Journal of Engineering and Applied Sciences 12 (1): 69-77, 2017
ISSN: 1816-949X

Exploiting Noisy Data Normalization for Stock Market Prediction

[1]Assia Mezhar, [1]Mohammed Ramdani and [2]Amal El Mzabi
[1]Computer Science Laboratory of Mohammedia, Faculty of Sciences and Technologies,
Hassan II University of Casablanca BP 146 Mohammedia,
20650 Mohammedia, Casablanca, Morocco
[2]Team Economic Modeling-Lab PEL, Faculty of Law Economics and Social Sciences,
Hassan II University of Casablanca Mohammedia, Casablanca, Morocco

Abstract: Stock market prediction has grown to be an interesting and intriguing research area in the field of big data analytics, predictive analytics and statistical analysis. The field of stock prediction has employed machine learning and artificial intelligence techniques to forecast the behavior of the financial market and to predict stock prices. Recently, social media has evolved to incorporate a massive amount and variety of textual data. The analysis of this information furthers the mining of public sentiment and opinions about real-time trends. In addition, the study of the inherently complex social media feeds promises new opportunities to discover empirical regularities to measure economic activity and analyze economic behavior at high frequency and in real-time. However, the noisy and short nature of social media feeds mask this information: unlike structured news content, social media content is characterized by the presence of metadata related to social media sites, (e.g., hashtags for Twitter) and the extensive usage of casual language, unstructured grammar, colloquial words, ad hoc multi-token nonstandard lexical items such as acronyms and abbreviations that need situational context to be interpreted and don't fit with traditional technical analysis simply based on forecasting models. Under those purposes and in order to meet the trading challenge in today's global market, technical analysis must be reconsidered. Before using any analysis model, data need to be preprocessed and regularities must be reviewed. So, the precision of the forecasting and prediction systems of the financial market and stock prices will be optimized and improved, also the accuracy of the data analysis models will be higher than state-of-art models. In this context, this study introduces the challenges of the noisy information overload from social media, gives a brief description of stock market prediction and its methodologies. Then, we discuss some of the current methods of stock prediction methodologies and emphasis the need of new improved ones which are more adapted to the context of noisy data. Finally, we present a new approach for the financial market forecasting and prediction which uses data preprocessing and normalization from noisy data in Twitter. The strong influence of the proposed data normalization model on the proposed approach's precision and accuracy can lead to a better results than traditional ones.

Key words: Stock market, prediction, natural language processing, social media, data mining, normalization

INTRODUCTION

Stock market prediction or stock price forecasting is a major and important economic task in the planning of business activity. It has been an attractive and an appealing topic of research in several research fields such as engineering, mathematics, finance and computer science. Stock market prediction is a challenging task because of the noisy and intricate nature of the market. Its dependence to various factors such as unpredictable political events, product releases and the mood of the society based on its behavior and emotions (Hung, 2011; Layyinaturrobaniyah and Sekartadjie, 2016; Amelia, 2016). Finally, due to the complexity of the market dynamics modeling present in different formats either structured numeric data or unstructured textual data which provide both quantitative and qualitative information about the stock price movement: quarterly and annual financial reports, news articles or internet. Therefore, as much as a large set of those factors and dynamics is exploited into stock market analysis, the more its prediction will be improved (Fama, 1965). With the tremendous emergence of social media, information about neutral and public opinions or feelings is abundant and can be incorporated

Corresponding Author: Assia Mezhar, Computer Science Laboratory of Mohammedia, Faculty of Sciences and Technologies, Hassan II University of Casablanca BP 146 Mohammedia, 20650 Mohammedia, Casablanca, Morocco

to clarify the public investment behavior. Furthermore, exploiting the social mood provided by social media can improve the accuracy of the stock market prediction because it might be one of the important influencing factors on its stock price (Zhang *et al*., 2011; Liew and Chou, 2016).

In the recent times, social media has received a great deal of attention from researchers which looked at their exploitation in the daily stock price movement prediction's process. Social media sites especially microblogging ones allow people to share their personal thoughts or express their feelings and opinions about real-daily world events (Cohen, 2005). Twitter for instance is the best corpus of valuable, freely available and rapidly updated data which provide researchers with early warnings about daily rise and fall in stock prices and concise peeks into the public future purchasing behavior of consumers. Having all into consideration, exploiting social media textual data in addition to numeric stock data is very useful to increase the stock market prediction's accuracy and quality.

