Journal of Applied Economics and Business

OPEN ACCESS ARTICLE

Stock Movement Prediction Based on Social Media Sentiment Analysis

Aleksandar Dodevski, Natasa Koceska, Saso Koceski

Faculty of Computer Science, University Goce Delchev, Stip, Macedonia

aleksandar.dodevski@hotmail.com, {natasa.koceska, saso.koceski}@ugd.edu.mk

Abstract

Stock prediction is attracting a lot of attention, mainly due the financial gains that may be obtained. However, this is very difficult task and it is very challenging problem that intrigues multiple scientific disciplines such as finance, computer sciences as well as engineering and mathematics. There are many approaches and theories regarding stock movement prediction. In the era of social network and big data, as well as huge speed of news spreading, probably the most interesting are those theories based on social media sentiment analysis. This approach is taking as input non quantifiable data such as financial news articles, social networks posts, or tweeter messages, about a company and predicting its future stock trend with news sentiment classification. Sentiment classification is performed using artificial intelligence algorithms. Main aim of this paper is to give a comprehensive review of current state of the art related to stock movement prediction based on social media sentiment analysis with an emphasis on a Twitter platform.

Keywords

Stock prediction; Social media; Tweeter; Sentiment analysis, Artificial intelligence.

INTRODUCTION

The stock market usually refers to the collection of markets and exchanges where the issuing and trading of equities or stocks of publicly held companies, bonds, and other classes of securities take place. This trade is either through formal exchanges or over-the-counter (OTC) marketplaces.

Also known as the equity market, the stock market is one of the most vital components of a free-market economy. It provides companies with access to capital in exchange for giving investors a slice of ownership.

The stock market is a playground where investors can make profit or loss. During time investors have formed various strategies with the intention to beat the market. The outcomes and final results of these strategies differ widely, yet overall profits should correspond to the outcome of zero-sum game, i.e. that the total sum of winnings and losses of various market players is always equal to zero. There is no way to create a perfect deterministic mathematical

model that will help in finding the best time to buy or sell the stocks, because there are too many independent variables and other numerous factors influencing stock prices, that simply can't be taken into account.

Stock prediction has remained a very difficult task and it is still very challenging problem that intrigues multiple scientific disciplines such as finance, computer sciences as well as engineering and mathematics. Due to its financial gain, it has attracted much attention both from academic and business community. They analyze the market movements and plan their trading strategies on top of it. Considering the fact that the stock market delivers huge amount of data on a daily basis, it is becoming almost impossible task for an individual to consider all the current and past information in order to forecast future behavior of a stock. In the literature, two different approaches for market trends prediction, are published up to date. The first one is named "technical analysis" and other one "fundamental analysis". In order to predict the future trends technical analysis takes into account past price and volume data. In contrary, fundamental analysis is performing detailed analysis of the business itself including its financial data in order to get some insights. There is a theory and hypothesis that argues about the results obtained by both types of analysis, and this hypothesis is called efficient-market hypothesis (EMH). Following this hypothesis, securities markets are extremely efficient in reflecting information about individual stocks and about the stock market as a whole (Marwala and Hurwitz, 2017). The main idea upon which this hypothesis is based, starts from the fact that when information arises, the news travels quickly and it influencing the prices of stocks and securities almost without any delay. Therefore, neither technical analysis, nor even fundamental analysis, would enable an investor to achieve returns greater than those that could be obtained by holding a randomly selected portfolio of individual stocks, at least not with comparable risk (Marwala and Hurwitz, 2017). The EMH is following the so called "random walk" patterns, which as a word is used in the economy and finance sector to describe economic data series where all subsequent data variations represent random departures from previous. The logic of the random walk idea is that if the flow of information is unimpeded and information is immediately reflected in stock prices, then tomorrow's price change will reflect only tomorrow's news and will be independent of the price changes today (Marwala and Hurwitz, 2017). But, news as a concept and their pathways of spreading is by definition are unpredictable, and, therefore, consequent price changes should also be unpredictable and random. Due to this, market prices should completely reflect all known information, and even uninformed investors buying a diversified portfolio at the tableau of prices given by the market will obtain a rate of return as generous as that achieved by the experts (Marwala and Hurwitz, 2017).

