the importance of each order extracted within Step 1. This is done following an out of sample 80/20 validation. The hyperparameters k_2 to k_d are selected based on the model that produces the lower RMSE.
- Within Step 3, the model parameters of the final dth-order FM model (with $k_2, \ldots, k_d$) are optimized based on the whole data set $(\boldsymbol{X}, \boldsymbol{y})$. In this step, all data available during the model training are used.

This algorithm for adaptive-order FMs pursues two objectives. First, it attempts to hold the highest included order as low as possible because the time complexity of model parameter optimization increases with the order. Second, it tries to minimize the number of model parameters because the optimization of more model parameters is more time-consuming.

2.4. Learning methods

Some learning methods are described for training second-order FM models, such as stochastic gradient descent, stochastic gradient descent with adaptive regularization, coordinate descent, or Markov Chain Monte Carlo (MCMC) (Rendle, 2012). In this article, we focus on the application of the MCMC approach because it does not require the determination of any further hyperparameters, such as learning rate or regularization value. The intuition behind this learning method is to estimate each model parameter with a Gibbs-sampler based on a normal gamma hyperprior with unknown mean and unknown precision.

3. Data and Software

Our back-testing framework appropriates two data sources for predicting future stock market returns based on financial information in social media from January 2014 to December 2015.

Today, the internet provides an amazing number of information sets depicting consumer behavior. Twitter data especially reflect real-time sentiments and perceptions about future price trends. Due to their topicality with respect to market development, we derive social media data from Twitter, a free social networking and micro-blogging service with a total of 1.3 billion accounts, over 500 million posts per day, and more than 40 supported languages (Twitter, 2017). Users of this social network interact via "tweets", which are messages constricted to 140 characters per posting. This strict limit and a well-defined markup vocabulary (e.g., RT stands for re-tweet) lead to an above-average information density. Our data set is directly obtained from Twitter (2017) and contains all

tweets about S&P 500 companies from January 2014 to December 2015, resulting in approximately 10 million tweets. Concentrating on the official company names prevents the inclusion of tweets which are not related to the stock market, e.g., requesting the corporation "Amazon.com Inc." avoids the tweet "The Amazon is a large river in South America". Additionally, the acquired data set provides language, date, exact time, and further information for each tweet.

Figure 1 reports the number of tweets over the analyzed time period. On average, we observe approximately 15,000 tweets per day over time. However, there still exist a few outliers caused by changes in the stock market, e.g., we find a peak with over 50,000 tweets per day on the 19th of May 2014 – not surprising since AT&T Inc. presents a takeover offer for the DirecTV Corporation on this day. These strong reactions via Twitter can be explained by the fact that DirecTV Corporation supplies a daily commodity service which is used by a large audience.

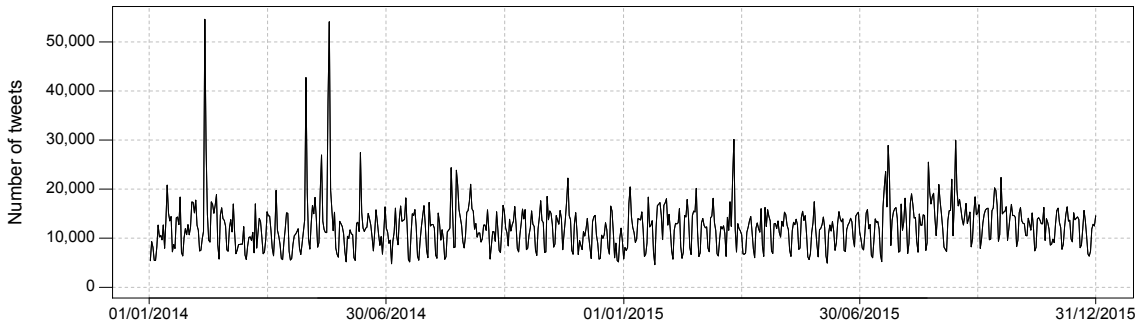

Figure 1: Total number of analyzed tweets in the years 2014 and 2015.

Financial data originate from QuantQuote (2016), which is a leading provider of high resolution historical intraday stock data. We obtain minute-by-minute stock prices from all companies listed on the S&P 500 from January 2014 to December 2015. The corresponding exchange is opened from 9.30 am to 4 pm Eastern time, Monday through Friday. The S&P 500 index, a highly liquid subset of the U.S. stock market, covers about 80 percent of available market capitalization (S&P 500 Dow Jones Indices, 2015). Given thorough investor scrutiny as well as analyst coverage, this market segment serves as a true acid test for any potential capital market anomaly. Following Stübinger and Bredthauer (2017), we perform a two-stage procedure with the objective of eliminating any survivor bias. First, we create a constituent list of the S&P 500 for all stocks that have been part of the S&P 500 for the period of January 2014 to December 2015. This information is further consolidated into a binary matrix, in which each element features a "1" if the stock is constituent

of the index on the current day and a "0" otherwise. Second, we receive the historical minute-by-minute data for all stocks from QuantQuote (2016). Prices are adjusted for dividends, stock splits, and additional corporate actions. By applying the described two-stage procedure, we are able to replicate the S&P 500 constituency and the corresponding prices over time.

The presented methodology in this paper and all relevant analyses are conducted in the programming language `R` (R Core Team, 2017). Furthermore, we use the packages `Matrix` by Bates and Maechler (2016) and `tm` by Feinerer and Hornik (2017) in the field of text mining applications. Handling of time-based data classes is predicated on the packages `TTR` by Ulrich (2016) and `xts` by Ryan and Ulrich (2014). We rely on the package `PerformanceAnalytics` by Peterson and Carl (2014) for performance and risk analysis. The flexible package `FactoRizationMachines` by Knoll (2016a), which serves as central component of our implementation, enables us to compare different FM approaches (see section 4). Furthermore, the package provides an implementation of the MCMC optimization method which allows us to conduct our simulation study without an extensive search for hyperparameters, such as learning rate and regularization values. Moreover, the package contains the adaptive-order algorithm described in section 2.

4. Methodology

For our empirical application, we opt for all tweets about the S&P 500 stock constituents and their associated minute-by-minute prices from January 2014 to December 2015 (see section 3). The entire data set is divided into 473 overlapping study periods, each shifted by one day (see figure 2). In the spirit of Knoll et al. (2017) and Stübinger and Endres (2017), each study period covers a 30-day formation period (subsection 4.1) and a consecutive 1-day trading period (subsection 4.2). While the former estimates the model parameters and identifies the most suitable stocks based on in-sample training, the latter conducts out-of-sample predictions on the corresponding trading sets. In the following, we provide a detailed description of both the formation period and the trading period.

