Expert Systems With Applications 136 (2019) 171–186

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Buzzwords build momentum: Global financial Twitter sentiment and the aggregate stock market☆

Axel Groß-Klußmann[a,*], Stephan König[b], Markus Ebner[a]

[a] Quoniam Asset Management GmbH, Westhafen Tower, Frankfurt 60327, Germany
[b] University of Applied Sciences, Hannover, Faculty IV: Business and Computer Science, Ricklinger Stadtweg 120, Hannover 30459, Germany

ARTICLE INFO

Article history:
Received 27 February 2019
Revised 10 May 2019
Accepted 14 June 2019
Available online 14 June 2019

Keywords:
Text mining in finance
Social media data
Sentiment extraction
Trend-Following
Alternative data

ABSTRACT

We examine the long-term relationship between signals derived from nine years of unstructured social media microblog text data and financial market developments in five major economic regions. Employing statistical language modeling techniques we construct directional sentiment metrics and link these to aggregate stock index returns. To address the noise in finance-related Twitter messages we identify expert users whose tweets predominantly focus on finance topics. We document that expert users are the main drivers behind an interdependence between Twitter sentiment and financial markets. The direct prediction value of expert sentiment metrics for stock index returns, however, is found to be elusive and short-lived. Yet, we detect significant predictive gains over benchmark models in times of negative market returns. In consequence, the relation between expert sentiment metrics and stock indices is sufficient to devise hypothetically profitable cross-sectional as well as time series momentum investment strategies for futures based on Twitter signals that survive basic transaction cost assumptions. In this context, our results show that expert sentiment signals can yield higher risk-adjusted returns than classical price-based signals.

1. Introduction

Much research is devoted to processing the user-generated information shared through social media venues like Twitter in order to gain deeper understanding of the behavior of its users. Even though a large part of the information can typically be seen as irrelevant noise, the potential value of these data lies in the sheer volume of both users and messages which gives rise to the hope of capturing the wisdom of a large crowd. In light of this potential, numerous attempts in the financial domain exist to extract the author sentiment from tweets to predict future stock price returns. By now, financial data vendors like Bloomberg routinely incorporate tweets and predictive support functions based on textual data into their trading data systems.[1] However, no consensus exists so far with regard to the implementation of sentiment signals into investment strategies.

Our study is ultimately motivated by the lack of an expert system based on Twitter sentiment assisting decision making in investment strategies for stock index futures. Due to their low transaction costs in contrast to single stocks, futures on regional stock indices are popular derivative instruments for trend-following investment strategies requiring fast trading and high turnover. Systematic strategies for futures exploiting trends constitute an important part of the financial industry, managing ~$293 bn. in assets as of 2014 (see Baltas & Kosowski (2017)). Several empirical findings indicate that futures-based strategies are well-suited as investment framework around Twitter sentiment signals. First, Checkley, Higón, and Alles (2017) find the information dissemination and signal decay of Twitter sentiment to be very fast which requires quick trading. Second, textual sentiment-driven investment strategies in the extant literature such as Yang, Yu, and Almahdi (2018) are trend-following in nature and induce high turnover. Third, Garcia (2013) finds that stock predictability using sentiment measures is concentrated in recessions. As futures conveniently allow to hold short-selling positions, negative sentiments potentially offer profit opportunities to be exploited in

☆ **Disclaimer** Among other things, Quoniam invests in momentum strategies. The views expressed here are solely those of the authors and not necessarily those of affiliated institutions.

* Corresponding author.
E-mail addresses: axel.grossklussmann@gmail.com (A. Groß-Klußmann), stephan.koenig@hs-hannover.de (S. König), markus.ebner@quoniam.com (M. Ebner).

[1] https://www.bloomberg.com/company/announcements/bloomberg-launches-twitter-feed-optimized-trading/

https://doi.org/10.1016/j.eswa.2019.06.027

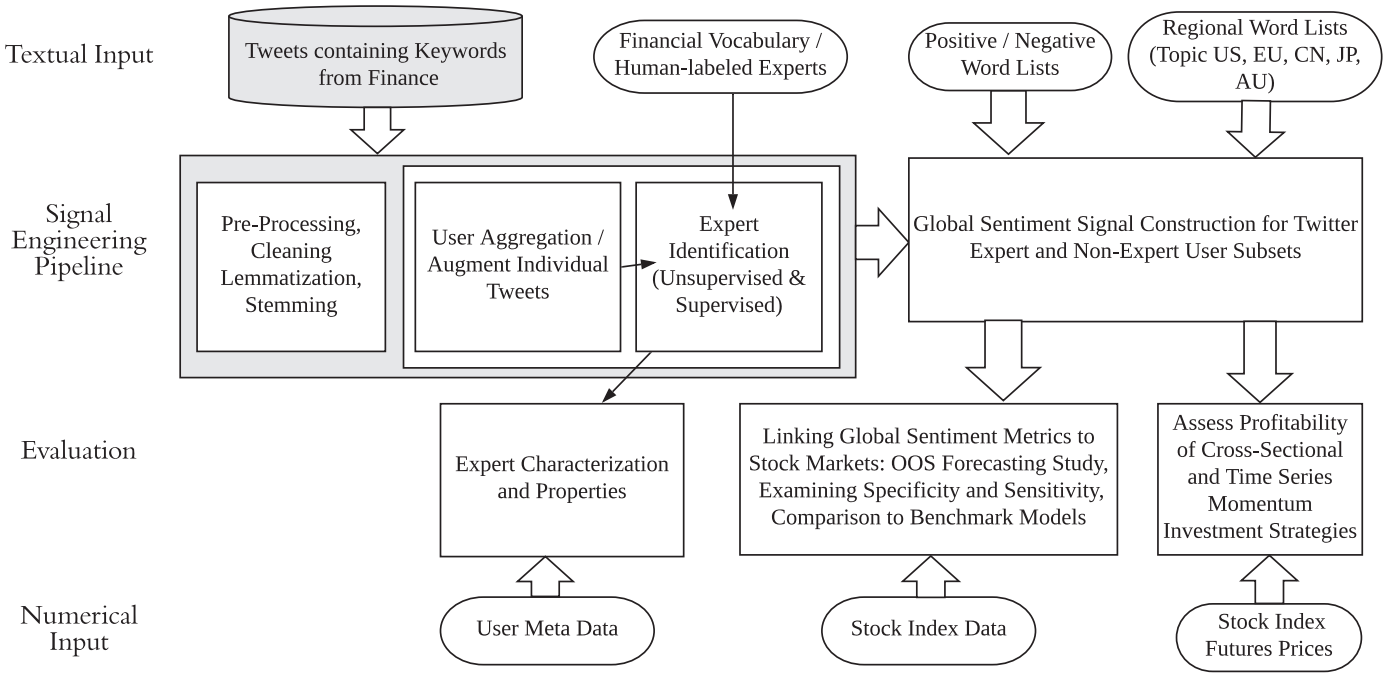

Fig. 1. Flowchart of the proposed analysis system.

adverse market environments. Finally, while the sparse tweet volume and news coverage for single stocks can drop to zero (see Ranco, Aleksovki, Caldarelli, Grcar, & Mozetic (2015)), the daily Twitter information stream for major stock index regions is rich and continuous, i.e. covers every time of a year. This permits constant positioning irrespective of news arrivals.

In the following we link Australian, Chinese, European, Japanese and US (AU, CN, EU, JP, US henceforth) stock markets to signals based on tweets. The stock market indices chosen are globally relevant and, importantly, liquidly tradable futures contracts exist for all indices. A central feature of our study is the nine year sample of tweets starting in 01/2010, which considerably increases the prediction sample from a previous high of 439 days (Oliveira, Cortez, & Areal (2017)). Most extant studies on Twitter data focus on the statistical significance of point forecasts of stock returns via sentiment signals, where results are mostly in line with the elusive and time-varying predictability of stock returns reported by Timmermann (2008). However, no existing study explores Twitter sentiment signals in the context of trend following strategies originally devised for stock index returns. This is surprising in light of the more recent observation that signals capturing textual news data can be more informative for future financial market states than end-of-day price signals (see Atkins, Niranjan, & Gerding (2018)). Moreover, trading strategies typically pose weaker requirements on signals than direct point forecasts. Most trading applications focus on the return decomposition discussed in, e.g., Christoffersen and Diebold (2006),

$$r_t = \text{sign}(r_t) \cdot |r_t|, \tag{1}$$

where both righthandside components feature attractive properties. While Leung, Daouk, and Chen (2000) argue the prediction of signs of returns, $\text{sign}(r_t)$, is superior to the prediction of return levels, the absolute return $|r_t|$ is a measure of return volatility exhibiting rich time series structure which results in significant predictability of realized variance (see Andersen, Bollerslev, Diebold, & Labys (2003)). Hence, a standardization or scaling with predictions of return variance offers a natural way to control the volatility of investment strategies based on forecasts of signs of returns. To the best of our knowledge, no study so far explores strategies for directional return forecasts from sentiments in combination with a volatility scaling derived from the decomposition (1). The potential long-term *economic* gains from Twitter signals are thus unclear. To close this gap we employ regionally aggregated sentiment measures in time series and cross-sectional momentum investment strategies for stock index futures. In addition, the high textual data volume available per region context makes it possible to consider more finance-orientated subsets of Twitter users (experts) for sentiment construction. We draw on works of Bar-Haim, Dinur, Feldman, Fresko, and Goldstein (2011), Si et al. (2014) and Ghosh et al. (2013) on social media expert identification and use techniques from computational linguistics to automatically identify network users with a distinct focus on financial contexts. The resulting user sets are used to construct 'expert' and 'non-expert' sentiment metrics.

Using $\sim$102 million tweets from a finance-related subset of all tweets posted from January 01, 2010 through September 30, 2018 we derive indicators capturing the sentiment among users on a given day. Our study draws on a dictionary approach to sentiment extraction where word polarity is measured according to positive and negative word lexicons. To enhance the signal precision we employ supervised and unsupervised statistical learning methods to group all users into expert and non-expert sets. Hand-crafted topic lists for five regions are introduced in order to assign sentiment measures to the regions. Next we assess the out of sample accuracies of Twitter sentiment. Finally, we put signals derived from Twitter expert sentiment measures to use in hypothetical daily trend-following investment strategies and explore their profitability. Fig. 1 details the signal construction and analysis framework alongside the data input.

In short, our main contributions to several strands of the literature are as follows.

1. We propose a new decision support framework for social media expert sentiment-driven macro strategies based on stock index futures. Next to its empirical foundation, a key building block of the framework is the partition of Twitter users into experts and non-experts. See result Section 4.1.

2. In Section 4.2 our analysis uncovers a better accuracy of directional expert sentiment metrics for predicting *negative* market return directions when compared to non-expert sentiment metrics, past returns and return classifications based on a SVM. The Diebold and Mariano (1995) test confirms the statistical significance of this finding.
3. We introduce distinct and novel features to investment strategies for sentiment in Section 4.3. Importantly, the use of futures allows to realize the significant accuracy increases of expert sentiment in negative markets with short-selling positions. Further, an ex-ante volatility scaling mitigates the over-reliance of sentiment trading gains on volatile periods.
4. The evaluation of the trading strategies for expert sentiments show that large differences in the prediction accuracy for negative returns directly result in corresponding hypothetical trading profits. Notably, the gains are obtained without employing the sentiment signals as features in statistical learning models.

The remainder of the paper is organized as follows. After a discussion of related literature in Section 2 we describe the data processing as well as NLP and machine learning approaches used for the modeling of the textual data in Section 3. In Section 4, we outline predictive properties of Twitter signals for stock market developments. Finally, we investigate momentum investment strategies based on sentiment signals. Section 5 gives conclusions.

