Expert Systems With Applications 73 (2017) 125–144

Contents lists available at ScienceDirect

Expert Systems With Applications

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

The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices

Nuno Oliveira[a,*], Paulo Cortez[a], Nelson Areal[b]

[a] ALGORITMI Centre, Department of Information Systems, University of Minho, 4804-533 Guimarães, Portugal
[b] School of Economics and Management, Department of Management, University of Minho, 4710-057 Braga, Portugal

ARTICLE INFO

Article history:
Received 10 October 2016
Revised 6 December 2016
Accepted 26 December 2016
Available online 27 December 2016

Keywords:
Stock market
Twitter
Data and text mining
Regression

ABSTRACT

In this paper, we propose a robust methodology to assess the value of microblogging data to forecast stock market variables: returns, volatility and trading volume of diverse indices and portfolios. The methodology uses sentiment and attention indicators extracted from microblogs (a large Twitter dataset is adopted) and survey indices (AAII and II, USMC and Sentix), diverse forms to daily aggregate these indicators, usage of a Kalman Filter to merge microblog and survey sources, a realistic rolling windows evaluation, several Machine Learning methods and the Diebold-Mariano test to validate if the sentiment and attention based predictions are valuable when compared with an autoregressive baseline. We found that Twitter sentiment and posting volume were relevant for the forecasting of returns of S&P 500 index, portfolios of lower market capitalization and some industries. Additionally, KF sentiment was informative for the forecasting of returns. Moreover, Twitter and KF sentiment indicators were useful for the prediction of some survey sentiment indicators. These results confirm the usefulness of microblogging data for financial expert systems, allowing to predict stock market behavior and providing a valuable alternative for existing survey measures with advantages (e.g., fast and cheap creation, daily frequency).

1. Introduction

Due to the growth of the Internet and Web 2.0 phenomenon, social media is an important big data source (Fan & Gordon, 2014). Users spend a significant part of their time on social media services. Thus, the analysis of these social media data may allow a deeper understanding of users' behavior that can be utilized for various purposes, including the financial domain. For instance, Thomson Reuters Eikon and Bloomberg are examples of financial services that include sentiment analysis of tweets.[1,2]

In effect, the usage of sentiment and attention indicators for stock market behavior modeling and prediction is an active research topic. As shown in Table 1, there is a large list of related works that can be distinguished in terms of several dimensions. The sentiment and attention indicators can be created using distinct sources (column *Source*), sentiment analysis method (*Meth.*) and combination method used to merge distinct sources (*Comb.*). The financial analysis assumes a periodicity (*Per.*) of the applied variables (e.g., daily, monthly), type of stock (*Stocks*, e.g., individual or portfolios), methods (*Meth.*) used to model or predict (e.g., multiple regression) and data (*Data*) used to fit the models (e.g., four months). Some of the most recent studies (after 2011), perform a prediction that is characterized by its data (*Data*) period (e.g., nineteen days) and the statistical tests used to verify the statistical (*St.*) significance of the sentiment and attention based predictions when compared to baseline models, such as the Diebold-Mariano (DM) test (Diebold & Mariano, 2002). None of the related works attempts to predict survey sentiment indices (*Sur.*), which is addressed in this study. In the next few paragraphs, we detail some of these dimensions and explain the novelty of this paper when compared with the related works.

The earlier studies, from 1988 to 2010, adopted surveys, financial data, message boards (e.g., ragingbull.com) and news (e.g., Wall Street Journal) to create the sentiment and attention indicators. After 2011, Web 2.0 services, such as microblogs (e.g., Twitter, StockTwits) and Google searches, have also been adopted. Some financial measures (e.g., closed-end fund discount) and survey values, such as American Association of Individual Investors (AAII)

* Corresponding author.
E-mail addresses: nunomroliveira@gmail.com (N. Oliveira), pcortez@dsi.uminho.pt (P. Cortez), nareal@eeg.uminho.pt (N. Areal).

[1] http://thomsonreuters.com/en/press-releases/2014/thomson-reuters-adds-unique-twitter-and-news-sentiment-analysis-to-thomson-reuters-eikon.html

[2] http://www.bloomberg.com/company/announcements/trending-on-twitter-social-sentiment-analytics/

http://dx.doi.org/10.1016/j.eswa.2016.12.036
0957-4174/

Table 1
Summary of related work.

Study	Sentiment			Attent.	Financial analysis				Prediction		
	Source[a]	Meth.[b]	Comb.[c]	Source[a]	Per.[d]	Stocks[e]	Meth.[f]	Data[g]	Data[g]	St.[h]	Sur.[i]
(Solt & Statman, 1988)	S				w	Ix	MR	22y			
(Lee, Shleifer, & Thaler, 1991)	F				m	Pf	MR	20y			
(Neal & Wheatley, 1998)	F				m,q,a	Pf	MR	60y			
(Fisher & Statman, 2000)	S				m	Ix,Pf	MR	13y			
(Tumarkin & Whitelaw, 2001)	MB			MB	d	I	VAR	11m			
(Lee et al., 2002)	S				w	Ix	GARCH	22y			
(Antweiler & Frank, 2004)	MB	ML		MB	d	I	MR	1y			
(Brown & Cliff, 2004)	F,S		KF,Pca		m,w	Pf	VAR	33y			
(Brown & Cliff, 2005)	S				m	Pf	MR	19y			
(Das, Martínez-Jerez, & Tufano, 2005)	MB,N	ML		MB,N	d	I	MR	7m			
(Baker & Wurgler, 2006)	F		Pca		m	Pf	MR	38y			
(Qiu & Welch, 2006)	F,S				m,q	Pf	MR	38y			
(Schmeling, 2007)	S				w	Ix	MR	4y			
(Das & Chen, 2007)	MB	ML		MB	d	Ix,I	MR	2m			
(Tetlock, 2007)	N	GL			d	Am,Ix,Pf	VAR	15y			
(Ho & Hung, 2009)	S		Pca		m	I	MR	41y			
(Schmeling, 2009)	S				m	Am,Pf	MR	21y			
(Kurov, 2010)	F,S		Pca		d	Ix,I	MR	14y			
(Yu & Yuan, 2011)	F		Pca		m	Am	MR	42y			
(Bollen et al., 2011)	M	GL			d	Ix	NN	11m	19d		
(Deng et al., 2011)	N	GL		N	d	I	2ML,RW	32m	2y		
(Groß-Klußmann & Hautsch, 2011)	N	P			i	I	VAR	18m			
(Mao et al., 2011)	G,M,N,S	FL,K			d,w	Ix	MR	15m	30d,20w		
(Oh & Sheng, 2011)	M	ML		M	d	I	8ML	4m	10d		
(Sabherwal et al., 2011)	MB	ML		MB	d,i	I	MR	13m			
(Sheu & Wei, 2011)	F				d	Am	MR,TR	4y	59d		
(Zhang, Fuehres, & Gloor, 2011)	M	K			d	Ix	Cor	7m			
(Baker et al., 2012)	F		Pca		m	Am,Pf	MR	25y			
(Schumaker et al., 2012)	N	GL			i	I	SVM	23d	23d		
(Stambaugh, Yu, & Yuan, 2012)	F,S		Pca		m	Pf	MR	42y			
(Chen & Lazer, 2013)	M	GL			d	Am	MR,TR	97d	25-33d		
(Corredor, Ferrer, & Santamaria, 2013)	F,S		Pca		m	Pf	MR	18y			
(Garcia, 2013)	N	FL			d	Ix,Pf	MR	100y			
(Hagenau et al., 2013)	N	ML			d	I	TR	14y	12y		
(Oliveira et al., 2013)	M	K		M	d	I	MR,RW	28m	305-505d	DM	
(Smailović, Grčar, Lavrač, & Žnidaršič, 2013)	M	ML			d	I	GC	10m			
(Yu, Duan, & Cao, 2013)	B,M,MB,N	ML		B,M,MB,N	d	I	MR	3m			
(Sprenger et al., 2014)	M	ML		M	d	I	MR	6m			
(Al Nasseri et al., 2015)	M	ML			d	Ix	TR	13m	1y	ST	
(Nguyen et al., 2015)	MB	ML,GL		MB	d	I	SVM	13m	78d		
This study	M,S	MFL	KF	M	d	Ix,Pf	5ML,RW	35m	350-439d	DM	2S

[a] Sentiment and attention sources: B – blogs, F – financial data, G – Google searches, M - microblogs, MB – message boards, N – news, S – surveys.
[b] Sentiment analysis method: FL – financial lexicon, GL – generic lexicon, K – keywords, ML – supervised machine learning, MFL – microblog financial lexicon, P – sentiment analysis product
[c] Combination method: KF - kalman filter, Pca - principal component analysis
[d] Periodicities: a – annual, d – daily, i – intraday, m – monthly, q – quarterly, w – weekly
[e] Stocks: Am – aggregated market, I – individual stocks, Ix - indices, Pf - Portfolios
[f] Financial analysis method: Cor – correlation, GARCH – generalized autoregressive conditional heteroskedasticity, GC – granger causality, MR – multiple linear regression, *n*ML - *n* machine learning methods, NN - neural networks, RW – rolling windows, SVM – support vector machine, TR – trading rules, VAR – vector auto-regression
[g] Data Period: d – days, m – months, w – weeks, y – years
[h] Statistical Test for Out of Sample Evaluation: DM – Diebold-Mariano test, ST – Student's *t*-test
[i] Prediction of Surveys: *n*S - *n* survey sentiment indices

and Investors Intelligence (II), are often used as proxy for sentiment. AAII and II are popular sentiment tools that are created from polls to investors and newsletters created by market professionals (Brown & Cliff, 2004; Fisher & Statman, 2000). However, the indicators extracted from texts (e.g., Twitter) have many advantages when compared with survey sentiment indices. The creation of text based sentiment indicators, as executed in this work, is faster and cheaper, permits greater periodicities (e.g., daily) and may be targeted to a more restrict set of stocks (e.g., stock market indices or individual stocks).

There are two main approaches for the extraction of sentiment indicators from text: supervised and unsupervised. Some studies use supervised machine learning, such as Naive Bayes or Support Vector Machines (SVM) (Antweiler & Frank, 2004; Hagenau, Liebmann, & Neumann, 2013) but it requires labeled training data that is often difficult to obtain, since social media often do not provide classified data and their manual labeling is costly and impractical. Thus, other studies use an unsupervised approach based on lexicons or keywords (Bollen, Mao, & Zeng, 2011; Mao, Counts, & Bollen, 2011). Most of the applied lexicons are domain independent (e.g., General Inquirer, MPQA, SentiWordNet). Only two studies use the financial lexicon created by Loughran and McDonald (2011). Yet, as recently shown in Oliveira, Cortez, and Areal (2016), generic domain independent lexicons are ineffective for assessing the sentiment of stock market messages. For instance, the term "explosive" is often negative in generic contexts but can be positive within the financial domain ("explosive rise"). Moreover, the financial lexicon of Loughran and McDonald (2011) was created using large text reports and it obtains low recall values for short microblogging messages (Oliveira et al., 2016). As such, in this paper

we use a recent and extensive lexicon, adapted to microblogging stock market conversations and that should permit a more reliable sentiment classification of tweets.

Furthermore, the various existing investor sentiment indicators are measured by different approaches (e.g., surveys, social media) and have distinct characteristics (e.g., monthly, weekly or daily frequencies). Therefore, they usually have different values and may contain some noise. The Kalman Filter (KF) procedure allows the aggregation of several observed variables with distinct frequencies (e.g., daily, weekly, monthly) to extract a latent variable. Thus, the application of KF may allow the production of a less noisy and more representative investor sentiment indicator than their individual constituents.

A KF procedure was already applied to extract a weekly and a monthly sentiment indicator from financial data and surveys (Brown & Cliff, 2004). However, to the best of our knowledge, there were no previous attempts in this topic to produce sentiment indicators by combining social media with other sources, particularly for higher frequencies than the weekly. Therefore, in this paper we experiment the extraction of an unique daily indicator from microblogging data and diverse weekly and monthly survey sentiment indicators by applying a KF procedure.

As shown in Table 1, the size of the datasets used in this topic is usually high for studies using sentiment based on financial data, surveys or news but it is low for sentiment extracted from social media data. These latter studies use less than two years of data, with the exception of our previous study using StockTwits data (Oliveira, Cortez, & Areal, 2013). In this study, we use almost three years of Twitter data containing around 31 million messages.

The evaluation of sentiment on portfolios based on some characteristics (e.g., size, book to market, volatility) is frequent on financial data or surveys studies. However, this analysis is very scarce for text based sentiment and for higher periodicities than the weekly. For instance, Tetlock (2007) and Garcia (2013) analyzed the influence of sentiment created from news on some variables based on portfolios formed on size. We did not find any analysis of the impact of sentiment extracted from social media on portfolios of any type. Thus, we evaluate the influence of sentiment created from Twitter data on returns of portfolios formed on size and industries.

To evaluate the predictive value of sentiment for stock market variables, the majority of the studies apply Multiple Regression (MR) and Vector Auto-Regression (VAR) methods, which are generalized for survey and financial based sentiment and very frequent for text based sentiment. The usage of more flexible learning ML models, such as SVM (Deng, Mitsubuchi, Shioda, Shimada, & Sakurai, 2011; Nguyen, Shirai, & Velcin, 2015; Schumaker, Zhang, Huang, & Chen, 2012) or Neural Networks (NN) (Bollen et al., 2011), is more scarce. Also, very few studies compared the predictive accuracy of different Machine Learning (ML) models for stock market variables (Deng et al., 2011; Oh & Sheng, 2011). We compare five regression models, MR, NN, SVM, Random Forest (RF) and Ensemble Averaging (EA), using a more robust rolling window validation scheme that was only used in two related works (Deng et al., 2011; Oliveira et al., 2013).

Most forecasting studies that use text extracted indicators present limitations in terms of lack of a robust evaluation. For instance, an out of sample evaluation was not used in almost all papers using financial data or surveys and in the majority of the studies using textual contents. Other works applied very short test sets: 10 predictions (Oh & Sheng, 2011); 19 forecasts (Bollen et al., 2011); 20 and 30 forecasts (Mao et al., 2011); 25 and 35 predictions (Chen & Lazer, 2013); and 79 forecasts (Nguyen et al., 2015). Moreover, the utilization of statistical tests to evaluate the predictive accuracy is very limited (Table 1). The DM test for predictive accuracy that we use in this study allows the evaluation of the statistical significance of the predictions.

The prediction of survey sentiment indices can be very useful for investors. It may permit a valuable anticipation of their values or constitute a cheap alternative measure. Though there are regressions of survey sentiment using contemporaneous values of other sources (e.g., financial measures, other surveys) in some studies (Brown & Cliff, 2004; Schmeling, 2007), we did not find their prediction based on lagged values of other sentiment proxies, particularly social media sentiment. In this study, we predict two popular survey sentiment indicators (AAII and II) using Twitter and KF sentiment indicators.