However, some research has successfully challenged the efficient market hypothesis by investigating stock markets in some countries, e.g. (Butler and Malaikah, 1992) and attempted to extract patterns in the way stock markets behave and respond to external stimuli.

In the year 2008, the last and the most violent stock market crash took place in United States. On September 20, 2008, the bank bailout bill (Emergency Economic Stabilization Act of 2008) to Congress. The DJIA fell 777.68 in intra-day trading. That was the largest point drop in any single day in history. The month started with the news of Lehman Brothers being declared bankrupt, which took place on September 15, 2008. The Dow dropped 504.48 points.

The Dow hit its pre-recession high on October 9, 2007, closing at 14164.43. Less than 18 months later, it had dropped more than 50 percent to 6594.44 on March 5, 2009 (Temin, 2010).

There had been a lot of speculation and predictions about this great recession by the economists. Apart from the economists, people can now discuss about these issues over the internet and share ideas and feelings. Behavioral economics suggests that the decision-making process of individual mostly depends on his/her emotional or sentimental state.

Information has always been important playing a role building market perception and market investment. In digital era of information and social networks, we can access all sorts of information from any sources easily. Today one of the big information sources is social media that include Twitter.

Twitter, which also features the tag "microblogging site", is such a social networking site in action since 2006. 100 million active Twitter users update nearly 500 million tweets every day. Users express their opinion, decisions, feeling etc. through these tweets. As time passed, Twitter is being used as a tool to share important information rather than casual tweets. Twitter as a social media has a huge implication with rapid and sensitive reaction to political, social, and economic issues. Sharing tweeter messages is involving huge number of people that are tweeting stock related messages that contain information about stocks market and news about updates. As a consequence, Twitter as one of popular social media, can be a valuable source of information in predicting stock market trends based on people's sentiment.

Main aim of this paper is to give a comprehensive review of current state of the art related to stock movement prediction based on social media sentiment analysis with an emphasis on a Twitter platform.

SOURCES OF DATA

As stated by (Blume, Easley and O'hara, 1994) technical analysts believe that future price movements can be predicted by analyzing price and volume data. (Pring, 1991) details the use and the role of technical analysis in his work. However, the accuracy of this technique is discussed controversially by economists and is considered to be less successful by (Gidófalvi and Elkan, 2001).

Financial news articles are considered as the major source of market information for investors. (Cecchini et. al., 2010) found that financial news can have the power to predict financial events (i.e. bankruptcy and fraud). Thus, many researchers have tried to figure out if one can transfer this predictive power to the stock markets. (Wüthrich et. al., 1998) managed to achieve an accuracy of 45% on predicting the direction of the Dow Jones Industrial Average using online

financial news. They created a system that uses a three class prediction (up, down or steady) based on old financial news texts combined with previous stock values as a training set.

(Schumaker and Chen, 2009) achieved a predictive power of 58% on values of the S&P 500 stock index using a similar training set as described by Wüthrich et. al. to train Support Vector Machines. They found that extracting nouns from news texts leads to a more accurate result compared to a bag-of-words representation. (Ammann, Frey and Verhofen, 2014) examined the Handelsblatt, a leading German financial newspaper, and stated that newspaper content in general do have the power to predict future price movements. They used the DAX as a target for their predictions. Although, all these researchers stated that news most certainly have influence to the stock market, we must not underrate the effect of public mood stages and sentiment.

Psychological researchers like (Dolan, 2002) state that emotion biases decision making, particularly in making decisions under risk. Naturally, this also affects investors, traders and other financial decision-makers as pointed out by (Nofsinger, 2005). From his investigations, he concluded that the stock market itself is a measure of social mood, whereupon a high (low) stock market level indicates a high (low) social mood. Since the appearance of social media platforms like Twitter it has never been easier to analyze public mood. Nowadays, almost everybody can share his opinion online and researchers are then able to use this data.

