Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1349–1354
DOI: 10.15439/2015F230
ACSIS, Vol. 5

Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction

Michał Skuza
Email: michalskuza@hotmail.com

Andrzej Romanowski
androm@kis.p.lodz.pl

Lodz University of Technology, Institute of Applied Computer Science, Poland.

***Abstract*— This paper covers design, implementation and evaluation of a system that may be used to predict future stock prices basing on analysis of data from social media services. The authors took advantage of large datasets available from Twitter micro blogging platform and widely available stock market records. Data was collected during three months and processed for further analysis. Machine learning was employed to conduct sentiment classification of data coming from social networks in order to estimate future stock prices. Calculations were performed in distributed environment according to Map Reduce programming model. Evaluation and discussion of results of predictions for different time intervals and input datasets proved efficiency of chosen approach is discussed here.**

***Keywords*:** Sentiment Analysis, Big Data Processing, Social Networks Analysis, Stock Market Prediction.

I. Introduction & Rationale

It is believed that information is the source of power. Recent years have shown not only an explosion of data, but also widespread attempts to analyse it for practical reasons. Computer systems operate on data measured in terabytes or even petabytes and both users and computer systems at rapid pace constantly generate the data. Scientists and computer engineers have created special term "big data" to name this trend. Main features of big data are volume, variety and velocity. Volume stands for large sizes, which cannot be easily processed with traditional database systems and single machines. Velocity means that data is constantly created at a fast rate and variety corresponds to different forms such as text, images and videos.

There are several reasons of a rise of big data. One of them is the increasing number of mobile devices such as smartphones, tablets and computer laptops all connected to the Internet. It allows millions of people to use web applications and services that create massive amounts of logs of activity, which in turn are gathered and processed by companies. Another reason is that computer systems started to be used in many sectors of the economy from governments and local authorities to health care to financial sector. The analyses of information that is a by-product of different business activities by companies can lead to better understanding the needs of their customers and prediction future trends. It was previously reported in several research papers that precise analysis of trends could be used to predict financial markets [1]

Big size of data and the fact it is generally not well structured result in situation that conventional database systems and analysis tools are not efficient enough to handle it. In order to tackle this problem several new techniques ranging from in-memory databases to new computing paradigms were created.

Besides big size, the analysis and interpretation are of main concern and application for big data perspective stakeholders. Analysis of data, also known as data mining, can be performed with different techniques such as machine learning, artificial intelligence and statistics. And again it is important to take into consideration the size of data to be processed that in turn determines if a given existing algorithm or approach is applicable.

A. Big data

There are several definitions what Big data is, one of them is following: "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse." [2] This definition emphasizes key aspects of big data that are volume, velocity and variety [3]. According to IBM reports [4] everyday "2.5 quintillion bytes of data" is created. These figures are increasing each year. This is due to previously described ubiquitous access to the Internet and growing number of devices. Data is created and delivered from various systems operating in real-time. For example social media platforms aggregate constantly information about user activities and interactions e.g. one of most popular social sites Facebook has over 618 million daily active users [5]. Output rate of the system can be also important when nearly real-time analyses are needed. Such an on-the-fly analysis is required in recommendations systems when the user's input affects content provided by web site; a good examples are online retail platforms such as Amazon.com. This aspect requires various ways of storing the data to maximize speed and sometimes using column-

oriented database or one of schema-less systems (NoSQL) can do the job, since big data is rarely well structured
But big data is not only challenging but primarily creates opportunities. They are, among the others: creating transparency, optimization and improving performance, generation of additional profits and nothing else than discovering new ideas, services and products.