Regarding the obtained results, text based sentiment was considered useful to make trading decisions (Al Nasseri, Tucker, & de Cesare, 2015; Schumaker et al., 2012) or predict useful stock market variables, such as: daily or intraday values of stock prices (Bollen et al., 2011), price directions (Nguyen et al., 2015), returns (Sabherwal, Sarkar, & Zhang, 2011; Tetlock, 2007), volatility (Sabherwal et al., 2011) and trading volume (Antweiler & Frank, 2004). The analysis of the informative value of sentiment extracted from text has been almost exclusively focused on the prediction of daily variables of individual stocks, indices or aggregate market. However, sentiment collected from surveys and financial data has been mainly applied for lower periodicities (e.g., monthly) and portfolios formed based on diverse characteristics (e.g., firm dimension, age, volatility, book-to-market ratio). Sentiment seems to have more effect on returns of some portfolios having extreme values on these characteristics (Baker, Wurgler, & Yuan, 2012; Brown & Cliff, 2005; Neal & Wheatley, 1998). To the best of our knowledge, there are no studies using microblogging data to predict variables based on portfolios, as executed in this work. Posting volume on social media services (e.g., microblogs and message boards) has also been applied to predict volatility (Antweiler & Frank, 2004; Das & Chen, 2007) or trading volume (Oliveira et al., 2013; Sprenger, Tumasjan, Sandner, & Welpe, 2014). Yet, these findings are not consensual and there are studies that find scarce evidence of the predictive power of sentiment for stock prices or returns (e.g., (Antweiler & Frank, 2004; Brown & Cliff, 2004; Oliveira et al., 2013; Timmermann, 2008; Tumarkin & Whitelaw, 2001)) and posting volume for volatility (Antweiler & Frank, 2004; Tumarkin & Whitelaw, 2001).

The main goal of this paper is to assess the value of microblogging data to the forecasting of stock market variables. To achieve this, we propose the methodology that is presented in Fig. 1 and detailed in Section 2. Our predictive models are fitted using microblog sentiment and attention indicators. These indicators are based on Twitter (TWT) and on weekly (AAII, II) and monthly (UMSC, Sentix) survey indices retrieved using the Datastream service. A fast unsupervised sentiment analysis is performed by using a specialized financial microblog lexicon (Oliveira et al., 2016). This is the first paper that adopts a specialized stock market lexicon adapted to microblogging messages. The individual tweet and survey sentiment indicators are aggregated into daily values using diverse forms (e.g., bullish ratio). We also use KF to merge weekly and monthly survey indices with the daily microblog sentiment indicators, which is a new approach in this context, and test its forecasting ability. Using rolling windows, we fit five distinct ML models to predict daily stock indices and portfolio values (e.g., SP500, RMRF) that were collected from Datastream and Prof. Kenneth French Webpage. We also fit five ML models to predict weekly sentiment survey indices (AAII and II). To our knowledge, this is the first attempt to predict these survey indices. Finally, we assume that predictions using sentiment and attention indicators are valuable when they are statistically better, accordingly to the DM test, than an autoregressive baseline model. We extend and improve our previous work (Oliveira et al., 2013) by: using an ex-

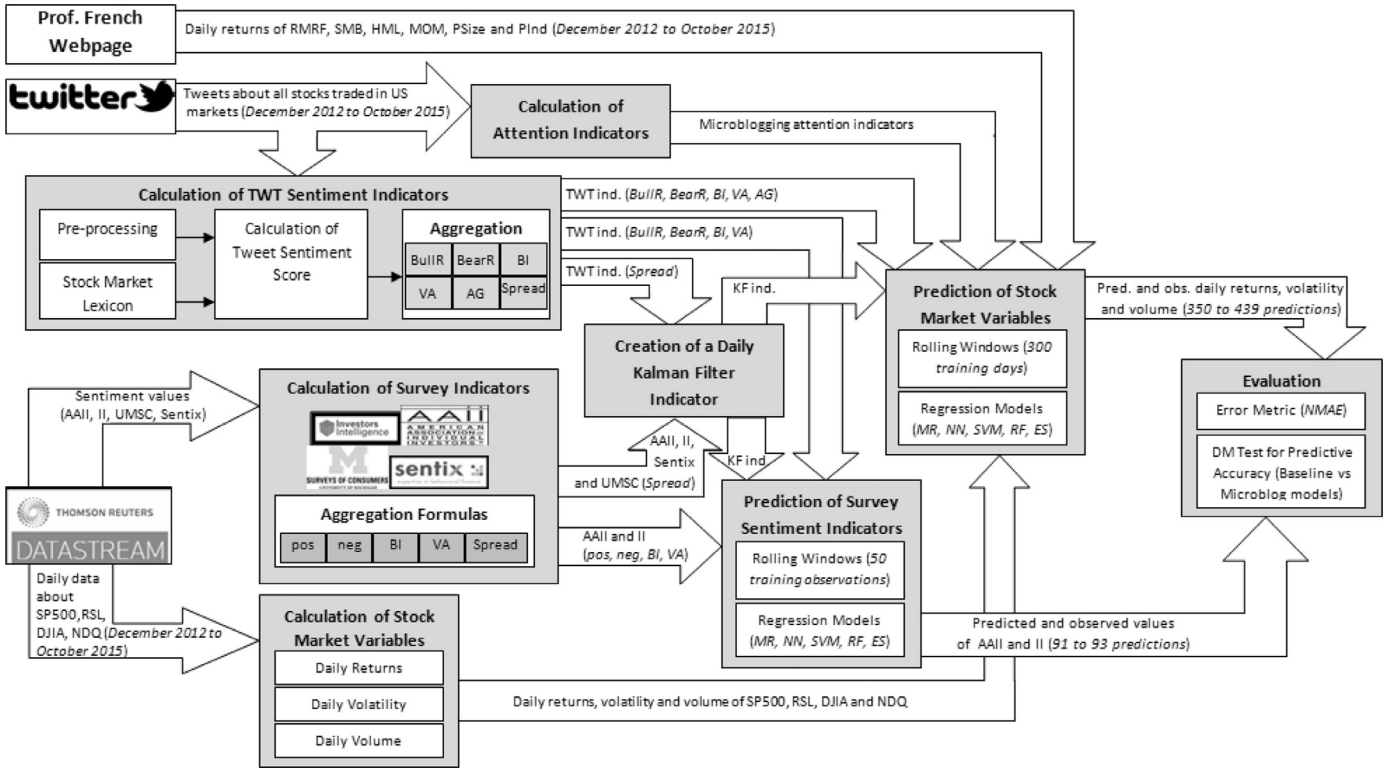

Fig. 1. Schematic of the proposed methodology.

tensive financial microblog lexicon (Oliveira et al., 2016), instead of the simplistic "bullish" and "bearish" keywords); analyzing stock indices and portfolios; applying 5 ML methods, instead of the simpler MR; and predicting also survey indexes. We further note that in Oliveira et al. (2016) we proposed a new method for the automatic creation of a specialized financial lexicon but such method was validated using StockTwits labeled sentiment messages and no financial prediction was attempted. Our main contributions are summarized as follows:

(i) we propose a robust methodology to assess the value of text based sentiment and attention indicators when forecasting stock market variables (Fig. 1). The methodology assumes a realistic rolling window evaluation (with 300 training days for stock market variables and 50 observations for survey indicators), the DM test for predictive accuracy and five ML regression models (MR, NN, SVM, RF and EA), under two main strategies: baseline model (without microblog features) and microblog based (with such features). This methodology can be easily adapted in future research on this topic;

(ii) we generate new Twitter sentiment indicators based on a recent lexicon (Oliveira et al., 2016) that contains more than 20,000 entries and that is specifically adjusted to financial microblogs. These indicators were extracted from a recent and very large Twitter based dataset, containing around 31 million messages from December 2012 to October 2015 related with all stocks traded in US markets (about 3800 stocks);

(iii) we propose a less noisy KF sentiment indicator that combines measures of different periodicities: daily Twitter indicators, weekly AAII and II values and monthly UMSC and Sentix indices. We also explore diverse sentiment aggregation formulas (e.g., bullish ratio, variation, agreement);

(iv) we conduct a large set of experiments, predicting daily stock market variables (returns, trading volume and volatility) of diverse indices, such as Standard & Poor's 500 (SP500) and Nasdaq 100 (NDQ), and portfolios formed on size and industries;

(v) we also predict two popular survey sentiment indicators (AAII and II) using Twitter and KF sentiment indicators, which may permit a satisfactory anticipation of AAII and II values or a decent alternative whenever they are unavailable.

We consider that this study may support further advances in financial expert systems. For instance, real time financial systems can be enhanced byproviding personalized sentiment and prediction values related to specific stocks (e.g., individual stocks, indices) or surveys (e.g., AAII, II) for diverse periodicities (e.g., intraday, daily, weekly). Thus, it may contribute to the creation of more flexible financial systems that are more capable to satisfy user needs.

This paper is structured as follows. Section 2 presents the applied methodology. Section 3 shows and analyzes the results of the prediction of returns, volatility, trading volume and survey sentiment values. Finally, conclusions are presented in Section 4.

2. Material and methods

2.1. Microblogging data

Microblogging data have characteristics that may indicate potential value to the forecasting of stock market behavior. Services such as Twitter and StockTwits have large communities of investors sharing information about stock market. These users frequently interact during the day and react readily to events. Messages are usually very objective due to the character limit. Microblogging users usually apply cashtags in stock market conversations to refer to the involved stocks. Cashtags are composed by

a "$" character and the respective ticker (e.g., $AAPL) and its presence means that the message is related with that stock. The utilization of cashtags permits an easy and less noisy selection of messages related to specific stocks. The extraction of attention and sentiment from microblogging is faster and cheaper than from traditional sources (e.g., surveys) because data is promptly available at very low cost.

The sentiment and attention indicators created in this study were extracted from Twitter, which is a large microblog platform with more than 300 million active users.[3] Using Twitter REST API (https://dev.twitter.com/docs/api), we collected all messages (around 31 million) from 22nd of December 2012 to 29th of October 2015 holding cashtags of all stocks traded in US markets (nearly 3800 stocks). R tool (http://www.r-project.org) was used in all processing tasks and Mongodb (https://www.mongodb.org) was applied to store Twitter data.

2.2. Microblogging sentiment and attention indicators

The number of tweets were applied to produce the attention indicators. We opted to use the first difference of the posting volume because there is a visible growing number of tweets during the analyzed time period. To create the investor sentiment indicators, we used the sentiment scores produced by sentiment analysis of all tweets. The sentiment analysis applies a recently proposed lexicon (Oliveira et al., 2016) that is properly adapted to microblogging conversations about stock market and publicly available at https://github.com/nunomroliveira/stock_market_lexicon. This lexicon was automatically created using data from June, 2010 to March, 2013. It is an up-to-date lexicon because its training data ends nearly one year before the first prediction days of this study. It contains approximately 7000 unigrams, 13,000 bigrams and the respective sentiment scores for affirmative and negated contexts. For instance, a negative score indicates a bearish word and a positive sentiment value indicates a bullish word. Negated contexts are text segments starting with a negation expression while the affirmative contexts are all other segments. To identify affirmative and negated segments, we applied the same approach used in the lexicon creation. Negated segments begin with a negation item included in the Christopher Potts' sentiment tutorial (http://sentiment.christopherpotts.net/lingstruc.html) and end with one of the following punctuation marks: ",", ".", ":", ";", "!", "?". The sentiment score of each tweet corresponds to the sum of the sentiment score of all lexicon items present in the message. When lexicon bigrams are identified in the text, we only account the score of the bigrams and do not consider the score of their individual components. In our opinion, bigrams scores are more precise than unigram scores because bigrams have a more defined context. For instance, the bigram "debt free" is usually bullish while its individual components may have distinct sentiment orientation (e.g., "greek debt", "more debt", "free fall"). To adequately verify the presence of lexicon elements, we executed the preprocessing tasks:

- replace all cashtags by the tag "tkr"; all numbers by the tag "NUM"; all mentions by "@user"; all URL addresses by "URL";
- execute tokenization, Part of Speech (POS) tagging and lemmatization by applying Stanford CoreNLP (Toutanova, Klein, Manning, & Singer, 2003).
- identify the affirmative and negated segments in order to apply the adequate score.

The sentiment indicators are created using the scores produced by the sentiment analysis. We created two major types of investor sentiment indicators: general and sectorial. The general indicators represent the sentiment of the whole investor community. Thus, we used all tweets in the construction of these indicators. The sectorial indicators measure the sentiment regarding specific sectors (e.g., industries). In the creation of these indicators, we applied tweets enclosing cashtags of stocks belonging to the respective sector. We selected the cashtags composing each sector based on their Standard Industrial Classification (SIC) code.

We also experimented diverse forms to calculate the daily sentiment values:

- Bullish Ratio (BullR) (Oliveira et al., 2013):

$$BullR_t = \frac{NBull_t}{NBull_t + NBear_t} \tag{1}$$

- Bearish Ratio (BearR):

$$BearR_t = \frac{NBear_t}{NBull_t + NBear_t} \tag{2}$$

- Bullishness Index (BI) (Antweiler & Frank, 2004; Sprenger et al., 2014):

$$BI_t = \ln \frac{NBull_t + 1}{NBear_t + 1} \tag{3}$$

- Variation (VA) (Oliveira et al., 2013):

$$VA_t = BullR_t - BullR_{t\text{-}1} \tag{4}$$

- Agreement (AG) (Antweiler & Frank, 2004; Sprenger et al., 2014):

$$AG_t = 1 - \sqrt{1 - \left(\frac{NBull_t - NBear_t}{NBull_t + NBear_t}\right)^2} \tag{5}$$

where $NBull_t$ and $NBear_t$ are the bullish and bearish score of day t. We did not found any paper applying a sentiment measure computed exactly like BearR, however there are papers using similar measures, such as Tetlock (2007). The tested sentiment aggregation measures are distinct formulas applied in diverse studies in this research topic. In this study, we applied BullR, BI and VA indicators in the prediction of returns, volatility, trading volume and survey sentiment values. Since dispersion of expectations is considered to be related with trading volume (Antweiler & Frank, 2004; Shalen, 1993) and volatility (Shalen, 1993), we added AG indicators in the forecasting of these stock market variables. Additionally, we applied BearR indicators in the prediction of negative values of survey sentiment values, because we used Twitter indicators counterparts in the forecasting of each survey sentiment value (e.g., BullR indicators applied in the prediction of AAII positive values or BearR indicators in the forecasting of AAII negative values).

2.3. Survey sentiment indicators

Survey sentiment indicators are frequently applied in studies about the analysis of the sentiment impact on stock market behavior. For instance, AAII provides weekly values of the votes of their members to a poll questioning their sentiment (bullish, bearish, neutral) on the stock market for the next six months. Research works such as Fisher and Statman (2000) and Verma and Soydemir (2009) used AAII index as a sentiment measure. Also, UMSC is a monthly sentiment index constructed from a consumer confidence survey answered by a random group of five hundred continental US households. Despite being considered less related to investor sentiment, UMSC index has been applied in stock market studies (e.g., (Fisher & Statman, 2000)). Moreover, Sentix (www.sentix.de) creates sentiment indices for various stock markets (e.g., US, Japan, Germany, Euro zone) from surveys answered by more than 3500 participants (Schmeling, 2007). Another example is the II weekly

[3] http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

index based on over a hundred independent market newsletters, with each newsletter being categorized as bullish, bearish or correction. II measures may be more correlated to institutional sentiment than AAII, because many of the newsletters' authors are market professionals (Brown & Cliff, 2004). The II index is widely applied in behavior finance literature (e.g., (Fisher & Statman, 2000; Solt & Statman, 1988; Verma & Soydemir, 2009)).