(Bollen and Mao, 2011) were able to bring the mood of people posting on Twitter in relation with the behavior of the stock market. They used different tools to classify about ten million collected Twitter messages by approximately 2.7 million users into six stages of mood, namely calm, alert, kind, vital, sure and happy. Then they used the classified Twitter messages combined with the corresponding stock market values (Dow Jones Industrial Average) as training set for a neural network. Their model achieved an accuracy of 86.7%.

DATA VARIABLES

Dependent variables

Stock data have several interesting values to investigate. In most cases, several dependent variables are observed such as stock's closing price or adjusted closing price, which is the closing price amended to include dividends and other distributions that occurred before. Some authors tried to use different market levels, considering either individual companies or sectors (Liu, Mao, Wang & Wei, 2012).

The adjusted price is the former price, deducted of the aforementioned dividends and distributions. Among others, Sandner, Sprenger, Tumasjan and Welpe (2010) also have a look at traded volumes, which are more likely to grow with the number of tweets mentioning the stock. Then, some authors try to predict returns and volatility.

Returns are actually natural logarithm returns R of stock prices S(t) over a time interval of one day. This additional operation has several advantages such as normalization of the variations:

$$R_t = \ln(S_{(t)}) - \ln(S_{(t-1)}) \quad (1)$$

Volatility is then calculated on a defined period of time, usually ranging from 10 to 50 days. It takes the standard deviation (sd) of the last n days, pondered by the square root of the total number of days N in the time series. For each data point X_t and the average $\overline{X}$ the standard deviation is computed as follows:

$$sd = \sqrt{\frac{\sum_{t=1}^{n}(X_t - \overline{X})^2}{n-1}} \quad (2)$$

Then naturally, volatility at time t is obtained as below for the whole time series at the disposition.

$$Volatility_t = sd * \sqrt{N} \quad (3)$$

The meaning behind volatility is the uncertainty or risk about the scale of change in a stock's value. A high volatility refers to the likelihood for this stock to change significantly in a short period of time in either direction.

Incidentally, the binary variable representing the movements of the stock are rarely compared with twitter data in articles. Earlier studies strictly focused on establishing a relationship between different time series but prediction models are harder to set for non-binary variables as they can take a very wide range of values. Krauss, Nann and Schoder (2013) address these issues by describing returns on investment based on transactions predicted by several sources. However, they mention several limitations which prove that the application to prediction is still experimental in this field.

Independent variables

If the predominance of Twitter can be explained by its research-friendly attributes, some authors still mention other web sources as their text data (Kraus & Nann, 2013; Bollen, Counts & Mao, 2011). Many authors find it reasonable to consider a delay of one day between twitter and stock data but Sandner, Sprenger, Tumasjan and Welpe (2014) also tried a longer delay of two days.

With concern for a methodological approach, studies involving Twitter data usually begin by omitting classification through opinion or sentiment and only count the number of tweets. Liu, Mao, Wang and Wei (2012) investigated the daily number of tweets related to the S&P500 stocks. The advantage of such choice is the simplification of operations by removing a machine learning or manual annotation step of the different tweets at hand. Message volume is obtained

after taking the natural logarithm of the number of tweets. A slight variant of the message volume is the number of users that tweeted and is used notably by Castillo, Gionis, Hristidis, Jaimes and Ruiz (2012).

Nevertheless, such variables are a bit too simple and are not expected to yield robust results in the case of stock movements prediction. Moreover, they are not suited for instigating a relation with bullish or bearish market as they do not provide any opinionated information. Therefore, authors tend to implement specific public mood state in their correlation analysis. The approach can differ in diverse ways but the basic principle remains the same. The point is to use an indicator which depicts the reality with a certain precision.

The most classical time series in this regard is the overall sentiment of tweets. It consists of point data of two or three classes in which tweets are assigned. For instance, a tweet can be positive, negative or eventually neutral regarding the stock studied. Bollen, Mao and Zeng (2011) even enrich this simple reduction of human mood structure by adding other mood dimensions2. Another article by Bollen, Mao and Pepe (2011) tries with other mood states3. But deriving bullishness or bearishness from overall sentiment might be too simplistic for some authors.