B. *Social media*

One of the trends leading to rise of big data is Web 2.0. It is a major shift from static websites to interactive ones with user-generated content (UGC). Popularization of Web 2.0 resulted in many services such as blogging, podcasting, social networking and bookmarking. Users can create and share information within open or closed communities and by that contributes to volumes of big data.
Web 2.0 led to creation of social media that now are means of creating, contributing and exchanging information with others within communities by electronic media. Social media can be also summarized as "built on three key elements: content, communities and Web 2.0" [6]. Each of those elements is a key factor and is necessary for social media. One of the most important factors boosting social media is increasing number of always Internet-connected mobile devices such as smartphones and tablets.
Twitter is a micro blogging platform, which combines features of blogs and social networks services. Twitter was established in 2006 and experienced rapid growth of users in the first years of operations. Currently it has over 500 million registered users and over 200 million active monthly users [7]. Registered users can post and read messages called "tweets"; each up to 140 Unicode characters long – originated from SMS carrier limit. Unregistered users can only view tweets. Users can establish only follow or be-followed relationships. A person who subscribes to other user is referred as "follower" and receives real-time updates from that person. However users do not have to add people who are their followers. Twitter can be accessed from various services such as official Twitter web page, mobile applications from third parties and SMS service. As Twitter is an extremely widespread service, especially in US and as the data structure is compact so it forces users to post short comments authors of this paper believe this is a good source of information in the sense of snapshots of moods and feelings as well as for up-to-date events and current situation commenting. Moreover, Twitter is a common PR communication tool for politicians and other VIPs shaping, or having impact on the culture and society of large communities of people. Therefore Twitter was chosen for experimental data source for this work on predicting stock market.

II. PREDICTING FUTURE STOCK PRICES

A. *Experimental System Design and Implementation*

Main goal of this section is to describe implementation of a system predicting future stock prices basing on opinion detection of messages from Twitter micro blogging platform. Unlike the authors of [12] we chose Apple Inc. – a well known consumer electronics company – a producer of Mac computers, iPod, iPad, iPhone products and provider of related software platforms and online services just to name a few.
System design is presented on Figure 1 and it consists of four components: Retrieving Twitter data, pre-processing and saving to database (1), stock data retrieval (2), model building (3) and predicting future stock prices (4). Each component is described later in this text.

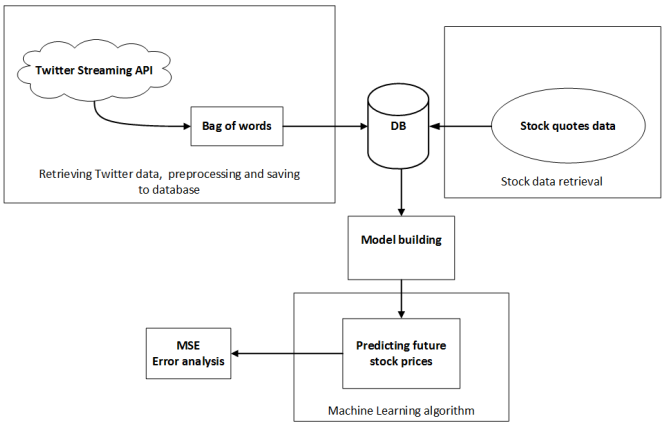

Figure 1: Design of the system

1. Retrieving Twitter data, pre-processing and saving to database.

This component is responsible for retrieving, pre-processing data and preparing training set. There are two labelling methods used for building training set: manual and automatic.

2. Stock data retrieval

Stock data is gathered on a per minute basis. Afterwards it is used for estimating future prices. Estimation is based on classification of tweets (using sentiment analysis) and comparing with actual value by using Mean Squared Error (MSE) measure.

3. Model building.

This component is responsible for training a binary classifiers used for sentiment detection.

4. Predicting future stock prices

This component combines results of sentiment detection of tweets with past intraday stock data to estimate future stock values.

B. *Twitter data acquisition and pre-processing*

Twitter messages are retrieved in real time using Twitter Streaming API. Streaming API allows retrieving tweets in quasi-real time (server delays have to be taken into consideration). There are no strict rate limit restrictions, however only a portion of requested tweets is delivered. Streaming API requires a persistent HTTP connection and authentication. While the connection is kept alive, messages are posted to the client. Streaming API offers possibility of

filtering tweets according to several categories such as location, language, hashtags or words in tweets. One disadvantage of using Streaming API is that it is impossible to retrieve tweets from the past this way.