2. Related literature

2.1. The link between sentiment metrics and financial markets

Nassirtoussi, Aghabozorgi, Wah, and Ngo (2014), Kumar and Ravi (2016) as well as Xing, Cambria, and Welsch (2018) review numerous attempts to capture the sentiment in textual sources like tweets in order to predict future stock price returns. Broad empirical support exists for the hypothesized interdependence between financial time series and text data like, e.g., earnings press releases or news articles for companies. This is covered in the works of Davis, Piger, and Sedor (2012), Garcia (2013), Loughran and McDonald (2011), Li (2008), Tetlock (2007) as well as Tetlock, Saar-Tsechansky, and Macskassy (2008). Feuerriegel and Gordon (2018) and Kraus and Feuerriegel (2017) find significant predictive content in sentiment measures of company releases for stock returns. While company news are examples of signals occuring at discrete (and possibly only few) dates, Antweiler and Frank (2004) and Das and Chen (2007) are early analyses linking financial market activity to signals based on a more continuous textual data stream from internet stock message boards. More recently, attention has been drawn to the Twitter platform, with Oliveira et al. (2017) and Checkley et al. (2017) among others documenting little added predictive value of tweet sentiment measures over benchmarks for stock index return forecasts. Bollen, Mao, and Zeng (2011) and Zhang, Fuehres, and Gloor (2011) report attractive prediction properties of Twitter sentiment and mood measures, albeit on non-representative samples. In contrast to aggregated sentiment measures for stock index topics, considerably lower textual data volume is available per individual sentiment construction for single stocks. In this respect, Sprenger, Tumasjan, Sandner, and Welpe (2014) as well as Ranco et al. (2015) find only comparably low correlations of Twitter sentiment metrics to single stock returns. Diverting from Twitter, Nguyen, Shirai, and Velcin (2015) and Chen, De, Hu, and Hwang (2014) highlight strong associations of social media sentiment from the StockTwits platform and www.seekingalpha.com to future stock returns. As the platforms considered in these works speak mostly to finance experts we interpret the results as indicating the usefulness of finance expert identification. Given the data limitations in most studies, the debate about the long-term benefits of social media signals for financial decision making remains inconclusive so far.

2.2. Momentum investing based on price and textual signals

Our study is related to recent works on systematic investment strategies for futures. In this context, Moskowitz, Ooi, and Pedersen (2012), Hurst, Ooi, and Pedersen (2013) and Baltas and Kosowski (2017) describe successful time series momentum strategies (*TSM* henceforth). For the TSM, signals for assets are constructed relative to their own past return performance. A different approach is taken in works like Asness, Liew, and Stevens (1997), Miffre and Rallis (2007) and Asness, Moskowitz, and Pedersen (2013) who devise profitable cross-sectional momentum strategies (*CSM* henceforth). Unlike the TSM, the CSM constructs signals for assets based on their past over- or underperformance within the cross-section of assets.

In the context of textual signals, Renault (2017) is a recent example for an intraday trading strategy for the S&P 500 ETF based on Twitter signals. Apart from Twitter, Feuerriegel and Prendinger (2016), Yang, Mo, Liu, and Kirilenko (2017) and Yang et al. (2018) introduce explicit day-trading strategy systems based on signals derived from company news texts. Nasseri, Tucker, and de Cesare (2015) outline an investment strategy built on word mentions on the StockTwits platform. The studies show gains over purely price-based strategies. However, the improvements are significant mostly for combinations of price return and sentiment signals in statistical learning models. Thus, it is hard to disentangle the role of the raw text-based signals from qualities of the statistical models. The reliance on statistical learning models is in stark contrast to the above futures strategies where signals are exclusively based on past returns and hence widely *model-free*. A further difference of the sentiment momentum strategies to their price-based counterparts lies in the ruling out of short-selling. The sentiment strategies are in general cast as long-only TSM strategies with investment exposures (asset weights) either 0% or 100%. In consequence, the fixed weight grid means the volatility of the sentiment strategies is uncontrolled for which can lead to artificially time-varying trading gains dependent on the ups and downs of market volatility. Our study aims to address these issues of trading exercises for text-based signals news in introducing novel CSM and TSM investment frameworks for Twitter sentiment.

3. Materials and methods

3.1. The tweet sample

The textual data used for the present study consists of microblog messages with up to 140 characters on the Twitter social media platform. To concentrate more on the finance topic and reduce the amount of data to be processed, we focus on tweets containing at least one of the words in the search term list {'stocks', 'stock market', 'economy' }. In addition to the pure text, we obtain non-personal meta information on the user sharing the message as well as the exact time and language.

Using the Java Twitter streaming API we automatically downloaded tweets from November 30, 2013 to September 30, 2018 in real time.[2] The public stream is constrained to 1% of all available messages on a given day. This is achieved through random sampling of the tweets containing one of the search terms. However, due to the search term list restriction we conjecture the constraint is seldom relevant to our analysis. Even the highest daily number

[2] An exception is a 30-day period from June, 25th to July, 24th 2015 where no Twitter data was downloaded.

of tweets collected by us, 974,201 on 06/24/2016, stays way below the 2016 average number of tweets per day given by Twitter, 500 million. In addition to the API we augment the data by historic tweets during January 01, 2010 to November 29, 2013 obtained from the Gnip, Inc., company, now a subdivision of Twitter. In this second subsample the daily amount of tweets is similarly constrained to 1% of the overall tweet volume. Our 3,158-day sample of size 83.2 Gigabytes is comprised of 102,737,864 tweets, with 10,151,869 tweets posted before Nov. 2013 and 92,585,995 tweets stemming from the streaming API.

Liu, Kliman-Silver, and Mislove (2014) and Morstatter, Pfeffer, Liu, and Carley (2013) highlight data consistency issues with the streaming API. However, the study of Gonzalez-Bailon, Wang, Rivero, Borge-Holthoefer, and Moreno (2014) implies that lower daily aggregation frequencies as employed in our study mitigate possible biases.

3.2. Finance experts on Twitter

To improve topic accuracy and identify subgroups of finance tweets we aim to discover finance 'experts', i.e. users who predominantly share tweets on the finance subtopic. Several websites portray such twitter experts for finance. Screening these websites we identified 190 experts and additionally inspected individual tweets to confirm the selection. The appendix A.1 gives details on the process.

3.3. Sentiment and domain-specific topic word lists

Dictionary-based sentiment measures for documents can utilize positive and negative word lists for the construction of sentiment signals. The following results are based on the positive and negative word lists put forward in Hu and Liu (2004) and published at https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. It contains 6800 words with 2006 positive and 4783 negative. We further consider the Harvard General Inquirer IV-4 (HGIV-4) psychological dictionaries based on Stone, Dunphy, Smith, and Ogilvie (1966) available at http://www.wjh.harvard.edu/~inquirer/.

To account for negations and reversals in sentiment polarities we follow Oliveira et al. (2017) and use a negation list inspired by the sentiment tutorial of Christopher Potts, http://sentiment.christopherpotts.net/lingstruc.html. The exact list of negation terms is given in the appendix A.2.

The HGIV-4 dictionary covers a broad range of topics beyond positive and negative word lists. In the course of the study, we make use of word lists from the HGIV-4 on the topic 'economy' to establish similarity of the textual data to this context.

To refine aspects (or targets) of the sentiment measures we hand-craft topic word lists for the five regions under considerations. Each topic word list is composed according to the following protocol based on four rules applied per topic AU, CN, EU, JP and US.

1. Include country, region and currency names and denominations and corresponding international abbreviations.
2. Include names and financial market mnemonics of major financial instruments tracking local stock markets, FX cross rates and fixed income.
3. Include important political and economic entities, recurring events and agents like, e.g., the Fed and its chairmen for the US topic.
4. Include single stock constituent names (without legal form abbreviation) of a major regional equity index and corresponding official mnemonics.

Data for rules 1, 3 and 4 can be easily obtained via the Wikipedia knowledge base. To prevent hindsight biases we access the page as of Dec. 2009, i.e. before our sample starts. An exception is rule 3, where we include all government and central bank heads covering the sample span as all candidates were known long before elections and appointments. For rule 2 we collect name and mnemonic data from Yahoo! Finance, Twitter (cashtags) and the Bloomberg system. The regional equity indices for which we pull single stock constituent names and mnemonics are the ASX200 (AU), the HangSeng (CN), the EURO STOXX 50 (EU), Nikkei225 (JP) and the S&P 100 (US). Finally we clean the lists from single letters and items with ambivalent meaning like, e.g. 'news', a company whose stock is listed on US and Australian stock exchanges.

The five region topic lists can be downloaded as.csv-file from our web appendix at https://www.dropbox.com/s/284yatbsc1nco09/TopicLists.csv?dl=0.

3.4. Financial time series

The major stock market indices representing important developments around the world are taken to be the S&P ASX 200 index for the region 'AU', the Hang Seng Index for CN, the Euro Stoxx 50 index for EU, the Topix for JP and the S&P 500 index for US.

We simulate trading based on signals for the regional indices with prices of the corresponding generic futures. The futures track the development of the aforementioned stock market indices closely via liquidly tradable contracts. All data are taken from the Bloomberg system. Corresponding identifiers are given in appendix A.3. Log returns r_t are always constructed from prices p_t at the end of trading day t as $r_t = \log(p_t) - \log(p_{t-1})$.

3.5. Processing of the textual data and software tools

All software implementations for textual data processing in this study are based on the Python programming language. Later statistical learning algorithms are taken from the Python scikit-learn library put forward by Pedregosa et al. (2011). For the initial processing and word normalization we employ the Python Natural Language Toolkit (http://www.nltk.org). We remove punctuation and stop words like 'the'. Hashtags, i.e. links to trending topics on Twitter, are treated as common words after stripping the '#'-sign. All characters are changed to lower cases. Moreover, to reduce the fraction of noise induced by informal language on Twitter we discard all words not contained in the union of all word lists from 3.3. With the exception of the final sentiment construction via word lists we further use the Porter (1980) stemming algorithm and operate on the set of stemmed terms.

3.6. Addressing data sparsity and noise in microblog messages

Three main strands of strategies exist to deal with the data sparsity and noise of tweets. First, Hong and Davison (2010) evaluate topic models for Twitter and propose to aggregate or pool tweets by user or hashtag to end up with bigger document sizes. This approach is also taken by Mehrotra, Sanner, Buntine, and Xie (2013) and Quan, Kit, Ge, and Pan (2015).

Second, sensible clusters of tweet sets can be found based on external information rather than the information given by the short tweets alone. Jin, Liu, Zhao, Yu, and Yang (2011) use small auxiliary texts to group tweets, similar to Phan, Nguyen, and Horiguchi (2008) and Phan et al. (2011) who propose to model latent topics with the help of outside information. Yildirim, Üsküdarli, and Özgür (2016), Tran, Tran, Hadgu, and Jäschke (2017) as well as Gattani et al. (2013) link tweets to external topics taken from the knowledge base Wikipedia.

Third, to deal with the large amount of uncoordinated postings and general noise, Si et al. (2014), Ghosh et al. (2013) and Bar-Haim

et al. (2011) among others uncover expert networks on Twitter to achieve precision gains in, e.g., topic and sentiment extraction.

In line with the above approaches we make use of external domain-specific word lists to model topics as well as author sentiments. Moreover, we use outside information to detect experts posting finance-related messages. The small set of pre-selected known 'experts' (see Section 3.1) allows to utilize supervised learning methods for the classification of all remaining Twitter users into an expert and non-expert group. To avoid dependence on the human element in the pre-selected expert sample we further employ unsupervised learning methods and identify twitter user groups as clusters in the data.

3.7. The vector space document representation

Under the standard notation in the information retrieval literature we formally define a vocabulary (or lexicon or word list) $\mathcal{V}$ as consisting of V terms, $\mathcal{V} = (w_1, w_2, \ldots, w_V)$. A tweet is now identified with a *document* D, $D = (w_1, w_2, \ldots, w_M)$, formed by a tuple of M terms (words). Further, a collection of several tweets like all N tweets on a day is denoted a *corpus* $\mathcal{C}$, $\mathcal{C} = \{D_1, D_2, \ldots, D_N\}$. Our setting follows the so-called bag-of-words model, where information on order of terms or grammar is discarded.

A numerical representation of text is given by the popular tf-idf scheme outlined in Salton and McGill (1986). Tabulating tf-idf-weighted occurrences of terms contained in the V words of the vocabulary $\mathcal{V}$ for each document on a day we obtain a *term document matrix* (tdm) M per day,

$$M = (M_{i,j}) = \text{tf\_idf}\ (D_i, w_j), \quad D_i \in \mathcal{C}, w_j \in \mathcal{V}, \quad i \in \mathbb{N}_N\, j \in \mathbb{N}_V \tag{2}$$

The column index j stands for the position of a term in the word list $\mathcal{V}$. In our case, each row vector $M_{i\cdot} \in \mathbb{R}^V$ represents a tweet. To minimize effects of different tweet lengths when comparing documents, all rows of M are normalized.

Documents are considered more similar the smaller the Euclidean distance between them or the smaller the angle between their tdm rows. In the latter case the similarity between documents D_i, D_{i^*} in M can be computed based on the cosine similarity

$$\text{sim}(D_i, D_{i*}) = \frac{||M_{i\cdot} \cdot M_{i*\cdot}||^2}{||M_{i\cdot}|| \cdot ||M_{i*\cdot}||} \in [0, 1]. \tag{3}$$

3.8. Unsupervised learning methods for language modeling

Unsupervised machine learning methods like K-Means Clustering can be used to detect hidden structure in the tdm M by grouping text data into clusters of similar topics. However, the typically extremely large dimension of M ($\mathbb{R}^V$ with $V \gg 100$ for most vocabularies) requires to first address the 'curse of dimensionality' problem.