This paper tests the prediction of weekly sentiment indicators (i.e., AAII, II) by using Twitter sentiment measures. This procedure may allow an acceptable anticipation of AAII and II values or a satisfactory alternative whenever they are unavailable. We did not experiment the forecasting of monthly sentiment indicators because there is not sufficient data to perform a feasible analysis. However, we also used monthly sentiment measures (i.e., UMSC, Sentix for US) in the creation of a sentiment indicator using the Kalman Filter measure described in the next subsection. AAII, II, UMSC and Sentix values were obtained from Thompson Reuters Datastream (http://online.thomsonreuters.com/datastream/).

2.4. Sentiment indicators created by Kalman filter procedure

KF permits the combination of diverse observed variables in order to extract a latent variable. Existing investor sentiment indicators are measured by different approaches at different frequencies (e.g., surveys or social media interactions) and they usually produce distinct values. These observed values are related to sentiment but they contain some noise. The utilization of KF may permit the creation of a sentiment indicator that is more representative and less noisy than their individual components. Moreover, KF allows the usage of indicators with distinct frequencies (e.g., daily, weekly, monthly). Thus, we can produce a daily sentiment indicator from the combination of daily, weekly and monthly values. The linear dynamical model can be represented as follows:

$$\begin{aligned} Y_t &= F_t\theta_t + v_t, \quad v_t \sim N(0, V_t) \\ \theta_t &= G_t\theta_{t-1} + w_t, \quad w_t \sim N(0, W_t) \end{aligned} \tag{6}$$

where the first equation describes the observed variables (e.g., survey or social media sentiment indicators), the second equation represents the latent variable and V, W are parameter matrices. θ_0 is assumed to be normally distributed with mean 0 and variance $1e7$. The model is estimated by maximum likelihood allowing the observation noises, v_t, to be cross-correlated. To reduce the complexity of the optimization problem and ensure that the variance-covariance matrix is positive semi-definite we use the approach suggested by Pinheiro and Bates (1996) and followed by Petris et al. (2010), parametrizing the covariance matrix, V, in terms of the elements of its log-Cholesky decomposition and the system variance, W, using its log.

We created a daily sentiment indicator by applying the KF procedure to five different sentiment indicators: AAII, II, UMSC, Sentix and the daily Twitter sentiment indicator (TWT) created in this work. AAII, II and TWT values correspond to the Bull-Bear spread (e.g., (Verma & Soydemir, 2009)):

$$SP_t = \frac{NBull_t - NBear_t}{NBull_t + NBear_t} \tag{7}$$

All indicators were normalized by calculating their standard score. The model parameters were estimated using the first training rolling window also applied in the forecasting of stock market variables (i.e., first 300 days). Then, we created the sentiment indicators for the entire time period by filtering the series using the estimated model.

KF values were tested as sentiment indicators in forecasting models to assess eventual improvements compared to the utilization of TWT as proxy for sentiment. As an example, Fig. 2 shows the overall (for all analyzed stocks) KF sentiment indicator values and its individual components (AAII, II, UMSC, Sentix, TWT) when considering the period that ranges from 1st January of 2014 to 30th June of 2014. The plot confirms that KF indicator presents smoother values than TWT.

2.5. Stock market data

Various studies have analyzed the influence of sentiment on portfolios formed on different characteristics (e.g., market capitalization, book-to-market ratio, industries). For example, some papers refer that sentiment has more impact on returns of stocks with lower market capitalization (Baker & Wurgler, 2006; Baker et al., 2012). However, few of these studies apply sentiment measures extracted from social media and higher frequencies than the monthly periodicity. Therefore, in this study we explored a comprehensive set of stocks and portfolios having distinct characteristics:

- Standard & Poor's 500 (SP500): index composed by 500 large companies.
- Russell 2000 (RSL): index of the smallest 2000 companies belonging to Russell 3000 index.
- Dow Jones Industrial Average (DJIA): constituted by 30 large companies listed on NYSE and NASDAQ.
- Nasdaq 100 (NDQ): includes the 100 of the largest non-financial stocks traded on NASDAQ.
- Excess return on the market (RMRF): return of the market minus the risk-free return rate. The return of the market corresponds to the value-weight return of all CRSP companies integrated in the US and traded on the NYSE, AMEX, or NASDAQ while the risk-free return rate is the one-month Treasury bill rate.
- Small Minus Big (SMB): a Fama and French factor corresponding to the difference in returns between small and large firms.
- High Minus Low (HML): a Fama and French factor that is equal to the difference in returns between value (i.e., high book-to-market ratios) and growth (i.e., low book-to-market ratios) stocks.
- Momentum Factor (MOM): spread in returns between high prior return portfolios and low prior return portfolios.
- CBOE Volatility Index (VIX): measures the implied volatility of SP500 index options. It is often considered the market's "fear gauge" because it intends to represent investors expectation of future 30 days volatility.
- Portfolios formed on size (PSize): contains return values of portfolios constructed on the market capitalization, namely bottom 30% (Lo30), middle 40% (Mid40), top 30% (Hi30) and quintiles (Lo20, Qnt2, Qnt3, Qnt4, Hi20).
- 10 Industry Portfolios (PInd): returns of ten different industries, namely Consumer Non-Durables (NoDur), Consumer Durables (Durbl), Manufacturing (Manuf), Energy (Enrgy), Business Equipment (HiTec), Telecommunications (Telcm), Shops (Shops), Health (Hlth), Utilities (Utils) and Other (Other).

Thus, we analyze the effect of sentiment and attention on stocks having distinct size (e.g., PSize, SMB), industries (e.g., PInd), momentum (e.g., MOM) and book-to-market ratios (e.g., HML).

Several studies defend that investor sentiment and attention may influence stock market variables. For instance, sentiment may have predictive value for returns (e.g., (Bollen et al., 2011; Deng et al., 2011; Sabherwal et al., 2011; Sprenger et al., 2014)), volatility (e.g., (Antweiler & Frank, 2004; Sabherwal et al., 2011)) and trading volume (e.g., (Antweiler & Frank, 2004; Sabherwal et al., 2011; Tetlock, 2007)). Posting volume on social media (e.g., microblogs and message boards) may also add information for the forecasting of returns (e.g., (Antweiler & Frank, 2004; Wysocki, 1998)), trading

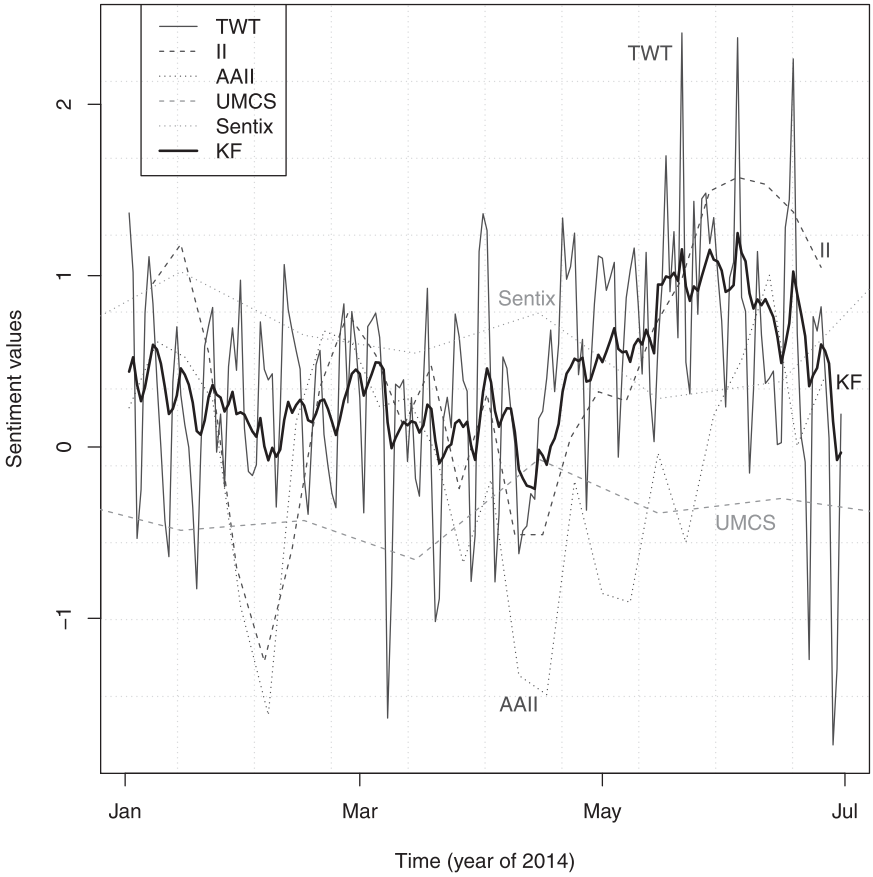

Fig. 2. Sentiment indicator values (KF, AAII, II, UMSC, Sentix and TWT).

volume (e.g., (Oliveira et al., 2013; Sprenger et al., 2014; Wysocki, 1998)) and volatility (e.g., (Antweiler & Frank, 2004)). We have focused on the prediction of these three different stock market variables:

- Daily returns of SP500, RSL, DJIA, NDQ, RMRF, SMB, HML, MOM, PSize and PInd. Returns measure changes in the asset value. We calculated the returns of SP500, RSL, DJIA and NDQ using the total return index (RI datatype) retrieved from Datastream as follows:

$$r_t = \frac{RI_t - RI_{\text{t-1}}}{RI_{\text{t-1}}} * 100 \quad (8)$$

  where RI_t and $RI_{\text{t-1}}$ are the total return index values of day t and $t-1$. The returns of the remaining stocks were directly collected from Professor Kenneth French webpage (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html). All values are in percentage.
- Daily trading volume of SP500 and DJIA collected from Datastream. Trading volume is the number of shares traded in a given period of time (values in thousands).
- Daily volatility is measured using the model free estimate given the VIX index, and also the realized volatility measure given by a realized kernel for SP500, RSL, DJIA and NDQ. Since the realized kernel values are very small, we converted them into annualized realized volatility as (Areal & Taylor, 2002):

$$av_t = \sqrt{rk_t} * \sqrt{252} * 100 \quad (9)$$

  where av_t is the annualized realized volatility and rk_t is the realized kernel value. Volatility provides a measures of total risk associated with an investment. VIX data is available at CBOE webpage (http://www.cboe.com/micro/vix/historical.aspx) and the realized kernel of the referred indices were collected from Oxford-Man Institute of Quantitative Finance (http://realized.oxford-man.ox.ac.uk/data/download).

2.6. Models

In this work, we tested five different regression methods (Hastie, Tibshirani, & Friedman, 2008) to predict stock market variables: Multiple Regression (MR), Neural Network (NN), Support Vector Machine (SVM), Random Forest (RF) and an Ensemble Averaging (EA) method.

The classical MR model assumes a linear relationship between several independent variables and a dependent target. This model is very fast to learn, easy to interpret and has been extensively applied in finance.

NN is a system of interconnected neurons whose functioning is inspired by biological neural networks. The numeric weights linking the neurons are calibrated during the learning process. In this study we applied the multilayer perceptron having one hidden layer with logistic functions. The output node applies a linear function for regression (e.g., returns, volatility, trading volume). The final output is dependent of the selection of initial weights. To address this problem, we apply an ensemble of NNs and calculate the average of the individual predictions (Hastie et al., 2008). The number of nodes in the hidden layer (H) was the only hyperparameter we had to tune in this work.

The SVM model was initially proposed to perform classification tasks and then extended to regression by adopting an ϵ-insensitive loss function. SVM can execute a nonlinear mapping by projecting the inputs into high-dimensional space using kernel functions. When compared with multilayer perceptron, the SVM al-

gorithm has the advantage of always converging to the optimal set of weights. In this work, we utilize the popular gaussian kernel and tuned γ, C and ϵ hyperparameters.

RF is an ensemble model that generates a larger number of unpruned decision trees during the training process. The individual trees are based on a random feature selection, using bootstrap training samples. The final RF predictions are built by averaging the outputs of the individual trees.

Ensembles can be used to combine multiple prediction models. In this work, we assume a simple and popular EA approach that often leads to good results and that consists in producing a new predictive response based on the averaging predictions of the MR, NN, SVM and RF models.

Each regression method has its own learning capabilities and advantages. MR is more easy to interpret, does not require the tunning of hyperparameters and has less tendency to overfit. Yet, MR is a rather rigid model and it has limited capacity to extract nonlinear associations. All other methods (NN, SVM, RF and EA) are more flexible and suited to deal with nonlinear complex variable relationships. Often, these nonlinear methods lead to complex data-driven models. However, these complex models (including the ensemble) can still be understood by humans by using a sensitivity analysis and visualization techniques (Cortez & Embrechts, 2013). Functional nonlinear methods, such as NN and SVM, are more prone to overfitting and thus several hyperparameters need to be tunned (e.g., hidden nodes, kernel parameter). In contrast, RF tends to provide good results with its default parameters but for several small or medium sized datasets the computational effort to fit the model is often larger than a single NN or SVM training due to the usage of a large number of trees. When compared with NN, SVM present theoretical advantages, such as the absence of local minima in the model optimization phase. Ensembles often achieve better performances than their individual prediction models (Oztekin, Kizilaslan, Freund, & Iseri, 2016). The tested ensemble (EA) is quite simple and does not demand a significant extra computation but requires that all of its individual models are previously trained.

In order to evaluate the relevance of microblogging data to predict stock market behavior, we test all five regression models under two main strategies: with and without microbloging data. The baseline model is the Auto-Regressive (AR) model of order p (past time lags) of the predicted target, while the microblog based models are described next. When comparing the baseline and microblog based models, we select the best regression method (among MR, NN, SVM, RF and EA). Given the extensive number of experiments conducted in this paper, we fixed the number of adopted time lags to $p = 5$ for both baseline, i.e., AR(5), and microblog based models. We note that several other related works also used a similar short and fixed number of time lags. For instance, the same five past trading days were used in Tetlock (2007), Deng et al. (2011) and Oh and Sheng (2011).

All predictive experiments were conducted using the R tool and rminer package, which facilitates the application of data mining methods in real-world tasks (Cortez, 2010). The same methods (MR, NN, SVM and RF) were executed for both baseline and microblog based strategies. The default parameters (e.g., 500 trees for RF, ensemble with 3 multilayer perceptrons for NN, tolerance termination criterion of 0.001 for SVM) were used for the all methods except for the hyperparameters (H for NN and γ, C or ϵ for SVM). These hyperparameters were set using a grid search and internal 10-fold cross-validation considering the first training window (e.g., first 300 trading days for the daily stock market variables). The grid search values were set to $H \in \{1, 2, 3, 4\}$, $\gamma \in \{2^{-9}, 2^{-7}, \ldots, 2^{3}\}$, $C \in \{2^{-3}, 2^{-1}, \ldots, 2^{9}\}$ and $\epsilon \in \{0, 0.01, 0.1, 0.2, 0.4, 0.6\}$.