Tweets were collected over 3 months period from 2nd January 2013 to 31st March 2013. It was specified in the query that tweets have to contain name of the company or hashtag of that name. For example in case of tweets about Facebook Inc. following words were used in query 'Apple', '#Apple', 'AAPL' (stock symbol of the company) and '#AAPL'. Tweets were retrieved mostly for Apple Inc. (traded as 'AAPL') in order to ensure that datasets would be sufficiently large for classifications. Retrieved data contains large amounts of noise and it is not directly suitable for building classification model and then for sentiment detection. In order to clean twitter messages a program in Python programming language was written. During processing data procedure following steps were taken. Language detection information about language of the tweet is not always correct. Only tweets in English are used in this research work. Duplicate removal - Twitter allows to repost messages. Reposted messages are called retweets. From 15% to 35% of posts in datasets were retweets. Reposted messages are redundant for classification and were deleted. After pre-processing each message was saved as bag of words model – a standard technique of simplified information representation used in information retrieval.

C. *Sentiment Analysis*

Unlike classical methods for forecasting macroeconomic quantities [13,15,16] prediction of future stock prices is performed here by combining results of sentiment classification of tweets and stock prices from a past interval. Sentiment analysis [8,14] - also known as opinion mining refers to a process of extracting information about subjectivity from a textual input. In order to achieve this it combines techniques from natural language processing and textual analysis. Capabilities of sentiment mining allow determining whether given textual input is objective or subjective. Polarity mining is a part of sentiment in which input is classified either as positive or negative.

In order to perform a sentiment analysis classification a model has to be constructed by providing training and test datasets. One way of preparing these datasets is to perform automatic sentiment detection of messages. This approach was used in several works such as [9]. Another possibility of creating training and test data is to manually determine sentiment of messages, which means it is a standard, supervised learning approach. Taking into consideration large volumes of data to be classified and the fact they are textual, Naïve Bayes method was chosen due to its fast training process even with large volumes of training data and the fact that is it is incremental. Considered large volumes of data resulted also in decision to apply a map reduce version of Naïve Bayes algorithm. In order to perform sentiment analysis on prepared bags of words a model has to be constructed by providing training and test datasets for classification. These datasets were created using two different methods. One was applying an automatic sentiment detection of messages. It was achieved by employing SentiWordNet [10] which is a publicly available resource aimed to support performing sentiment and opinion classifications. The other method was a manual labelling of sentiment of tweets. Each message was marked as positive, negative or neutral. There were two training datasets. First one consisted of containing of 800 hundred tweets. The other dataset consisted of 2.5 million messages. Only 90% of each dataset was used directly as a training set the other 10% was used for testing.

As a result of two classifiers were obtained using manually labelled dataset. First classifier determines subjectivity of tweets. Then polarity classifier classifies subjective tweets, i.e. using only positive and negative and omitting neutral ones. In order to use classification result for stock prediction term: 'sentiment value' (denoted as a $\boldsymbol{\varepsilon}$) was introduced - it is a logarithm at base 10 of a ratio of positive to negative tweets (Eq. 1).

$$\varepsilon = \log_{10} \frac{number\_of\_positive\_tweets}{number\_of\_negative\_tweets} \quad (1)$$

If $\boldsymbol{\varepsilon}$ is positive then it is expected that a stock price is going to rise. In case of negative $\boldsymbol{\varepsilon}$ it indicates probable price drop. In order to estimate price of stock, classification results are combined with a linear regression of past prices where one weight is a sentiment value. Predictions for a specific time point are based on analysis of tweets and stock prices from a past interval - Eq. 2 shows the formula for the relationships of past value of stock taken into analysis.

$$y_i = \alpha + (\beta + \varepsilon_i) x_i \quad (2)$$

where: y_i is a past value of a stock at a given time of x_i, x_i is time variable, ε_i is a sentiment value calculated for a given time of x_i, $i = 1, \ldots, n$, β is a linear regression coefficient defined as (Eq. 3)

$$\beta = \frac{\Sigma x_i y_i - \frac{1}{n} \Sigma x_i \Sigma y_i}{\Sigma x_i^2 - \left(\frac{1}{n} \Sigma x_i \right)^2} \quad (3)$$

and α coefficient is given by (Eq. 4):

$$\alpha = \bar{p} - \beta \bar{t} \quad (4)$$

where $\bar{p}$ and $\bar{t}$ are mean values of price of stock over a period of t.