3.8.1. Dimension reduction based on Latent Semantic Analysis

Introduced by Deerwester, Dumais, Landauer, and Harshman (1990), the LSA or latent semantic indexing is based on a singular value decomposition (SVD) of the transposed tdm, M',

$$M' = S\Lambda^{\frac{1}{2}}R', \tag{4}$$

where $SS' = I$ and $R'R = I$. $\Lambda^{1/2}$ contains square roots of eigenvalues from S in descending order.

To lower the dimensionality of the tdm a *reduced* SVD is computed to approximate M while preserving its characteristics. For the reduced SVD transformation based on (4) we first keep only the top k eigenvalues of Λ and obtain $\Lambda_k^{1/2}$. Second, we retain the k corresponding rows of S to get a reduced matrix S_k. Next we transform the tdm according to $M_k' := S_k\Lambda_k^{1/2}R_k'$, where $R_k = (\Lambda_k^{-1/2}S_k'M')'$. Mitigating the curse of dimensionality we now have dimension $N \times k$ for M_k with $k \ll V$, where V is the length of the full vocabulary.

3.8.2. K-Means Clustering

The K-Means algorithm of MacQueen (1967) is a heuristic method to partition data into K clusters such that the within-cluster data are maximally close to the cluster centers according to a distance measure. After dimension reduction we apply the K-Means to the tdm in order to find user clusters with similar shared Twitter content.

Clusters for normalized rows of M are based on the squared Euclidean distance measure $||M_{i\cdot} - M_{l\cdot}||^2 = \sum_{j=1}^{n}(M_{ij} - M_{lj})^2$. The i-th document vector is assigned to its closest cluster with mean m_k via the mapping $i \to C(i) = \arg\min_{1\le k\le K} ||M_{i\cdot} - m_k||^2$. Hastie, Tibshirani, and Friedman (2009) describe how the converging K-Means algorithm finds an optimal cluster assignment function C by repeatedly minimizing the total cluster variance.

3.8.3. Topic modeling via the Latent Dirichlet Allocation

Topic models are popular methods to compactly represent textual data with a few topics. In this context, topics are given as (weighted) collections of important terms in the documents at hand. However, as most topic models are unsupervised, the resulting topics typically need to be labeled and characterized further.

In a seminal paper, Blei, Ng, and Jordan (2003) put forward the Latent Dirichlet Allocation model (LDA). The LDA belongs to the class of hierarchical Bayesian models and describes documents as a mixture of topics. Given the number of topics, k, each document is assumed to be generated by its topics z, a $(k \times 1)$ multinomial random variable, and the words which have conditional multinomial probabilities.

After training via an variational EM algorithm, the LDA yields document topic distributions, i.e. the proportion of each topic in a document as well as topic word distributions, i.e. the proportions of terms within the topics.

3.9. Supervised learning methods for document classification

Given the pre-selected expert data we learn the unknown true function f: $M_{i\cdot} \to \{1, -1\}$ assigning expert and non-expert group labels to individual tweets. The task at hand represents a supervised document classification problem and our approach follows the strategies reviewed in Mironczuk and Protasiewicz (2018).

3.9.1. TDM dimension reduction with the χ^2-Statistic

To select terms t with high discriminatory power between two groups, Schütze, Hull, and Pedersen (1995) propose to use a χ^2-test for independence of the events $\{t \in D\}$ and $\{D \in G\} :=$ 'D issued from a user of group G'. High deviations resulting in large values of the χ^2-statistic indicate a dependency of term and group occurrence and thus a high discriminatory power of the term for group classification. Ultimately, the χ^2 test statistic is used to rank the features and keep the top k features, hence resulting in a dimension reduction of the tdm for $k \ll V$.

3.9.2. Support Vector Machines

Support vector machines were put forward by, e.g., Vapnik (1995) and Schoelkopf, Burges, and Vapnik (1995) as classifiers for a wide range of applications. Joachims (1998) and Burges (1998) document the good performance of linear SVMs in text classification exercises. The fundamental idea behind SVMs is to find a hyperplane in the data vector space such that the labeled training data are separated according to the class labels.

We operate on the document space in $\mathbb{R}^n$ spanned by the N rows of M. Let $y_i \in \{-1, 1\}$ represent the expert or non-expert class

of the tweet given by the row $M_{i.}$. The classification itself is based on the set $\{x \in \mathbb{R}^n \mid g(x) = \omega' x + b = 0\}$ representing a hyperplane. Individual data points (tweets) x are classified into $\{-1, 1\}$ according to the decision function $f(\cdot) = \text{sign}(g(\cdot))$, which gives the location of features x with respect to the hyperplane. In an ideal scenario, the hyperplane perfectly separates the classes, while in general, classes are allowed to overlap in the SVM.

The optimal weight vector $\omega \in \mathbb{R}^n$ and bias $b \in \mathbb{R}$ are given as solutions to an optimization problem where the margin between the separating hyperplane and the two classes is maximized. The rationale for the margin maximization is to ultimately reduce the classification error on unseen data after training the SVM on a labeled training sample. After numerical optimization, the estimated group of an unlabeled tweet x, represented by a row in M, is given by $\widehat{f}(x) = \text{sign}(\widehat{\omega}' x + \widehat{b})$, where the hats denote optimal parameters.

3.10. Extracting sentiments from microblogs

For sentiment construction we rely on measuring word polarity via the established positive and negative word lists outlined in subsection 3.3. Following Antweiler and Frank (2004) we compute sentiment measures as the normalized difference of positive and negative mentions,

$$S = \frac{\#(\text{positive terms}) - \#(\text{negative terms})}{\#(\text{positive terms}) + \#(\text{negative terms})}. \tag{5}$$

By construction, we expect high values of (5) to reflect positive developments on financial markets, resulting in a positive correlation between returns and twitter sentiments. The period of the sentiment formation, i.e. the data used for computation of S, will be a trading day. Taboada, Brooke, Tofiloski, Voll, and Stede (2011) show that the word list approach is a robust way of obtaining opinions from text while keeping the computational burden moderate. An excellent overview of more refined sentiment analysis is given by Liu (2015) as well as Ravi and Ravi (2015).

To account for sentiment polarity inversion of terms due to negations we define as negative context the text segments following a term from the negation word list (A.2) up to a punctuation mark. In the negated contexts we simple invert the polarity of the terms occurring, i.e. positive terms become negative and vice versa.

4. Results

4.1. Characterizing expert clusters in the tweet haystack

In the following we establish daily 'expert' and 'non-expert' user sets on Twitter. As shown by Si et al. (2014), Ghosh et al. (2013), Bar-Haim et al. (2011) and Yang, Mo, and Liu (2015) among others, the precision of social media sentiment extraction can be improved substantially with expert networks.

We derive expert user candidates in two ways based on both supervised and unsupervised statistical learning techniques. The latter serves to substantiate results and to address potential problems like hindsight biases often experienced with ex-post human-labeled training data for supervised approaches. Ultimately, the two approaches permit to draw conclusions independent of the varying degrees of outside information, parameter calibration and the human element employed.

In both supervised and unsupervised cases we mitigate the data sparseness in tweets (see the overview 3.6) and first pool tweets per user to enrich the individual data history. The pooled document corpus $\mathcal{C}_t^U$ on each day t is comprised of user-aggregated tweets,

$$\begin{aligned} &\mathcal{C}_t^U = \{D_{u_1}, D_{u_2}, \ldots, D_{u_n}\}, \quad \text{with} \\ &D_{u_i} = \{\text{Document made of all tweets up to day } t \\ &\text{of user } u_i \text{ tweeting at } t.\} \end{aligned} \tag{6}$$

The supervised learning of daily expert sets is based on the 190 handpicked expert users described in Section 3.2. Unfortunately, the number of users from this sample posting tweets per day can be very low such that sentiment measures from the 190 experts alone are typically very volatile. Further, handpicking an ever larger set of experts becomes increasingly infeasible. In this context, supervised statistical learning methods can be employed to automatically increase the small initial expert sample. The problem represents a classification problem where unseen data is classified into positive (experts) and negative (non-experts) groups after training the classifier with the help of the preselected expert sample (see Section 3.9).

Supervised classification models require both a pre-labeled positive *and* a pre-labeled negative sample. To obtain a non-expert sample we rely on the similarity of tweets to the HGIV-4 'economy' topic word list. We choose users associated with tweets having a cosine similarity measure smaller than the lowest empirical 15%-percentile as the non-expert set. Due to the up to 180,000 tweets collected per day, the portion of such chosen non-experts is substantially greater than its expert counterpart. Drawing on He and Garcia (2009) and Haerdle, Lee, Schaefer, and Yeh (2009) we downsample the non-expert set by selecting thrice as many non-expert users as expert users per day to facilitate the statistical learning by a better data balance. The reason for still keeping more non-expert data lies in the generally larger texts in the expert sample.

The pre-labeled expert and non-expert data is split into daily training (two thirds) and held out test samples (one third). To maintain large test and training sets per day we include the complementary users from past training and test samples whenever these have not posted tweets on a specific day. We reduce the dimension of the document vector space to the 50 most important terms using the χ^2-feature selection procedure based on the expert/non-expert training sample (see 3.9.1).

We train the SVM on the daily user-aggregated training sets based on $\mathcal{C}_t^U$. Hyperparameters are calibrated daily through maximization of the F1-Score using 10-fold cross-validation. After obtaining a daily estimate $\widehat{f}$ for the decision rule f: $M_{i.} \to \{1, -1\}$ assigning expert group labels to tdm rows we classify the remaining daily tweet data. The daily classification performance of the SVM is evaluated on the held out test samples. In the next step we compute standard classification metrics like the ROCAUC, accuracy, precision and specificity averaged across the sample. We observe consistently high values (above 0.9) for all metrics considered, which indicates good overall classifier performance. We conclude that statistical features of the expert and non-expert sets are quite distinct and hence the classes well separable.

The main building block of our unsupervised approach to expert identification is the K-Means algorithm (3.8.2) which is a major workhorse in clustering analyses. Due to its simplicity and the guaranteed convergence the K-Means is especially well suited for our use-case of an application on a daily frequency (i.e. over 3000 iterations). The unsupervised learning methods are applied to each user-aggregated daily corpus $\mathcal{C}_t^U$. We begin with a dimension reduction via the LSA (Section 3.8.1) and cut each daily document vector space to dimension 50. Clusters of similar users can be obtained via the K-Means applied to the reduced document spaces. Elbow plots for the K-Means indicate that three clusters are a conveniently low number of clusters capturing a large portion of the variance in the data. After iterating over the days with the K-Means algorithm we identify three clusters, i.e. three distinct sets of similar users, per day.

Table 1
Average characteristics for user sets. Abbreviations are: cor.Econ. - similarity to the HGIV-4 economy word list, tw.-length - length of tweets, tw./ user - tweets per user, exp. overlap - overlap of expert groups in per cent, predef. exp. coverage - percentage of handpicked experts included in group. * denotes significance of the mean vs. the 'all' -group at the 10% level. ** is the 5% level.

user group	cor. Econ.	follower	tw.-length	tw./user	nr. of users	exp. overlap	predef. exp. coverage
KMexp	0.11**	9807**	3.59**	4.83**	5523	0.42	0.81
KMnonexp	0.05**	2771**	3.82**	4.86**	7744	-	0.19
SVMexp	0.11**	9507**	3.77**	5.34**	4596	0.36	0.95
SVMnonexp	0.04**	2331**	3.70**	4.59**	8672	-	0.05
all	0.07	5052	3.70	4.82	13,273	-	-

Table 2
Top three LDA topics with weights (LDA topic distribution) and LDA word distributions extracted from aggregation of all tweets per user group.

user group	topic-Nr.	weight	content
KMexp	1:	(36.48%)	economy (26.04%) + equity (2.83%) + business (1.37%) + china (1.24%)
KMexp	2:	(22.43%)	economy (27.75%) + equity (3.01%) + market (0.87%) + business (0.77%)
KMexp	3:	(14.01%)	economy (25.69%) + equity (2.62%) + market (1.64%) + stock (1.53%)
KMnonexp	1:	(17.93%)	economy (16.15%) + stock (8.84%) + equity (5.4%) + home (1.33%)
KMnonexp	2:	(15.8%)	economy (16.47%) + equity (9.79%) + stock (2.56%) + business (1.59%)
KMnonexp	3:	(14.23%)	economy (18.49%) + equity (7.06%) + stock (2.38%) + business (1.56%)
SVMexp	1:	(28.77%)	economy (24.88%) + equity (4.53%) + market (1.2%) + stock (1.02%)
SVMexp	2:	(19.53%)	economy (16.02%) + equity (9.15%) + stock (4.09%) + market (1.64%)
SVMexp	3:	(14.43%)	economy (22.24%) + equity (5.07%) + stock (2.78%) + market (2.35%)
SVMnonexp	1:	(20.91%)	economy (25.61%) + equity (3.02%) + business (1.65%) + money (0.75%)
SVMnonexp	2:	(18.3%)	economy (20.27%) + equity (7.03%) + business (1.62%) + stock (1.46%)
SVMnonexp	3:	(14.51%)	economy (20.12%) + stock (7.35%) + equity (3.26%) + home (1.02%)
all	1:	(28.74%)	economy (22.77%) + equity (5.42%) + stock (1.48%) + business (1.43%)
all	2:	(18.61%)	economy (20.01%) + stock (5.81%) + equity (4.47%) + home (1.02%)
all	3:	(17.91%)	economy (21.41%) + equity (6.01%) + stock (1.78%) + market (1.45%)

To obtain meaningful representatives for the user clusters we re-label the clusters in an order corresponding to their average tweet similarity to the economy word list of the HGIV-4 dictionary outlined in Section 3.3. After the re-labeling, the users from the clusters labeled '3', i.e. the maximum cluster number, are from the cluster with overall highest cosine similarity to the 'economy' topic given by the external word list. We take these to be the expert candidates while clusters '1' and '2' form the non-expert group.