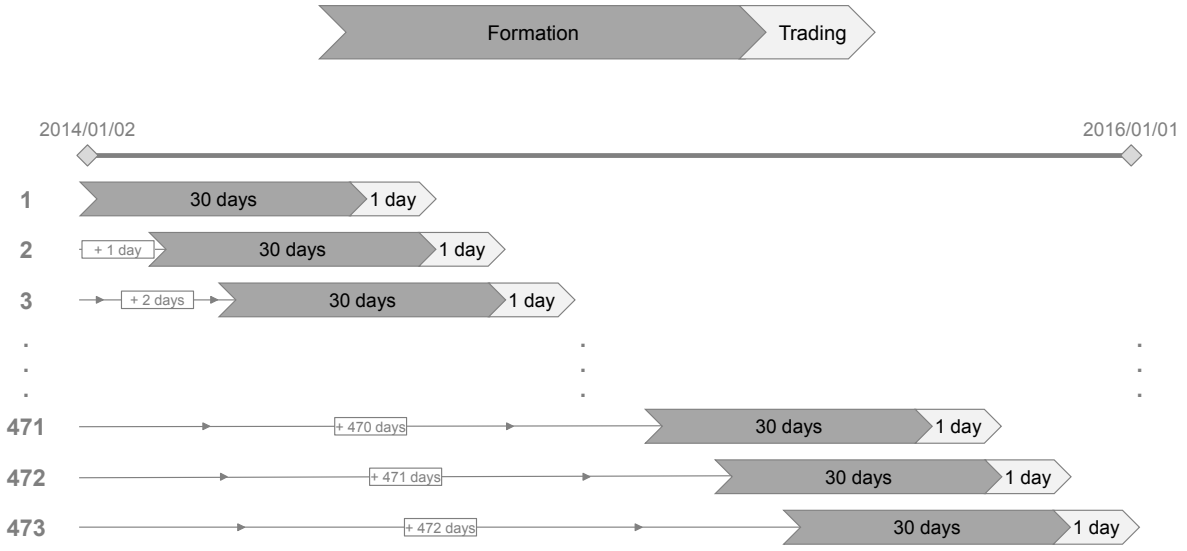

Figure 2: The empirical application consists of 473 overlapping study periods from January 2014 to December 2015. Each study period is built up of a 30-day formation and a 1-day trading period.

4.1. Formation period

In the 30-day formation period, we (i) construct the document-term matrix and the corresponding future returns, (ii) connect the created document-term matrix and future returns by employing FMs with different complexity, and (iii) select the top stocks for the subsequent trading period. This subsection describes the 3-step logic outlined above in detail.

In the first step, we build the document-term matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, which poses as the feature matrix in our back-testing framework. Therefore, we proceed along the lines of Knoll et al. (2017) and perform a six-step procedure with the objective of extracting high-quality information from the tweets. First, we extract the body, time stamp, and associated language of each tweet from the primary data set (see section 3). Second, we restrict the data set to English tweets and convert them to lower-case form. Third, we remove uniform resource locators, numbers, and punctuation marks as well as stop words based on the system for the mechanical analysis and retrieval of text by Salton (1971). Forth, we focus on tweets published between 4 pm and 9.30 am before trading days to exploit the overnight social media data. Fifth, we remove common morphological and inflectional endings from the tweets using Porter's stemming algorithm (Porter, 1980). Sixth, the tweets are transformed into a document-term matrix $\boldsymbol{X}$ in which rows describe the tweets and columns represent all stemmed terms. We apply binary weights for specifying the term frequency

counts in our collection of the tweets, i.e., cell $(i, j) \in \boldsymbol{X}$ takes the value "1" in the presence of term j in tweet i ($i \in \{1, \ldots, n\}, j \in \{1, \ldots, p\}$). The complete document-term matrix $\boldsymbol{X}$ comprises $p = 463,278$ columns for all unique terms. For each study period, we exclusively consider columns related to terms arising in the formation period. This procedure avoids any look-ahead bias since our trading algorithm only converts information which has been realized.

Afterwards, we adapt the tweets to the respective future returns, which serve as the target vector $\boldsymbol{y} \in \mathbb{R}^n$ in our simulation study. For this purpose, we assign each tweet to at least one S&P 500 company using pattern matching since a tweet can mention several stock corporations in one post. Then, the appropriated return is calculated by the percentage change of the price from 4 pm of the last day to 9.45 am of the current day. We choose the target time 9.45 am following Kim et al. (1997) and Visaltanachoti and Yang (2010), who pointed out that prices incorporate information from news within 15 minutes on average after the opening. Concluding, the ith row of the document-term matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ characterizes the stemmed terms $x_1, \ldots, x_p$ of the ith tweet, and the ith element of the target vector $\boldsymbol{y} \in \mathbb{R}^n$ describes the respective return from 4 pm to 9.45 am.

In the second step, we combine the document-term matrix with the respective future returns. In line with Knoll et al. (2017), our simulation study employs models based on support vector machines, second-order FMs, third-order FMs, and adaptive-order FMs. In the following, we depict the key facts of the four approaches.

Support vector machines (SVM). This naive approach describes the relationship between stemmed terms and respective future returns by means of one global intercept and weights for each term, i.e., any ties between the terms are disregarded. To be more specific, the support vector machine with a linear kernel results in the model $\hat{y}(\boldsymbol{x}) := v^{(0)} + \sum_{j_1=1}^{p} x_j v_{j_1}^{(1)}$, where $v^{(0)} \in \mathbb{R}$, $\boldsymbol{v}^{(1)} \in \mathbb{R}^p$.

Second-order FMs (SFM). Extending the baseline approach by a term for pairwise interactions between terms results in the second-order FM model given by equation (1). We set $k_2 = 10$ motivated by the examination of the dimensionality of FMs by Knoll et al. (2017) who found that most of the chosen models in their study contained between 6 and 10 second-order factors.

Third-order FMs (TFM). We enlarge the second-order FM model by supplementarily considering third-order interactions between the terms. As such, we obtain the third-order FM model described in equation (3). Similar to SFM, we set $k_2 = 10$. Furthermore, we set $k_3 = 3$ because once more, most of the models Knoll et al. (2017) selected during their study included 2 to 3 third-order factors.

Adaptive-order FMs (AFM). All interactions between terms up to the dth order are gathered in the higher-order FM model in equation (2). Following the algorithm outlined in section 2, we specify the highest involved order d ($d \in \mathbb{N}$) and the number of factors $k_2, \ldots, k_d$ ($k_2, \ldots, k_d \in \mathbb{N}_0$).

In the third step, we follow Gatev et al. (1999, 2006) and select the most suitable stocks for the out-of-sample trading period. Our algorithm aims at identifying stocks with a minimal error between predicted and actually observed returns. Therefore, we select the top s target stocks per strategy possessing the lowest root relative squared error ($s \in \mathbb{N}$). Additionally, top stocks have to depict a quorum of 25 tweets averaged per day in the formation period. This filter ensures that we pick stocks with sufficient information on the basis of considerable social media activity. We transfer the top stocks to the trading period (section 4.2).

