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