The tested models to predict returns were:

$$
\begin{aligned}
&\hat{R}_t = f(R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}) && (\textbf{MRet1}, \text{baseline})\\
&\hat{R}_t = f(S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}) && (\textbf{MRet2})\\
&\hat{R}_t = f(R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5},\\
&\qquad S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}) && (\textbf{MRet3})\\
&\hat{R}_t = f(S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5},\\
&\qquad Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) && (\textbf{MRet4})\\
&\hat{R}_t = f(R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5},\\
&\qquad S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5},\\
&Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) && (\textbf{MRet5})\\
&\hat{R}_t = f(Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) && (\textbf{MRet6})\\
&\hat{R}_t = f(R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5},\\
&\quad Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) && (\textbf{MRet7})
\end{aligned}
\tag{10}
$$

where S_t refers to the sentiment value of day t (VA, BI, BR or KF indicators) and Nt_t refers to the first difference of posting volume of day t.

To forecast volatility, we experimented the following models:

$$
\begin{aligned}
&\hat{Vt}_t = f(Vt_{t-1}, Vt_{t-2}, Vt_{t-3}, Vt_{t-4}, Vt_{t-5}) && (\textbf{MVt1}, \text{baseline})\\
&\hat{Vt}_t = f(S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}) && (\textbf{MVt2})\\
&\hat{Vt}_t = f(Vt_{t-1}, Vt_{t-2}, Vt_{t-3}, Vt_{t-4}, Vt_{t-5},\\
&\qquad S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}) && (\textbf{MVt3})\\
&\hat{Vt}_t = f(S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5},\\
&\qquad Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) && (\textbf{MVt4})\\
&\hat{Vt}_t = f(Vt_{t-1}, Vt_{t-2}, Vt_{t-3}, Vt_{t-4}, Vt_{t-5},\\
&\qquad S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5},\\
&Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) && (\textbf{MVt5})\\
&\hat{Vt}_t = f(Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) && (\textbf{MVt6})\\
&\hat{Vt}_t = f(Vt_{t-1}, Vt_{t-2}, Vt_{t-3}, Vt_{t-4}, Vt_{t-5},\\
&\qquad Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) && (\textbf{MVt7})
\end{aligned}
\tag{11}
$$

where Vt_t is the volatility value of day t.

In the prediction of trading volume, we tested the following models:

$$
\begin{aligned}
&\hat{Vo}_t = f(Vo_{t-1}, Vo_{t-2}, Vo_{t-3}, Vo_{t-4}, Vo_{t-5}) && (\textbf{MVol1}, \text{baseline})\\
&\hat{Vo}_t = f(S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}) && (\textbf{MVol2})\\
&\hat{Vo}_t = f(Vo_{t-1}, Vo_{t-2}, Vo_{t-3}, Vo_{t-4}, Vo_{t-5},\\
&\qquad S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}) && (\textbf{MVol3})\\
&\hat{Vo}_t = f(S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5},\\
&\qquad Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) && (\textbf{MVol4})\\
&\hat{Vo}_t = f(Vo_{t-1}, Vo_{t-2}, Vo_{t-3}, Vo_{t-4}, Vo_{t-5},\\
&\qquad S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5},\\
&Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) && (\textbf{MVol5})\\
&\hat{Vo}_t = f(Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) && (\textbf{MVol6})\\
&\hat{Vo}_t = f(Vo_{t-1}, Vo_{t-2}, Vo_{t-3}, Vo_{t-4}, Vo_{t-5},\\
&\qquad Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) && (\textbf{MVol7})
\end{aligned}
\tag{12}
$$

where Vo_t is the trading volume of day t.

In this study, we also used Twitter and KF sentiment indicators to predict survey sentiment indicators. AAII and II are popular survey sentiment indicators already used in diverse research studies about stock market behavior (e.g., (Fisher & Statman, 2000; Solt & Statman, 1988; Verma & Soydemir, 2009)). These indicators have distinct publication days. AAII values are released each Thursday comprising data from previous Thursday until Wednesday. II indicators are published each Wednesday and they are related to analysis performed from previous Wednesday to last Tuesday. To properly compare Twitter and KF indicators to each survey measure, we created different weekly microblogging indicators corresponding to the time interval used by each survey indicator. Additionally,

we tested the utilization of the previous 7 daily (one week) Twitter and KF indicators.

We predicted four different values for each survey indicator: VA, BI, negative and positive percentage. In the forecasting of each survey value, we used the equivalent Twitter indicator (i.e., VA, BI, BearR and BullR). To compute the weekly Twitter values, we also experimented two different approaches. The first approach (AA) calculates the weekly value using the total number of positive and/or negative messages of the week while the second approach (MA) computes the average of the seven daily indicators that compose the week. The production of the weekly KF values applies the second approach.

We applied seven different models by exploring five lags of the target survey indicator, five lags of the weekly microblogging indicators and seven lags (one week) of the daily microblogging indicators. The explored forecasting of survey sentiment models are:

$$
\begin{aligned}
&\hat{Sv}_t = f(Sv_{t-1}, Sv_{t-2}, Sv_{t-3}, Sv_{t-4}, Sv_{t-5}) && \textbf{(MSv1}\text{, baseline)}\\
&\hat{Sv}_t = f(Sw_{t-1}, Sw_{t-2}, Sw_{t-3}, Sw_{t-4}, Sw_{t-5}) && \textbf{(MSv2)}\\
&\hat{Sv}_t = f(Sd_{t-1}, Sd_{t-2}, Sd_{t-3}, Sd_{t-4}, Sd_{t-5},\\
&\quad Sd_{t-6}, Sd_{t-7}) && \textbf{(MSv3)}\\
&\hat{Sv}_t = f(Sw_{t-1}, Sw_{t-2}, Sw_{t-3}, Sw_{t-4}, Sw_{t-5},\\
&Sd_{t-1}, Sd_{t-2}, Sd_{t-3}, Sd_{t-4}, Sd_{t-5}, Sd_{t-6}, Sd_{t-7}) && \textbf{(MSv4)}\\
&\hat{Sv}_t = f(Sv_{t-1}, Sv_{t-2}, Sv_{t-3}, Sv_{t-4}, Sv_{t-5},\\
&Sw_{t-1}, Sw_{t-2}, Sw_{t-3}, Sw_{t-4}, Sw_{t-5}) && \textbf{(MSv5)}\\
&\hat{Sv}_t = f(Sv_{t-1}, Sv_{t-2}, Sv_{t-3}, Sv_{t-4}, Sv_{t-5},\\
&Sd_{t-1}, Sd_{t-2}, Sd_{t-3}, Sd_{t-4}, Sd_{t-5}, Sd_{t-6}, Sd_{t-7}) && \textbf{(MSv6)}\\
&\hat{Sv}_t = f(Sv_{t-1}, Sv_{t-2}, Sv_{t-3}, Sv_{t-4}, Sv_{t-5},\\
&Sw_{t-1}, Sw_{t-2}, Sw_{t-3}, Sw_{t-4}, Sw_{t-5},\\
&Sd_{t-1}, Sd_{t-2}, Sd_{t-3}, Sd_{t-4}, Sd_{t-5}, Sd_{t-6}, Sd_{t-7}) && \textbf{(MSv7)}
\end{aligned}
\tag{13}
$$

where Sv_t corresponds to the weekly survey sentiment values (AAII or II) of day t, Sw_t corresponds to the weekly microblogging sentiment values (BI, VA, BullR, BearR or KF) of day t and Sd_t corresponds to the daily microblogging values (AAII or II) of day t.

2.7. Evaluation

There is empirical evidence that good forecasting methods provide consistent results across multiple metrics (Crone, Hibon, & Nikolopoulos, 2011). Given the large number of experiments conducted in this work, we opted for a single error metric when evaluating the quality of the predictions. In this work, we selected an absolute error based metric, which is a common approach in the forecasting domain (Hyndman & Koehler, 2006). For instance, in (Armstrong, 2001; Armstrong & Collopy, 1992) it is argued that squared error metrics, such as Root Mean Square Error (RMSE), are not reliable due to their sensitivity to outliers and should be replaced by absolute error metrics when comparing across time series. We note that other related works also have adopted absolute error metrics, such as: Mean Absolute Error (MAE) (Deng et al., 2011) and Mean Absolute Percentage Error (MAPE) (Bollen et al., 2011; Deng et al., 2011; Mao et al., 2011).

Using any absolute error measure should lead to the same ranking differences when comparing distinct forecasting models, thus the particular choice of such measure affects mostly its range of values and interpretation. In this paper, we selected the Normalized Mean Absolute Error (NMAE) metric that is calculated as (Goldberg, Roeder, Gupta, & Perkins, 2001):

$$
\begin{aligned}
MAE &= \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|\\
NMAE &= \frac{MAE}{y_H - y_L}
\end{aligned}
\tag{14}
$$

where y_i is the target value for time i, y_H is the highest target value, y_L is the lowest target value, $\hat{y}_i$ is the predicted value and N corresponds to the number of predictions considered. When compared with other absolute based metrics, the NMAE presents several advantages. First, it is easier to interpret than MAE, since it expresses the error as a percentage of the full target scale. The lower the NMAE values, the better are the forecasts. Second, it is scale independent, which is particularly useful in this work since we predict variables with distinct scales. Third, it does not contain the limitations of other scale independent measures. For instance, MAPE can produce infinity values when the denominator (target values) is zero (Hyndman & Koehler, 2006), which might occur in several of the predicted variables (e.g., returns can have near zero values). While we only consider NMAE values, in the tables with predictive results we also show the target range value ($y_H - y_L$), thus the MAE values can be easily obtained by computing the inverse of Eq. 14.

To measure the generalization capability of the predictive models, we applied a fixed-size rolling windows scheme (Tashman, 2000). For each prediction (e.g., day t), the model is trained using a window of the previous W consecutive samples (e.g., from day $t-W$ to day $t-1$) and used to predict the next value (time t). Then, the training window is slided by discarding its oldest element and adding the value of t in order to retrain the model and predict the value at time $t+1$, and so on. Therefore, a dataset of length L will produce $L-W$ model trainings and their respective predictions. This rolling windows validation is realistic since it mimics the way a predictive model would be used in a real-environment, trained with a large number of past data and used to predict the next daily/weekly values. And it is robust, since it allows the training and testing of a large number of models. In this work, we applied a window size of $W = 300$ days for the prediction of the daily stock market variables and $W = 50$ observations for the forecasting of the weekly survey sentiment indicators. The number of predictions range from 392 to 439 for the daily variables and 92 to 93 for the weekly survey indices. We note that these numbers are much higher than most state of the art works (e.g., 8 and 30 predictions).

The prediction ability of the models was evaluated by the Diebold-Mariano (DM) test for predictive accuracy (Diebold & Mariano, 2002), under a pairwise comparison between the baseline and microblog based models. Thus, we assume that the microblogging data has predictive content if the respective model has a statistical significant DM test.

3. Results and discussion

3.1. Prediction of returns

This work analyzed the prediction of returns of diverse indices as well as portfolios formed on size and industries. We tested sentiment indicators created by three different formulas (BullR, BI and VA) and KF procedure. Table 2 presents the NMAE results produced by the four regression methods (MR, RF, NN and SVM) using sentiment indicators produced by BullR approach and the p-value calculated by DM test for predictive accuracy (Diebold & Mariano, 2002) for SP500, RSL, DJIA, NDQ, RMRF, SMB, HML and MOM. In the table, the baseline results are underlined, while the lowest NMAE value is in **bold**. The first column also shows in brackets the number of predictions ($L-W$) and the target range ($y_H - y_L$).

We summarize the results for BullR, BI, VA and KF indicators in Table 3. For each index, Table 3 identifies the baseline model, the lowest NMAE model and models generating statistically significant results in the DM test. Ten SVM models significantly improve the results of baseline models for the forecasting of SP500, DJIA, MOM, SMB and RMRF. Two of these models obtain p-value

Table 2
Predictive results for returns of SP500, RSL, DJIA, NDQ, RMRF, SMB, HML and MOM. Utilization of general sentiment indicators calculated using BullR formula and first difference of the posting volume. For each index, the baseline model is underlined and the lowest NMAE value is in **bold** (* – p-value < 10%, ** – p-value < 5%, *** – p-value < 1%, NMAE in %).

	Mtd	MRet1	MRet2	MRet3	MRet4	MRet5	MRet6	MRet7
DJIA	MR	8.12	8.11	8.24	8.20	8.33	8.11	8.25
(number of predictions: 414;	RF	8.38	8.41	8.43	8.44	8.42	8.38	8.45
returns range: 7.53)	SVM	8.19	8.09	**8.01***	8.07	8.07	8.05	8.05
	NN	8.28	8.42	8.16	8.29	8.44	8.56	8.26
	EA	8.14	8.2	8.16	8.15	8.23	8.13	8.22
HML	MR	10.49	10.37	10.54	10.36	10.54	10.34	10.52
(number of predictions: 392;	RF	10.42	10.56	10.39	10.49	10.49	10.62	10.56
returns range: 3.36)	SVM	10.29	10.31	**10.24**	10.26	10.25	10.26	10.33
	NN	10.92	10.84	10.66	10.62	10.73	10.99	10.47
	EA	10.55	10.32	10.37	10.33	10.37	10.35	10.4
MOM	MR	10.91	10.80	10.94	11.02	11.18	11.03	11.11
(number of predictions: 392;	RF	11.00	11.14	10.93	11.18	11.17	11.31	11.21
returns range: 4.63)	SVM	10.78	10.79	**10.73**	10.84	10.78	10.88	10.81
	NN	11.18	11.01	11.99	11.33	11.32	11.13	11.11
	EA	10.8	10.84	10.92	11.31	11.1	11.08	11.01
NDQ	MR	7.68	7.70	7.82	7.86	7.99	7.74	7.85
(number of predictions: 439;	RF	7.85	7.90	7.88	7.97	7.96	7.88	7.93
returns range: 9.35)	SVM	7.61	**7.59**	7.61	7.78	7.62	7.64	7.69
	NN	7.99	8.38	7.94	8.93	7.81	7.97	8.04
	EA	7.71	7.74	7.7	7.81	7.85	7.7	7.88
RMRF	MR	8.31	8.32	8.45	8.42	8.58	8.36	8.44
(number of predictions: 392;	RF	8.54	8.59	8.61	8.68	8.81	8.56	8.74
returns range: 7.58)	SVM	8.27	8.35	8.28	**8.24**	**8.24**	**8.24**	8.26
	NN	8.95	8.46	8.47	8.65	8.54	9.57	8.61
	EA	8.36	8.38	8.38	8.43	8.41	8.35	8.53
RSL	MR	11.10	11.31	11.35	11.46	11.49	11.28	11.30
(number of predictions: 439;	RF	11.17	11.47	11.19	11.58	11.40	11.44	11.33
returns range: 7.02)	SVM	11.15	11.71	11.20	11.24	11.19	11.13	11.13
	NN	11.21	11.60	11.21	11.46	11.64	12.57	13.85
	EA	**11**	11.37	11.44	11.36	11.32	11.34	11.3
SMB	MR	12.44	12.44	12.59	12.50	12.76	12.40	12.59
(number of predictions: 392;	RF	12.46	12.72	12.54	12.74	12.50	12.72	12.42
returns range: 3.36)	SVM	12.44	12.41	12.40	**12.27***	12.46	12.48	12.44
	NN	13.49	13.09	13.05	13.41	12.70	12.55	12.89
	EA	12.5	12.66	12.48	12.4	12.46	12.44	12.4
SP500	MR	7.90	7.88	8.01	7.97	8.11	7.92	8.01
(number of predictions: 413;	RF	8.32	8.20	8.32	8.23	8.35	8.13	8.32
returns range: 7.85)	SVM	7.87	7.83	7.88	7.92	7.86	7.83	**7.80****
	NN	8.59	8.17	7.95	8.00	8.02	7.92	8.40
	EA	7.92	7.99	7.94	8	7.99	8.11	7.97

Table 3
Predictive results for returns of SP500, RSL, DJIA, NDQ, RMRF, SMB, HML and MOM. Utilization of general sentiment indicators (BullR, BI and VA approaches), KF indicators and first difference of the posting volume. NMAE values of baseline model, lowest NMAE model and models producing statistical significant results in DM test (* – p-value < 10%, ** – p-value < 5%, *** – p-value < 1%, sentiment indicators in parenthesis, NMAE in %, NMAE values lower than baseline are in **bold**).