4.2. Trading period

The top s target stocks with the lowest root relative squared error are considered in the 1-day trading period. For every top stock, we calculate the overnight return y^o, i.e., the percentage change of the price from 4 pm of the last day to 9.30 am of the current trading day (see figure 3). Furthermore, we observe n tweets ($n \in \mathbb{N}_0$) about the corresponding company during the overnight period, i.e., tweets are posted between 4 pm and 9.30 am. For each tweet i ($i \in \{1, \ldots, n\}$), the corresponding return $\hat{y}_i(\boldsymbol{x})$ from 4 pm to the target time 9.45 am is predicted using the estimated set of parameters. Therefore, we merge these predictions by calculating the average prediction $\hat{\mu}(\boldsymbol{x})$ and the respective standard deviation $\hat{\sigma}(\boldsymbol{x})$:

$$\hat{\mu}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i(\boldsymbol{x}) \tag{4}$$

$$\hat{\sigma}(\boldsymbol{x}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\hat{y}_i(\boldsymbol{x}) - \hat{\mu}(\boldsymbol{x})\right)^2} \tag{5}$$

Concluding, $\hat{\mu}(\boldsymbol{x})$ characterizes the average predicted return from 4 pm of the last day to 9.45 am of the current trading day (see figure 3). Our framework is based on a two-stage logic. First, overnight tweets about stocks contain information that have an essential effect on stock returns in the future. Second, FMs capture the causality between overnight tweets and future returns. If our assumption holds, we are in a position to take advantage of these market inefficiencies, i.e., the market prices of common stocks are not always exactly priced and tend to deviate temporarily from the true discounted value of their future cash flows. To be more specific, the back-testing framework aims to predict the return from 9.30 am to the target time 9.45 am $(1+\hat{\mu}(\boldsymbol{x}))/(1+y^o)-1$

based on the observed overnight return y^o and the average predicted return from 4 pm of the last day to 9.45 am of the current trading day $\hat{\mu}(\boldsymbol{x})$.

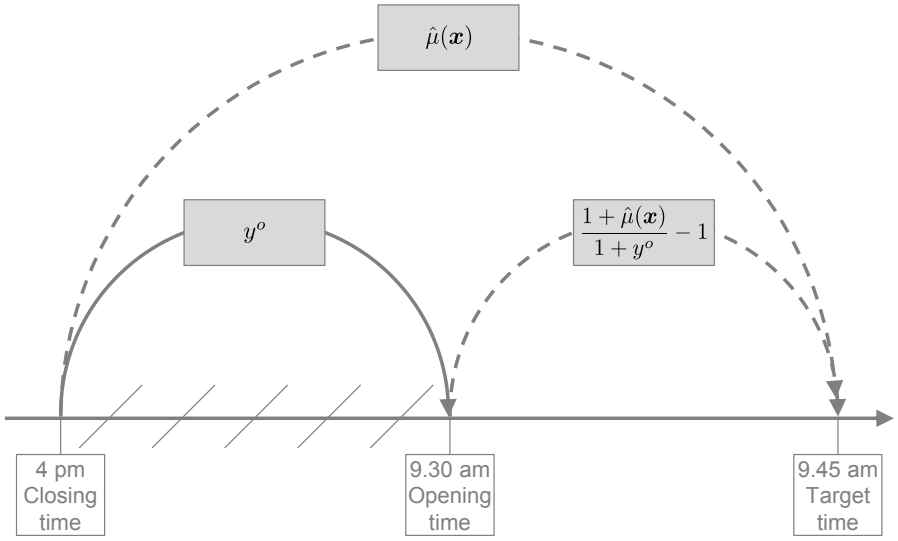

Figure 3: The back-testing framework aims at predicting the return from 9.30 am to 9.45 am $((1+\hat{\mu}(\boldsymbol{x}))/(1+y^o)-1)$ based on the observed overnight return (y^o) and the average predicted return from 4 pm of the last day to 9.45 am of the current trading day $(\hat{\mu}(\boldsymbol{x}))$.

Using the realized overnight return y^o and the predicted returns from 4 pm to 9.45 am, we are able to capture mispricings and generate profits. Therefore, we define transaction costs c and the risk parameter b $(c, b \in \mathbb{R}_0^+)$.

If we do not observe any tweets related to the top stock during the night, our algorithm does not execute any trade. If we observe at least one tweet related to the top stock during the night, we apply the following trading rules:

- $\frac{1+\hat{\mu}(\boldsymbol{x})}{1+y^o} - 1 > c + b \cdot \hat{\sigma}(\boldsymbol{x})$, i.e., the stock is undervalued. Consequently, we go long in the stock and reverse the trade at 9.45 am.
- $\frac{1+\hat{\mu}(\boldsymbol{x})}{1+y^o} - 1 < -c - b \cdot \hat{\sigma}(\boldsymbol{x})$, i.e., the stock is overvalued. Consequently, we go short in the stock and reverse the trade at 9.45 am.
- $-c - b \cdot \hat{\sigma}(\boldsymbol{x}) \leq \frac{1+\hat{\mu}(\boldsymbol{x})}{1+y^o} - 1 \leq c + b \cdot \hat{\sigma}(\boldsymbol{x})$, i.e., the stock is in its 'normal' region. In consequence, we do not execute any trades.

Since we still aim for a classic long-short investment strategy in the sense of Gatev et al. (2006), we follow Avellaneda and Lee (2010) and hedge market exposure day-by-day with corresponding

capital expenditures in the S&P 500 index. In accordance with Liu et al. (2017) and Stübinger and Endres (2017), we set $b = 3$.

Following Gatev et al. (2006), return computation is based on committed capital, the more common metric in trading literature. Thus, we divide the sum of net profits by the number of active stocks. A stock is called active if it exhibits at least one round-trip trade during the respective trading period. Following Prager et al. (2012), we depict $c = 2$ bps per share per half-turn. This assumption is deemed feasible in light of our minute-by-minute data in a highly liquid investment universe from 2014 to 2015.

5. Results

Following Krauss and Stübinger (2017)'s approach, we run a holistic performance analysis for the top five stocks ($s = 5$) of the strategies SVM, SFM, TFM, and AFM, compared to a naive buy-and-hold strategy of the S&P 500 index (MKT). Therefore, we analyze risk-return characteristics as well as trading statistics (subsection 5.1), present statistical measures of the performance (subsection 5.2), examine the profitability varying the target time (subsection 5.3), and discuss the returns in light of market frictions (subsection 5.4). Finally, we focus on AFM and investigate the exposure of the daily returns to common systematic sources of risk (subsection 5.5), perform a bootstrap trading (subsection 5.6), and conduct a deep dive on the dimensionality of the FMs (subsection 5.7).

