Eurasian Journal of Economics and Finance, 3(4), 2015, 1-7
DOI: 10.15604/ejef.2015.03.04.001

EURASIAN JOURNAL OF ECONOMICS AND FINANCE

http://www.eurasianpublications.com

THE POWER OF MICRO-BLOGGING: HOW TO USE TWITTER FOR PREDICTING THE STOCK MARKET

Francesco Corea
Corresponding Author: LUISS Guido Carli University, Italy. Email: fcorea@luiss.it

Enrico Maria Cervellati
University of Bologna, Italy. Email: enrico.cervellati@unibo.it

Abstract

The availability of new data and techniques enriched the existing extensive literature on the importance of investors' sentiment and on his impact of the stock price oscillations. The purpose of this paper is to exploit micro-blogging data in order to construct a new index-tracking variable that may be used to earn some insights on the Nasdaq-100's future movements. The results are promising: the models augmented with the newly created variable show an incremented explanatory power with respect to the benchmark.

Keywords: Micro-Blogging, Sentiment Analysis, Forecasting, Twitter, Stock Index

1. Introduction and Literature Review

Nowadays, financial markets are increasingly volatile and difficult to understand. In order to reach a greater comprehension of the mechanisms that regulate the markets, it is thus necessary to take into account new and unexpected source of data in the creation of pricing and forecasting models. New technologies and the capacity to gather and store a huge amount of data (i.e., big data) allow us to take the cue from data which were not available ten years ago, but that are now extremely relevant and with strong explanatory power. The most probably innovative kind of data we can exploit is the one coming from the social networks, e.g., Facebook, Twitter, Instagram. They all are able to give us some extra insights and further information in order to build more efficient trading models (Asur and Huberman, 2010). Recently, the use of social networks and micro-blogging platforms is becoming extremely popular. A huge spectrum of distinct applications may be found in completely different fields, e.g., predictions of presidential elections (Tumasjan *et al.* 2010), music albums release (Dhar and Chang, 2009), epidemics and disease spread (Culotta, 2010), movie revenues (Mishne and Glance, 2006) or commercial sales (Choi and Varian, 2012) forecasting. Although the use of these new sources of data may be extremely clear for some of the above-mentioned applications, it might not be so well defined for financial markets. The fundamentals may in fact not be able to explain everything, but we could gain some insights from news and information (Nofsinger, 2005; Peterson, 2007). We should then investigate how we could use the socials to build new efficient forecasting trading models and what kind of information they provide us with. Even though a wide knowledge could be inferred from them, what we focus on is how to extrapolate the investors' sentiment from their opinions on the web.

The literature on how embedding the investors' sentiment into trading, pricing or forecasting models is quite varied: Da *et al.* (2012) propose first a direct measure of investor demand for attention using search frequency in Google for a five-year sample of Russell 3000

stocks, while in a second work (2015) they use daily Internet search volume from millions of households to reveal market-level sentiment and to build a new index as a new measure of investor sentiment. Furthermore, Tetlock (2007), and Tetlock *et al.* (2008) showed how different financial languages in the financial news affect the stock returns. More in details, it seems that i) the impact of the negative news is larger for the stories focused on the fundamentals, ii) negative words in firm-specific news predicts better low firm earnings, and iii) the prices shortly under-react to this information. Fisher and Statman (2000) instead dealt with the investors' sentiment in relation with tactical allocation strategies, while Baker and Wurgler (2006, 2007) studied incorporate to some extent some behavioral biases into the stocks selection process.

Regarding instead the social networks and micro-blogging uses for financial markets, a robust literature is arising in the last few years. Agarwal *et al.* (2011) and Ruiz *et al.* (2012) showed how to use micro-blogging data in the stock market and how they are correlated with financial time series. However, one of the probably most popular works in this field is the one from Bollen *et al.* (2011). Several other works (Bollen and Mao, 2011; Mao *et al.* 2011), and Mittal and Goel in another study (2012), used Twitter to forecast the stock prices for general index such as Dow Jones Industrial Average based on different investors' mood.A very recent paper (Mao *et al.* 2015) analyzed instead the tweets predicting power for international financial markets, obtaining very good results for countries such as United States, United Kingdom, and Canada. Zhang (2013) and Brown (2012) followed the same trend –with slight variations– while Oliveira *et al.* (2013) found a positive effect of the posting volume on robust forecasting. Instead, Sprenger and Welpe (2010) proved how the sentiment of the tweets is indeed associated with abnormal stock returns and message volume. Finally, Oh and Sheng (2011) provided a model for irrational investor sentiment. Nevertheless, there also exists a vast literature on alternative sources of data relevant for financial modelling purposes, such as blogs (De Choudhury *et al.* 2008), security analyst recommendations (Barber *et al.* 2001), web search queries (Bordino *et al.* 2012), stock message boards (Antweiler and Frank, 2004; Koski *et al.* 2008), or simply financial news (Schumaker and Chen, 2009; Lavrenko *et al.* 2000). It then seems clear that it may be valuable to try to integrate this extra information into our forecasting models, and this is indeed the purpose of this paper. The structure of the work will be as follows: in section 2 we present the data and the methodology used, in the section 3 we show some empirical results, and in section 4 we conclude.