Index	Baseline	Lowest NMAE	Statistical significant results
DJIA (n° predictions: 414; returns range: 7.53)	MR: 8.12	SVM MRet2 (KF): **7.98***	SVM MRet3 (BR): **8.01*** SVM MRet2 (KF): **7.98***
HML (n° predictions: 392; returns range: 3.36)	SVM: 10.29	SVM MRet3 (BR): **10.24**	
MOM (n° predictions: 392; returns range: 4.63)	SVM: 10.78	SVM MRet2 (KF): **10.69***	SVM MRet2 (KF): **10.69***
NDQ (n° predictions: 439; returns range: 9.35)	SVM: 7.61	SVM MRet7: **7.58**	
RMRF (n° predictions: 392; returns range: 7.58)	SVM: 8.27	SVM MRet3 (KF): **8.19***	SVM MRet3 (KF): **8.19***
RSL (n° predictions: 439; returns range: 7.02)	EA: 11.02	EA MRet1: 11.02	
SMB (n° predictions: 392; returns range: 3.36)	MR: 12.44 SVM: 12.44	SVM MRet4 (BR): **12.27***	SVM MRet4 (BR): **12.27***
SP500 (n° predictions: 439; returns range: 7.85)	SVM: 7.87	SVM MRet3 (KF): **7.79****	SVM MRet7: **7.80**** SVM MRet6: **7.81*** SVM MRet4 (VA): **7.81*** SVM MRet5 (VA): **7.81*** SVM MRet3 (KF): **7.79****

Table 4
Predictive results for returns of portfolios formed on size. Utilization of general sentiment indicators (BullR, BI and VA approaches), KF indicators and first difference of the posting volume. NMAE values of baseline model, lowest NMAE model and models producing statistical significant results in DM test (* – p-value $<$ 10%, ** – p-value $<$ 5%, *** – p-value $<$ 1%, sentiment indicators in parenthesis, NMAE in %, NMAE values lower than baseline are in **bold**, 392 predictions).

Index	Baseline	Lowest NMAE	Statistical significant results
Lo30 (returns range: 5.96)	MR: 11.75	SVM MRet4 (VA): **11.57***	SVM MRet4 (BR): **11.59*** SVM MRet4 (BI): **11.59*** SVM MRet4 (VA): **11.57*** SVM MRet5 (VA): **11.60***
Med40 (returns range: 6.87)	EA: 11.04	SVM MRet4 (BR): **11.01** SVM MRet5 (BR): **11.01** SVM MRet7: **11.01** SVM MRet5 (BI): **11.01** SVM MRet5 (KF): **11.01**	
Hi30 (returns range: 7.34)	SVM: 8.71	SVM MRet7 (BI): **8.68** SVM MRet4 (VA): **8.68** SVM MRet5 (VA): **8.68** SVM MRet6: **8.68**	
Lo20 (returns range: 5.64)	MR: 11.94 EA: 11.94	SVM MRet4 (VA): **11.66*****	SVM MRet4 (BR): **11.74*** SVM MRet4 (BI): **11.75*** SVM MRet5 (BI): **11.70**** SVM MRet7: **11.75**** SVM MRet2 (VA): **11.75*** EA MRet2 (VA): **11.75*** SVM MRet3 (VA): **11.78*** SVM MRet4 (VA): **11.66***** SVM MRet5 (VA): **11.74**** SVM MRet6: **11.71**** SVM MRet2 (KF): **11.74*** SVM MRet4 (KF): **11.72****
Qnt2 (returns range: 7.31)	EA: 11.56	SVM MRet6: **11.53**	
Qnt3 (returns range: 6.94)	MR: 10.83 EA: 10.83	SVM MRet5 (BR): **10.77**	
Qnt4 (returns range: 6.72)	SVM: 10.31	SVM MRet2 (KF): **10.28**	
Hi20 (returns range: 7.33)	SVM: 8.63	SVM MRet6: **8.57** SVM MRet7: **8.57**	

inferior to 5%, both for the forecasting of SP500. There are no microblogging models having significantly higher predictive accuracy than baseline models for the remaining indices (i.e., HML, NDQ and RSL). However, the lowest NMAE values are produced by models containing microblogging features for all items, except for RSL. The utilization of KF sentiment indicators lowered the NMAE results for four indices: DJIA, MOM, RMRF and SP500. Furthermore, models applying KF indicators are significantly more accurate than baseline for SP500, DJIA, MOM and RMRF, while models using TWT indicators obtain this statistical significant results only for DJIA, SMB and SP500. Moreover, SVM MRet3 model using KF indicators produces statistically better forecasts of the SP500 than the baseline. The posting volume features were important for some indices. The lowest NMAE values of the prediction of NDQ and SMB were produced by models using the first difference of the number of tweets. Additionally, there are diverse models applying these features significantly more accurate than baseline for SP500 and SMB. The majority of the most accurate models applies the SVM method.

Some studies refer that there is a distinct sentiment effect in prices of stocks with large and small market capitalization (e.g., (Baker & Wurgler, 2006; Baker et al., 2012; Lee, Jiang, & Indro, 2002; Neal & Wheatley, 1998)). To analyze the influence of sentiment on stocks with different size, we predicted returns of portfolios formed on size. Table 4 shows a summary of all tested models for the prediction of portfolios formed on size. There are sixteen models significantly more accurate than baseline in the forecasting of portfolios of lower market capitalization (i.e., Lo20 and Lo30). The prediction of Lo20 returns has one model obtaining p-value less than 1% in the pairwise DM test and five other models generating p-value inferior to 5%. These results may indicate that sentiment is more informative to the prediction of stocks of smaller capitalization, which is consistent with previous findings (e.g., (Baker & Wurgler, 2006; Baker et al., 2012)). The most accurate microblogging models are SVM. Posting volume seems informative for the prediction of these portfolios. The majority of the models obtaining statistical significant results in the pairwise DM test and the lowest NMAE values apply the first difference of posting volume. KF indicators seem less useful for the forecasting of portfolios formed on size than for the prediction of indices. Models applying KF indicators outperformed models using Twitter indicators only for Qnt4. Nevertheless, in the prediction of Lo20 there is one model applying KF indicators that produced p-value less than 5% in the DM test and another model obtained p-value less than 10%.

We also tested the prediction of portfolios of 10 different industries. The respective evaluation results are summarized in Table 5. There are several microblogging models significantly more accurate than baseline for Enrgy, HiTec and Other. Seven of these models obtain p-value less than 5% in the pairwise DM test for the prediction of Enrgy and three models have p-value less than 5% for the forecasting of HiTec. Therefore, microblogging features are particularly informative for Enrgy and HiTec. A possible explanation for these results is the unusual high number of microblogging messages related to stocks of these sectors, mainly for HiTec. Thus, the general sentiment indicators may be biased toward these industries. For demonstration purposes, the forecasted returns of Enrgy sector by the NN MRet2 (VA) are presented in the Fig. 3. The utilization of posting volume was also important. The majority of models obtaining statistical significant results in the pairwise DM test use posting volume features. Moreover, the utilization of KF indicators is beneficial in some situations. KF indicators permit lower NMAE values than TWT indicators for the HiTec and Utils sectors, and they are the most applied sentiment indicator in models which are significantly more accurate than baselines.

Table 5
Predictive results for returns of portfolios formed on industries. Utilization of general sentiment indicators (BullR, BI and VA approaches), KF indicators and first difference of the posting volume. NMAE values of baseline model, lowest NMAE model and models producing statistical significant results in DM test (* – p-value < 10%, ** – p-value < 5%, *** – p-value < 1%, sentiment indicators in parenthesis, NMAE in %, NMAE values lower than baseline are in **bold**, 350 predictions).

Index	Baseline	Lowest NMAE	Statistical significant results
Durbl (returns range: 5.71)	MR: 13.90	EA MRet2 (BR): **13.74** SVM MRet6: **13.74**	
Enrgy (returns range: 21.12)	SVM: 7.18	NN MRet2 (VA): **7.07****	SVM MRet3 (BI): **7.10**** SVM MRet5 (BI): **7.10**** SVM MRet6: **7.13*** SVM MRet7: **7.12*** SVM MRet2 (VA): **7.08**** NN MRet2 (VA): **7.07**** SVM MRet3 (VA): **7.11*** SVM MRet5 (VA): **7.10**** SVM MRet2 (KF): **7.10**** SVM MRet3 (KF): **7.10**** SVM MRet4 (KF): **7.11*** SVM MRet5 (KF): **7.12***
HiTec (returns range: 5.90)	MR: 12.79 EA: 12.79	SVM MRet5 (KF): **12.55****	SVM MRet3 (BR): **12.64*** SVM MRet4 (BR): **12.56**** SVM MRet4 (BI): **12.56**** SVM MRet2 (KF): **12.58*** SVM MRet3 (KF): **12.64*** SVM MRet5 (KF): **12.55****
Hlth (returns range: 7.10)	NN: 12.94	SVM MRet6: **12.85**	
Manuf (returns range: 5.17)	SVM: 13.66	SVM MRet7: **13.62**	
NoDur (returns range: 4.55)	SVM: 12.77	SVM MRet3 (BR): **12.72**	
Other (returns range: 4.04)	SVM: 13.51	SVM MRet6: **13.34***	SVM MRet6: **13.34***
		SVM MRet4 (VA): **13.34***	SVM MRet3 (VA): **13.37***
		SVM MRet4 (KF): **13.34***	SVM MRet4 (VA): **13.34***
			SVM MRet4 (KF): **13.34***
Shops (returns range: 4.76)	SVM: 13.85	SVM MRet4 (BI): **13.67***	SVM MRet4 (BI): **13.67***
Telcm (returns range: 5.52)	SVM: 14.15	SVM MRet2 (BR): **13.98***	SVM MRet2 (BR): **13.98***
Utils (returns range: 5.77)	SVM: 11.18	RF MRet5 (KF): **11.04**	

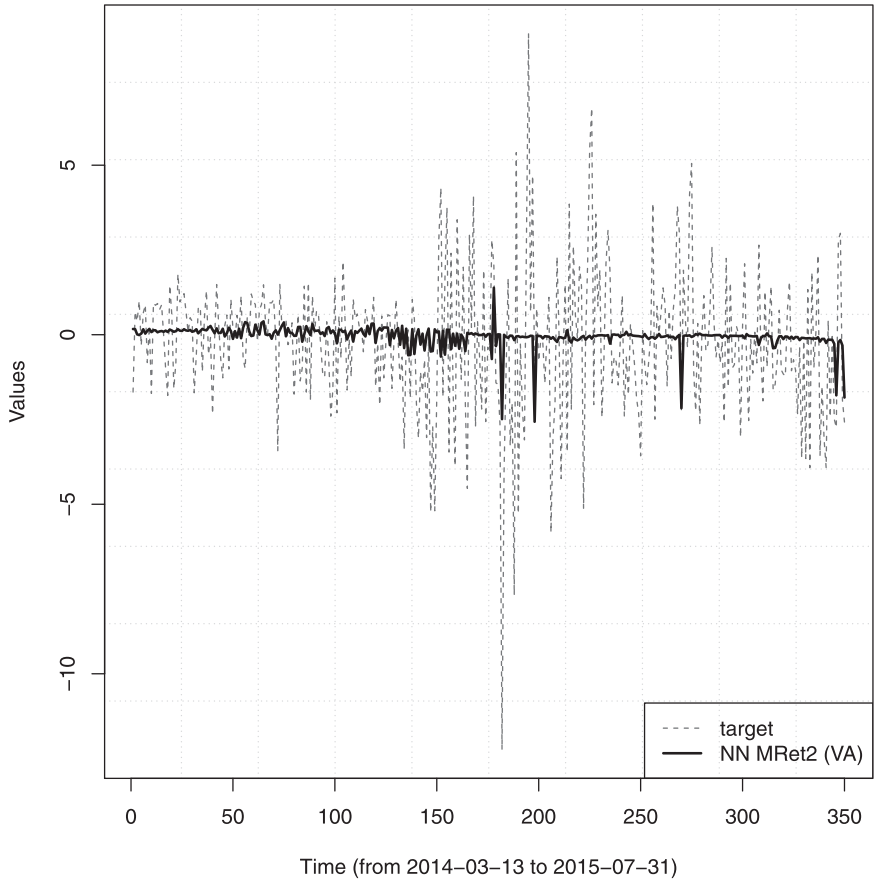

Fig. 3. Predicted and real values for Enrgy.

Table 6
Predictive results for returns of portfolios formed on industries. Utilization of sectorial sentiment indicators (BullR, BI and VA approaches) and first difference of the posting volume. NMAE values of baseline model, lowest NMAE model and models producing statistical significant results in DM test (* – p-value < 10%, ** – p-value < 5%, *** – p-value < 1%, sentiment indicators in parenthesis, NMAE in %, NMAE values lower than baseline are in **bold**, 350 predictions).