5.1. Risk-return characteristics and trading statistics

Table 1 depicts daily risk-return measures for the analyzed period both before and after the incorporation of transaction costs. The majority of the metrics can be found in Bacon (2008). Irrespective of the FM model employed, we observe positive returns after transaction costs ranging between 2 bps per day for SVM and 5 bps per day for AFM, compared to 3 bps for a naive buy-and-hold strategy of the S&P 500. From a statistical perspective, the returns of AFM are also significant with Newey-West(NW) t-statistics of 2.49 after transaction costs. A straight forward investment in the general market leads to a standard deviation of 0.0085, approximately 2-times higher than the key figures of the four strategies based on FMs. Moreover, SFM, TFM, and AFM possess high values for the kurtosis as well as positive skewness – a pleasant property for investors (Cont, 2001). We depict the historical Value at Risk (VaR) as propounded by J.P. Morgan's RiskMetrics approach in Mina and Xiao (2001). The tail risk of FMs is at a very low level by contrast with

the general market, e.g., the historical VaR (1 %) after transaction costs is -0.0084 for AFM versus -0.0212 for MKT. This picture barely changes considering the maximum drawdown – the decline from a historical peak is greatly reduced for FMs (3.87 percent – 6.72 percent), compared to the benchmark (12.63 percent). The strategy AFM produces the highest hit rate, i.e., the percentage of days with positive returns, with approximately 62 percent after transaction costs – complexity pays off. In summary, AFM achieves meaningful risk-return characteristics, even considering transaction costs. However, we have to investigate the robustness to systematic sources of risk.

	Before transaction costs				After transaction costs				
	SVM	SFM	TFM	AFM	SVM	SFM	TFM	AFM	MKT
Mean return	0.0004	0.0005	0.0006	0.0008	0.0002	0.0003	0.0003	0.0005	0.0003
Standard error (NW)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0004
t-Statistic (NW)	2.2856	2.2894	2.8456	3.6245	1.2395	1.4451	1.5465	2.4931	0.6259
Minimum	-0.0338	-0.0169	-0.0202	-0.0169	-0.0342	-0.0173	-0.0207	-0.0173	-0.0402
Quartile 1	-0.0005	-0.0007	-0.0007	-0.0007	-0.0006	-0.0007	-0.0007	-0.0008	-0.0041
Median	0.0001	0.0002	0.0005	0.0003	0.0001	0.0001	0.0002	0.0001	0.0003
Quartile 3	0.0010	0.0013	0.0017	0.0017	0.0007	0.0010	0.0012	0.0012	0.0048
Maximum	0.0246	0.0418	0.0419	0.0400	0.0242	0.0414	0.0414	0.0396	0.0383
Standard deviation	0.0038	0.0040	0.0043	0.0040	0.0038	0.0040	0.0042	0.0040	0.0085
Skewness	-0.5159	2.4861	1.6000	2.1865	-0.6412	2.4495	1.5002	2.2486	-0.2907
Kurtosis	22.8953	26.2757	21.7965	21.9991	24.5109	27.5101	22.6053	23.9771	2.3528
Historical VaR 1 %	-0.0115	-0.0103	-0.0144	-0.0080	-0.0119	-0.0107	-0.0149	-0.0084	-0.0212
Historical CVaR 1 %	-0.0191	-0.0142	-0.0171	-0.0136	-0.0195	-0.0146	-0.0176	-0.0140	-0.0300
Historical VaR 5 %	-0.0030	-0.0046	-0.0046	-0.0041	-0.0030	-0.0050	-0.0052	-0.0042	-0.0144
Historical CVaR 5 %	-0.0084	-0.0079	-0.0091	-0.0074	-0.0086	-0.0082	-0.0097	-0.0076	-0.0200
Maximum drawdown	0.0414	0.0334	0.0547	0.0343	0.0632	0.0422	0.0672	0.0387	0.1263
Share with return ≥ 0	0.5897	0.6060	0.6341	0.6545	0.5201	0.5518	0.5887	0.6180	0.5243

Table 1: Daily characteristics of returns for the top five stocks of SVM, SFM, TFM, and AFM compared to a long-only S&P 500 benchmark (MKT) in the time frame between March 2014 and December 2015. NW as acronym for Newey-West standard errors with one-lag correction, and CVaR refers to the Conditional Value at Risk.

Table 2 portrays statistics about the trading behavior of SVM, SFM, TFM, and AFM. All FMs are substantially similar in their trading activity during the trading period. Across all models, we observe approximately 1.60 traded top stocks and a corresponding standard deviation at approximately 0.80. This relatively small number is based on the fact that the trading rule is only applied once per day. The resemblance among the compared models is most likely driven by the same underlying trading logic as outlined in section 4 – however, the respective characteristics of the returns differ vastly depending on the information level extracted by the various complex FMs.

	SVM	SFM	TFM	AFM
Average number of target stocks traded per 1-day period	1.4755	1.4608	1.5324	1.6371
Standard deviation of number of target stocks traded per 1-day period	0.7584	0.7710	0.8005	0.8488

Table 2: Daily trading volume for the top 5 stocks of SVM, SFM, TFM, and AFM.

In table 3, we present advanced annualized risk-return measures for all strategies. The annualized returns after transaction costs range between 5.96 percent for SVM and 13.52 percent for AFM, compared to the general market with 5.63 percent. Across all strategies, the mean return equals the mean excess return owing to the fact that the risk free rate amounts to zero during the analyzed period. The standard deviation proves to be roughly homogeneous among the models at around 0.06, while the long-only S&P 500 benchmark leads to a standard deviation of 0.14. The Sharpe ratio, i.e., the excess return per unit of deviation, exceeds 2 in case of AFM – the excess return clearly overcompensates the risk.

	Before transaction costs				After transaction costs				
	SVM	SFM	TFM	AFM	SVM	SFM	TFM	AFM	MKT
Mean return	0.1146	0.1344	0.1668	0.2079	0.0596	0.0822	0.0853	0.1352	0.0563
Mean excess return	0.1146	0.1344	0.1668	0.2079	0.0596	0.0822	0.0853	0.1352	0.0563
Standard deviation	0.0605	0.0642	0.0681	0.0643	0.0597	0.0632	0.0673	0.0628	0.1355
Downside deviation	0.0395	0.0350	0.0402	0.0332	0.0406	0.0365	0.0424	0.0343	0.0968
Sharpe ratio	1.8946	2.0922	2.4483	3.2347	0.9978	1.3011	1.2667	2.1518	0.4158
Sortino ratio	2.9012	3.8357	4.1494	6.2649	1.4695	2.2543	2.0098	3.9406	0.5822

Table 3: Annualized characteristics of returns for the top 5 stocks of SVM, SFM, TFM, and AFM compared to a long-only S&P 500 benchmark (MKT) during the time frame between March 2014 and December 2015.

Table 4 describes advanced drawdown metrics. Both the Sterling ratio and Calmar ratio divide the annualized return by the maximum drawdown in absolute terms. Additionally, the denominator within the Sterling ratio is augmented by a 10 percent excess risk buffer. With respect to the Calmar ratio, we observe that the annual returns of the FMs are at more than 2 times the magnitude of maximum drawdown, compared to a value of 0.44 for the general market. The pain index, which measures the depth, duration, and frequency of losses, ranges between 0.0094 for AFM and 0.0156 for SVM by contrast with 0.0207 for the naive buy-and-hold strategy. The highest pain ratio is generated by AFM (34.15), which is not surprising since it exhibits the highest mean excess return (table 3) and the lowest pain index.