After the application of the SVM and K-Means algorithm we further inspect the resulting user groupings. The user groups will henceforth be denoted 'all' for the full sample of users, 'KMexp' and 'KMnonexp' for the K-Means clusters and 'SVMexp' and 'SVMnonexp' for the SVM user sets. Table 1 characterizes the sets of users along several meta data dimensions with the following findings emerging. Two-Sample t-tests against the 'all' -group confirm the significance of differences of the averages at the 5% level. First, we observe that our expert candidates feature a higher number of followers on average than both full sample and non-expert sets.[3] This is in line with Weng, Lim, Jiang, and He (2010) who find that the follower number of a user is positively related to her influence and importance. Second, all expert groups exhibit higher cosine similarity to the HGIV-4 'economy' topic than both full sample and non-expert groups.[4] Third, computing the overlap of expert user groupings (column 'exp. overlap') we find that 42% of Kmeans experts are also in the SVM expert sample while 36% of the SVM experts overlap to the KMeans experts. Further, the predefined set of 192 hand-picked experts allows to explore the quality of the unsupervised expert clustering method. In this respect we observe that 85% of the pre-defined experts are correctly contained in the KMeans expert clusters (column 'predef. exp. coverage'). Taken together we interpret the findings as indicative of two distinct user groupings (due to the low overlap) which both exhibit expert properties (high coverage of pre-defined experts).

We employ the Latent Dirichlet Allocation (3.8.3) to back-engineer the dominant topics, i.e. the characteristic talk, in the candidate expert and non-expert document groupings for the full sample. Before applying the LDA we form a corpus of daily documents made of all tweets aggregated by user group per day. Next, the LDA with $k = 10$ topics is trained on this corpus. The trained LDA is subsequently applied to the two documents comprised of the aggregated full sample of tweets, with aggregation again by user sets. Table 2 gives the major topics uncovered for each user cluster under consideration. The table shows the weights for the top three topics according to the document *topic distribution* as well as weights for the top four words within these topics according to the topic *word distributions*.

We identify two dominant themes or sub-topics reflected in the LDA-based word distributions. The 'economy' theme is characterized by the corresponding term, 'economy'. We further interpret the occurrence of the terms 'equity' or 'stock' together with 'market' as the sub-topic 'stock market'. Subtle differences among topic word distributions are visible across user sets. Comparing the top three topics for expert and non-expert users based on the KMeans we find the expert word distributions dominated by the economy-theme as reflected by the weight for the corresponding term. Conversely, the KMnonexp word distributions substantially down-weight the economy-theme compared to the full sample word distributions ('all') and emphasize the stock and equity topic more. In addition, the term 'china' reveals a more international focus in topic 1 of the KMexp set compared to a domestic focus ('home') for topic 1 of the complementary set KMnonexp. In the supervised case (SVMexp), the three expert topics assign markedly higher weights to the stock market topic compared to both the SVMnonexp group and the full user sample group. Taken together we interpret the findings as indicative of a split of the user set along the economy topic for the KMeans and along the stock market topic for the SVM classification. Differences in expert

[3] The average given is based on a representative list of followers for 314,936 users retrieved through 07/2014.

[4] Note that only in case of the K-Means clusters this is a direct consequence of our cluster re-labeling.

Table 3
Descriptives for stock indices in 2010–2018. 'ret.' gives the annualized return in per cent. 'vol.' gives the annualized standard deviation in per cent. 'dwn.dev.' and 'up.dev.' give the annualized downside and upside deviation, respectively. Down days and up days refer to the number of days with negative and positive returns, respectively.

Name	ret.%	vol.%	dwn.dev.%	up.dev.%	down days	up days
Euro Stoxx 50	2.0	20	15	14	1132	1181
S&P 500	11.0	15	12	10	1021	1246
Hang Seng	2.7	18	13	11	1076	1146
Topix	7.6	20	15	13	1038	1172
ASX 100	3.0	14	10	09	1083	1197

and non-expert topics are further amplified by the topic distributions itself (column 'weight'). In both the supervised and unsupervised case we observe leptokurtic topic distributions concentrated at the topics 1 and 2 for the two expert sets compared to platykurtic distributions for the non-experts.

In summary, the two approaches uncover two distinctly different candidate expert sets. Both the unsupervised and the supervised approach are able to detect external hand-picked experts with more than 80% accuracy in the KMeans case and more than 95% accuracy in the SVM case. Twitter users from all candidate expert sets have significantly more followers than their non-expert counterparts. In case of the SVM split the experts tend to chat more about the stock market topic than the non-experts. In contrast, the topics for KMeans experts are dominated by the 'economy' topic.

4.2. Directional expert sentiment for major global equity markets

Collecting users according to the expert and non-expert groups per day we compute group-specific sentiment measures S from subsection 3.10 for both SVM and KMeans user splits. As aspects (or *targets*) for which user-group-specific sentiments are formed we consider sentiments capturing financial market developments for separate regions like Australia, China, Europe, Japan and the USA. In addition to providing robustness and more precision over an analysis for just one region context, such regional sentiments will allow to devise and test trading strategies exploiting cross-sectional stock market differences.

In light of the difficulties encountered in topic modeling for short texts we follow the strategy outlined in Section 3.6 and employ prior domain knowledge for the region topic extraction.[5] We use distinct word lists of terms from all five region topics and require a message to contain at least a term from either of these lists in addition to one of the finance terms. The five topic word lists constructed according to the rule protocol given in subsection 3.3 represent external domain knowledge to efficiently infer topics from short tweets. Major parts of the word lists are comprised of alternative region designations, subregion names, financial index and currency identifiers from the two regions under consideration. A tweet used in the construction of the 'US' sentiment measure, for example, will necessarily contain a finance term and a term from the US word list. Notably, Zhao et al. (2011) find that single tweets are typically about a single topic, thus supporting the word list approach for short microblog entries over more advanced multi-topic models such as the LDA (3.8.3). Our approach is also taken by, e.g., Oliveira et al. (2017), Ranco et al. (2015) as well as Sprenger et al. (2014) among numerous others in the topic modeling for individual stocks. In these works, cash-tag symbols $ and company name mentions are required for a tweet to be included in a certain stock or stock market sentiment. Moreover, in the formation of regional daily sentiment series we respect time zone differences around the world to match the stock price series recorded at the end of local trading days. This means we define days according to an $\sim$UTC+8h time shift for the regions CN, JP and AU compared to an $\sim$UTC-6h shift for the US.

We link sentiment metrics for the five regions to the stock indices outlined in Section 3.4. While Fig. 2 illustrates the development of regional SVM-expert sentiment measures over the sample spanning 01/01/2010 to 09/30/2018, Table 3 gives key descriptive statistics of the daily stock index returns. Our sample is characterized by generally positive annual returns and a higher number of days with positive returns than negative. However, most strikingly, the downside deviation, i.e. the standard deviation for negative returns always outweighs the corresponding volatility for positive returns (columns dwn.dev. vs up.dev.). The negative skewness in stock index returns is a long-standing stylized fact of financial markets (see Albuquerque (2012)) and of great importance to our analysis: A return predictor focusing on positive returns would achieve a high but unbalanced accuracy and still fail to capture a large portion of the volatility in the returns.

Since we aim to connect our analysis of the sentiment-to-return predictability to momentum investment strategies for futures we refrain from computing point forecasts and focus on predicting the sign of daily returns. As argued in the introduction, investment strategies typically draw on pure binary direction signals from the set $\{-1, 1\}$. In this regard, we construct binary signals based on the sentiment change with respect to a lookback period. Letting $\bar{S}_{t_1,t_2}$ denote the sentiment average from t_1 to t_2 we compute the sentiment change

$$\Delta S_t = S_t - \bar{S}_{t_1,t_2} \tag{7}$$

and construct (binary) sentiment directions $S_t^{s(\Delta)}$ according to its sign, $S_t^{s(\Delta)} := \text{sign}(\Delta S_t)$. The sentiment metrics (7) are established and also called sentiment shocks in the extant literature like, e.g., Yang, Song, Mo, Datta, and Deane (2015) and Song, Liu, and Yang (2017). Further, the difference between a shorter and longer trend bears resemblance to classical moving average crossovers used in trend-following strategies (see Banga and Brorsen (2019) for an overview). In the following we set $t_1 = t - 1$ and $t_2 = t - 12$ and compare $S_t^{s(\Delta)}$ with the stock market return polarity, $\text{sign}(r_t)$.

Table 4 shows mean directional accuracies $1/T \sum_t \mathbb{1}\{S_t^{s(\Delta)} = \text{sign}(r_t)\}$ for contemporaneous and lagged $S^{s(\Delta)}$. We further report true positive (sensitivity) and true negative rates (specificity), denoted + and - , i.e. the portion of correctly positive predicted values per total count of true positive returns and the portion of correctly negative predicted per total count of truly negative return. The table gives both individual numbers per region context and expert group as well as averages across regions. Most importantly, the contemporaneous evaluation (column $t \leftrightarrow t$) shows that directional expert sentiment metrics capture contemporaneous market directions well. The accuracies for both expert sets surpass 65% on average and are thus in line with an expected positive relationship

[5] See also the strategies in Chen and Liu (2014), Andrzejewski, Zhu, Craven, and Recht (2011) and Chen et al. (2013).

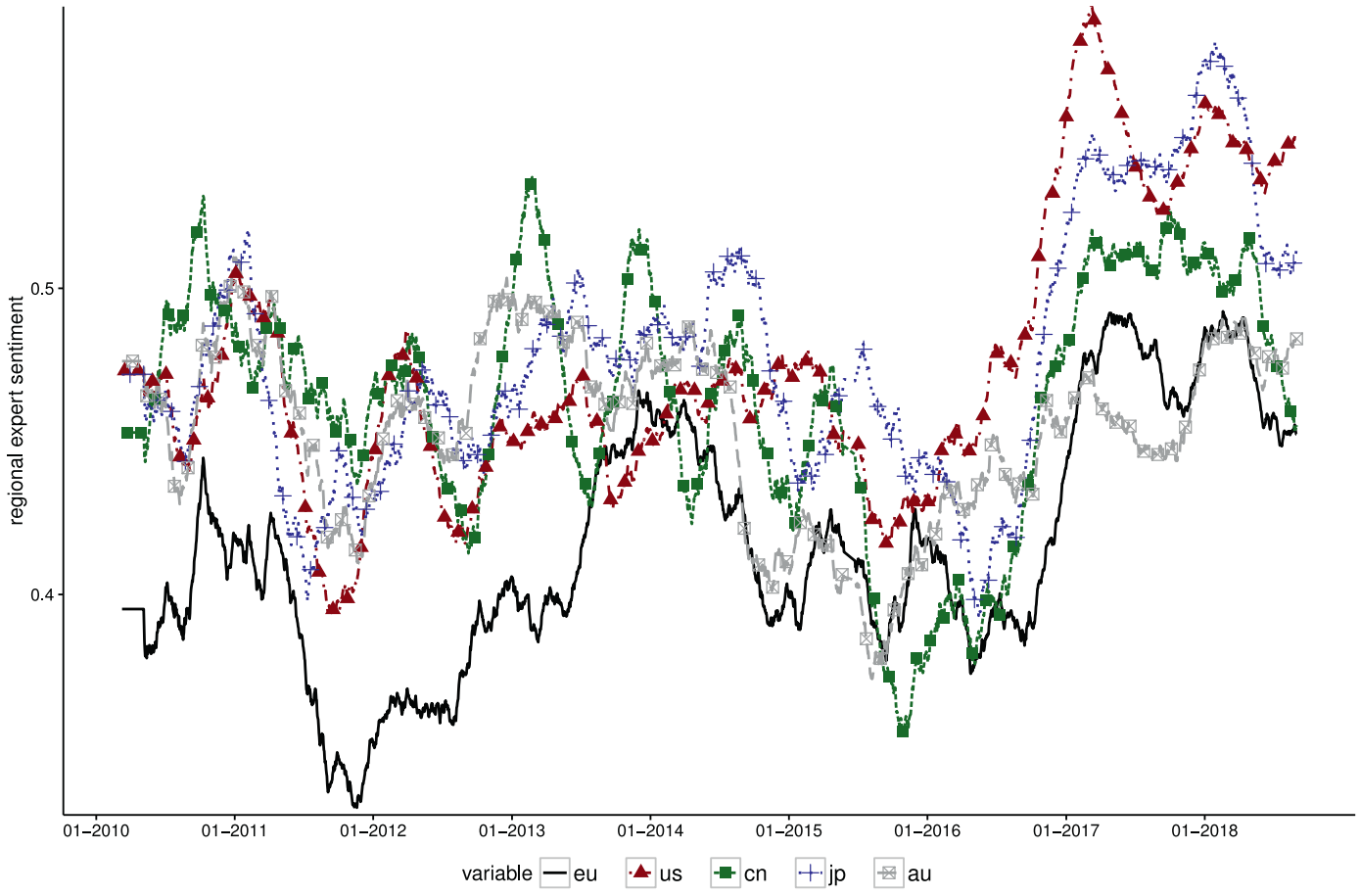

Fig. 2. The figure shows the regionally pooled daily SVM-expert sentiment over the sample. The daily sentiments are averaged over the past 6 months to improve visibility.