2. Data and Methodology

The micro-blogging data are nowadays widely used and available to the research community. They can be obtained through several means, and they usually only report basic information. In our analysis we use Twitter data, for which the basic information that may be exploited are the text of the tweet itself, the username, the hour and location from which it has been posted, and the gender of the user. We decided to obtain these data from the DataSift provider, mainly because in addition to the usual information it also assigns a sentiment score to each tweet. This scoring system is built in such a way that an algorithm can consistently rate the positivity/negativity of a particular text. Within DataSift, this score may typically swing between -20 and +20, even if particular topics/texts require sometimes a higher/lower evaluation. In order to build our estimator for the Nasdaq-100 Index, we collected tweets for a two-month period, for three major technology stocks belonging to the index, i.e., Apple, Google, and Facebook. We decide to select only three out of 100 stocks comprised in the index in order to perform a sort of an exante feature selection: not every stock in the bucket is useful for prediction purposes, but on the principal component analysis design fashion, some stocks have a greater explanatory power with respect to others, and are the only worthy to be used. We therefore skimmed the data to take into account only the English-speaker users, and we took into account only the tweets that showed a pre-existing user's financial knowledge. We only selected the tweets containing the company's ticker, respectively AAPL, GOOG, and FB.1Hence, for the period that

[1] This filter was easily constructed through the CSDL coding environment - a specific language provided by DataSift for this purpose The following codes have been used for filtering the data:

- twitter.symbols contains "AAPL" ;
- twitter.symbols in "GOOG, GOOGL" ;
- twitter.symbols contains "FB".

goes from September 24th to November 21st 2014, we were able to gather about 88,000 tweets for Apple, 43,600 for Facebook, and almost 32,000 for Google, as illustrated in Figure 1.[2]