	Before transaction costs				After transaction costs				
	SVM	SFM	TFM	AFM	SVM	SFM	TFM	AFM	MKT
Sterling ratio	0.8103	1.0075	1.0781	1.5487	0.3652	0.5784	0.5103	0.9743	0.2490
Calmar ratio	2.7691	4.0232	3.0478	6.0691	0.9436	1.9499	1.2703	3.4899	0.4461
Burke ratio	1.9171	2.1598	2.4271	3.8165	0.9879	1.2365	1.1993	2.4265	0.2795
Pain index	0.0097	0.0081	0.0101	0.0061	0.0156	0.0125	0.0140	0.0094	0.0207
Ulcer index	0.0149	0.0118	0.0170	0.0099	0.0226	0.0166	0.0222	0.0144	0.0322
Pain ratio	11.7769	16.5395	16.5642	34.1546	3.8198	6.5926	6.1032	14.3895	2.7163
Martin ratio	7.7019	11.3922	9.8297	21.0474	2.6394	4.9506	3.8477	9.3803	1.7479

Table 4: Drawdown measures after transaction costs for the top 5 stocks of SVM, SFM, TFM, and AFM compared to a long-only S&P 500 benchmark (MKT) during the time frame between March 2014 and December 2015.

Following Do and Faff (2010) and Bowen and Hutchinson (2016), we perform a sub-period analysis in an effort to discern possible fluctuating tendencies of the strategies within the portrayed time span. Figure 4 depicts the development of an investment of 1 USD for the analyzed strategies both before transaction costs (left) and after transaction costs (right). After transaction costs, we observe a growth to approximately 1.15 USD for TFM, SFM, SVM, and MKT. As expected, AFM outperforms the other strategies over the sample period with a final value of 1.27 USD. Moreover, the investments using SVM, SFM, TFM, and AFM exhibit steady growth from March 2014 until December 2015. In stark contrast, the cumulative return of the general market shows strong swings and large drawdowns.

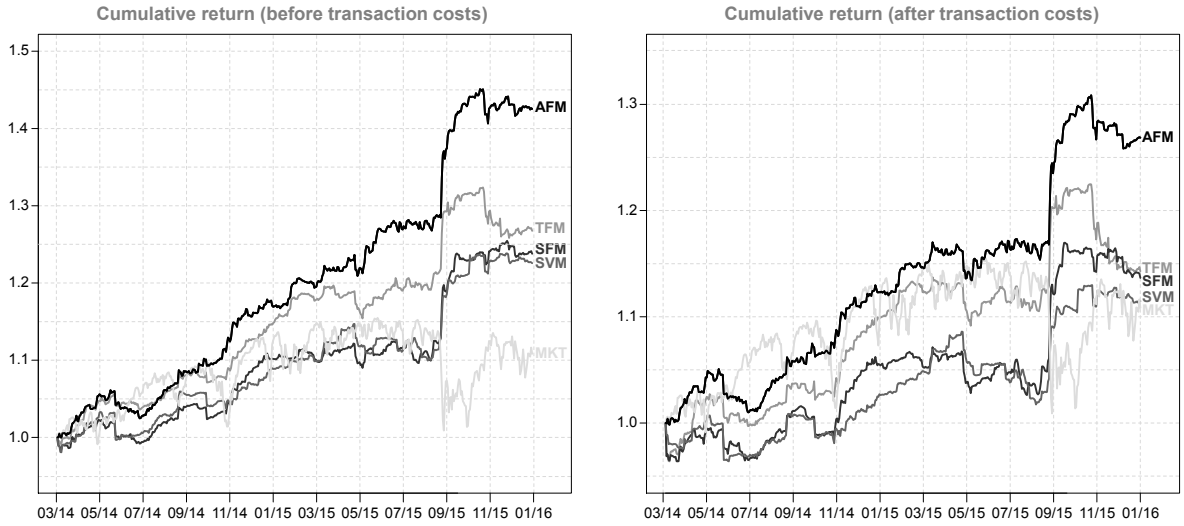

Figure 4: Investment of 1 USD for SVM, SFM, TFM, and AFM compared to a long-only S&P 500 benchmark (MKT) during the time frame between March 2014 and December 2015 both before transaction costs (left) and after transaction costs (right).

5.2. Statistical measures of the performance

Table 5 measures the degree of closeness of the predicted returns to the actually observed values. Specifically, we build an advanced contingency table based on each trading prediction and alongside it present additional ratios derived from the former. As expected, SVM does not perform well since predictions are wrong in the majority of the trades, i.e., the true positive rate as well as the true negative rate are below 50 percent. Since this strategy still achieves positive returns (see subsection 5.1), we conclude that fewer correct predictions generate high gains. Only AFM manages to return both a true positive rate and true negative rate above 50 percent showing that complexity pays off. The diagnostic odds ratio measures the effectiveness of a strategy and exceeds 1 if we observe more correct decisions than wrong decisions. In contrast to SVM (0.46) and SFM (1.02), the more extensive models TFM (1.33) and AFM (2.58) exceed clearly the desirable threshold. The total accuracy, i.e., the proportion of true results among the total number of cases examined, is greater 50 percent for SFM, TFM, and AFM. The strategy AFM still outperforms strongly the other strategies by scoring 61.76 percent, a value similar to the percentage of days with positive returns (table 1).

	SVM	SFM	TFM	AFM
True positive rate	0.4902	0.5658	0.6375	0.6729
False positive rate	0.5937	0.5606	0.5694	0.4433
False negative rate	0.5098	0.4342	0.3625	0.3271
True negative rate	0.4062	0.4394	0.4306	0.5567
Positive likelihood ratio	0.8256	1.0092	1.1195	1.5179
Negative likelihood ratio	1.2549	0.9882	0.8419	0.5876
Diagnostic odds ratio	0.6579	1.0213	1.3297	2.5834
Accuracy	0.4578	0.5070	0.5395	0.6176

Table 5: Statistical measures of the performance for SVM, SFM, TFM, and AFM during the time frame between March 2014 and December 2015.

5.3. Time-varying profitability

In subsection 4.2, the target time 9.45 am is motivated based on the existing literature (Kim et al., 1997; Busse and Green, 2002; Visaltanachoti and Yang, 2010). Figure 5 investigates the annualized returns for SVM, SFM, TFM, and AFM for varied target times before transaction costs (left) and after transaction costs (right). Across all models, we observe a 3-step behavior over time with upwards shifts for the more complex approaches. To be more specific, the strategies based on FMs generate annualized returns of approximately 0 percent after transaction costs considering target times close to the opening time. Second, the annualized returns increase strongly until 9.45 am – the peaks range between 11.46 percent for SVM and 20.79 percent for AFM before transaction costs. Incorporating transaction costs leads to returns from 5.96 percent p.a. (SVM) to 13.52 percent p.a. (AFM). Finally, the returns decline to the break-even point with increasing target time – TFM and SFM cross this threshold at around 9.55 am.