Actually, while positive or negative sentiment is linked with past or present transactions, bullishness and bearishness represent belief about the future (Bar-Haim, Dinur, Feldman, Fresko & Goldstein, 2011). This characteristic is far more elusive to determine because the level of significance of tweets relating to the future can change tremendously but if tweets can correctly be assigned, this method is a more realistic view of the relevant variables. Rao and Srivastava (2012) define bullishness B at time t with this equation which comes from the continuation of a previous work from Antweiler and Frank (2004). Where, $M_t^{Positive}$ is the number of positive tweets on day t and $M_t^{Negative}$ the number of negative tweets.

$$B_t = \ln\left(\frac{1+M_t^{Positive}}{1+M_t^{Negative}}\right) \tag{4}$$

Another interesting indicator, which they consider, is the level of agreement A among positive and negative tweets. It takes the value 1 when the tweets are either all bullish or bearish for a time unit measure t and is computed as follows.

$$A_t = 1 - \sqrt{1 - \frac{M_t^{Negative} - M_t^{Positive}}{M_t^{Negative} + M_t^{Negative}}} \tag{5}$$

Finally, Bar-Haim, Dinur, Feldman, Fresko and Goldstein (2011) also speak of the influence of the user's expertise. They assume that similar information does not have the same weight depending on the emitter's expertise in the financial field.

AI METHODOLOGIES

The growing popularity of the Artificial Intelligence (AI) and its application in various fields (Loshkovska and Koceski, 2015), starting from tourism (Koceski and Petrevska, 2012), through medicine (Trajkovik et al., 2014), (Stojanov and Koceski, 2014), (Koceski and Koceska, 2016), biology (Stojanov et al., 2012), education (Koceski and Koceska, 2013), robotics (Koceski et al., 2012), (Koceski et al., 2014), (Serafimov et al., 2012), (Koceska et al., 2013), and also in economy (Koceski and Koceska, 2014), is mainly due to the apparatus i.e. the models and techniques used to mimic the human reasoning, learn and improve during time.

Although there are highly accurate methods to analyze and extract relevant knowledge from structured data such as tables or databases, the task of extracting useful information from unstructured data like social media data still remains a major challenge (Montoyo et al., 2012). Sentiment analysis, also known as opinion mining, involves the use of natural language processing, text analysis and computational linguistics to identify and extract subjective messages. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to a known topic or the overall contextual polarity of being a positive or negative contexts of a document.

According to (Vohra and Teraiya, 2013) there are three main categories of sentiment classification approaches:

1. Supervised classification, used for predicting the polarity of sentiments based on a training set. Among the machine learning techniques, classifiers based on Naïve Bayes (NB), maximum entropy (ME), and support vector machines (SVM) usually exhibit the best performance [Vohra and Teraiya, 2013, Kim and Hwang, 2014).

2. Lexicon-based approach. This does not need any prior training in order to mine the data. It uses a predefined list of words, where each word is associated with a specific sentiment.

3. Hybrid approach which combines both machine learning and lexicon based approaches. Mudinas et. al. combined the lexicon based and the machine learning based approaches.

Forecasting performance evaluation

To evaluate the performance of their prediction tools, authors choose to compute different metrics. Deng et al. (2013) plainly compare accuracies of prediction of stock movements from their different VAR models. Many authors also opt for a simple accuracy measure (Krauss, Nann & Schoder, 2013; Liu, Mao, Wang & Wei, 2012; Rao & Srivastava, 2012). While other authors use it in combination with the mean absolute percentage error (MAPE) (Bollen, Counts & Mao, 2011):

$$MAPE \frac{\sum_i^n \left| \frac{a_i - \hat{a}_i}{a_i} \right|}{n} \times 100 \quad (6)$$

where $\hat{a}_i$ is the predicted value of the stock and a_i is the actual one.

CONCLUSION

The sector of exploiting microblogging for stock predictions is still quite recent and experimental. As a result, we identified some weaknesses among the various articles treating the subject.