Index	Baseline	Lowest NMAE	Statistical significant results
Durbl (returns range: 5.71)	MR: 13.90	EA MRet2 (BI): **13.70**	
Enrgy (returns range: 21.12)	SVM: 7.18	SVM MRet3 (BR): **7.10***	SVM MRet3 (BR): **7.10***
		SVM MRet3 (BI): **7.10****	SVM MRet4 (BR): **7.11***
		SVM MRet5 (BI): **7.10****	SVM MRet3 (BI): **7.10****
			SVM MRet5 (BI): **7.10****
HiTec (returns range: 5.90)	MR: 12.79	EA MRet2 (BR): **12.55****	SVM MRet2 (BR): **12.57****
	EA: 12.79		SVM MRet4 (BR): **12.58***
			EA MRet2 (BR): **12.55****
			SVM MRet6: **12.56****
			SVM MRet2 (BI): **12.61***
			EA MRet2 (BI): **12.59***
			SVM MRet4 (BI): **12.62***
			SVM MRet4 (VA): **12.56****
Hlth (returns range: 7.10)	NN: 12.94	NN MRet1: **12.94**	
Manuf (returns range: 5.17)	SVM: 13.66	SVM MRet5 (VA): **13.60**	
NoDur (returns range: 4.55)	SVM: 12.77	MR MRet2 (VA): **12.72**	
Shops (returns range: 4.76)	SVM: 13.85	SVM MRet5 (BI): **13.63***	SVM MRet5 (BR): **13.64***
			SVM MRet6: **13.67***
			SVM MRet5 (BI): **13.63***
			SVM MRet4 (VA): **13.65***
Telcm (returns range: 5.52)	SVM: 14.15	SVM MRet4 (BR): **13.92***	SVM MRet4 (BR): **13.92***
			SVM MRet5 (BR): **13.95***
			SVM MRet5 (BI): **13.93****
Utils (returns range: 5.77)	SVM: 11.18	SVM MRet6: **11.14**	

The application of sectorial sentiment indicators may allow a better analysis of the effect of sentiment in each sector. We had to exclude the "Other" sector from this analysis because there was no information about its SIC codes, so we were unable to create its sectorial indicator. Additionally, there are no tweets for Durbl sector for 37 days, so we excluded those days from the analysis. Table 6 present the results for these models.

Sectorial indicators are particularly useful for Shops, HiTec and Telcm industries. There are more models significantly more accurate than baseline using sectorial indicators than applying general indicators in these sectors. The SVM model using previous values of returns, posting volume and BI sentiment (SVM MRet5 (BI)) obtains a p-value less than 5% for the prediction of Telcm. Moreover, the lowest NMAE values for Telcm and Shops are obtained by models utilizing sectorial indicators.

In summary, microblogging sentiment and attention indicators were particularly informative for the forecasting of returns of SP500, portfolios of smaller market capitalization (Lo20 and Lo30) and some sectors such as HiTec, Enrgy and Telcm. In the prediction of the returns of the mentioned stocks, there are diverse microblogging models obtaining p-value less than 5% in the pairwise DM test. Many of these models apply both sentiment and posting volume indicators. These results may suggest that sentiment and attention have more impact on future returns of stocks of inferior capitalization and technology related companies. The impact of sentiment on these type of stocks may be explained by a higher concentration of irrational investors. Small stocks are considered to be mostly held by individual investors and less attractive to rational investors (Baker & Wurgler, 2006; Baker et al., 2012). Additionally, a considerable part of technology companies are young and have a small track record. So, they are less appealing to professional investors that prefer easier to value stocks (Baker & Wurgler, 2006). The unpredictability of irrational traders adds risk on prices and makes arbitrage strategies more difficult to implement (Long, Shleifer, Summers, & Waldmann, 1990). Thus, stock prices may differ from fundamental values because arbitrage may be insufficient to instantly eliminate mispricing generated by investor sentiment. In these situations, there is margin to predict future prices and sentiment indicators can be informative.

The utilization of KF indicators was important in some situations. For instance, models applying KF indicators have p-value less than 5% for the prediction of SP500, Lo20, HiTec and Enrgy. Moreover, the usage of KF sentiment indicators decreased the NMAE values for DJIA, MOM, RMRF, SP500, Qnt4, HiTec and Utils. The sectorial indicators were useful for the prediction of returns of Shops and Telcm industries.

Regarding the tested methods, SVM was clearly the most effective. Therefore, the relationship between sentiment, attention and returns may be nonlinear.

3.2. Prediction of volatility

In this paper, we forecasted VIX and the annualized realized volatility of SP500, RSL, DJIA and NDQ. Table 7 outlines the results produced by models applying posting volume and general sentiment indicators computed by BullR, BI, VA, AG and KF approaches.

The utilization of KF sentiment values and previous volatility values significantly improves the forecasting of DJIA realized volatility compared to the AR(5) model. However, the lowest p-value of DM test is not inferior to 5%. AG indicators were applied in the model that obtained the lowest NMAE value for DJIA and SP500. However, this model is not significantly more accurate than baseline. The inclusion of the number of tweets do not seems to benefit the forecasting of volatility. There are no models utilizing posting volume that significantly outperform the respective baseline models. Also, the most accurate models using posting volume features do not improve the results of models without posting volume information. MR, EA and SVM are the most effective methods.

3.3. Prediction of trading volume

In this paper, we predicted the trading volume of SP500 and DJIA. Table 8 digests the results for models using general sentiment indicators, KF indicators and first difference of the number of tweets.

Table 7
Predictive results for VIX and annualized realized volatility of SP500, RSL, DJIA and NDQ. Utilization of general sentiment indicators (BullR, BI, VA and AG approaches), KF indicators and first difference of the posting volume. NMAE values of baseline model, lowest NMAE model and models producing statistical significant results in DM test (* – p-value < 10%, ** – p-value < 5%, *** – p-value < 1%, sentiment indicators in parenthesis, NMAE in %, NMAE values lower than baseline are in **bold**).

Index	Baseline	Lowest NMAE	Statistical significant results
DJIA (n° predictions: 413; realized volatility range: 92.41)	MR: 2.91	SVM MVlt3 (AG): **2.79**	MR MVlt3 (KF): **2.85***
NDQ (n° predictions: 413; realized volatility range: 57.18)	MR: 3.89	SVM MVlt3 (VA): **3.87**	
RSL (n° predictions: 412; realized volatility range: 39.56)	MR: 5.71	MR MVlt1: 5.71	
SP500 (n° predictions: 413; realized volatility range: 67.90)	EA: 3.34	EA MVlt1: **3.34** EA MVlt3 (AG): **3.34**	
VIX (n° predictions: 413; VIX range: 30.42)	EA: 3.26	SVM MVlt3 (BR): **3.25**	

Table 8
Predictive results for trading volume of SP500 and DJIA. Utilization of general sentiment indicators (BullR, BI, VA and AG approaches), KF indicators and first difference of the posting volume. NMAE values of baseline model, lowest NMAE model and models producing statistical significant results in DM test (* – p-value < 10%, ** – p-value < 5%, *** – p-value < 1%, sentiment indicators in parenthesis, NMAE in %, NMAE values lower than baseline are in **bold**).

Index	Baseline	Lowest NMAE	Statistical significant results
DJIA (n° predictions: 414; volume range: 310804)	SVM: 6.00	SVM MVlt5 (BR): **5.84*** SVM MVlt5 (BI): **5.85***	SVM MVlt5 (BR): **5.84***
SP500 (n° predictions: 413; volume range: 1636036)	SVM: 4.98	SVM MVlt1: 4.98	

The application of microblogging sentiment values and the first difference of the number of tweets significantly improves baseline results for the prediction of DJIA trading volume in two situations: SVM MVlt5 (BR) and SVM MVlt5 (BI). However, the lowest NMAE values for SP500 are obtained by the AR(5) model. These results add some evidence that posting volume and sentiment may have predictive content for forecasting of trading volume (e.g., (Antweiler & Frank, 2004; Oliveira et al., 2013; Sabherwal et al., 2011; Sprenger et al., 2014; Tetlock, 2007; Wysocki, 1998)). The utilization of AG indicators do not result in statistical significant DM tests. Therefore, we do not have evidence that disagreement is associated with trading volume as found in some studies (e.g., (Antweiler & Frank, 2004; Shalen, 1993)). The most accurate models also did not applied KF indicators. The SVM method is clearly the best performing approach for the forecasting of trading volume.

3.4. Prediction of survey sentiment indicators using Twitter and KF sentiment indicators

In this subsection, we apply Twitter and KF indicators to forecast AAII and II indicators. The applied procedure is similar to the applied in the forecasting of stock market variables. However, we reduced the size of the rolling windows to 50 because the weekly periodicity downsized the data set. We tested the prediction of four different survey values: VA, BI, negative and positive values. Table 9 presents the evaluation results of the forecasting of AAII and II indicators using weekly Twitter indicators produced by AA approach (aggregating positive and negative messages of the week). For each prediction, the table also shows the number of predictions and target range ($y_L - y_H$).

The utilization of Twitter indicators produced by AA approach was important for the prediction of VA and negative AAII values. The combined utilization of weekly Twitter indicators and previous AAII values (MSv5) produced lower NMAE results than baseline for VA values. However, the relevance of Twitter indicators is higher for AAII negative values because there are nine models producing lower NMAE results than baseline, four of which use only Twitter indicators. These nine models apply daily Twitter indicators. Furthermore, SVM MSv7 is significantly more accurate than baseline according to the pairwise DM test. In the prediction of II indicators, there is only one model more accurate than AR(5) model. The utilization of daily Twitter indicators (i.e., MSv3) produces lower NMAE values than baseline for the forecasting of II values calculated by VA formula. For demonstration purposes, the prediction of negative values of AAII by the SVM MSv4 model (using only Twitter indicators) is present in the left of Fig. 4.

Table 10 shows the results of the prediction of survey indicators using Twitter indicators created by MA approach (average of the daily indicators). The most accurate models using Twitter indicators computed by MA approach produce slightly lower NMAE values than models applying Twitter indicators calculated by AA approach for the forecasting of AAII BI and VA values. However, there are no models significantly more accurate than baseline for the prediction of AAII negative values. Moreover, models utilizing these Twitter indicators do not outperform AR(5) models for the forecasting of any II value.

Table 11 presents the forecasting of survey values using KF indicators. The usage of KF indicators produced worse results than Twitter indicators in the forecasting of AAII indicators. Yet, KF indicators were more informative for the prediction of II VA values. While there is only one model applying Twitter indicators outperforming the baseline model for II values, there are twelve models using KF values more accurate than the best AR(5) model for II values calculated by VA approach. These II forecasts are presented in the right of Fig. 4.

Overall, Twitter sentiment indicators proved to be useful for the prediction of negative values of AAII but less important for the forecasting of II values. Nevertheless, KF indicators are informative for the prediction of II computed by VA formula. Contrary to the forecasting of stock market variables, SVM is not the dominant regression model and the most accurate AAII and II prediction models are obtained by distinct learning models (e.g., RF is the best

Table 9
Prediction of weekly AAII and II values using Twitter sentiment indicators calculated by AA approach. For each survey sentiment indicator, the baseline model is underlined and NMAE values lower than baseline are in **bold** (* – p-value < 10%, ** – p-value < 5% and *** – p-value < 1%, NMAE values in %).

Mtd	MSv1	MSv2	MSv3	MSv4	MSv5	MSv6	MSv7	Mtd	MSv1	MSv2	MSv3	MSv4	MSv5	MSv6	MSv7
Panel A: Prediction of AAII values calculated by BI formula (93 predictions; range: 58.21)															
MR	14.35	18.10	17.72	19.42	15.52	16.19	16.98	RF	14.61	17.59	17.92	16.95	14.71	15.18	14.69
SVM	14.30	16.92	16.84	16.56	14.41	16.42	14.73	NN	16.11	21.58	21.71	20.61	17.15	17.81	17.39
EA	14.79	17.34	17.41	17.68	14.99	15.77	15.4								
Panel B: Prediction of AAII values calculated by VA formula (92 predictions; range: 25.2)															
MR	18.74	19.36	21.7	22.42	20.12	21.96	24.01	RF	19.11	19.35	19.67	19.74	19.11	19.77	19.43
SVM	18.97	21.24	21.54	19.20	**18.60**	18.93	18.96	NN	19.76	27.84	23.57	22.77	24.14	21.08	23.79
EA	19.11	19.82	20.41	20.21	19.35	20.35	20.66								
Panel C: Prediction of AAII positive values (93 predictions; range: 37.89)															
MR	12.52	19.26	19.43	20.18	13.76	14.52	15.83	RF	13.52	18.50	19.61	18.58	13.03	13.74	13.52
SVM	14.54	18.31	19.49	18.29	14.19	13.55	14.36	NN	13.41	18.63	21.80	21.39	14.50	17.32	19.07
EA	12.96	18.65	19.09	18.62	13.22	13.86	14.19								
Panel D: Prediction of AAII negative values (93 predictions; range: 25.65)															
MR	16.42	18.26	17.17	18.77	17.62	16.56	18.07	RF	16.64	17.86	**16.23**	**15.90**	16.74	**15.57**	**15.44***
SVM	16.77	18.30	**16.33**	**15.88**	17.18	17.09	**16.37**	NN	18.05	22.61	18.34	18.05	17.59	19.81	22.63
EA	16.51	18.04	17.04	16.74	16.63	**15.97**	**16.20**								
Panel A: Prediction of II values calculated by BI formula (92 predictions; range: 55.8)															
MR	5.85	12.56	12.12	12.43	6.35	6.49	7.29	RF	8.60	12.14	11.18	11.06	8.37	7.78	8.13
SVM	6.33	14.64	13.10	13.57	6.80	6.79	7.50	NN	8.43	13.18	11.83	12.50	10.95	7.82	9.19
EA	6.22	12.49	11.38	11.02	7.17	6.58	7.39								
Panel B: Prediction of II values calculated by VA formula (91 predictions; range: 19.4)															
MR	16.20	17.48	17.69	19.00	18.12	17.43	19.17	RF	16.7	17.76	16.44	16.78	17.06	16.23	16.59
SVM	16.20	20.27	17.08	16.82	17.63	17.23	16.58	NN	18.16	22.27	**15.91**	16.95	20.61	17.83	18.97
EA	16.06	17.76	16.2	16.86	16.94	16.43	17.64								
Panel C: Prediction of II positive values (92 predictions; range: 37.9)															
MR	8.31	14.61	12.69	13.31	8.76	8.48	9.11	RF	11.12	14.31	12.74	12.49	10.78	10.24	10.43
SVM	8.76	19.64	13.97	14.04	11.10	8.56	10.39	NN	8.73	16.89	14.11	13.89	12.73	9.67	11.84
EA	8.84	14.71	12.76	12.88	9.17	8.36	8.97								
Panel D: Prediction of II negative values (92 predictions; range: 21.8)															
MR	4.70	11.97	11.93	12.12	5.05	5.53	5.80	RF	6.17	11.31	11.18	10.86	6.23	6.12	6.47
SVM	6.38	13.86	12.26	11.59	9.25	7.00	7.70	NN	6.07	11.42	10.85	11.58	7.39	7.26	8.08
EA	4.87	11.99	10.94	11.66	5.23	5.32	6.04								

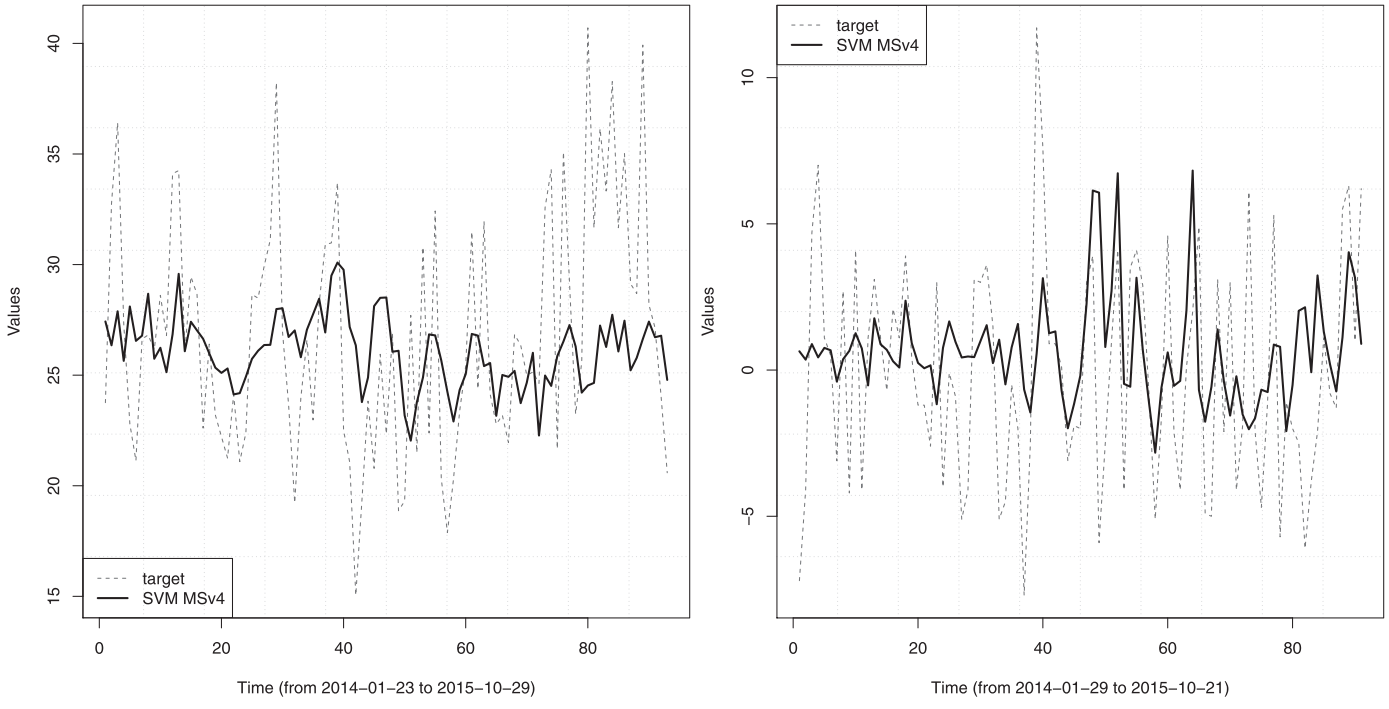

Fig. 4. Predicted results for negative values of AAII (left) and values of II calculated by VA formula using KF indicators (right).