Our findings are well in line with the literature explaining the behavior based on the effects of market efficiency and market turmoils during the opening minutes. The first effect results from the fact that stock prices deviate temporarily from the true discounted value (Rosenberg et al., 1985; Stout, 2002). Visaltanachoti and Yang (2010) examined the speed of convergence using multivariate regressions and find that on average US stock prices take approximately 30 minutes to achieve market efficiency. The profitability of trading strategies decreases over time because more and more prices are adjusted to incorporate information about the corresponding firms. The second effect is based on high market turmoils during the first minutes of the trading day. Brooks (2011) and Wendell (2017) pointed out that the opening range is the easiest time span to lose money since there are a huge number of irrational and unpredictable events. Therefore, meaningful trading

thresholds can only be defined after market makers have squared their offsetting positions. The effect of this above-average volatility, caused by gamble trades, lessens over time and disappears after around 15 minutes. Therefore, the interaction of both effects explains the return characteristics for varying target times.

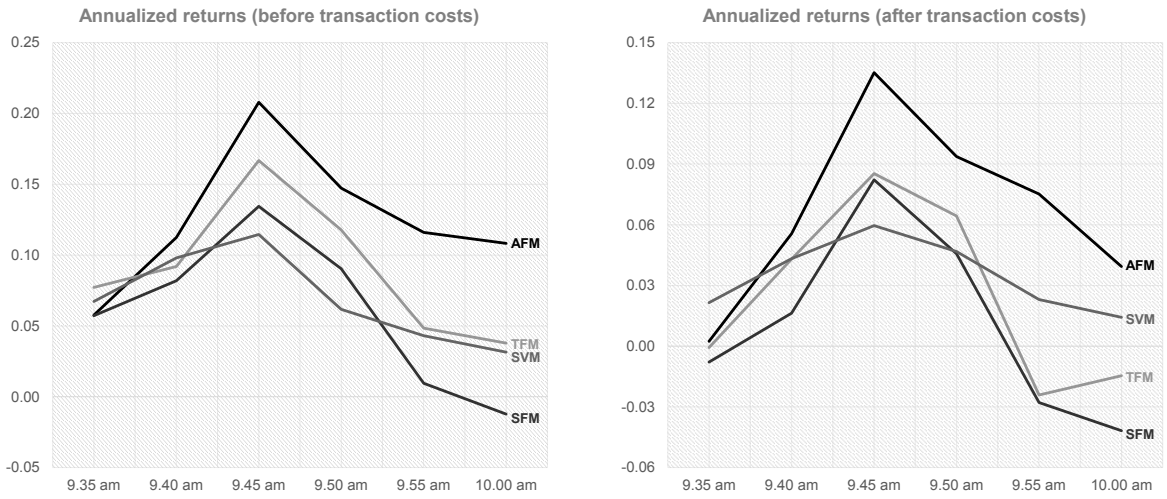

Figure 5: Annualized returns for varying target times for SVM, SFM, TFM, and AFM before transaction costs (left) and after transaction costs (right).

5.4. Market frictions

In this subsection, we evaluate the robustness of our statistical arbitrage strategies in light of market frictions. Most notably, we evaluate the impact of bid-ask spreads, i.e., the amount by which the bid price falls below the ask price for an asset, on our results (Abadir and Rockinger, 2003; Jondeau et al., 2015; Krauss and Herrmann, 2017). Voya Investment Management (2016) pointed out that the use of algorithmic trading, changes in the exchange landscape, and decimalization are the main drivers for falling bid-ask spreads over time. Following Voya Investment Management (2016), we apply 3 bps per share per half-turn for the considered trading years. In contrast to table 3, table 6 depicts annualized risk-return characteristics in light of market frictions. Across all strategies, we observe positive annualized returns ranging between 2.06 percent for SVM and 8.72 percent for AFM, compared to 5.63 percent for the general market. The standard deviation of MKT (0.14) is approximately two-times higher than the standard deviation of the 4 approaches based on FMs. As expected, the Sharpe ratios of SFM (0.53), TFM (0.72), and AFM (1.37) exceed the simple buy-and-hold strategy of the S&P 500 index (0.42). To summarize, these strategies produce

positive results in light of market frictions. The return time series of AFM especially remains statistically and economically significant even with bid-ask spreads, posing a severe challenge to the semi-strong form of market efficiency.

	SVM	SFM	TFM	AFM	MKT
Mean return	0.0206	0.0343	0.0483	0.0872	0.0563
Standard deviation	0.0621	0.0643	0.0667	0.0635	0.1355
Sharpe ratio	0.3317	0.5334	0.7241	1.3732	0.4158

Table 6: Annualized characteristics of returns for the top 5 stocks of SVM, SFM, TFM, and AFM including bid-ask spreads compared to a long-only S&P 500 benchmark (MKT) during the time frame between March 2014 and December 2015.

5.5. Common risk factors

In table 7, we analyze the systematic risk exposure for the top 5 stocks of AFM after transaction costs. In this respect, we employ three types of regression, namely Fama-French 3-factor model (FF3), Fama-French 3+2-factor model (FF3+2), and Fama-French 5-factor model (FF5). The standard model FF3, introduced by Fama and French (1996), explains the sensitivity to the general market, small minus big capitalization stocks (SMB), and high minus low book-to-market stocks (HML). The second regression FF3+2, in the spirit of Gatev et al. (2006), extends FF3 by adding the factors momentum and short-term reversal. Following Fama and French (2015), FF5 enhances the first model by two supplemental factors, i.e., portfolios of stocks with robust minus weak profitability (RMW) and conservative minus aggressive investment behavior (CMA). All data related to these models are downloaded from Kenneth R. French's website[1].

Across all employed Fama-French models, we observe statistically significant daily alphas of 0.05 percent after transaction costs – similar to the raw returns. Exposure to the general market, SMB, HML, momentum, RMA, and CMA are statistically insignificant and close to zero due to the fact that our strategy is dollar neutral. Most interestingly, the short-term reversal factor shows a statistical positive loading implying that we buy short-term losers and short short-term winners. The FF3+2 model presents the highest explanatory power provoked by the short-term reversal factor. In short, AFM produces statistically significant and economically remarkable returns after transaction costs, does not show loadings on any systematic risk exposure, and outperforms the less complex benchmarks.

[1]We thank Kenneth R. French for supplying all required data to these models on his website.

	FF3	FF3+2	FF5
(Intercept)	0.0005**	0.0005**	0.0005**
	(0.0002)	(0.0002)	(0.0002)
Market	−0.0083	−0.0105	−0.0082
	(0.0210)	(0.0225)	(0.0223)
SMB	0.0099	0.0011	
	(0.0353)	(0.0359)	
HML	−0.0757	−0.0546	
	(0.0408)	(0.0492)	
Momentum		0.0371	
		(0.0307)	
Reversal		0.0879**	
		(0.0330)	
SMB5			0.0229
			(0.0398)
HML5			−0.0674
			(0.0564)
RMW5			0.0545
			(0.0735)
CMA5			−0.0352
			(0.1029)
R^2	0.0376	0.0525	0.0388
Adj. R^2	0.0315	0.0423	0.0285
Num. obs.	473	473	473
RMSE	0.0039	0.0039	0.0039

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 7: Exposure to systematic risk factors for daily returns after transaction costs of the top 5 stocks of AFM during the time frame between March 2014 and December 2015. Standard errors are depicted in parentheses.