One major research task for tracking public opinions and sentiments about particular real-time trends from social media is sentiment analysis. On account of its evolving importance, sentiment analysis application in finance has attracted many researchers of computational linguistics (Linoff and Berry, 2011). The valuable Twitter data when equipped with text mining and sentiment analysis will offer a great intelligence of future public expectations and speculations about stock prices movements. Over the past few years, many research studies have approached the incorporation of Twitter with sentiment analysis or data mining techniques such as Ruiz used graphs which constrain time series to correlate the Twitter data activity with the stock price ups and downs. Aramaki *et al.* (2011) exploited Twitter data to catch the flu outbreaks. All these works emphasized the utility of Twitter as a massive source of fresh and valuable data that need to be interpreted in order to predict stock market and to forecast stock price movements.

Although, the studies above have gained great performance, their prediction model accuracy is hampered by the noisy and short nature of data extracted from Twitter: Unlike structured, clean and formal news content, social media users often prefer communicating unconventionally with poor spelling and unstructured grammar such as misspelled words, (e.g., thnk is a variant of thank), abbreviations, (e.g., tmh is an abbreviation of too much hate), slang, (e.g., nuh is a common slang for no) or phonetic attributes of words, (e.g., l8tr for later) that need situational context to be interpreted before being exploited in the stock market analysis. One possible solution to this problem is normalization, in which the informal text is converted into a more standard canonical form which until now, cannot be done by traditional stock market prediction approaches (Clark and Araki, 2011).

Under those purposes, the primary aim of this research paper is to investigate the influence of noisy data preprocessing and normalization on the optimization of stock market prediction models accuracy. It can be significantly improved by the exploitation of noisy data preprocessing enhanced by sentiment analysis from the social media. Our main contribution is that we propose a new approach for the financial stock market forecasting and prediction which uses data normalization and sentiment analysis from noisy data in Twitter. This proposed model exploit the strong influence of data normalization on the stock market prediction accuracy so as it can lead to a better results than traditional stock price movements prediction existing approaches.

Litterature review: Stock market prediction is a very challenging topic of research. Many scientific researchers have addressed the capability of predicting the stock prices ups and downs. Below, we divide the state-of-art methods into two categories: Works that tried to solve the problem by incorporating fundamental and technical factors of the stock market and traditional tools. Second, works that exploit social media as a new tool and which use sentiment analysis techniques for tracking public mood in order to predict the stock prices rise and fall.

Traditional stock market prediction works: In general, the stock market prediction models usually incorporates two types of indicators: technical or fundamental ones. Simple or exponential moving average extracted from structured data are the best historical data researchers exploited as a quantitative measure for the stock price movement forecasting. On the other hand, macroeconomic dynamics extracted from unstructured data are the best fundamental indicators researchers incorporate as a non-historical quantitative measure for stock market prediction. Here after, we will discuss some works that performed stock market prediction with traditional approaches.

Wanjawa and Muchemi (2014) have used artificial neural networks for stock market prediction. It uses an artificial intelligent agent that learns the knowledge from the past events. The intelligent agent learns these past facts without any human supervision. Such artificial intelligence based approaches proved that stock prices forecasting can give performant results. Also, Schumaker

and Chen (2006) have predicted the stock prices by exploiting financial news articles and stock price quotes. The proper nouns were extracted from news articles and combined with stock quotes in order to build a machine-learning algorithm with support vector regression. They proved that financial text mining can give great results but still one limitation found is that they only took into consideration the articles that were published during the time when the stock market was open: this is not realistic because other key information can be extracted before or after the opening of the stock market and can have a strong impact on the stock analysis. In addition, Ayodele *et al.* (2012) and Maizir *et al.* (2016) have used fundamental and technical indicators to predict the stock prices by a hybrid model which combines fundamental and technical analysis. The technical indicators included opening and closing price, high and low price per day and were combined with fundamental indicators which included sell and buy rumors of the company. The joint approach results outperform the technical analysis results. Otherwise, the best known work for stock market prediction is the one conducted by Bollen *et al.* (2011) which proved a very strong correlation between the public mood and the stock price movements when compared with fundamental and technical indicators impact on the stock. The public mood and emotions are present in unstructured sources such as microblogging sites. Therefore, it is essential to extract the information from unstructured sources and perform the analysis to make use of them in the prediction research. Hence, we present works that are based on opinions mining from unstructured data in order to exploit them in the stock price movement analysis.