III. Results and Analysis

Predictions were prepared using two datasets for several different time intervals, i.e. time differences between the moment of preparing the prediction and the time point for which the forecast was prepared. Predictions were conducted

for one hour, half an hour, 15 minutes and 5 minutes ahead of the moment being forecasted. Two tweet datasets were used: one with messages containing company stock symbol 'AAPL', and the other dataset included only tweets containing name of the company, i.e. 'Apple'. Training datasets consisted of 3 million tweets with stock symbols and 15 million tweets with company name accordingly. Tweets used for predictions were retrieved from 2nd to 12th of April 2013. Approximately 300 000 tweets were downloaded during New York Stock Exchange trading hours each day via Twitter Streaming API.

Experiments were conducted using two models of classifiers, first was built using manually labelled, dataset-based trained classifier and the other was trained with automatically labelled tweet training datasets. Experiments were conducted in the following manner. For each of the following prediction time intervals: 1 hour, 30 min and 15 min all four models (permutations of manual or auto-classifiers coupled with AAPL or Apple keywords) were used. For 5 minutes prediction small number of tweets with stock symbol (AAPL) per time interval resulted in limiting predictions only to models trained with messages with company name (Apple).

Time axis shown on Fig 2 and following figures shows NASDAQ trading hours converted to CEST 1 time zone. Results of predictions are also compared to actual stock prices using Mean Square Error (MSE) measure that are presented in table 1. Sample predictions are presented for 1 hour, 30 minutes, 15 minutes and 5 minutes, while following figures are presented only for 1 hour, 30 min and 5 min.

TABLE 1: MEAN SQUARE ERROR VALUES OF PREDICTED AND ACTUAL STOCK PRICES.

MSE	1 Hour	30 Min	15 Min	5 Min
Manual & 'AAPL'	1.5373	0.6325	0.3425	-
Auto & 'AAPL'	0.947	0.3698	0.2814	-
Manual & 'Apple'	1.9287	1.5152	0.9052	0.5764
Auto & 'Apple'	1.8475	1.4549	0.8325	0.3784

One-hour predictions

A first objective was to test a performance of predictions for 1-hour intervals. Example results of 1-hour interval predictions are presented on Fig. 2.

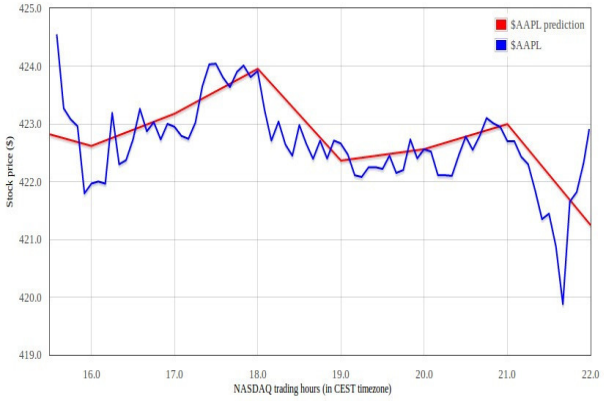

Figure 2: One-hour prediction. Manually labelled 'AAPL' training dataset.

Blue line (fluctuated) corresponds to actual stock prices and red line (steadily changing) shows predicted values. Predictions in this time intervals would not provide accurate result but they can be used to evaluate if the method correctly estimates trends. Predictions for the same day are presented below using 4 different models of classifier.

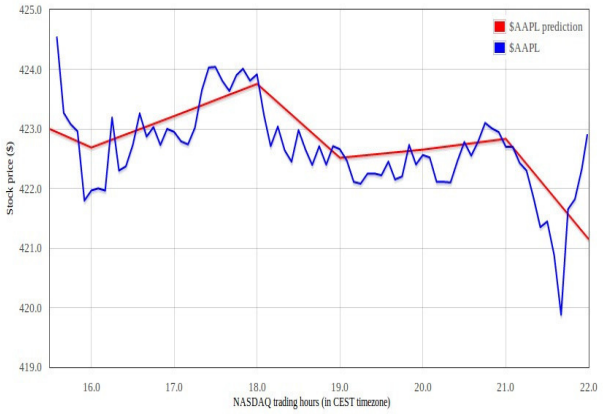

Figure 3: One-hour prediction. Automatically labelled 'AAPL' training dataset.