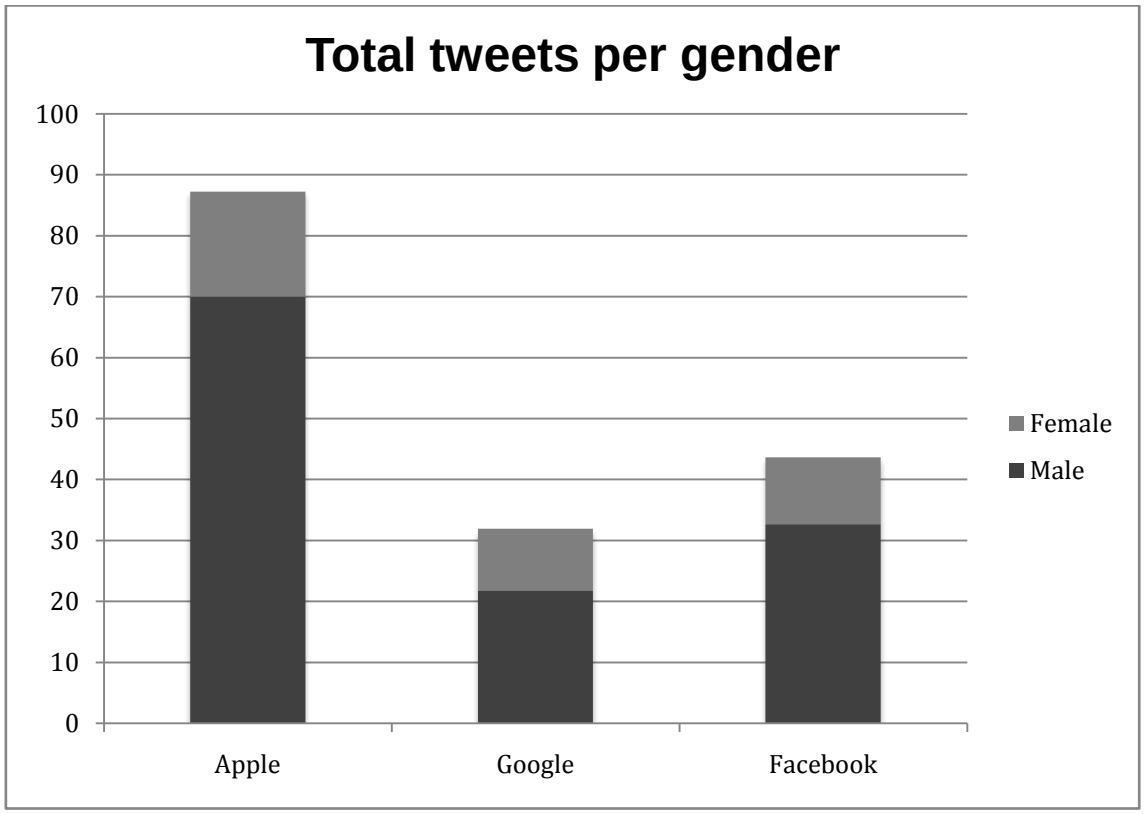

Figure 1. Number of Total Tweets per Gender for Each Stock

We construct two variables: a *sentiment mean* variable for each stock, averaging the sentiment score on a daily basis, and a *tweets volume*, a simple counting variable of the number of tweets for a certain stock on a certain date. Hence, we test the predictive power of individual stocks on the Nasdaq price and volume, with respect to a benchmark autoregressive process:

$$\boldsymbol{M1}: \qquad \boldsymbol{P_t = \alpha + \varphi P_{t-1} + \epsilon_t} \tag{1}$$

where P_t is the Nasdaq price, P_{t-1} is the lagged Nasdaq Price, and ϵ_t is the error.

$$M2: \qquad P_t = \alpha + \varphi P_{t-1} + \beta_1 SM_{Apple} + \beta_2 SM_{Google} + \beta_3 SM_{Facebook} + \epsilon_t \tag{2}$$

where SM_i, for i = Apple, Google, Facebook, represents the sentiment mean related to that specific company.

The first model represents the benchmark, while the second one is augmented for the sentiment mean variable for each stock.
Similarly, we implemented the same models for the volume as well:

$$\boldsymbol{M3}: \qquad \boldsymbol{Volume_t = \alpha + \varphi\, Volume_{t-1} + \epsilon_t} \tag{3}$$

$$M4: \quad Volume_t = \alpha + \varphi\, Volume_{t-1} + \beta_1 TV_{Apple} + \beta_2 TV_{Google} + \beta_3 TV_{Facebook} + \epsilon_t \tag{4}$$

where *Volume* is the Nasdaq transactional volume, and TV_i the tweets volume for a certain

[2] The data for the price and volume of the Nasdaq-100 have been instead obtained through Yahoo!Finance.

stock in a certain day.
We then construct our *sentiment index-tracking* (SIT) variable, as the average of the sentiment mean for each stock in a certain day weighted for the respective tweets volume:

$$SIT_t = \frac{SM_{Apple\ ,t}TV_{Apple\ ,t} + SM_{Google\ ,t}TV_{Google\ ,t} + SM_{Facebook\ ,t}TV_{Facebook\ ,t}}{3}. \quad (5)$$

We therefore augment the autoregressive benchmark models for the SIT variable so constructed (*M5* for the price and *M6* for the volume):

$$\boldsymbol{M5}: \quad \boldsymbol{P_t = \alpha + \varphi_1 P_{t-1} + \varphi_2 SIT_{t-1} + \epsilon_t} \quad (6)$$

$$\boldsymbol{M6}: \quad \boldsymbol{Volume_t = \alpha + \varphi_1 Volume_{t-1} + \varphi_2 SIT_{t-1} + \epsilon_t}. \quad (7)$$

3. Empirical Results

We perform an ordinary least square regression for every model previously exposed. In particular, the estimations obtained are shown in the Table 1.

Table 1. OLS Regressions Results for Model 1-6

Model	M1	M2	M3	M4	M5	M6
Nasdaq price$_{(t-1)}$	0.983***	0.914***			0.949***	
	(20.21)	(14.90)			(19.18)	
Apple SM$_{(t-1)}$		35.86				
		(1.67)				
Google SM$_{(t-1)}$		8.272				
		(0.35)				
Facebook SM$_{(t-1)}$		26.31				
		(0.92)				
SIT$_{(t-1)}$					0.0431**	-251,659.0*
					(2.11)	(-1.95)
Nasdaq volume$_{(t-1)}$			0.704***	0.738***		0.606***
			(6.17)	(6.37)		(5.01)
Apple TV$_{(t-1)}$				59,933.1		
				(1.23)		
Google TV$_{(t-1)}$				-235,036*		
				(-1.71)		
Facebook TV$_{(t-1)}$				35,732.7		
				(0.48)		

Notes: T-statistics in parenthesis, * p<0.1, ** p<0.05, *** p<0.01.