Social media based works for stock market prediction: In recent years, many works have been oriented towards a new concept: social media. Many techniques have been applied to sentiment analysis for knowledge discovery in different domains including stock market prediction. With the emergence of social media the majority of works are based on Twitter as a major source for data-driven investigation to predict stock prices daily ups and downs. Below, we will discuss some social media based works for stock market prediction.

Bing *et al.* (2014) studied the tweets and concluded the predictability of stock prices based on the type of industry. And, Dickinson and Hu (2015) proved the correlation of sentiments of public with stock rises and falls using Pearson correlation coefficient for stocks. Furthermore, Mao used a random baseline of public sentiment to decide every tweet as "bullish" and "bearish" the stock market. They showed a significant performance in the prediction process. But, Haim *et al.* (2011) who exploited Twitter to identify expert investors to predict stock price ups by using a support vector machine to classify each stock related message to two polarities "bullish" and "bearish", proved that unsupervised approaches for identifying experts outperform random baseline approaches in precision. However, they didn't offer a high performance in term of prediction. In addition, Ding *et al.* (2013) have incorporated time series data as well as sentiments obtained from Twitter data in order to forecast stock price movements by extracting data from Yahoo! Finance. They trained their model with support vector machine which outperforms other training models in term of stock prediction's accuracy. The Twitter data and time series data joint yielded to a performant prediction. The sentiment analysis method they used was based on just the keywords without the analysis of the situational context of the entire tweet. They also proposed that a more sophisticated tool could outperform sentiment analysis with better accuracy. Finally, Si *et al.* (2014) have used data from Twitter to propose a new approach of Semantic Stock Network (SSN) where the network nodes are the companies and the edges are the correlation between them. The proposed stock network shows a significant improvement on sentiment analysis based stock market prediction.

For stock market prediction, recently a lot of Twitter data is gathered in order to be mined to get relevant information relating to the prediction of stock prices and their daily rises and falls. There have been a plethora of research works that try to propose performing tools to improve the mining of social media data and its analysis in order to get accurate results. Those tools when applied to Twitter doesn't perform with a simple list of positive and negative words but also with respect to superfluous, noisy words as well which provides an extra processing for social media analytics in the field of stock market prediction. Besides the fact that there is a lot of improvement in terms of the accuracy of stock prediction. This research field still needs a more accurate results: The presence of noisy words hamper the accuracy of analysis results. Social media text or noisy corpus is characterized by the presence of metadata related to social media sites, (e.g., hashtags for Twitter) and the extensive usage of casual language, unstructured grammar, colloquial words, ad hoc multi-token non-standard lexical items such as acronyms and abbreviations that need situational context to be interpreted (Mezhar *et al.*, 2016) which until now cannot be done by existing approaches. Although the majority of social media based approaches incorporate data normalization in their prediction model, they don't pay a great deal of attention to the context and the area of