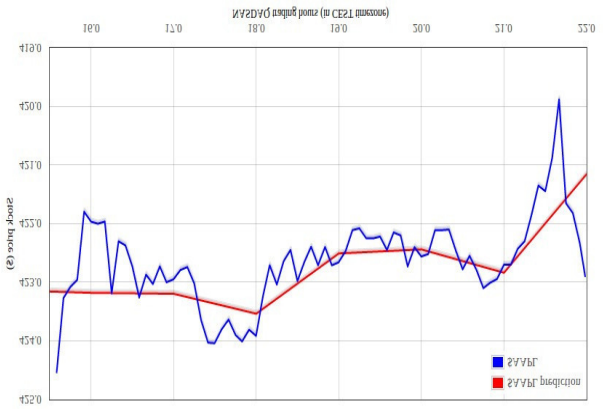

Figure 4: One-hour prediction. Manually labelled 'Apple' training dataset.

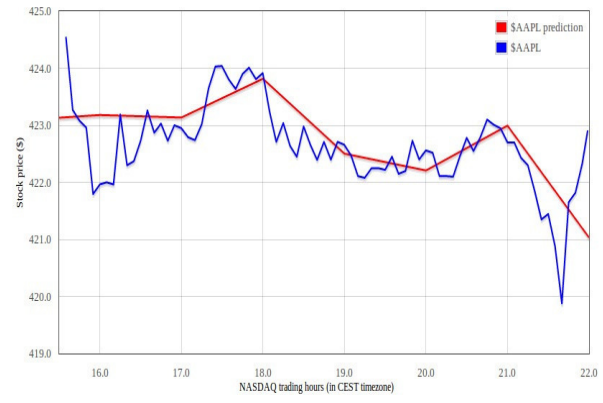

Figure 5: One-hour prediction. Automatically labelled 'Apple' training dataset.

As it can be observed on Fig. 2-5 predictions of each model are similar. They correctly forecast trends however due to big time interval, i.e. 1 hour it is not possible to

determine whether models can predict sudden price movements. Furthermore when comparing results for two datasets, predictions using models trained with tweets with stock symbol perform better that those trained with tweets containing company name.

30 minutes predictions

This subsection describes 30 minutes predictions. It is expected for these predictions not only to forecast trends but sudden price movements as well; this expectation is due to the fact of smaller time interval between time of prediction and forecasted moment. Results are shown on Fig. 6-7 graphs. First model (‘manual’ for AAPL keyword) predictions are less accurate in comparison with second model (auto/AAPL). Yet, significant difference between predicted and actual prices from 17 to 21 in both first models is still there (for further analysis no figures for AAPL is shown from this point on since APPLE gives better performance). Models using dataset with actual company name (plots at Fig. 6 and Fig. 7) perform much better than two first ones. Predictions of prices follow actual ones. However in all cases price forecasting is less accurate when there are several dynamic changes of price movement trend. It is especially visible in all figures for periods from 18 to 20 hours that predictions do not show correlations with actual prices. It may result, among the others, from too long time intervals in comparison to rapid price movements.

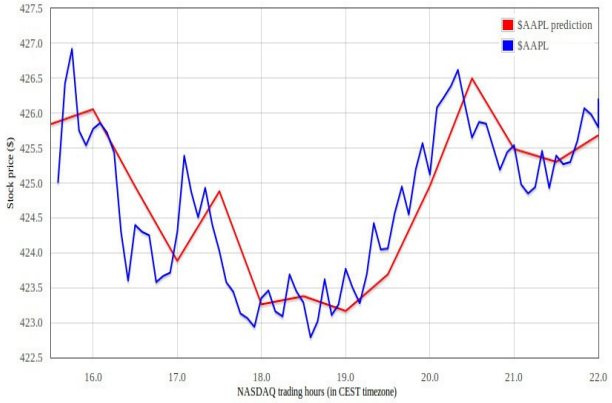

Figure 6: 30 minutes prediction. Manually labelled ’Apple’ training dataset.