To start, all of the studies reviewed are focused on American stock markets. These stocks are highly volatile in nature and it is easier to crawl relevant tweets for them as they can be affected by many actors across the world. This might not hold true for stocks from another origin and especially since there are far more U.S. users than from the rest of the world. American users are expected to be more active as well. In other words, representation of the public mood might be way less accurate for other geographical areas such as Europe.

Another limitation of some articles is the short time span. The longest period considered is 9 months, but most authors limit their analysis to not more than 3 months of data. However, some results of a Granger causality analysis with individual companies' stocks are indeed better with a longer time span.

This could indicate that results with shorter periods are actually pessimistic. Then, the next limitation is about the selection of the time span. The correlation between stock data and Twitter data is clearer when the stocks are submitted to strong variations. It is then tedious to use predictions when the market is not subject to changes. To use tweets relatively safely as an input for stock movements prediction, the investor must be sure that the stocks will move dramatically enough. This raises the question of significance of such predictions.

Finally, there is no systematic testing of correlation between variables and the accuracy measures can be too simplistic. The subject is very experimental in nature and authors usually just check the accuracy of their models without validating formally their assumptions on data. Plain accuracy can be misleading as a bad model can still provide good results in some cases. As a matter of fact, this can lead to biased predictions. Moreover, a model applied to a particular set might give good results but be inefficient in other conditions as the variables such as public mood or stock movements are derived from very complex phenomena difficult to translate into mathematical models.

REFERENCES

[1] Ammann, Manuel, Roman Frey, and Michael Verhofen. "Do newspaper articles predict aggregate stock returns?." Journal of behavioral finance 15, no. 3 (2014): 195-213.

[2] Blume, Lawrence, David Easley, and Maureen O'hara. "Market statistics and technical analysis: The role of volume." The Journal of Finance 49, no. 1 (1994): 153-181.

[3] Bollen, J., Counts, S. & Mao, H. (2011). Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv preprint arXiv:1112.1051.
[4] Butler, Kirt C., and S. Jamal Malaikah. "Efficiency and inefficiency in thinly traded stock markets: Kuwait and Saudi Arabia." Journal of Banking & Finance 16, no. 1 (1992): 197-210.
[5] Cecchini, Mark, Haldun Aytug, Gary J. Koehler, and Praveen Pathak. "Making words work: Using financial text as a predictor of financial events." Decision Support Systems 50, no. 1 (2010): 164-175.
[6] Deng, X., Li, H., Li, Q., Liu, B., Mukherjee, A. & Si, J. (2013). Exploiting Topic based Twitter Sentiment for Stock Prediction. ACL (2), 2013, 24-29.
[7] Dolan, Raymond J. "Emotion, cognition, and behavior." science 298, no. 5596 (2002): 1191-1194.
[8] Gidofalvi, Gyozo, and Charles Elkan. "Using news articles to predict stock price movements." Department of Computer Science and Engineering, University of California, San Diego (2001).
[9] Kim, Sukjoong, and Byung-Yeon Hwang. "Propensity analysis of political attitude of twitter users by extracting sentiment from timeline." Journal of Korea Multimedia Society 17, no. 1 (2014): 43-51.
[10] Koceska, Natasa, Saso Koceski, Francesco Durante, Pierluigi Beomonte Zobel, and Terenziano Raparelli. "Control architecture of a 10 DOF lower limbs exoskeleton for gait rehabilitation." International Journal of Advanced Robotic Systems 10, no. 1 (2013): 68.
[11] Koceska, Natasa, and Saso Koceski. "Financial-Economic Time Series Modeling and Prediction Techniques–Review." Journal of Applied Economics and Business 2, no. 4 (2014): 28-33.
[12] Koceski, Saso, and Biljana Petrevska. "Empirical evidence of contribution to e-tourism by application of personalized tourism recommendation system." Annals of the Alexandru Ioan Cuza University-Economics 59, no. 1 (2012): 363-374.
[13] Koceski, Saso, and Natasa Koceska. "Evaluation of an assistive telepresence robot for elderly healthcare." Journal of medical systems 40, no. 5 (2016): 121.
[14] Koceski, Saso, and Natasa Koceska. "Challenges of videoconferencing distance education-a student perspective." International Journal of Information, Business and Management 5, no. 2 (2013): 274.
[15] Koceski, Saso, Natasa Koceska, and Ivica Kocev. "Design and evaluation of cell phone pointing interface for robot control." International Journal of Advanced Robotic Systems 9, no. 4 (2012): 135.
[16] Koceski, Saso, Stojanche Panov, Natasa Koceska, Pierluigi Beomonte Zobel, and Francesco Durante. "A novel quad harmony search algorithm for grid-based path finding." International Journal of Advanced Robotic Systems 11, no. 9 (2014): 144.
[17] Krauss, J., Nann, S. & Schoder, D. (2013). Predictive Analytics On Public Data-The Case Of Stock Markets. In ECIS (p. 102).
[18] Liu, B., Mao, Y., Wang, B. & Wei, W. (2012). Correlating S&P 500 stocks with Twitter data. In Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research (pp. 69-72). ACM.