Table 10
Prediction of weekly AAII and II values using Twitter sentiment indicators calculated by MA approach. For each survey sentiment indicator, the baseline model is underlined and NMAE values lower than baseline are in **bold** (* – p-value < 10%, ** – p-value < 5% and *** – p-value < 1%, NMAE values in %).

Mtd	MSv1	MSv2	MSv3	MSv4	MSv5	MSv6	MSv7	Mtd	MSv1	MSv2	MSv3	MSv4	MSv5	MSv6	MSv7
Panel A: Prediction of AAII values calculated by BI formula (93 predictions; range: 58.21)															
MR	14.35	18.81	17.72	19.86	15.93	16.19	17.05	RF	14.61	17.80	17.96	17.71	**14.26**	15.14	14.85
SVM	14.30	16.44	16.84	17.35	14.40	16.42	15.36	NN	16.11	17.35	17.92	22.71	16.23	17.87	17.58
EA	14.79	17.41	17.12	18.27	14.85	16.53	15.22								
Panel B: Prediction of AAII values calculated by VA formula (92 predictions; range: 25.2)															
MR	18.74	19.65	21.70	23.09	19.89	21.96	23.12	RF	19.11	19.30	19.70	19.58	19.33	19.87	19.34
SVM	18.97	18.92	20.96	19.49	19.11	**18.44**	18.97	NN	19.76	20.21	23.10	21.53	25.50	22.37	22.15
EA	19.11	19.24	20.11	20.84	19.8	20.35	20.58								
Panel C: Prediction of AAII positive values (93 predictions; range: 37.89)															
MR	12.52	20.65	19.43	20.98	14.15	14.52	16.19	RF	13.52	19.41	19.59	19.67	13.32	13.71	14.01
SVM	14.54	19.38	18.64	19.39	14.07	14.36	14.60	NN	13.41	21.28	20.71	21.11	16.37	15.34	18.72
EA	12.96	19.82	19.25	19.9	14	13.99	14.69								
Panel D: Prediction of AAII negative values (93 predictions; range: 25.65)															
MR	16.42	18.51	17.17	18.86	18.11	16.56	18.06	RF	16.64	16.97	**16.35**	**16.06**	16.56	**15.45**	**15.60**
SVM	16.77	17.19	**16.33**	**16.30**	17.36	17.07	**16.19**	NN	18.05	18.26	17.80	18.79	17.70	18.70	17.73
EA	16.51	17.33	**16.25**	**16.33**	17.08	**16.21**	**15.53**								
Panel A: Prediction of II values calculated by BI formula (92 predictions; range: 55.8)															
MR	5.85	12.82	12.12	12.65	6.26	6.49	7.31	RF	8.60	11.87	11.16	10.74	8.14	7.66	8.08
SVM	6.33	14.80	13.32	12.27	10.37	7.11	7.82	NN	8.43	14.61	12.51	12.84	7.37	7.19	10.07
EA	6.22	12.44	11.15	10.98	6.99	6.57	7.65								
Panel B: Prediction of II values calculated by VA formula (91 predictions; range: 19.4)															
MR	16.20	18.14	17.69	18.52	17.96	17.43	17.97	RF	16.70	17.73	16.38	16.92	17.00	16.39	16.58
SVM	16.20	17.78	17.37	19.29	16.85	17.23	18.70	NN	18.16	18.90	16.27	19.17	19.43	20.04	19.48
EA	16.06	18.54	16.33	18.11	17.51	16.78	17.28								
Panel C: Prediction of II positive values (92 predictions; range: 37.9)															
MR	8.31	14.44	12.69	13.45	8.84	8.48	9.25	RF	11.12	14.41	12.80	13.14	10.83	10.12	10.77
SVM	8.76	20.15	14.54	14.24	11.44	8.68	10.51	NN	8.73	15.21	14.42	14.83	9.78	9.94	11.73
EA	8.84	14.55	12.69	12.97	9.64	8.55	9.39								
Panel D: Prediction of II negative values (92 predictions; range: 21.8)															
MR	4.70	11.98	11.93	12.05	4.88	5.53	5.82	RF	6.17	11.11	11.06	10.74	6.12	6.07	6.43
SVM	6.38	12.91	12.25	11.47	9.13	5.47	7.79	NN	6.07	11.43	10.90	11.51	6.39	6.97	8.38
EA	4.87	11.87	11.27	11.51	5.5	5.4	6.22								

model for negative AAII predictions; NN obtains the lowest NMAE values for the prediction of II calculated by VA formula).

3.5. Summary of the predictive results

Table 12 summarizes the results obtained in the prediction of stock market variables and survey sentiment indices by indicating the number of models that are significant in the pairwise DM test. For each predicted variable, it shows the number of these models having TWT (*TWT*), KF (*KF*), posting volume (*Nt*) information and the respective methods (*Method*). To facilitate the analysis, the non-significant predictive models are not shown in the table.

Microblogging sentiment (*TWT*) indicators and posting volume information (*Nt*) were especially informative for the forecasting of S&P 500 index, portfolios of lower market capitalization (i.e., Lo20 and Lo30) and some industries (e.g., High Technology, Energy and Telecommunications). There were diverse models that were significant in the DM test for the prediction of these variables. These results support previous findings about the informative content of social media sentiment and posting volume for the forecasting of returns (e.g., (Bollen et al., 2011; Sabherwal et al., 2011; Sprenger et al., 2014)) and the higher impact of sentiment on smaller stocks (e.g., (Baker & Wurgler, 2006; Baker et al., 2012)). We highlight that we predicted short term returns of portfolios formed on size, while most studies forecasted medium to long term returns and used other sentiment sources (e.g., financial data, surveys).

The sectorial indicators seem to benefit the prediction of the returns of some industries. Several models applying these indicators for the prediction of returns from Shops, Telecommunications and High Technology portfolio sectors were significantly better, considering the DM test, when compared against using general sentiment indicators.

The utilization of KF indicators was particularly useful for the prediction of returns of portfolios of small size stocks and the High Technology and Energy sectors. When forecasting returns, there are several models that use KF values, that provide statistically better results, accordingly to the DM test, than the base model (e.g., Lo20 (2 models), Enrgy (4 models), HiTec (3 models)).

Microblogging data was less relevant for the prediction of trading volume and volatility. For these variables there is only evidence of better forecasting ability for the DJIA index.

Regarding the forecasting of survey sentiment values, microblogging sentiment was particularly informative for the prediction of the negative values of AAII. A RF model using Twitter sentiment calculated by AA approach was significant in the DM test for the forecasting of these survey values. Additionally, there are diverse models using Twitter sentiment and KF indicators that produced lower NMAE values than the most accurate AR(5) model for the prediction of the negative values of AAII and II values computed by VA formula. Therefore, we consider that Twitter and KF indicators can be informative to predict AAII and II values and produce an adequate anticipation or a satisfactory alternative of these survey sentiment indices with some advantages (e.g., cheaper, faster, higher periodicities, targeted to specific stocks).

4. Conclusions

In this work, we proposed a robust methodology that allows to assess the usefulness of microblogging data to the prediction of stock market variables. The methodology (shown in Fig. 1 and

Table 11
Prediction of weekly AAII and II values using KF sentiment indicators. For each survey sentiment indicator, the baseline model is underlined and NMAE values lower than baseline are in **bold** (* – p-value < 10%, ** – p-value < 5% and *** – p-value < 1%, NMAE values in %).

Mtd	MSv1	MSv2	MSv3	MSv4	MSv5	MSv6	MSv7	Mtd	MSv1	MSv2	MSv3	MSv4	MSv5	MSv6	MSv7
Panel A: Prediction of AAII values calculated by BI formula (93 predictions; range: 58.21)															
MR	14.35	18.96	17.77	19.1	15.22	15.69	17.73	RF	14.61	17.34	18.73	17.68	14.36	15.84	15.27
SVM	14.3	18.12	16.36	17.69	14.58	14.88	15.32	NN	16.11	20.77	17.11	19.14	17.11	17.38	19.64
EA	14.79	18.29	16.86	16.87	14.66	14.83	15.58								
Panel B: Prediction of AAII values calculated by VA formula (92 predictions; range: 25.2)															
MR	18.74	20.48	20.82	23.76	20.55	20.86	23.55	RF	19.11	19.67	20.75	20.59	18.95	20.2	19.87
SVM	18.97	20.93	21.79	26.9	19.27	21.25	20.07	NN	19.76	21.2	21.12	24.34	23.24	19.61	21.89
EA	19.11	20.13	19.88	21.07	19.44	19.34	20.34								
Panel C: Prediction of AAII positive values (93 predictions; range: 37.89)															
MR	12.52	20.47	19.94	21.87	13.6	14.16	15.94	RF	13.52	19.41	20.35	20.14	12.8	14.06	13.83
SVM	14.54	18.79	21	20.15	13.42	15.43	14.38	NN	13.41	23.29	20.46	25.15	15.23	15.86	15
EA	12.96	19.83	19.44	19.85	13.24	13.68	14.39								
Panel D: Prediction of AAII negative values (93 predictions; range: 25.65)															
MR	16.42	18.63	16.76	17.67	18.11	**16.39**	18.13	RF	16.64	17.34	17.04	16.45	16.55	16.48	**16.33**
SVM	16.77	18.88	**16.13**	16.56	17.85	18.41	16.59	NN	18.05	18.94	16.86	19.99	20.05	20.21	20.17
EA	16.51	18.11	**15.81**	**16.09**	17.75	**15.95**	**15.81**								
Panel A: Prediction of II values calculated by BI formula (92 predictions; range: 55.8)															
MR	5.85	11.74	11.83	11.44	6.12	6.44	6.81	RF	8.6	10.95	10.55	10.13	8.06	7.71	7.88
SVM	6.33	14.05	11.56	10.88	8.61	6.55	7.02	NN	8.43	13.09	11.62	12.41	6.81	7.28	7.93
EA	6.22	11.42	10.56	10.51	6.96	6.79	6.53								
Panel B: Prediction of II values calculated by VA formula (91 predictions; range: 19.4)															
MR	16.2	**15.70**	16.72	17.19	**15.99**	17.44	17.99	RF	16.7	**15.67**	16.77	16.18	**15.75**	16.55	16.27
SVM	16.2	16.36	16.57	**15.43**	16.58	**15.74**	**16.00**	NN	18.16	17.36	19.4	16.46	16.73	17.4	17.87
EA	16.06	**15.33**	16.53	**15.95**	**16.00**	**15.89**	**15.56**								
Panel C: Prediction of II positive values (92 predictions; range: 37.9)															
MR	8.31	14.54	13.39	13.41	8.65	8.71	9.36	RF	11.12	13.85	12.99	12.45	11.2	10.34	10.7
SVM	8.76	16.19	13.39	12.78	9.77	8.96	9.02	NN	8.73	16.13	14.6	13.81	10.08	9.85	11.21
EA	8.84	14.4	12.85	12.37	9	8.35	9.17								
Panel D: Prediction of II negative values (92 predictions; range: 21.8)															
MR	4.7	10.81	10.95	10.58	5.09	5.21	5.45	RF	6.17	10.16	9.59	9.41	6.12	6.03	6.18
SVM	6.38	11.33	10.91	10.47	5.74	7.31	7.99	NN	6.07	13.07	9.22	10.6	5.9	6.12	6.62
EA	4.87	10.09	9.91	10.07	5.31	5.46	5.96								

Table 12
Summary of prediction results (number of models with a p-value< 10% in DM test for each predicted variable with text based indicators).