5.6. Bootstrap trading

In view of our remarkable returns of AFM, we compare the financial performance with one million random bootstraps of monkey trading. In the sense of Malkiel (2007), we randomly combine the top stocks and the corresponding entry and exit signals for each of the trading days. Figure 6 illustrates the daily return characteristics of the bootstrapped monkey trading before transaction costs. As anticipated, the average daily return of the random trading is zero prior transaction costs. Most importantly, the best performing bootstrap, with an average daily return of 0.04 percent, achieves a weaker result than AFM (see table 1).

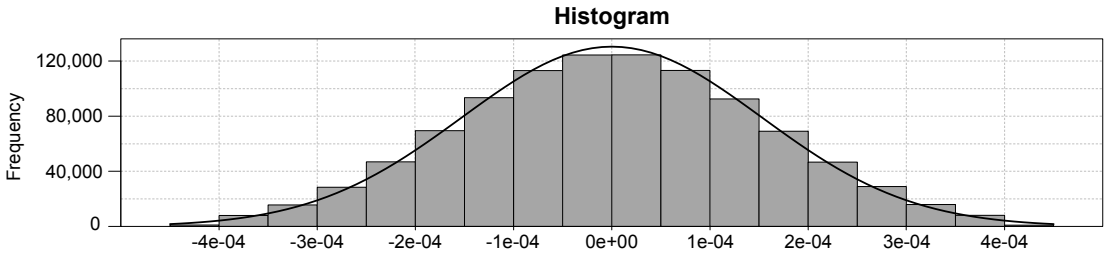

Figure 6: Empirical distribution of daily mean returns of 1,000,000 bootstrap tradings during the time frame between March 2014 and December 2015 (before transaction costs).

5.7. Dimensionality of Factorization Machines

In order to analyze the impact of the factors per order of the AFM model, Figure 7 shows a summary of the particular FM models optimized for the 473 trading periods. The left plot expresses the characteristics of the highest included order d among the FM models chosen, while the right diagram presents the number of FM models which contained a specific dimensionality (k_l) corresponding to order l.

The left plot reveals that the maximum value of the highest included order is 8. The values decline from 3, followed by 4 and 5, while 6, 7, and 8 are barely chosen. In this context, we note that the highest included order cannot be 2 because the adaptive-order algorithm, which chooses the order of the AFM models, only selects from higher-order FM models. The decreasing frequency of d reflects the design of the adaptive-order algorithm to use the more time-efficient lower orders rather than higher ones.

The plot on the right-hand side shows that the most frequently chosen number of factors is 1 for all orders except the second. Consequently, the second-order factors seem to be well suited for representing the information extracted within the formation period. The mode of 1 factor for all other orders is probably caused by the design of the adaptive-order algorithm which sets the minimum number of factors for each order lower than d to 1. Hence, the large bar for 1 third-order factor is a result of the algorithm design as well: if the FM model with only second-order factors already produces good results, 1 third-order factor is added to the FM model to fulfill the higher-order requirement. Furthermore, there is a tendency for lower rather than higher number of factors, which also reflects the idea of the adaptive-order algorithm and saves computational resources.

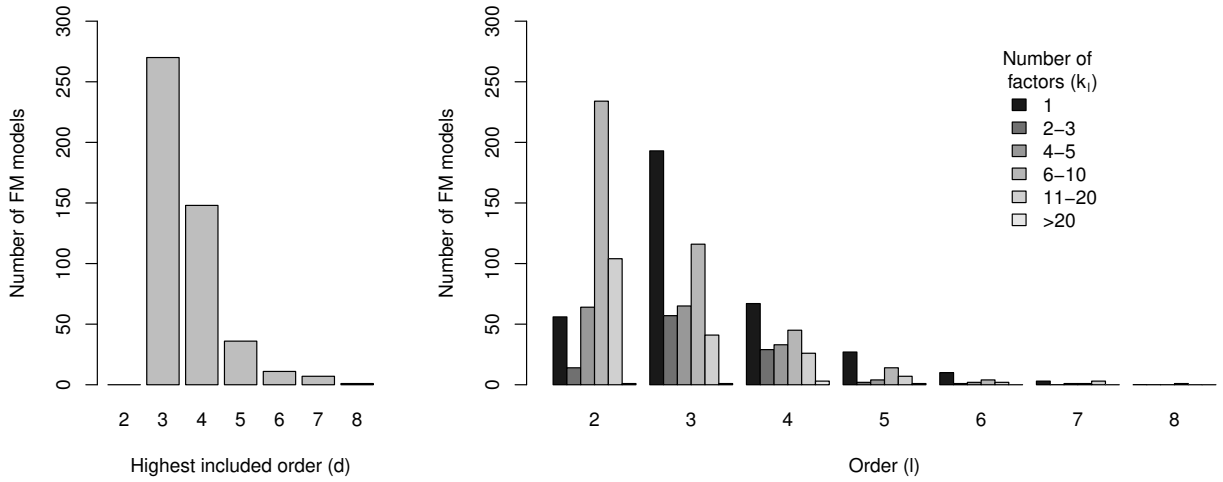

Figure 7: Analysis of the highest included order and the factors per order obtained by running the adaptive-order algorithm for the 473 trading periods.

6. Conclusions

In this paper, we introduce a statistical arbitrage strategy using FMs to exploit overnight social media data and deploy it on the S&P 500 constituents from January 2014 to December 2015. Across all strategies based on FMs, we observe remarkable annualized returns after transaction costs for the top 5 stocks demonstrating the efficiency of our strategy. Moreover, we observe that increasing complexity of the FMs leads to higher profitability – annualized returns after transaction costs range between 5.96 percent for SVM and 13.52 percent for AFM. Additionally considering the risk-component, AFM outperforms the benchmarks with a Sharpe ratio of 2.15 after transaction costs. Furthermore, AFM achieves the highest accuracy rate with a value of 61.76 percent; the corresponding returns are robust against the impact of bid-ask spreads and do not load on any systematic risk exposure.

For future research in this field, we could incorporate the time structure of the tweets during the overnight period assigning different weights to data at varying positions in the time frame. In addition, terms with synonymous meaning might be aggregated using a dictionary, e.g., the terms “automobile” and “car” belong to the category “vehicle”. Third, the feature matrix could be extended by financial data and economic data with the objective of more precise predictions about the future price changes.

Acknowledgments: This research was supported by the GfK Verein e. V., which funded the purchase of the Twitter data set. We are especially grateful to Raimund Wildner and Holger Dietrich for their commitment and effort over the course of this project. Furthermore, we would like to thank Ingo Klein and the participants of the CEQURA Conference on Advances in Financial and Insurance Risk Management 2017 for many helpful discussions.

Bibliography

Abadir, K. M., Rockinger, M., 2003. Density functionals, with an option-pricing application. Econometric Theory 19 (5), 778–811.