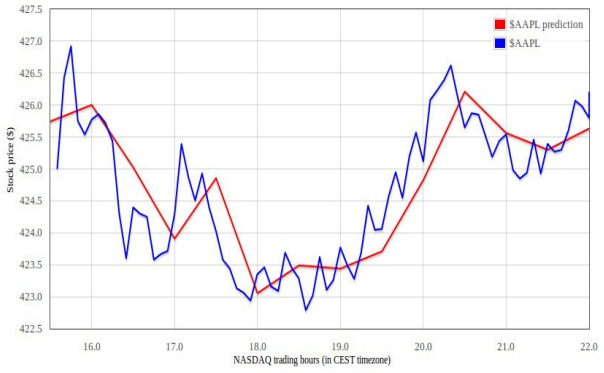

Figure 7: 30 minutes prediction. Automatically labelled ’Apple’ training dataset.

For 15 minutes dynamic price movements are somehow reflected in prediction, although it is not any significant indication in a sense of preserving real nature and amplitude of those fluctuations. Rapid price movements are not easily indicated due to chosen time interval; still too long for better accuracy. Only for this 15 min interval classification based on tweets with stock symbol yield better results. Furthermore it is important to note that using automatically trained training with bigger number of records strongly affects result of prediction.

5 minutes predictions

Last experiment was to perform 5 minutes predictions. Due to short time interval it was expected that prediction would the most accurate. In this part only dataset build with messages with actual company name was used. This is because the number of messages from datasets with stock symbol per time interval was very small and the results of predictions were not reliable. Results are shown on fig. 8-9.

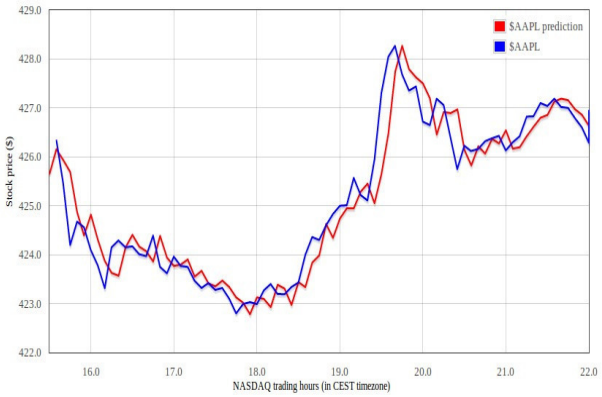

Figure 8: 5 minutes prediction. Manually labelled ’Apple’ training dataset.

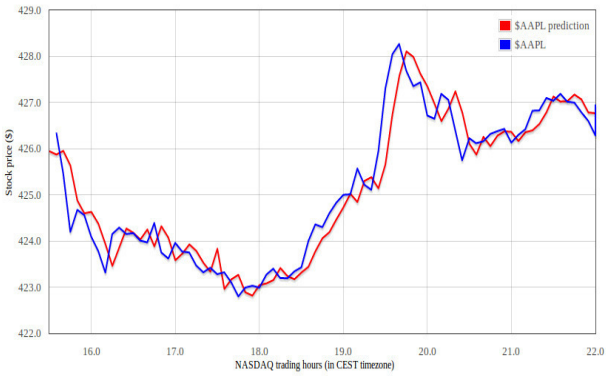

Figure 9: 5 minutes prediction. Automatically labelled ’Apple’ training dataset.

IV. DISCUSSION OF RESULTS

As it can be observed from presented results, predictions of stock prices depend strongly on choice of training dataset, their preparation methods and number of appearing messages per time interval. Predictions conducted

with models trained with datasets with messages containing company stock symbol performs better. It can be explained by the fact that these messages refer to stock market. Tweets with company name may just transfer information, which does not affect financial results. Another important factor is a choice of preparation of training set. Two methods were used. One of the methods was a manual labeling sentiment value of messages. This method allows to more accurately label training data but is not effective for creating large training sets. The other method was applying SentiWordNet, which is a lexical resource for sentiment opinion mining. It enabled to create bigger training datasets, which resulted in building more accurate models. Last factor that is important for prediction is number of appearing messages per time interval. Although model trained with datasets with company name were not accurate in comparison to the other datasets, there is bigger number of tweets per time interval. It allowed performing prediction for shorter time intervals, which were not possible for dataset with messages containing company stock symbol. Described methods can be also used with other stock predictions procedures in order to maintain higher accuracy. It is also important to note that stock prediction methods are not able to predict sudden events called 'black swans' [11].