Table 4
Main directional accuracy results. 'reg.' denotes the regional context. KMexp and SVMexp are the user groups from Section 4.1. The last two lines give pooled averages for KMexp and SVMexp across the regions. Columns with header 'Acc. (-,+)' give the mean accuracies for the return sign prediction as well as the proportion of correct negative return sign predictions (-) and the proportion of correct positive return sign predictions (+) in brackets. The 'Acc. (-,+)' metrics are given for contemporaneous directional sentiment measures (case $s(r_t) \leftrightarrow S_t^{s(\Delta)}$) as well as for one- and two-day lagged $S^{s(\Delta)}$, (cases $s(r_t) \leftrightarrow S_{t-1}^{s(\Delta)}$ and $s(r_t) \leftrightarrow S_{t-2}^{s(\Delta)}$). Asterisks indicate individual significance of the pooled average against all three competitor models in Table 5 as per DM-test: *** denotes significance at the 1%-level, ** and * at the 5% and 10%-level, respectively. The bold-faced numbers denote individual significant difference to the past return signal and the SVM-classification-signal at the 5%-level.

reg.	user group	Acc. (-/+) $s(r_t) \leftrightarrow S_t^{s(\Delta)}$	Acc. (-/+) $s(r_t) \leftrightarrow S_{t-1}^{s(\Delta)}$	Acc. (-/+) $s(r_t) \leftrightarrow S_{t-2}^{s(\Delta)}$
AU	KMexp	**66** (**67**/64)	53 (**54**/53)	52 (52/51)
AU	SVMexp	**66** (**69**/64)	53 (**54**/53)	53 (53/52)
CN	KMexp	**62** (**66**/59)	51 (**52**/50)	54 (**54**/53)
CN	SVMexp	**62** (**67**/58)	51 (**53**/50)	54 (**54**/54)
EU	KMexp	**66** (**66**/66)	51 (52/**50**)	50 (51/48)
EU	SVMexp	**65** (**65**/65)	51 (51/**52**)	51 (52/50)
JP	KMexp	**64** (**66**/63)	50 (**49**/51)	50 (**50**/51)
JP	SVMexp	**65** (**67**/63)	51 (**50**/51)	50 (**50**/50)
US	KMexp	**70** (**72**/69)	51 (**51**/52)	50 (**49**/51)
US	SVMexp	**71** (**73**/69)	51 (**50**/51)	51 (**50**/51)
av.	SVMexp	**66.0***** (**68.2*****/63.9)	51.4* (**51.7****/51.0)	51.7(**51.9*****/51.5)
av.	KMexp	**65.7***** (**67.3*****/64.3)	51.4(**51.7***/51.3)	51.0(**51.1**/51.0)

between sentiment measures and financial market movements. However, reflecting market efficiency at a daily frequency we find predictive accuracies for sentiment signal lags 1 and 2 to be scarcely higher than 51%. Notably, for all lags and the contemporaneous case we observe that expert sentiment polarities are better in predicting negative signed returns than positive. With only few exceptions this finding can be observed for each individual region. Moreover, it holds for both the unsupervised and supervised expert sets.

To better gauge the relevance of the observations we first draw comparisons to recent studies covering a substantial range of daily data. Yang et al. (2018) for instance report an accuracy of 54% for the best performing model with a true positive rate of 81% and a true negative rate of 23% in predicting return directions of the S&P 500 during 2008 to 2015. In a forecasting study of daily returns for 30 stocks of the Dow Jones Industrial index spanning 2004–2017 Becker and Leschinski (2018) give an average accuracy of 52% with a true positive rate of 56% and a true negative rate of 48% for the

Table 5
Prediction accuracy for main competitor signals. All values are pooled averages across the regions. Non-experts comprised of experts from K-Means and SVM split. Notation as in Table 4.

Signal	Acc. (-/+) $sign(r_t) \leftrightarrow S^*_t$	Acc. (-/+) $sign(r_t) \leftrightarrow S^*_{t-1}$	Acc. (-/+) $sign(r_t) \leftrightarrow S^*_{t-2}$
Non-Experts	61.9 (62.3/61.6)	50.9 (50.2/51.6)	51.1 (50.1/51.8)
SVM classifier for sign(r_t)	58.2 (41.1/73.0)	51.1 (34.0/66.0)	51.1 (34.0/66.0)
r_t	- (-/-)	49.4 (46.7/51.8)	50.2 (47.5/52.5)

best model. While the results in Table 4 for the out-of-sample prediction based on expert sentiment measures are in line with these numbers we note a much more balanced accuracy breakdown into sensitivity and specificity.

To directly test the significance of the high true negative rates in the Twitter expert cases we look at competitive benchmark return prediction models. We consider three approaches. First, natural competitors are given by the sentiment metrics computed for non-expert sets resulting from the supervised and unsupervised learning of experts. Second, as past returns are the main ingredients in forecasting models for returns we analyze the prediction value of the sign of the past return, sign(r_{t-1}). Third, in light of good forecasting performance in both Becker and Leschinski (2018) and Oliveira et al. (2017) we compute rolling window out of sample forecasts from the linear SVM introduced in subsection 3.9.2. In the notation of subsection 3.9.2 we derive forecasts $\widehat{f}(r_{t-1}, \ldots, r_{t-10}) = \text{sign}(\widehat{g}(r_{t-1}, \ldots, r_{t-10}))$ for sign(r_t), where g specifies the hyperplane. The features for the classification of the time t return are thus based on the return lag series 1 to 10. The reason for the slightly higher number of return lags compared to the 5 lags used in Oliveira et al. (2017) lies in the better F1-Score achieved for 10 lags in a model selection exercise covering lags 1 to 10. We set the cost parameter of the linear SVM to one after calibration on the grid {0.01, 0.05, 0.1, 0.1, 1, 10, 50, 100} using the full sample. In the next step we start at the 20th day of the sample and loop over the dates t to train the SVM in each iteration on a 252-day window from $\max(20, t-252)$ up to t. Finally we collect all forecasts after each training procedure. For the two-step ahead and contemporaneous cases of Table 5 we adjust the return lags used as features accordingly. The information set used in prediction of the signed returns sign(r_t), sign(r_{t+1}) and sign(r_{t+2}) is thus comparable to the main Table 4.

Table 5 gives average accuracy as well as average true negative and true positive rates for the benchmark signals averaged across all regions. For sake of brevity we tabulate the average of the K-Means and SVM non-expert as one sentiment signal. Several results emerge from the comparison of Tables 4 and 5. Comparing expert averages in Table 4 to the non-expert average we observe that accuracies and true negative rates for the expert user sentiments are higher than corresponding non-expert averages in the contemporanous case and both lags of the signals. Notably, this finding also holds for the comparison of expert sentiments to both the SVM classification and the past return as signal. The differences in accuracies and sensitivity and specificity can be tested via the test of Diebold and Mariano (1995) (DM). For the specificity we compute the DM test conditional on returns being negative and vice versa for the sensitivity. Table 4 shows joint significant differences to the past return signal and the SVM signal at the 5% level in bold faced numbers. Asterisks indicate signficant differences to all three competing signals, i.e. against returns, the SVM and the non-expert signal. In this respect we observe that the contemporaneous prediction accuracy and specificity for each region significantly surpasses the corresponding number for the return and SVM signal. For the signal lags 1 and 2 we find no joint significant differences in accuracy and sensitivity of expert forecasts within regions. However, the majority of true negative rates for both lags and both regional expert group sentiments is significantly different from the SVM and the past return signal. Inspecting the pooled averages we observe that for all signal lags the pooled true negative rates for the SVM experts are significantly superior to those of all three competitor signals at the 5% level (indicated by the asterisks). Interestingly, while accuracies of K-Means and SVM experts are close, the pooled true negative rates for the K-Means are not significantly different to those of the non-expert counterpart. The reason is simply the better performance of the K-Means non-expert sentiment compared to the SVM non-expert sentiment.

The following key findings can be summarized. First, both unsupervised and supervised expert sentiment measures exhibit the strongest link to market developments compared to the competing signals considered in terms of the contemporenous and predictive accuracy. Second, the high true negative rate emerges as a main characteristic of the expert sentiments. We find significant differences in the true negative rate against non-expert sets and competing signals. Further, the rate is higher than reported in closely related studies in the extant literature. Given the generally higher volatility for down-days (see Table 3) we consider the observation a highly desirable feature of the expert sentiment metrics.

4.3. Investment strategies based on sentiment scores

While past stock returns typically do not have direct predictive power for future returns, empirical evidence still supports the existence of profitable momentum strategies for stocks. In light of this we assess the economic meaning of the close relation between micro-blog sentiment and returns and employ sentiment signals in momentum strategies taken from the extant literature on stock market investing.

The hypothetical investments in regional stock markets will be taken in the respective tradable index futures from the list given in subsection 3.4. In the following, let $i \in \{\text{AU, CN, EU, JP, US}\}$ denote the region. Inspired by the introductory decomposition (1), $r_t = \text{sign}(r_t) \cdot |r_t|$, the long and short positions, i.e. the directions of trades, will be given by binary up- and down signals from $\{-1, 1\}$. The size of each position is intended to be inversely proportional to its risk. Specifically, in line with the approach in Moskowitz et al. (2012) and Hurst et al. (2013) the absolute size of a position in the instrument for region i is $\frac{1}{\hat{\sigma}_{i,t-l}}$ with $l \geq 1$. The ex ante annualized variance $\hat{\sigma}^2_{i,t-l}$ stands for the volatility estimate $\sum_{\tau \leq t-l}(r_{i\tau} - \bar{r}_{i\tau})^2$ for σ^2_{it} with $l \geq 1$. Due to the pronounced autocorrelation in (realized) variance it can be taken as forecast of future volatility. We compute $\hat{\sigma}^2_{it}$ based on 250 days. Per construction of the scaling we hence expect the ex ante risk budget in terms of the annualized volatility for each underlying to be equal at all times.

4.3.1. Expert sentiment time series momentum

Moskowitz et al. (2012), Hurst et al. (2013) and Baltas and Kosowski (2017) describe profitable time series momentum strategies. In these strategies the individual signals for assets of the TSM are not constructed relative to other assets in the cross-section but rather relative to their own past. Specifically, the strategy looks at the direction (sign) of the asset's own return or change over a past time span and maintains a position of the same sign going forward. As such, the TSM is devised to capture over- and underreactions of returns to news over time.