Avellaneda, M., Lee, J.-H., 2010. Statistical arbitrage in the US equities market. Quantitative Finance 10 (7), 761–782.

Bacon, C. R., 2008. Practical portfolio performance: Measurement and attribution, 2nd Edition. John Wiley & Sons, Chichester, England.

Bates, D., Maechler, M., 2016. Matrix: Sparse and dense matrix classes and methods. R package.

Bowen, D. A., Hutchinson, M. C., 2016. Pairs trading in the UK equity market: Risk and return. The European Journal of Finance 22 (14), 1363–1387.

Brooks, A., 2011. Trading price action reversals: Technical analysis of price charts bar by bar for the serious trader. John Wiley & Sons, Hoboken, USA.

Busse, J. A., Green, T. C., 2002. Market efficiency in real time. Journal of Financial Economics 65 (3), 415–437.

Chatrath, A., Miao, H., Ramchander, S., Villupuram, S., 2014. Currency jumps, cojumps and the role of macro news. Journal of International Money and Finance 40, 42–62.

Cont, R., 2001. Empirical properties of asset returns: Stylized facts and statistical issues. Quantitative Finance 1 (2), 223–236.

Do, B., Faff, R., 2010. Does simple pairs trading still work? Financial Analysts Journal 66 (4), 83–95.

Fama, E. F., French, K. R., 1996. Multifactor explanations of asset pricing anomalies. The Journal of Finance 51 (1), 55–84.

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116 (1), 1–22.

Feinerer, I., Hornik, K., 2017. tm: Text mining package. R package.

Gatev, E., Goetzmann, W. N., Rouwenhorst, K. G., 1999. Pairs trading: Performance of a relative value arbitrage rule. Working paper, Yale School of Management's International Center for Finance.

Gatev, E., Goetzmann, W. N., Rouwenhorst, K. G., 2006. Pairs trading: Performance of a relative-value arbitrage rule. Review of Financial Studies 19 (3), 797–827.

Hagenau, M., Liebmann, M., Neumann, D., 2013. Automated news reading: Stock price prediction based on financial news using context-capturing features. Decision Support Systems 55 (3), 685–697.

Jin, F., Self, N., Saraf, P., Butler, P., Wang, W., Ramakrishnan, N., 2013. Forex-foreteller: Currency trend modeling using news articles. In: Proceedings of the 19th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining. ACM, Chicago, USA, pp. 1470–1473.

Jondeau, E., Lahaye, J., Rockinger, M., 2015. Estimating the price impact of trades in a high-frequency microstructure model with jumps. Journal of Banking & Finance 61 (2), 205–224.

Kim, S. T., Lin, J.-C., Slovin, M. B., 1997. Market structure, informed trading, and analysts' recommendations. Journal of Financial and Quantitative Analysis 32 (4), 507–524.

Knoll, J., 2016a. FactoRizationMachines: Machine Learning with higher-order factorization machines. R package.

Knoll, J., 2016b. Recommending with higher-order Factorization Machines. In: Bramer, M., Petridis, M. (Eds.), Research and Development in Intelligent Systems XXXIII: Incorporating Applications and Innovations in Intelligent Systems XXIV. Springer, Cham, pp. 103–116.

Knoll, J., Stübinger, J., Grottke, M., 2017. Exploiting social media with higher-order factorization machines: Statistical arbitrage on high-frequency data of the S&P 500. FAU Discussion Papers in Economics (13), University of Erlangen-Nürnberg.

Krauss, C., Herrmann, K., 2017. On the power and size properties of cointegration tests in the light of high-frequency stylized facts. Journal of Risk and Financial Management 10 (1), 7.

Krauss, C., Stübinger, J., 2017. Non-linear dependence modelling with bivariate copulas: Statistical arbitrage pairs trading on the S&P 100. Applied Economics 49 (52), 5352–5369.

Liu, B., Chang, L.-B., Geman, H., 2017. Intraday pairs trading strategies on high frequency data: The case of oil companies. Quantitative Finance 17 (1), 87–100.

Malkiel, B. G., 2007. A random walk down Wall Street: The time-tested strategy for successful investing. W.W. Norton & Company, New York, USA.

Mina, J., Xiao, J. Y., 2001. Return to RiskMetrics: The evolution of a standard. RiskMetrics Group.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., Ngo, D. C. L., 2014. Text mining for market prediction: A systematic review. Expert Systems with Applications 41 (16), 7653–7670.

Peterson, B. G., Carl, P., 2014. PerformanceAnalytics: Econometric tools for performance and risk analysis. R package.

Porter, M. F., 1980. An algorithm for suffix stripping. Program 14 (3), 130–137.

Prager, R., Vedbrat, S., Vogel, C., Watt, E. C., 2012. Got liquidity? BlackRock Investment Institute.

QuantQuote, 2016. QuantQuote market data and software. https://www.quantquote.com.

R Core Team, 2017. R package stats: A language and environment for statistical computing. R package.

Rendle, S., 2010. Factorization Machines. In: 10th International Conference on Data Mining. IEEE, pp. 995–1000.

Rendle, S., 2012. Learning recommender systems with adaptive regularization. In: Proceedings of the 5h ACM International Conference on Web Search and Data Mining. pp. 133–142.

Rosenberg, B., Reid, K., Lanstein, R., 1985. Persuasive evidence of market inefficiency. The Journal of Portfolio Management 11 (3), 9–16.

Ryan, J. A., Ulrich, J. M., 2014. xts: eXtensible time series. R package.

Salton, G., 1971. The smart retrieval system: Experiments in automatic document processing. Prentice-Hall, Englewood Cliffs, USA.

Schumaker, R. P., Zhang, Y., Huang, C.-N., Chen, H., 2012. Evaluating sentiment in financial news articles. Decision Support Systems 53 (3), 458–464.

S&P 500 Dow Jones Indices, 2015. Equity S&P 500. http://us.spindices.com/indices/equity/sp-500.

Stout, L. A., 2002. The mechanisms of market inefficiency: An introduction to the new finance. Journal of Corporation Law 28 (4), 635.

Stübinger, J., Bredthauer, J., 2017. Statistical arbitrage pairs trading with high-frequency data. International Journal of Economics and Financial Issues 7 (4), 650–662.

Stübinger, J., Endres, S., 2017. Pairs trading with a mean-reverting jump-diffusion model on high-frequency data. FAU Discussion Papers in Economics (10), University of Erlangen-Nürnberg.

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

Twitter, 2017. Online news and social networking service. https://twitter.com.

Ulrich, J., 2016. TTR: Technical trading rules. R package.

Visaltanachoti, N., Yang, T., 2010. Speed of convergence to market efficiency for NYSE-listed foreign stocks. Journal of Banking & Finance 34 (3), 594–605.

Voya Investment Management, 2016. The impact of equity market fragmentation and dark pools on trading and alpha generation. https://investments.voya.com.

Wendell, B., 2017. How to trade the opening 15 minutes. Online Trading Academy.