V. CONCLUSIONS

This paper discusses a possibility of making prediction of stock market basing on classification of data coming Twitter micro blogging platform. Results of prediction, which were presented in previous section show that there is correlations between information in social services and stock market. There are several factors that affect accuracy of stock predictions. First of all choice of datasets is very important. In the paper two types of datasets were used one with name of the company and the other with stock symbol. Predictions were made for Apple Inc. in order to ensure that sufficiently large datasets would be retrieved. There were large differences in size between these two sets. This lead to situation that it was not possible to perform 5 minutes predictions basing on tweets with stock symbol due to too few messages. Additionally although dataset with company name was bigger it may not be accurate for predictions. This due to the fact that company name can be used as a household name and the messages do not refer to stock market. In case of tweets with stock symbol there is bigger probability that people who posted are relating to stock prices. Predictions can be improved by Adding analysis of metadata such as exact location of a person while posting message, number of retweets, number of followers etc. This information may be used to determine which users are more influential and creating a model of interactions between users. Number of messages posted by a user and its frequency may be used to discard spammers and automated Twitter accounts. It is also possible to employ different sources of information. Although Twitter is very popular and offers nearly time communications there exist other sources information such different social networks, blogs, articles in online newspapers. Adding analysis of other may contribute to more accurate predictions.

REFERENCES

[1] Z. Da, J. Engelberg, P. Gao: *In Search of Attention*, The Journal of Finance Volume 66, Issue 5, pages 1461–1499, October 2011, doi: 10.1111/j.1540-6261.2011.01679.x

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A.H. Byers. Big data: The next frontier for innovation, competition, and productivity, McKinsey, May 2011.

[3] Edd Dumbill. What is big data? : an introduction to the big data landscape. http://radar.oreilly.com/2012/01/what-is-big-data.html, 2012.

[4] P. Zikopoulos, C.Eaton, D. DeRoos, T. Deutch and G. Lapis, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media, 2011

[5] M. Zajicek. Web 2.0: hype or happiness? In Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A), W4A '07, pages 35–39, New York, NY, USA, 2007. ACM. doi: 10.1145/1243441.1243453

[6] T. Ahlqvist and Valtion teknillinen tutkimuskeskus. Social media roadmaps: exploring the futures triggered by social media. VTT tiedotteita. VTT, 2008.

[7] Twitter Statistics. http://www.statisticbrain.com/twitter-statistics/, 2013. [Online; accessed 2-January-2013].

[8] B. Pang and L. Lee. Opinion mining and sentiment analysis. Found. Trends Inf. Retr., 2(1-2):1–135, January 2008, doi: 10.1561/1500000011

[9] Y-W Seo, J.A. Giampapa, and K. Sycara. Text classification for intelligent portfolio management. Technical Report CMU-RI-TR 02-14, Robotics Institute, Pittsburgh, PA, May 2002.

[10] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06, pages 417–422, 2006. In In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06, pages 417–422, 2006, doi: 10.1155/2015/715730

[11] N.N. Taleb, Common Errors in the Interpretation of the Ideas of The Black Swan and Associated Papers (October 18, 2009)

[12] M. Paluch, L. Jackowska-Strumillo: The influence of using fractal analysis in hybrid MLP model for short-term forecast of closing prices on Warsaw Stock Exchange. Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 2, pages 111–118 (2014) doi: 10.15439/2014F358

[13] ·M. Marcellino, J. H. Stock, M.W. Watson, A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series, Journal of Econometrics Volume 135, Issues 1–2, November–December 2006, Pages 499–526 doi:10.1016/j.jeconom.2005.07.020

[14] Asur, S., Huberman, B.A., Predicting the Future with Social Media IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010, pp 492 - 499 doi:10.1109/WI-IAT.2010.63

[15] K-J. Kim, I. Han, Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, Expert Systems with Applications Volume 19, Issue 2, 2000, Pages 125–132 doi:10.1016/S0957-4174(00)00027-0

[16] E. J. Ruiz, V. Hristidis, C. Castillo, and A. Gionis, "Correlating Financial Time Series with Micro-Blogging activity," WSDM 2012. Doi: 10.1145/2124295.2124358