As it can be noticed, the models where the single sentiment means or volume are taken into account (i.e., *M2* and *M4*), are poorly explicative and statistically non-significant. It is thus clear that further specification is needed in order to create an efficient forecasting model able to identify the specific source of price variation. It is instead interesting how the SIT variable well captures some further useful information. Indeed, this variable is statistically significant for both price and volume forecasting (*M5* and *M6*), and above all it is able to improve both the adjusted R^2 and the root means square error (RMSE) of the model, as shown in Table 2.

Table 2. Adjusted R^2 and Root Mean Square Error for All the Models

Model	M1	M2	M3	M4	M5	M6
Adj-R^2	0.9065	0.9102	0.4690	0.4778	0.9137	0.5029
RMSE	43.382	42.51	2.6e+08	2.6e+08	41.672	2.5e+08

The models augmented with the *SIT* variable show an improvement both in term of R^2 and RMSE with respect to their own benchmark.

4. Conclusions

Our results are not fully conclusive, because the model cannot be generalized to any sector, stock or index yet. However, we can infer some important insights on how to incorporate new kinds of information into trading and forecasting model.

We created a dataset for a two-month period using data from Twitter, and we constructed some indicators in order to forecast the Nasdaq-100. We then compared our models to the benchmarks, and it seems that they are able to increase the explanatory power and to provide a better prediction of the Nasdaq price and volume. Both the adjusted R^2 and the RMSE improved as a consequence of the index we built, and even if we may improve the model in the future – we can test it for different sector and for a different time frame, as well as for a different frequency – it gives anyway great intuitions and superior advantage for a potential trading strategies built on it. It would be also interesting to assess how the sentiment might better (or worst) capture the stock market oscillations in crisis or boom periods, and finally to see whether other socials may be actually helpful in term of stock market predictions.

Acknowledgment: Part of this research was performed while the corresponding author was visiting the Institute for Pure and Applied Mathematics (IPAM), which is supported by the National Science Foundation. The author is grateful for the support and help, even if every statement or mistakes are on the author only.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R., 2011. Sentiment analysis of Twitter data. In: M. Nagarajan and M. Gamon, eds. 2011. *LSM '11 Proceedings of the Work-Shop on Languages in Social Media*. Stroudsburg: Association for Computational Linguistics. pp.30-38.

Antweiler, W. and Frank, M.Z., 2004. Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59(3), pp.1259-1294. http://dx.doi.org/10.1111/j.1540-6261.2004.00662.x

Asur, S. and Huberman, B., 2010. Predicting the future with social media. In: IEEE/WIC/ACM International Conferences, *Web Intelligence and Intelligent Agent Technology (WI-IAT).* Toronto, Canada, 31 August-3 September 2010. Toronto: IEEE/WIC/ACM. http://dx.doi.org/10.1109/wi-iat.2010.63

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

Baker, M. and Wurgler, J., 2007. Investor sentiment in the stock market. *Journal of Economics Perspectives*, 21(2), pp.129-151. http://dx.doi.org/10.1257/jep.21.2.129

Barber, B., Lehavy, R., McNichols, M., and Trueman, B., 2001. Can investors profit from the prophets? Security analyst recommendations and stock returns. *The Journal of Finance*, 56(2), pp.531-563. http://dx.doi.org/10.1111/0022-1082.00336

Bollen, J. and Mao, H., 2011. Twitter mood as a stock market predictor. *IEEE Computer*, 44(10), pp.91-94. http://dx.doi.org/10.1109/MC.2011.323

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

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., and Weber, I., 2012. Web search queries can predict stock market volumes. *PloS One*, 7(7), pp.1-17. http://dx.doi.org/10.1371/journal.pone.0040014

Brown, E.D., 2012. Will Twitter make you a better investor? A look at sentiment, user reputation and their effect on the stock market. In: *Proceedings of the Southern Association for Information Systems Conference*. Atlanta: SAIS. pp.36-42.