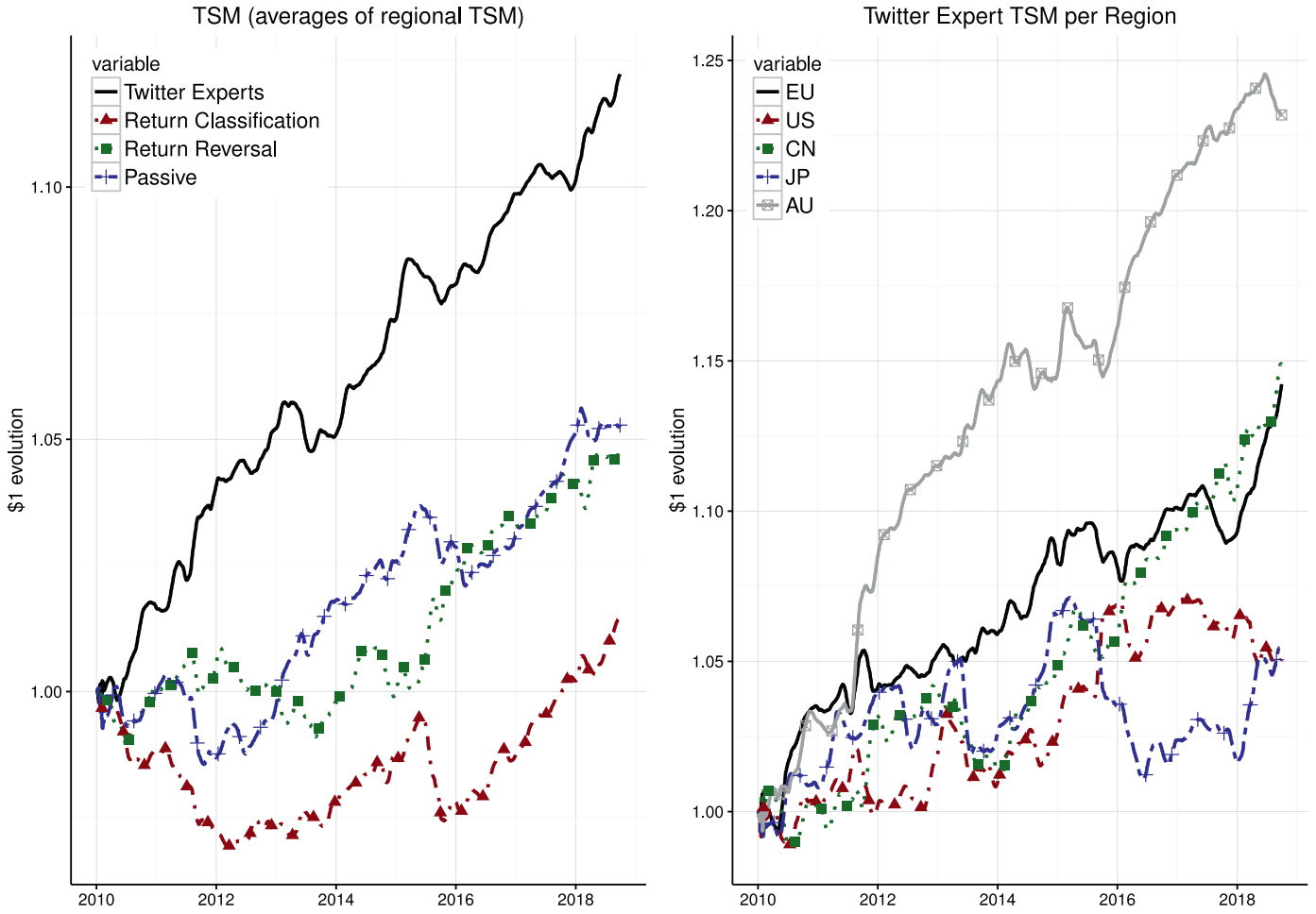

Fig. 3. Left panel shows the TSM wealth evolution based on the Twitter SVM expert and non-expert sentiment as well as those based on return classification and reversal signals. Zero transaction costs assumed. Passive BM denotes the long only strategy. The right panel breaks down the Twitter SVM expert TSM into the individual region TSM strategies. Performance index lines are averaged over 31 days.

Let $\mathbb{S}_{it}$ denote a signal for region i in t. Long and short positions in volatility-scaled futures are taken according to the sign of the one-day-lagged signals $\mathbb{S}_{i,t-1}$. The weights for daily trading at the end of day $t-1$ in order to realize $w_{i,t-1} \cdot r_{it}$ for region i are thus given by

$$w_{i,t-1} = \frac{1}{\hat{\sigma}_{i,t-1}} \text{sign}(\mathbb{S}_{i,t-1}). \tag{8}$$

Following our earlier approach we utilize the expert sentiment changes, $\Delta S_{it} = S_{it} - \bar{S}_{i,t-12,t-1}$ to construct the Twitter-based TSM. For sentiment signals, i.e. $\mathbb{S}_{i,t-1} = \Delta S_{i,t-1}$, this means we hold a long position during day t in the underlying future contract i if the past sentiment change $\Delta S_{i,t-1}$ is greater zero and vice versa for short positions. We report findings for the equal weighted strategy returns across all five region futures, i.e. for $r_t = 1/5 \cdot \sum_i w_{i,t-1} r_{it}$.

Note that the signals $\text{sign}(\mathbb{S}_{i,t-1})$ for the volatility-scaled futures returns do not necessarily sum to zero such that, e.g., persistent simultaneous long and short positions in equity markets can arise. A natural benchmark without market timing is thus a constant ('passive') equal-weighted long-only investment in the five volatility-scaled futures. Next to the expert and non-expert TSM strategies we consider the strategies based on the SVM classification for returns based on only past returns as features (denoted 'classification' in the following). Due to the short-term reversal effects documented by, e.g., Avramov, Chordia, and Goyal (2006) for daily returns we further include a reversal signal. In light of possible calibration and training errors in the SVM classification we compute the negative 12-day average return, $-\bar{r}_{i,t} := -1/12 \sum_{j=0}^{11} r_{i,t-j}$ as a model-free competing signal.

The cumulated TSM returns at 1% annualized volatility are plotted in the lefthand panel of Fig. 3. All lines give time-cumulated returns r_τ of an investment of \$1, $r_\tau = 1 + \sum_{t \le \tau} w_{t-1} r_t$. The plot shows a positive wealth development for the SVM-based Twitter expert TSM which is markedly stronger compared to its competitor strategies before transaction costs. While shown as equal-weighted average on the left panel, the right panel of Fig. 3 details the Twitter SVM-based expert TSM broken up by region. We observe that the TSM yields a positive performance in each of the regions individually, supporting the findings for the left-hand panel.

Table 6 gives the performance summary of the equal-weighted TSM. For sake of brevity we again omit non-expert user TSM strategies which are lower in all performance measures compared to their expert complement. The table shows that the daily Twitter sentiment TSM captures persisting market developments well. This is reflected in the risk-adjusted returns as given by the information ratio IR, i.e. the ratio of annualized strategy returns to the corresponding annualized standard deviation. The IRs of the Twitter sentiment are superior in comparison to average return and SVM classification signals. We attribute this to the out-of-sample accuracy above 51% of the directional sentiment signals $S^{s(\Delta)}$ with regard to regional index returns discussed in Section 4.2. Inspecting the return-based signal $-\bar{r}_t$ we observe that reversal effects dominate on the full sample as indicated by the positive IR of 0.56 for $-\bar{r}_t$. The passive strategy exhibits a IR slightly above the return

Table 6
TSM strategies with the right panel showing the statistics constrained to the dates with negative stock returns. Notations: ret.(s.d.) gives annualized return and standard deviation in per cent. IR denotes the information ratio, TO% gives the annual turnover as percentage. As before KMexp and SVMexp denote the expert groups.

signal	ret.% (s.d.)	IR	TO%	ret.% (s.d.) neg.returns	IR neg.returns	TO% neg.returns
ΔS_t^{KMexp}	1.30 (1.01)	1.29	2937	1.82 (1.07)	1.70	3014
ΔS_t^{SVMexp}	1.28 (.97)	1.31	2889	1.54 (1.03)	1.49	2941
$-\bar{r}_t$	0.55 (1.00)	0.56	716	-.04 (1.06)	-.04	715
classification	0.14 (.91)	0.16	2097	-4.09 (.94)	-4.34	2068
Passive	0.58 (.99)	0.59	18	-7.85 (1.15)	-6.79	21

reversal. The reason behind the latter are the in general upward-trending equity markets during 2010–2018.

The right panel of Table 6 highlights the strategy specifics conditional on dates with negative returns. During these times we report even higher risk-adjusted returns of the two Twitter TSM strategies than for the full sample. The performance on days with negative stock returns shows the stark contrast to the competing signals which predominantly gain based on their true positive rate characteristic and realize negative performances during periods of negative returns. In light of the results in Section 4.2, the elevated IR for the expert sentiment TSM during the dates of negative stock returns highlights the significantly higher true negative rate in the prediction of future return signs.

So far, the results reported are before transaction cost deductions. For each strategy, the costs are dominated by the rebalancing costs which are approximately given by the annual turnover times the transaction costs. Hurst, Ooi, and Pedersen (2017) postulate conservative transaction costs of 0.06% per notional traded. However, this estimate is given for an earlier time frame. As the five futures considered in our sample rank among the most liquid contracts available the costs are assumed to be closer to 0.01%-0.02%. The TSM for both Twitter expert sentiment signals incurs 2900% turnover per year such that the sentiment-based TSM strategies stay profitable under a 0.01%-0.02% rebalancing cost assumption per notional traded. Moreover, the ranking of the competing strategies is constant after transaction costs, too.

4.3.2. Cross-Sectional expert sentiment momentum

The classical cross-sectional momentum strategy outlined in Jegadeesh and Titman (1993) draws on short-selling individual stocks underperforming their peers in the past (past losers) while taking equal-sized long positions in stocks with positive past performance vs. their peers (past winners). The case of aggregated stock indices is covered by Asness et al. (1997). In the traditional CSM strategies, the past performance at t is measured by the past average 12 month excess return. The most recent month's return is typically excluded to account for reversal effects in monthly returns.

Synchronous cross-sectional strategies require to take into account different trading hours around the globe for, e.g. East Asia-based futures ($\sim$UTC+8) and US futures ($\sim$UTC-6). To make sure all signals are available for global trading at t we lag both the sentiment measures and the volatility estimates $\hat{\sigma}_{i,t}^2$ by 2 days. In line with the classical CSM strategies we further mitigate influences of first order autocorrelations in returns that way.

Numerous studies for a cross-section of stocks along the lines of Fama and French (1993) use so-called portfolio sorts to link signals to future stock returns. In these approaches, quintiles of the signals are used to form long portfolio based on the stocks corresponding to one or two top signal quintiles while short portfolios are based on the stocks from the bottom two signal quintiles. The formation process is typically based on the most recent signals and repeated according to a rebalancing schedule, e.g., on a monthly basis. The signal is thought to have predictive value if a portfolio comprised of the combined long and short positions generates positive returns over a benchmark investment. In this spirit, let $rank(\mathbb{S}_{it})$ denote the cross-sectional rank of a signal $\mathbb{S}_{it}$ for region i in t among all regional signals, ranked from lowest to highest. Hence, the rank with value {1, 2, 3, 4, 5} reflects the strength of a signal in the cross-section. We take long positions in stock market futures corresponding to the top two quintiles of the ranks, ranks 4 and 5. Conversely, short positions are taken for regions corresponding to the bottom two quintiles of the ranks (1 and 2). Respecting the inverse volatility scaling and the 2-day lag for signals, weights per futures contract for portfolio formation in $t-1$ are given by

$$w_{i,t-1} = \begin{cases} \dfrac{1}{\hat{\sigma}_{i,t-2}} & \text{if } rank(\mathbb{S}_{i,t-2}) \in \{4,5\}, \\ -\dfrac{1}{\hat{\sigma}_{i,t-2}} & \text{if } rank(\mathbb{S}_{i,t-2}) \in \{1,2\}. \end{cases} \tag{9}$$

The strategy thus realizes the return $w_{i,t-1} \cdot r_{it}$ per region at the end of trading day t. No direct exposure to general equity market movements is taken in this setup, i.e. the ex ante equity market risk for the long and short side sums to zero for all t.

Given the predictive content found in regional binary signals $S^{s(\Delta)}$ for signed returns (see Table 4) we construct cross-sectional signals based on sentiment changes $\Delta S_{i,t}$. As in subsection 4.2 we consider a lookback period of 12 days to compute the sentiment change $\Delta S_{i,t} := (S_{i,t} - \bar{S}_{i,t-1,t-12})$. This means the strategy will hold long positions in stock futures with the two highest past sentiment changes versus short positions in futures with the two lowest sentiment changes.