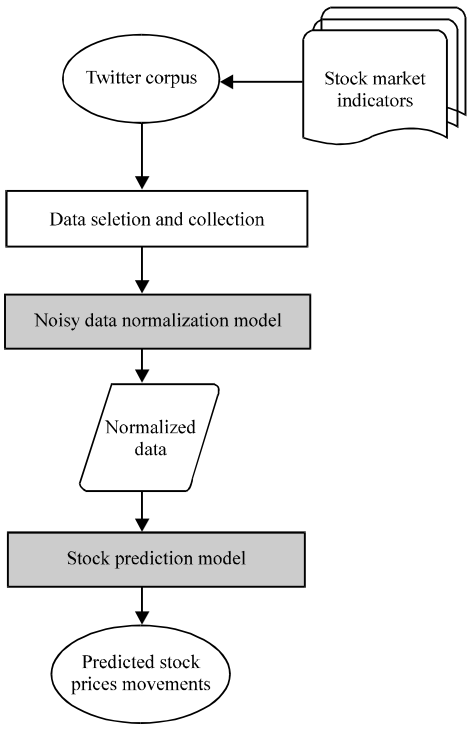

Fig. 1: Impact of data normalization model on stock prediction model

interest of noisy words by making a strong assumption that taking the first canonical form present in the correction dictionaries is the best one or they focus on spelling errors and orthographic noise. Which is not realistic owing to the fact that one noisy word can be restored to different standard forms on account of the context, the type of interest and the time period of the microblogging text and also because social media noisy data that affect predictive studies and analysis go beyond simple orthographic errors or misspelling ones. Hence, a well-studied and accurate data normalization model will have a strong impact on any stock prediction model which will be used after as seen in Fig. 1.

In the light of the above, this research paper aims at making it easier to predict stock by combining it with our context aware and type of interest, time period sensitive data normalization model (Mezhar *et al*., 2016). Also to construct an additional feature set from normalization of noisy data as well as technical, fundamental and sentiment analysis features in order to get a more accurate learning model. The proposed model is expected to lead to a better results than traditional stock price movement's prediction existing approaches.

MATERIALS AND METHODS

Only when normalization systems present accurate and high quality sources of data, prediction ones which exploit them will be able to give deep analysis and precise predictive studies. There by, we propose a novel approach for stock market prediction which exploit noisy data normalization and sentiment analysis. This approach is expected to lead to more accurate results and to be a more effective solution to the stock market prediction problem from noisy data. Below, we define the problem, then, we detail the proposed approach and its key steps Table 1.

Task definition: Our model tracks public emotions about companies and their offered products and services, then predicts their stock price daily rises and falls from a noisy real-time stream of tweets after normalizing it. Table 1 shows the comparative study that emphasizes the difference and added value of our proposed model for stock market prediction from noisy data when compared to other models.

The resulting pipeline of our proposed model is presented in Fig. 2. Given a raw stream of noisy tweets extracted with respect to a company's related keywords or stock indicators, it firstly filters out spam messages, secondly it tokenizes the stream, preprocess it and normalizes it by interpreting meanings with respect to situational context found by weakly learning from ABB, a freely non-standard dictionary (Mezhar *et al*., 2016). Then, we construct three levels of features: Financial features, Sentiment features and normalization features. Here after, all features are integrated together in order to generate weights for each one. Finally, we predict stock prices based on the generated weights. The strength of our model relies on the fact that it exploits a weakly supervised data normalization model based on situational context interpretation to reduce any noise and to extract additional features of sentiment analysis and predictive studies. This new model is expected to give a high accurate results than other state-of-art models because it uses an additional feature based on noisy data normalization. Therefore, it promises an accurate analysis of facts and an outperforming predictive studies of the stock market.

The proposed approach

Data collection: Unlike the traditional data sources such as blogs and forums, social media emerged as a way of communication over the last decade to be the best rapid and real time source of information. Hence, the data is gathered from the best microblogging site Twitter which

Table 1: Comparative study of stock market prediction models and the proposed model

	Key processes of stock market prediction		
Stock market prediction models	Data collection	Data preprocessing	Stock analysis and prediction
Technical analysis oriented models	Time series data	-	Technical analysis that includes support vector machines, artificial neural networks and logistic regression
Data analysis oriented models	News articles social media	Orthographic and misspelling errors correction	Mining the data obtained from textual sources such as social media sites like Twitter and financial news articles and analyze it
Hybrid models	Time series data news articles social media	Orthographic and misspelling errors correction Shortening elongating forms Removal of stop words and stemming	Considering technical, fundamental and sentiments analysis features
Proposed model	Time series data news articles social media noisy data	Orthographic and misspelling errors correction shortening elongating forms removal of stop words and stemming spam detection situational context interpretation and resolving slang and ad hoc abbreviations Named Entities Recognition Areas of interest inference	Considering technical, fundamental and sentiments analysis features Additional feature set constructed from situational context and inferred areas of interest of noisy data