[19] Loshkovska, Suzana, and Saso Koceski, eds. ICT innovations 2015: Emerging technologies for better living. Vol. 399. Springer, 2015.

[20] Marwala, Tshilidzi, and Evan Hurwitz. "Efficient Market Hypothesis." In Artificial Intelligence and Economic Theory: Skynet in the Market, pp. 101-110. Springer, Cham, 2017.

[21] Montoyo, Andrés, Patricio MartíNez-Barco, and Alexandra Balahur. "Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments." (2012): 675-679.

[22] Mudinas, Andrius, Dell Zhang, and Mark Levene. "Combining lexicon and learning based approaches for concept-level sentiment analysis." In Proceedings of the first international workshop on issues of sentiment discovery and opinion mining, p. 5. ACM, 2012.

[23] Nofsinger, John R. "Social mood and financial economics." The Journal of Behavioral Finance 6, no. 3 (2005): 144-160.

[24] Pring Martin, J. "Technical Analysis, Explained 'The Successful Investors' Guide to Spotting Investment Trends and Turning Points." (1991).

[25] Rao, T., & Srivastava, S. (2012). Twitter sentiment analysis: How to hedge your bets in the stock markets. In State of the Art Applications of Social Network Analysis (pp. 227-247). Springer International Publishing.

[26] Schumaker, Robert P., and Hsinchun Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system." ACM Transactions on Information Systems (TOIS) 27, no. 2 (2009): 12.

[27] Serafimov, Kire, Dimitrija Angelkov, Natasa Koceska, and Saso Koceski. "Using mobile-phone accelerometer for gestural control of soccer robots." In Embedded Computing (MECO), 2012 Mediterranean Conference on, Bar, Montenegro, pp. 140-143. 2012.

[28] Stojanov, Done, and Saso Koceski. "Topological MRI prostate segmentation method." In Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on, pp. 219-225. IEEE, 2014.

[29] Stojanov, Done, Aleksandra Mileva, and Sašo Koceski. "A new, space-efficient local pairwise alignment methodology." Advanced Studies in Biology 4, no. 2 (2012): 85-93.

[30] Temin, Peter. "The great recession & the great depression." Daedalus 139, no. 4 (2010): 115-124.

[31] Trajkovik, Vladimir, Elena Vlahu-Gjorgievska, Saso Koceski, and Igor Kulev. "General assisted living system architecture model." In International Conference on Mobile Networks and Management, pp. 329-343. Springer, Cham, 2014.

[32] Wüthrich, Beat, D. Permunetilleke, Steven Leung, W. Lam, Vincent Cho, and J. Zhang. "Daily prediction of major stock indices from textual www data." Hkie transactions 5, no. 3 (1998): 151-156.

[33] Vohra, S. M., and J. B. Teraiya. "A comparative study of sentiment analysis techniques." Journal JIKRCE 2, no. 2 (2013): 313-317.