As benchmark for sentiment-based signals we consider the SVM forecast for r_t. As before we further include the 12-day average return, $\bar{r}_{i,t} = 1/12 \sum_{j=0}^{11} r_{i,t-j}$ as a model-free trend signal. Fig. 4 traces out the daily CSM strategy evolution of a hypothetical \$1 investment based on expert sentiment signals from a SVM split as well as the SVM and average return signals. The lines give the time-cumulated returns r_τ of an investment of \$1, $r_\tau = 1 + \sum_{t \le \tau} \sum_i w_{i,t-1} r_{it}$. All strategies are scaled to have 1% annualized return volatility. We observe that Twitter SVM expert signals and competing signals capture persistent differences in stock index developments, thus generating positively sloping equity lines.

Table 7 gives detailed statistics for the CSM strategy. Looking at the risk-adjusted returns we find highest information ratios for the Twitter expert sentiment strategies based on the SVMexp set. The KMeans expert sentiment is slightly worse than the return classification signal, both signals, however, exhibit positive IRs. The finding reflects the strong contemporaneous and predictive link found for regional expert sentiment measures and the regional stock markets. Overall, expert sentiment strategies pick up short-term co-movements of regional markets well, translating into strategy returns which are competitive compared to past-return-based strategies.

The rightmost three columns of Table 7 give the performance analytics for the strategies strictly limited to the dates of negative market returns according to the S&P 500. We observe that the performance of all strategies deteriorates considerably during the event times. Still, we observe that the high true negative rate

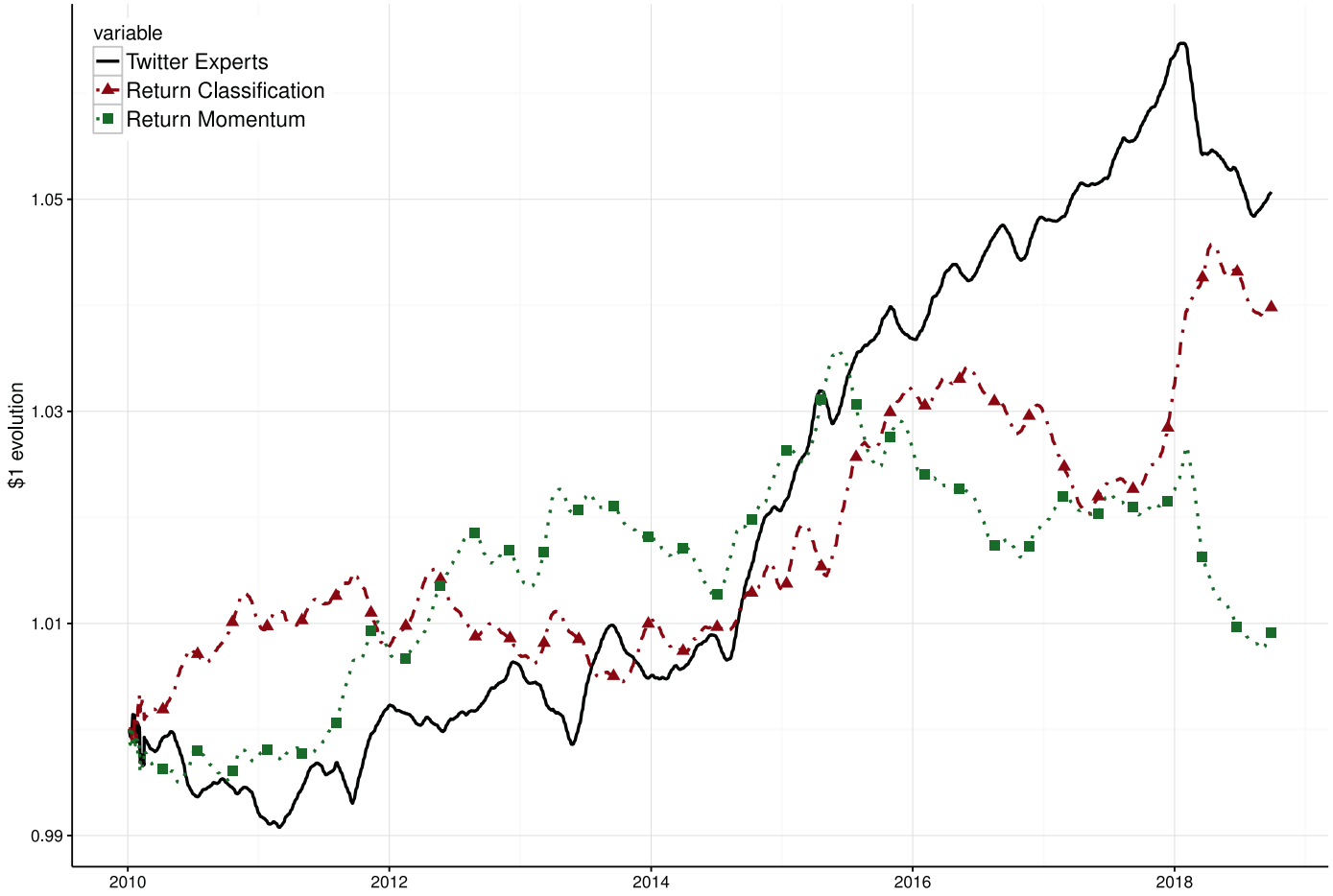

Fig. 4. Hypothetical cross-sectional strategies before transaction costs. Twitter expert and non-expert sets are based on the SVM separation. Performance index lines are averaged over 31 days.

Table 7
Table gives statistics for the CSM. Notation as in Table 6 for the TSM. The right panel gives the characteristics for negative stock return days.

signal	ret.% (s.d.)	IR	TO%	ret.% (s.d.) neg.returns	IR neg.returns	TO% neg.returns
ΔS_t^{KMexp}	0.41 (.99)	0.41	4250	0.08 (1.06)	0.08	4295
ΔS_t^{SVMexp}	0.57 (1.00)	0.57	4231	-.09 (1.07)	-.09	4328
$\bar{r}_t$	0.13 (1.00)	0.13	1697	-.62 (1.38)	-.45	1705
classification	0.43 (.99)	0.44	3679	-2.03 (1.03)	-1.97	3624

of the expert sentiment metrics is reflected in a near-zero performance of the Twitter sentiment strategies while both the return-classification signal and the average return signal exhibit negative strategy returns on the down days. We attribute the diminished returns during negative markets to abrupt breaks in cross-sectional stock market co-movements otherwise exploited by the CSM.

The CSM for both Twitter expert sentiment signals incur more than 4000% turnover per year. With the assumptions outlined in the discussion of the TSM we observe that the sentiment-based CSM strategies narrowly survive rebalancing costs of 0.01% per notional traded resulting in a deduction of $4000\% \cdot 0.01\% \approx 0.4\%$ from the returns, but no longer the 0.02%. The time series classification signal performs roughly equally after transaction costs as it incurs $\sim$3700% in turnover. Notably, the 0.01% transaction costs render the strategies based the average return signals already unprofitable.

4.4. Discussion

Given a still short time frame of 9 years of daily data, our study is faced with a trade-off of either exploring a high number of signals on daily frequency where markets are very efficient or resorting to slower, e.g., monthly data aggregation frequencies with possibly greater stock return prediction accuracies but far fewer data. Opting for the former we can investigate a high number (>2500) of prediction signals at the cost of a high turnover in the stylized trading strategies. However, we stress that two natural modifications of our framework exist to mitigate transaction costs. First, lowering the frequency of signals and returns will automatically reduce the times of transactions and lower the annual turnover. Second, thresholds could be introduced to focus on only extreme trading signals, thus reducing the number of trades.

To ultimately broaden the validity of results we consider variations of the external resources used. Importantly, handcrafting the region topic word lists by adding rule by rule of the construction protocol in 3.3 we observe virtually no change of the results. Table 9 in appendix A.4 tabulates corresponding statistics showing robust results even for the most basic topic list solely based on rule 1, yielding sparse lists of country and currency names per region. Table 9 further asserts that key conclusions also hold for two alternative sentiment lexicons, the Harvard General Inquirer IV-4

(HGIV-4) psychological dictionaries as well as the finance-related word lists of Loughran and McDonald (2011) (see Table 9). In addition, the results are widely robust under the *valence*-based sentiment measure system of Hutto and Gilbert (2014).

5. Conclusions

Making use of a unique nine year data set of microblog messages, this study presents novel results on the relationship between sentiment indicators mined from Twitter messages and global financial market developments. Our analysis of sentiment metrics based on user subgroups reveals that 'expert' sentiment measures capture broad financial developments in major economic regions well and can support profitable investment decisions for corresponding stock market index futures.

First, directional Twitter sentiment measures significantly reflect contemporaneous stock market index return directions. Second, contrasting expert and non-expert sentiment metrics we show that expert user sets are the main driver behind the interdependence of Twitter sentiment and financial markets. Compared to the complementary user set, candidate expert users exhibit higher follower numbers indicating higher network influence and issue messages closer to financial and economical topics. Importantly, the derivation of user sets with expert properties can be based on both unsupervised and supervised statistical methods and hence is independent of the external information used. Third, expert sentiment metrics predict signs of returns with more than 51% accuracy, surpassing the accuracy of all competing signals. This holds for both the supervised and unsupervised derivations of expert users who predominantly cover financial and economic topics, respectively. However, in line with the notion of stock market efficiency, we find no significant (as per the DM test) additive predictive value of direction forecasts based on sentiment measures beyond predictions based on a SVM benchmark model. Fourth, a main feature of our Twitter expert sentiment score is the high true negative rate (specificity) in predicting return directions, i.e. the prediction accuracy for signed *negative* returns. An application of the DM test shows that the true negative rate difference to all competitor predictions is significant. The finding is in stark contrast to non-expert sentiment measures and competing models in both our study and the extant literature where the accuracy predominantly draws on the accuracy in predicting positive returns. Fifth, expert sentiment measures can successfully be employed in traditional cross-sectional (CSM) and time series momentum (TSM) investment strategies adopted from the extant finance literature. Compared to related strategies based on past returns and forecasts from a SVM, both of the sentiment-based momentum strategies generate higher risk-adjusted returns before transaction costs. We attribute the positive performance of the sentiment-CSM strategies to regional sentiment measures capturing persisting regional financial differences. The TSM strategies on the other hand capitalize mostly on the market trends and reversals reflected in expert sentiment measures. A major drawback of the daily sentiment strategies is the high turnover induced by the signals which limits the hypothetical profitability of the CSM in the face of transaction costs. Sixth, due the generally higher stock market volatility on days with negative stock market returns we show that the high true negative rates for expert sentiment measures directly result in hypothetical strategy trading gains on negative market days. Ultimately, this leads to substantially better overall performances of the expert sentiment strategies in comparison to strategies for competing signals which are predominantly dependent on the true positive rates.

The findings are of interest for researchers and market participants alike. Our results show that past Twitter expert sentiment signals are competitive in comparison to signals combining past stock index returns and can be utilized in a similar fashion. Given the small prediction gains due to the efficiency of stock markets our study suggests to focus more on potential economic gains from classical investment strategies applied to social media expert sentiment. In this respect the analysis of the two basic daily trend-following strategies can serve as a methodology template for economic signal evaluation. Further, in light of the well-balanced specificity and sensitivity our study implies that the Twitter expert signals are sources of return with the potential to complement signals based on past returns which maximize their true positive rate.

Declaration of Competing Interest

None declared.

Appendix A

A1. Identification of Twitter users with finance focus

We collected finance expert users publicly listed on the below websites through December 31st, 2014. Any similar website will suffice. Altogether we identified 190 users and confirmed the selection by inspecting the individual tweets. An alternative is to completely hand-select experts without relying on outside information.

https://www.businessinsider.de/people-to-follow-on-twitter-from-finance-2017-6r=USIR=T.

http://blogs.sap.com/innovation/financial-management/top-50-financial-twitter-influencers-019897.

https://www.marketfolly.com/2012/06/top-finance-people-to-follow-on-twitter.html.

http://www.moneysense.ca/news/borzykowski-follow-these-finance-related-twitter-feeds/.

http://finansakrobat.com/blog2/2013/2/5/who-to-follow-on-twitter-the-finance-edition.

A2. Negation list

almost, aint, ain't, arent, aren't, contradictorily, cannot, can't, cant, can't, couldnt, couldn't, don't, dont, doesn't, doesnt, didn't, didnt, few, fewer, fewest, hasn't, hasnt, haven't, havent, hadn't, hadnt, hardly, invalidly, little, mightn't, mightnt, isn't, isnt, not, no, never, nobody, noone, none, nothing, nowhere, neither, nor, rarely, scarcely, seldom, shouldn't, shouldnt, weren't, werent, won't, wont, wasn't, wasnt, wouldn't, wouldnt.

A3. Financial data

Table 8.

Table 8
Bloomberg identifiers for financial data used. Generic Futures were downloaded with a 4 days to first notice date roll mechanism, i.e. suffix 'N:04_0_R Index'.