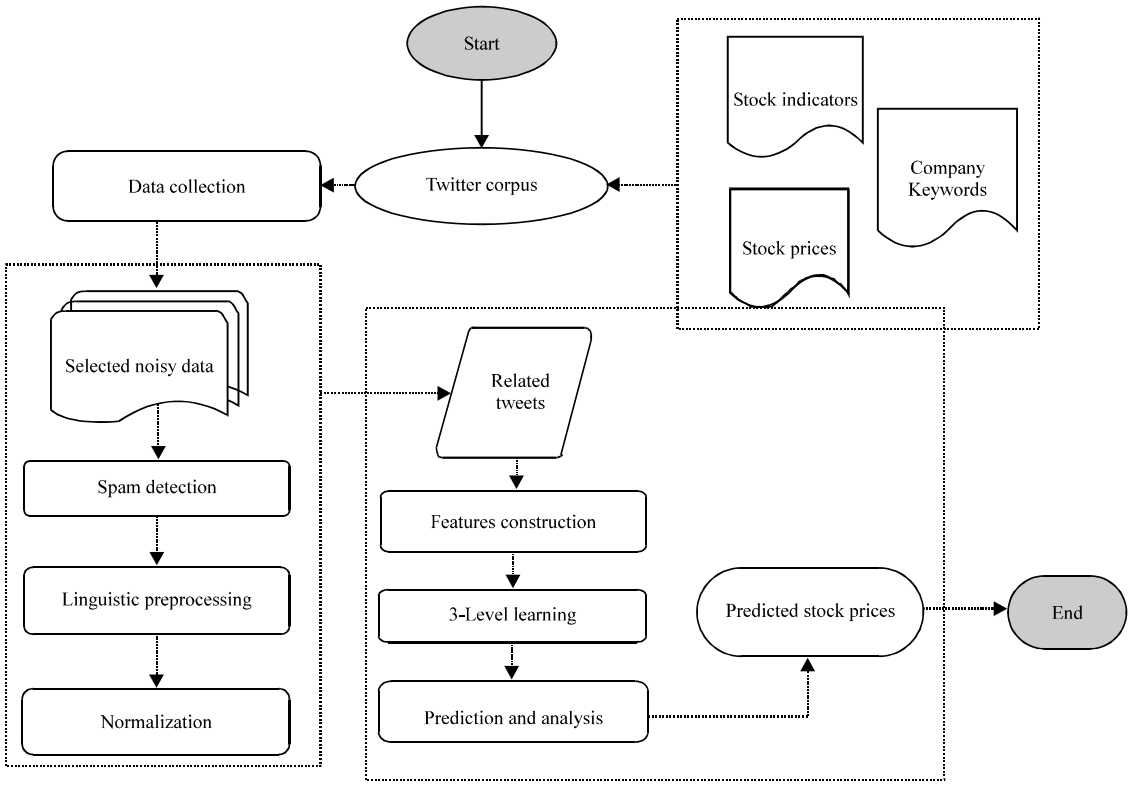

Fig. 2: The proposed stock market prediction model

contains a real-time data stream of short messages or tweets. The tweets are collected using Twitter API and filtered using keywords related to companies and the stock market indicators. We are interested in tracking not only the opinion of people about the company's stock but also their opinions about products and services offered by this company. The analysis of people's opinions about a company's products and services has a strong influence on its stock price movements forecasting. The collected data must be preprocessed and cleaned in order to get only relevant information.