	Indicators					Indicators			
Pred. Variables	TWT	KF	Nt	Method[a]	Pred. Variables	TWT	KF	Nt	Method[a]
Returns of indices									
DJIA	1	1	0	SVM(2)	MOM	0	1	0	SVM(1)
RMRF	0	1	0	SVM(1)	SMB	1	0	1	SVM(1)
SP500	2	1	4	SVM(5)					
Returns of portfolios formed on size									
Lo30	4	0	4	SVM(4)	Lo20	8	2	8	SVM(11),EA(1)
Returns of portfolios formed on industries (general sentiment)									
Enrgy	6	4	6	SVM(11),NN(1)	Hitec	3	3	3	SVM(6)
Other	2	1	3	SVM(4)	Shops	1	0	1	SVM(1)
Telcm	1	0	0	SVM(1)					
Returns of portfolios formed on industries (sectorial sentiment)									
Enrgy	4	0	2	SVM(4)	Hitec	7	0	4	SVM(6),EA(2)
Shops	3	0	4	SVM(4)	Telcm	3	0	3	SVM(3)
Trading volume									
DJIA	2	0	2	SVM(2)					
Volatility									
DJIA	0	1	0	MR(1)					
Survey Sentiment Indices									
AAII(neg)	1	0	–	RF(1)					

[a] $M_L(n)$, where M_L denotes the ML method (MR, NN, SVM, RF and EA) and n is the number of models using such method.

detailed in Section 2) assumes the usage of sentiment and attention indicators extracted from microblogs and survey indices, diverse forms of a daily aggregation of these indicators, usage of Kalman Filter (KF) to merge microblog and survey data sources, a realistic rolling windows evaluation, several Machine Learning (ML) methods and the Diebold-Mariano (DM) test to check if the sentiment and attention based predictions are useful when compared with an autoregressive baseline. In particular, a very recent and large Twitter dataset was collected and adopted, with around 31 million tweets from December 2012 to October 2015, related with around 3800 stocks traded in US markets. The Twitter sentiment indicators were extracted by considering a recent lexicon specifically adjusted to financial microblogging data (Oliveira et al., 2016). Within our knowledge, this is the first paper that uses sentiment indicators created from specialized financial microblogging lexicons. Moreover, it uses a much larger data period than the majority of studies using Twitter data to predict stock market behavior. Furthermore, we created a novel daily sentiment indicator based on the KF procedure that allows the combination of a daily Twitter indicator with weekly American Association of Individual Investors (AAII) and Investors Intelligence (II) values and monthly University of Michigan Surveys of Consumers (UMSC) and Sentix values. The predictive content of this KF indicator was compared with the content of Twitter indicator. We also explored different sentiment aggregation formulas, such as bullish ratio, variation and agreement. We predicted daily returns, trading volume and volatility of diverse indices, such as Standard & Poor's 500 (SP500), Russell 2000 (RSL), Dow Jones Industrial Average (DJIA) and Nasdaq 100 (NDQ), and portfolios (e.g., formed on size and industries). A fixed-sized rolling window of 300 training days was applied to predict the next day, allowing to perform a large number of model trainings and predictions (ranging from 392 to 439). Also, we explored five different ML methods: Multiple Regression (MR), Neural Network (NN), Support Vector Machine (SVM), Random Forest (RF) and Ensemble Averaging (EA). To analyze the predictive value of microblogging features, we considered the Normalized Mean Absolute Error (NMAE) and applied the Diebold-Mariano (DM) test between the most accurate AR(5) model (baseline) and similar models using microblogging extracted variables. In our opinion, this is a more robustmethodology when compared with most state of the art works.

Additionally, some state of the art studies have analyzed the influence of sentiment on portfolios formed on diverse characteristics (e.g., market capitalization, book-to-market ratio). Many of these works refer that the effect of sentiment is more evident on returns of some portfolios having extreme values (e.g., small market capitalization (Baker & Wurgler, 2006; Baker et al., 2012)). However, most of these studies apply low frequencies (e.g., monthly, annual) and do not use sentiment indicators extracted from social media. Therefore, we analyzed in this paper the predictive content of daily sentiment indicators extracted from Twitter on returns of portfolios formed on size and industries.

Considering that AAII and II are widely used by academics (Fisher & Statman, 2000; Solt & Statman, 1988; Verma & Soydemir, 2009) and practitioners, we also predicted these sentiment measures using Twitter and KF sentiment indicators. To the best of our knowledge, this is the first study that addresses such forecasting.

We found that microblogging sentiment and attention indicators were particularly useful for the prediction of returns of S&P 500 index, portfolios of lower market capitalization and some sectors such as High Technology, Energy and Telecommunications. In these situations, there are models obtaining p-value less than 5% in the pairwise DM test with the baseline model. The application of microblogging features were less convincing for the forecasting of trading volume and volatility. These results add some evidence about the predictive content of social media sentiment and posting volume for returns (e.g., (Bollen et al., 2011; Sabherwal et al., 2011; Sprenger et al., 2014)). Additionally, microblogging sentiment indicators have various advantages when compared to traditional sentiment measures (e.g., surveys). For instance, their creation is faster and cheaper, allows greater frequencies (e.g., daily) and may be targeted to a more limited group of stocks (e.g., individual stocks or indices). The obtained results also corroborate previous findings that sentiment has more impact on smaller stocks (e.g., (Baker & Wurgler, 2006; Baker et al., 2012)). We note that daily microblogging features were used in this paper while the majority of previous studies apply monthly sentiment indicators extracted from economic variables or surveys.

The utilization of sentiment indicators produced by KF were informative for the prediction of returns of some portfolios and indices. There are models using KF indicators having p-value less than 5% in the pairwise test for SP500, Lo20, HiTec and Enrgy. Also, the application of KF indicators decreased NMAE values for diverse indices and portfolios. However, KF indicators were less effective for the forecasting of trading volume and volatility.

Twitter sentiment values were specially informative for negative values of AAII. In this case, there were several models applying Twitter indicators producing lower NMAE results than the most accurate AR(5) model. One of these models is significantly more accurate than baseline according to the pairwise DM test, and uses exclusively Twitter indicators. KF indicators were particularly important for the prediction of II values calculated by VA formula. In this situation, there were twelve models more accurate than the baseline model. These results show that Twitter and KF indicators can be valuable to forecast AAII and II values. Therefore, they may permit a satisfactory anticipation of these sentiment indicators or an acceptable alternative whenever they are unavailable.

There are several finance studies that show that sentiment can explain medium to long term returns (e.g., (Baker & Wurgler, 2006; Baker et al., 2012)), but there is scarce evidence that sentiment can forecast short term returns. Our results are of interest for academics and practitioners alike. This research contributes to the literature by providing evidence that sentiment extracted from microblogging has short term predictive power for some series of financial returns. These results open future research avenues, for instance to test if firms are more prone to being affected by sentiment in times of greater ambiguity and risk (for instance around Initial Public Offerings). Moreover, the results obtained when forecasting sentiment surveys show that sentiment can be used to predict their future values. And they also justify trying to explore the applied methodology in other business areas, namely in marketing, as sentiment can be used as a proxy for consumer satisfaction.

Thus, the proposed methodology could be adopted by financial expert systems to support investors in their decisions by providing instant access to social media analytics, such as customized sentiment indicators or predictions. In the future, we also intend to identify influential microblog users and assess their contribution to the forecasting of specific stocks. For instance, social influence analysis has been applied for advertisement purposes (Gundecha & Liu, 2012) but has been scarcely explored for the stock markets. However, the identification of influent social media users may allow the creation of sentiment indicators of informed users and anticipate the overall sentiment of investors. Additionally, most studies apply only one source of Web data so the complementarity value of different data sources remains unclear. These sources have distinct characteristics that can be complementary and enable better predictions. For instance, blogs have more complete opinionated content, microblogging contents have greater objectivity, interactivity and posting frequencies and Google searches represent a superior number of users. The dynamic combination of diverse Web data sources may result in more informative financial indicators.

Acknowledgments

This work was supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013. We wish to thank the anonymous reviewers for their helpful comments.

References

Al Nasseri, 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.

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. doi:10.1111/j.1540-6261.2004.00662.x.

Areal, N., & Taylor, S. J. (2002). The realized volatility of ftse-100 futures prices. *Journal of Futures Markets, 22*(7), 627–648.

Armstrong, J. S. (2001). Evaluating forecasting methods. In *Principles of forecasting* (pp. 443–472). Springer.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting, 8*(1), 69–80.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance, 61*(4), 1645–1680. doi:10.1111/j.1540-6261.2006.00885.x.

Baker, M., Wurgler, J., & Yuan, Y. (2012). Global, local, and contagious investor sentiment. *Journal of Financial Economics, 104*(2), 272–287. doi:10.1016/j.jfineco.2011.11.002.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8. doi:10.1016/j.jocs.2010.12.007.

Brown, G. W., & Cliff, M. T. (2004). Investor sentiment and the near-term stock market. *Journal of Empirical Finance, 11*(1), 1–27. doi:10.1016/j.jempfin.2002.12.001.

Brown, G. W., & Cliff, M. T. (2005). Investor sentiment and asset valuation.

Chen, R., & Lazer, M. (2013). Sentiment analysis of twitter feeds for the prediction of stock market movement. *Stanford Education, 25*.

Corredor, P., Ferrer, E., & Santamaria, R. (2013). Investor sentiment effect in stock markets: Stock characteristics or country-specific factors? *International Review of Economics & Finance, 27*, 572–591.

Cortez, P. (2010). Data mining with neural networks and support vector machines using the R/rminer Tool. In P. Perner (Ed.), *Advances in data mining – applications and theoretical aspects, 10th industrial conference on data mining* (pp. 572–583). Berlin, Germany: LNAI 6171, Springer.

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences, 225*, 1–17.

Crone, S. F., Hibon, M., & Nikolopoulos, K. (2011). Advances in forecasting with neural networks? empirical evidence from the nn3 competition on time series prediction. *International Journal of Forecasting, 27*(3), 635–660.

Das, S., Martínez-Jerez, A., & Tufano, P. (2005). einformation: A clinical study of investor discussion and sentiment. *Financial Management, 34*(3), 103–137.

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

Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T., & Sakurai, A. (2011). Combining technical analysis with sentiment analysis for stock price prediction. In *Dependable, autonomic and secure computing (DASC), 2011 IEEE ninth international conference on* (pp. 800–807). IEEE.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 20*(1), 134–144.

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM, 57*(6), 74–81.

Fisher, K. L., & Statman, M. (2000). Investor sentiment and stock returns. *Financial Analysts Journal, 56*(2), 16–23. doi:10.2469/dig.v30.n4.771.

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

Goldberg, K., Roeder, T., Gupta, D., & Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval, 4*(2), 133–151.

Groß-Klußmann, A., & Hautsch, N. (2011). When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance, 18*(2), 321–340.

Gundecha, P., & Liu, H. (2012). Mining social media: A brief introduction. *Tutorials in Operations Research, 1*(4), 1–17.

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems, 55*(3), 685–697. doi:10.1016/j.dss.2013.02.006.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). NY, USA: Springer.

Ho, C., & Hung, C.-H. (2009). Investor sentiment as conditioning information in asset pricing. *Journal of Banking & Finance, 33*(5), 892–903. doi:10.1016/j.jbankfin.2008.10.004.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting, 22*(4), 679–688.

Kurov, A. (2010). Investor sentiment and the stock market's reaction to monetary policy. *Journal of Banking & Finance, 34*(1), 139–149. doi:10.1016/j.jbankfin.2009.07.010.

Lee, C., Shleifer, A., & Thaler, R. (1991). Investor sentiment and the closed-end fund puzzle. *The Journal of Finance, 46*(1), 75–109. doi:10.1111/j.1540-6261.1991.tb03746.x.

Lee, W. Y., Jiang, C. X., & Indro, D. C. (2002). Stock market volatility, excess returns, and the role of investor sentiment. *Journal of Banking and Finance, 26*(12), 2277–2299. doi:10.1016/S0378-4266(01)00202-3.

Long, J. B. D., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy, 98*(4), 703.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks.. *Journal of Finance, 66*(1), 35–65. doi:10.1111/j.1540-6261.2010.01625.x.

Mao, H., Counts, S., & Bollen, J. (2011). Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv preprintarXiv:1112.1051.

Neal, R., & Wheatley, S. M. (1998). Do measures of investor sentiment predict returns? *Journal of Financial and Quantitative Analysis, 33*(04), 523–547. doi:10.2307/2331130.

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.

Oh, C., & Sheng, O. R. L. (2011). Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. *ICIS 2011 proceedings*. Shanghai, China: AIS.

Oliveira, N., Cortez, P., & Areal, N. (2013). On the predictability of stock market behavior using stocktwits sentiment and posting volume. In *Progress in artificial intelligence*. In *Lecture notes in computer science: vol. 8154* (pp. 355–365). Springer Berlin Heidelberg. doi:10.1007/978-3-642-40669-0_31.

Oliveira, N., Cortez, P., & Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems, 85*, 62–73.

Oztekin, A., Kizilaslan, R., Freund, S., & Iseri, A. (2016). A data analytic approach to forecasting daily stock returns in an emerging market. *European Journal of Operational Research, 253*(3), 697–710.

Petris, G., et al. (2010). An r package for dynamic linear models. *Journal of Statistical Software, 36*(12), 1–16.

Pinheiro, J. C., & Bates, D. M. (1996). Unconstrained parametrizations for variance–covariance matrices. *Statistics and Computing, 6*(3), 289–296.

Qiu, L., & Welch, I. (2006). Investor sentiment measures. In *Seminar, No. 10794* (pp. 1–39). doi:10.3386/w10794.

Sabherwal, S., Sarkar, S. K., & Zhang, Y. (2011). Do internet stock message boards influence trading? evidence from heavily discussed stocks with no fundamental news. *Journal of Business Finance & Accounting, 38*(9–10), 1209–1237.

Schmeling, M. (2007). Institutional and individual sentiment: Smart money and noise trader risk? *International Journal of Forecasting, 23*(1), 127–145.

Schmeling, M. (2009). Investor sentiment and stock returns: Some international evidence. *Journal of Empirical Finance, 16*(3), 394–408. doi:10.1016/j.jempfin.2009.01.002.

Schumaker, R. P., Zhang, Y., Huang, C.-N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems, 53*(3), 458–464. doi:10.1016/j.dss.2012.03.001.

Shalen, C. T. (1993). Volume, volatility, and the dispersion of beliefs. *Review of Financial Studies, 6*(2), 405–434.

Sheu, H.-J., & Wei, Y.-C. (2011). Effective options trading strategies based on volatility forecasting recruiting investor sentiment. *Expert Systems with Applications, 38*(1), 585–596.

Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2013). Predictive sentiment analysis of tweets: A stock market application. In *Human-computer interaction and knowledge discovery in complex, unstructured, big data* (pp. 77–88). Springer.

Solt, M. E., & Statman, M. (1988). How useful is the sentiment index? *Financial Analysts Journal, 44*(5), 45–55. doi:10.2469/faj.v44.n5.45.

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. doi:10.1111/j.1468-036X.2013.12007.x.

Stambaugh, R. F., Yu, J., & Yuan, Y. (2012). The short of it: Investor sentiment and anomalies. *Journal of Financial Economics, 104*(2), 288–302. doi:10.1016/j.jfineco.2011.12.001.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting, 16*(4), 437–450. doi:10.1016/S0169-2070(00)00065-0.

Tetlock, P. C. (2007). Giving content to investor sentiment : The role of media in the stock market. *Journal of Finance, 62*(3), 1139–1168. doi:10.1111/j.1540-6261.2007.01232.x.

Timmermann, A. (2008). Elusive return predictability. *International Journal of Forecasting, 24*(1), 1–18.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology NAACL 03: 1 (June)* (pp. 173–180). doi:10.3115/1073445.1073478.

Tumarkin, R., & Whitelaw, R. F. (2001). News or noise? internet postings and stock prices. *Financial Analysts Journal, 57*(3), 41–51.

Verma, R., & Soydemir, G. (2009). The impact of individual and institutional investor sentiment on the market price of risk. *The Quarterly Review of Economics and Finance, 49*(3), 1129–1145. doi:10.1016/j.qref.2008.11.001.

Wysocki, P. D. (1998). Cheap talk on the web: The determinants of postings on stock message boards. *University of Michigan Business School Working Paper*, (98025).

Yu, J., & Yuan, Y. (2011). Investor sentiment and the mean-variance relation. *Journal of Financial Economics, 100*(2), 367–381.

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems, 55*(4), 919–926. doi:10.1016/j.dss.2012.12.028.

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*, 55–62.