Index	Index identifier	Generic future identifier
S&P ASX200 (AU)	SPTSX60	XP1
HangSeng (CN)	HSI	HI1
EURO STOXX50 (EU)	SX5E	VG1
TOPIX (JP)	TPX	TP1
S&P 500 (US)	SPX	ES1

A4. Variation of outside information

A4.1. The external resources

The top panel of Table 9 shows results for the Hu and Liu (2004) sentiment lexicon. The regional sentiments are constructed according to the rules given in Section 3.3, however we add rule

Table 9
Results for different external resources used. Columns 2 to 3 show the $t \leftrightarrow t$ and $t \leftrightarrow t-1$ correlations and accuracies. Column 4 gives the information ratio for the TSM, columns 5 and 6 give the IR for the CSM: CSM_{11} shows the IR for the long top quintile vs. short bottom quintile strategy while CSM_{22} shows the IRs for the long top *two* signal quintiles vs. short bottom two quintiles.

external information	$sign(r_t) \leftrightarrow S_t^{s(\Delta)}$ acc (−,+)	$sign(r_t) \leftrightarrow S_{t-1}^{s(\Delta)}$ acc (−,+)	*TSM*	CSM_{11}	CSM_{22}
BL(rule 1)	0.62 (0.63/0.61)	0.51 (0.51/0.50)	1.23	0.74	0.63
BL(rule 1&2)	0.64 (0.65/0.62)	0.51 (0.51/0.50)	1.13	0.70	0.76
BL(rule 1&2&3)	0.64 (0.65/0.63)	0.51 (0.51/0.50)	1.12	0.56	0.52
BL	0.66 (0.68/0.64)	0.51 (0.52/0.51)	1.31	0.29	0.57
VADER	0.62 (0.55/0.68)	0.51 (0.48/0.55)	1.24	0.13	−0.01
LMCD	0.64 (0.63/0.64)	0.51 (0.52/0.50)	1.08	0.41	0.05
HGIV-4	0.63 (0.67/0.60)	0.50 (0.52/0.50)	0.92	0.44	0.29

by rule here. BL(rule 1) gives results for regional sentiment constructed according to the basic rule 1 only, BL(rule 1 & 2) according to rules 1 and 2 while BL(rule 1 & 2 & 3) has rules 1 through 3. BL is the analysis outlined in the main part of the paper.

The bottom panel shows the analysis for the valence-based VADER sentiment tool of Hutto and Gilbert (2014). LMCD gives core statistics for the finance-orientated Loughran and McDonald (2011) lexicon. HGIV-4 denotes the Harvard General Inquirer word list.

References

Albuquerque, R. (2012). Skewness in stock returns: Reconciling the evidence on firm versus aggregate returns. *The Review of Financial Studies, 25*(5), 1630–1673.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica, 71*(2), 579–625.

Andrzejewski, D., Zhu, X., Craven, M., & Recht, B. (2011). A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *Proceedings of the twenty-second international joint conference on artificial intelligence - volume two*. In *IJCAI'11* (pp. 1171–1177). AAAI Press.

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.

Asness, C. S., Liew, J. M., & Stevens, R. L. (1997). Parallels between the cross-sectional predictability of stock and country returns. *The Journal of Portfolio Management, 23*(3), 79–87.

Asness, C. S., Moskowitz, T. J., & Pedersen, L. H. (2013). Value and momentum everywhere. *The Journal of Finance, 68*(3), 929–985.

Atkins, A., Niranjan, M., & Gerding, E. (2018). Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science, 4*(2), 120–137.

Avramov, D., Chordia, T., & Goyal, A. (2006). Liquidity and autocorrelations in individual stock returns. *The Journal of Finance, 61*(5), 2365–2394.

Baltas, N., & Kosowski, R. (2017). Demystifying time-series momentum strategies: Volatility estimators, trading rules and pairwise correlations. *Technical Report*. SSRN.

Banga, J. S., & Brorsen, B. W. (2019). Profitability of alternative methods of combining the signals from technical trading systems. *Intelligent Systems in Accounting, Finance and Management, 26*(1), 32–45.

Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., & Goldstein, G. (2011). Identifying and following expert investors in stock microblogs. In *Proceedings of the conference on empirical methods in natural language processing*. In *EMNLP-2011*.

Becker, J., & Leschinski, C. (2018). Directional predictability of daily stock returns. *Technical Report, 624*. Hannover Economic Papers.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(3), 993–1022.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121–167.

Checkley, M., Higón, D. A., & Alles, H. (2017). The hasty wisdom of the mob: How market sentiment predicts stock market behavior. *Expert Systems with Applications, 77*, 256–263.

Chen, H., De, P., Hu, Y. J., & Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*.

Chen, Z., & Liu, B. (2014). Mining topics in documents: Standing on the shoulders of big data. In *Kdd*.

Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Discovering coherent topics using general knowledge. In *Proceedings of the 22nd acm international conference on information & knowledge management*. In *CIKM '13* (pp. 209–218). New York, NY, USA: ACM.

Christoffersen, P. F., & Diebold, F. X. (2006). Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Management Science, 52*(8), 1273–1287.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science, 53*(9), 1375–1388.

Davis, A. K., Piger, J. M., & Sedor, L. M. (2012). Beyond the numbers: Measuring the information content of earnings press release language. *Contemporary Accounting Research, 29*(3), 845–868.

Deerwester, S., Dumais, S. T., Landauer, G. W. F. T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics, 13*(3), 253–265.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33*(1), 3–56.

Feuerriegel, S., & Gordon, J. (2018). Long-term stock index forecasting based on text mining of regulatory disclosures. *Decision Support Systems, 112*, 88–97.

Feuerriegel, S., & Prendinger, H. (2016). News-based trading strategies. *Decision Support Systems, 90*, 65–74.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance, 68*(3), 1267–1300.

Gattani, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X., Das, S., … Doan, A. (2013). Entity extraction, linking, classification, and tagging for social media: A wikipedia-based approach. *Proceedings of the VLDB Endowment, 6*(11), 1126–1137.

Ghosh, S., Zafar, M. B., Bhattacharya, P., Sharma, N., Ganguly, N., & Gummadi, K. (2013). On sampling the wisdom of crowds: Random vs. expert sampling of the twitter stream. In *Proceedings of the 22nd ACM international conference on information & knowledge management*. In *CIKM '13* (pp. 1739–1744). New York, NY, USA: ACM.

Gonzalez-Bailon, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks, 38*, 16–27.

Haerdle, W., Lee, Y.-J., Schaefer, D., & Yeh, Y.-R. (2009). Variable selection and oversampling in the use of smooth support vector machines for predicting the default risk of companies. *Journal of Forecasting, 28*(6), 512–534.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*. In *SOMA '10* (pp. 80–88). New York, NY, USA: ACM.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*. In *KDD '04* (pp. 168–177). New York, NY, USA: ACM.

Hurst, B., Ooi, Y. H., & Pedersen, L. H. (2013). Demystifying managed futures. *Journal of Investment Management, 11*(3), 42–58.

Hurst, B., Ooi, Y. H., & Pedersen, L. H. (2017). A century of evidence on trend-following investing. *Journal of Portfolio Management, 44*(1), 42–58.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international conference on weblogs and social media (ICWSM-14)*.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance, 48*(1), 65–91.

Jin, O., Liu, N. N., Zhao, K., Yu, Y., & Yang, Q. (2011). Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM international conference on information and knowledge management*. In *CIKM '11* (pp. 775–784). New York, NY, USA: ACM.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML* (pp. 137–142). Heidelberg: Springer.

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems, 104*, 38–48.

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems, 114*, 128–147.

Leung, M. T., Daouk, H., & Chen, A.-S. (2000). Forecasting stock indices: A comparison of classification and level estimation models. *International Journal of Forecasting, 16*(2), 173–190.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics, 45*(2–3), 221–247. Economic Consequences of Alternative Accounting Standards and Regulation

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments and emotions* (1st ed.). New York: Cambridge University Press.

Liu, Y., Kliman-Silver, C., & Mislove, A. (2014). The tweets they are a-changin: Evolution of twitter users and behavior. *ICWSM, 30*, 305–314.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance, 66*(1), 35–65.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of Berkeley symposium on mathematics, statistics and probability* (pp. 281–297). University of California Press.

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval*. In *SIGIR '13* (pp. 889–892). New York, NY, USA: ACM.

Miffre, J., & Rallis, G. (2007). Momentum strategies in commodity futures markets. *Journal of Banking & Finance, 31*(6), 1863–1886.

Mironczuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications, 106*, 36–54.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehose. *CoRR*. arXiv:1306.5204.

Moskowitz, T. J., Ooi, Y. H., & Pedersen, L. H. (2012). Time series momentum. *Journal of Financial Economics, 104*, 228–250.

Nasseri, A. A., Tucker, A., & de Cesare, S. (2015). Quantifying stocktwits semantic terms'trading behavior in financial markets: An effective application of decision tree algorithms. *Expert Systems with Applications, 42*(23), 9192–9210.

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.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications, 42*(24), 9603–9611.

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*(Complete), 125–144.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825–2830.

Phan, X. H., Nguyen, C. T., Le, D. T., Nguyen, L. M., Horiguchi, S., & Ha, Q. T. (2011). A hidden topic-based framework toward building applications with short web documents. *IEEE Transactions on Knowledge and Data Engineering, 23*(7), 961–976.

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on world wide web*. In *WWW '08* (pp. 91–100). New York, NY, USA: ACM.

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

Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015). Short and sparse text topic modeling via self-aggregation. In *Proceedings of the 24th international conference on artificial intelligence*. In *IJCAI'15* (pp. 2270–2276). AAAI Press.

Ranco, G., Aleksovki, D., Caldarelli, G., Grcar, M., & Mozetic, I. (2015). The effects of twitter sentiment on stock price returns. *PLOS ONE, 10*(9), 1–21.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis. *Knowledge-Based Systems, 89*(C), 14–46.

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

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York, NY, USA: McGraw-Hill, Inc..

Schoelkopf, B., Burges, C., & Vapnik, V. (1995). Extracting support data for a given task. In *Proceedings, first international conference on knowledge discovery & data mining, Menlo Park* (pp. 252–257). AAAI Press.

Schütze, H., Hull, D. A., & Pedersen, J. O. (1995). A comparison of classifiers and document representations for the routing problem. In *Proceedings of SIGIR* (pp. 229–237). ACM Press.

Si, J., Mukherjee, A., Liu, B., Pan, S. J., Li, Q., & Li, H. (2014). Exploiting social relations and sentiment for stock prediction. In *Proceedings of the conference on empirical methods in natural language processing*. In *EMNLP 2014*.

Song, Q., Liu, A., & Yang, S. Y. (2017). Stock portfolio selection using learning-to-rank algorithms with news sentiment. *Neurocomputing, 264*, 20–28. Machine learning in finance

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

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. Cambridge, MA: The MIT Press.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267–307.

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.

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.

Timmermann, A. (2008). Elusive return predictability. *International Journal of Forecasting*, 1–18.

Tran, A. T., Tran, N. K., Hadgu, A. T., & Jäschke, R. (2017). Semantic annotation for microblog topics using wikipedia temporal information. *CoRR*. arXiv:1701.03939.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer.

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on web search and data mining*. In *WSDM '10* (pp. 261–270). New York, NY, USA: ACM.

Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: a survey. *Artificial Intelligence Review, 50*(1), 49–73.

Yang, S. Y., Mo, S. Y. K., & Liu, A. (2015). Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance, 15*(10), 1637–1656.

Yang, S. Y., Mo, S. Y. K., Liu, A., & Kirilenko, A. A. (2017). Genetic programming optimization for a sentiment feedback strength based trading strategy. *Neurocomputing, 264*, 29–41. Machine learning in finance

Yang, S. Y., Song, Q., Mo, S. Y. K., Datta, K., & Deane, A. (2015). The impact of abnormal news sentiment on financial markets. *Journal of Business and Economics, 6*(10), 1682–1694.

Yang, S. Y., Yu, Y., & Almahdi, S. (2018). An investor sentiment reward-based trading system using Gaussian inverse reinforcement learning algorithm. *Expert Systems with Applications, 114*, 388–401.

Yildirim, A., Üsküdarli, S., & Özgür, A. (2016). Identifying topics in microblogs using wikipedia. *PLOS ONE, 11*(3), 1–20.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through twitter: I hope it is not as bad as i fear. *Procedia - Social and Behavioral Sciences, 26*(0), 55–62. The 2nd Collaborative Innovation Networks Conference - {COINs2010}

Zhao, W. X., Jiang, J., Weng, J., He, E.-P., Lim, J., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, & V. Mudoch (Eds.), *Advances in information retrieval* (pp. 338–349). Berlin, Heidelberg: Springer Berlin Heidelberg.