Choi, H. and Varian, H., 2012. Predicting the present with Google trends. *Economic Record*, Special Issue: Selected Papers from the 40th Australian Conference of Economists, 88(1), pp.2-9. http://dx.doi.org/10.1111/j.1475-4932.2012.00809.x

Culotta, A., 2010. Towards detecting influenza epidemics by analysing Twitter messages. In: P Melville, J. Leskovec, and F. Provost, eds. 2010. *Proceedings of the First Workshop on Social Media Analytics*. New York: ACM. pp.115-122. http://dx.doi.org/10.1145/1964858.1964874

Da, Z., Engelberg, J., and Gao, P., 2012. In search of attention. *The Journal of Finance,* 66(5), pp.1461-1499. http://dx.doi.org/10.1111/j.1540-6261.2011.01679.x

Da, Z., Engelberg, J., and Gao, P., 2015. The sum of all FEARS investor sentiment and asset prices. *Review of Financial Studies*, 28(1), pp.1-32. http://dx.doi.org/10.1093/rfs/hhu072

De Choudhury, M., Sundaram, H., John, A., and Seligmann, D.D., 2008. Can blog communication dynamics be correlated with stock market activity?. In: P. Brusilovsky and H. Davis, eds. 2008. *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*. New York: ACM. pp.55-60. http://dx.doi.org/10.1145/1379092.1379106

Dhar, V. and Chang, E.A., 2009. Does chatter matter? The impact of user-generated content on music sales. *Journal of Interactive Marketing*, 23(4), pp.300-307. http://dx.doi.org/10.1016/j.intmar.2009.07.004

Fisher, K.L. and Statman, M., 2000. Investor sentiment and stock returns. *Financial Analysts Journal*, 56(2), pp.16-23. http://dx.doi.org/10.2469/faj.v56.n2.2340

Koski, J.L., Rice, E.M., and Tarhouni, A., 2008. Day trading and volatility: Evidence from message board postings in 2002 vs. 1999. *Working paper,* under review by Management Science.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Allan, J., 2000. Language models for financial news recommendation. In: A. Agah, J. Callan, E. Rundensteiner, and S. Gauch, eds. 2000. *Proceedings of the Ninth International Conference on Information and Knowledge Management*. New York: ACM. pp.389-396. http://dx.doi.org/10.1145/354756.354845

Mao, H., Bollen, J., and Counts, S., 2011. Predicting financial markets: Comparing survey, news, Twitter and search engine data. *arXivpreprint,* arXiv:1112.1051.

Mao, H., Counts, S., and Bollen, J., 2015. Quantifying the effects of online bullishness on international financial markets. *ECB Statistics Paper Series*, No.9.

Mishne, G. and Glance, N., 2006. Predicting movie sales from blogger sentiment. In: AAAI (Association for the Advancement of Artificial Intelligence), *2006 AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*. California, USA, 27-29 March 2006. California: AAAI.

Mittal, A. and Goel, A., 2012. Stock prediction using Twitter sentiment analysis. *Working Paper*, Stanford University CS 229.

Nofsinger, J.R., 2005. Social mood and financial economics. *The Journal of Behavioral Finance*, 6(3), pp.144-160. http://dx.doi.org/10.1207/s15427579jpfm0603_4

Oh, C. and Sheng, O.R.L., 2011. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In: D.F. Galletta and T.-P. Liang, eds. 2011. *Proceedings from the 32nd International Conference on Information Systems (ICIS)*. Shanghai: Association for Information Systems.

Oliveira, N., Cortez, P., and Areal, N., 2013. On the predictability of stock market behaviour using stock twits sentiment and posting volume. *Progress in Artificial Intelligence, Lecture Notes in Computer Science*, 8154, pp.355-365. http://dx.doi.org/10.1007/978-3-642-40669-0_31

Peterson, R.L., 2007. Affect and financial decision-making: How neuroscience can inform market participants. *The Journal of Behavioural Finance*, 8(2), pp.70-78. http://dx.doi.org/10.1080/15427560701377448

Ruiz, E.J., Hristidis, V., Castillo, C., and Gionis, A., 2012. Correlating financial time series with micro-blogging activity. In: E. Adar, J. Teevan, E. Agichtein, and Y. Maarek, eds. 2012. *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. New York: ACM. pp.513-522. http://dx.doi.org/10.1145/2124295.2124358

Schumaker, R.P. and Chen, H., 2009. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2), pp.1-19. http://dx.doi.org/10.1145/1462198.1462204

Sprenger, T.O. and Welpe, I.M., 2010. Tweets and trades: The information content of stock microblogs. *Social Science Research Network Working Paper Series*, pp.1-89.

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

Tetlock, P.C., Saar-Tsechansky, M., and Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance*, 63, pp.1437-1467. http://dx.doi.org/10.1111/j.1540-6261.2008.01362.x

Tumasjan, A., Sprenger, T.O., Sandner, P.G., and Welpe, I.M., 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10, pp.267-288.

Zhang, L., 2013. *Sentiment analysis on Twitter with stock price and significant keyword correlation*. Honors Theses, The University of Texas.