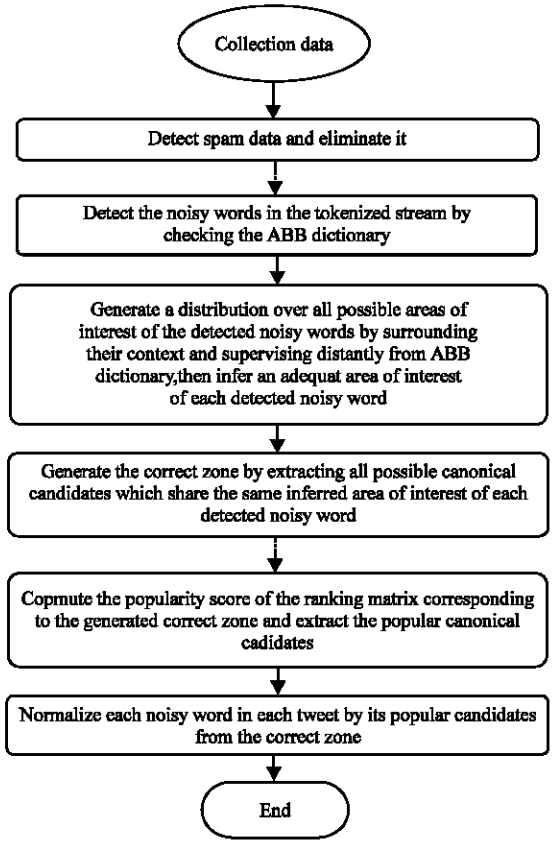

Fig. 3: Data normalization model flowchart

Noisy data normalization model: As a raw data, tweets are highly susceptible to noise. In other word0s, tweets consists of many acronyms, emoticons and mega elements that need situational context to be interpreted or even unnecessary data like pictures and URL's. The quality of raw data affects the sentiment mining results. In order to help improve the quality of the data and consequently of the mining results raw data is preprocessed in order to improve the efficiency and ease of the mining process. So tweets are preprocessed and normalized to represent correct emotions of public depending on the context. For preprocessing of tweets we employed three major processes. Below, we describe each process:

Spam Detection: Spam consist of messages which don't provide any relevant information or useless ones. Following Katsios *et al.* (2015) we exclude all tweets that link to untrusted web sites. Then, we compute the spam score to filter out higher scored spam messages than a learned threshold by:

$$S^* = \frac{|u| + |h| + |l| + |s| + |n|}{|t|}$$

Where:

$|u|$ = The number of user mentions
$|h|$ = The number of hashtags
$|l|$ = The number of web links
$|s|$ = The number of spam words (detected from a predefined list)
$|n|$ = The number of non-words character and is the total number of tokens

Preprocessing of the datastream: After filtering out spam tweets, we apply our linguistic preprocessing component (Mezhar *et al.*, 2016) to eliminate all Twitter's mega elements with a regex pattern tokenizer that breaks the text into tokens and reject all mega elements (hash-tags, user IDs), then we shorten any alphabet that is repeated more than three times to two letters (shoooow is shortened to show). Finally, we eliminate all the spelling errors and unintentional ones by one to one mapping with canonical forms from the english standard dictionary.

Normalization: All tweets outputted by the linguistic preprocessing component are additionally passed through our casual english normalizer (Mezhar *et al.*, 2016) in order to resolve multi-token non-standard items as seen in the flowchart in Fig. 3. We first, detect a tweet's noisy zone which is composed by multi-token non-standard items existing in that tweet as:

$$NZ(T_i) = \left\{ \bigcup_{j=1}^{p} w_j \,/\, w_j \in ABB \right\}$$

Where:

w_j = The jth multi-token non-standard item in the tweet T_i
ABB = A freely online non-standard dictionary

Then, we infer an adequate area of interest for elements in the noisy zone by constraining each detected element over areas of interest based on its set of possible areas of interest in ABB dictionary. Here after, based on the inferred area of interest, we extract each tweet's correct zone: a set of all canonical forms having the same area of interest inferred as:

$$CZ(T_i) = \left\{ \bigcup_{j=1}^{q} c_j \,/\, A_{c_j} = A_{w_j,i} \right\}$$

Where:

C_j = Canonical form
j^{th} = Multi-token non-standard item
A_{cj} = Its area of interest
$A_{wj,i}$ = The inferred area of interest of w_j

Finally, we normalize each element of the noisy zone by its popular correct element from the correct zone by computing the popularity score of the corresponding ranking matrix as:

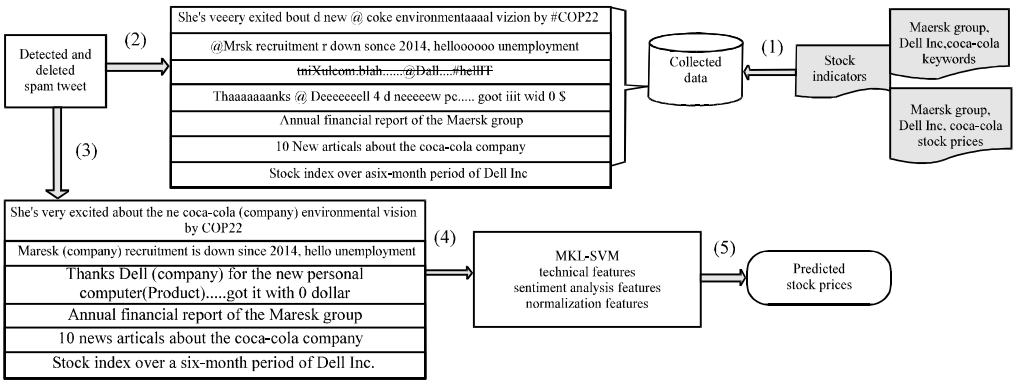

Fig. 4: Predicted stock prices process flow for Maesk group, Dell Inc. and coca-cola

$$P_{c_j}^{*} = \max_{n\in\{1,q\}}\left(\sum_{m=1}^{s} r_{nm}\right) / R = \left(r_{n,m}\right) \in \mathbb{N}^{q\times s}$$

Where:

S = The number of search engines

r_{nm} = The number of times

C_j = Is searched in a known time period

R = The corresponding matrix of ranking (Fig. 4)

Stock market prediction model: After normalizing noisy data and eliminating all non-relevant information, preprocessed data pass through our prediction model by three processes: Technical, fundamental features, sentimental features and normalization features construction, a three-level learning model that generates the optimal combination of the three sets of features and finally, the prediction of stock prices.

Features construction: The adopted features are indicators from technical, fundamental and sentiment analysis as the majority of previous research, the difference is that we combine them with an additional feature constructed from areas of interest and recognized named entities inferred by the normalization model to get more accurate results. The technical and fundamental features are constructed from historic data and sentiment features are constructed from social media as seen in (Devi and Bhaskaran, 2015; Mezhar *et al*., 2016).

Level learning: The three features sets constructed must give a key information to be learned. Here after, the features are fed to the three-level learner in order to be estimated. Following Devi and Bhaskaran (2015), we use a Multiple Kernel Learning Support Vector Machine (MKL-SVM) because it supports features from multiple sources rather than traditional Support Vector Machine (SVM) and its ability to select the optimal kernel and parameters from a larger set of kernels. Hence, the features constructed from technical and fundamental stock indicators extracted from financial documents, the features constructed from sentiment analysis from public tweets and the features constructed from areas of interest of noisy data inferred by the normalization model constitute kernels. The three-level learning component learns an optimal combination of the predefined kernels by:

$$K_{3-L}(a,b) = \sum_{i=1}^{n}\beta_i K_i(a,b) / \beta_i \geq 0, \sum_{i=1}^{n}\beta_i = 1$$

Where:

β_i = The vector of weights of each kernel composed by the three constructed features sets

$K_{3\text{-}L}$ = The estimates the weights for each of the three feature sets to find the optimum combined kernels

Prediction and analysis: After getting the optimized combination of kernels by the three-level learner, we predict the adequate prices of the next trading day and evaluate the three features sets by comparing actual prices and predicted ones.

RESULTS AND DISCUSSION

Our model track public emotions about companies and their offered products and services, then predict their stock price daily rises and falls from a noisy real-time stream of tweets after normalizing it. Figure 4 shows the stock market prediction pipeline for three companies Maersk group, Dell Inc. and coca-cola. Where at we collect the suitable data by extracting tweets from Twitter with respect to keywords related to the three companies and stock indicators from financial documents or news articles. Then, we eliminate spam data. Here after, we preprocess and normalize extracted data. Finally, we construct adequate features and generates predicted

stock prices for each company. We have to note that we are still in the earlier implementation stages of our model Fig. 4.

CONCLUSION

The stock market forecasting and prediction is a very challenging and highly complicated task because it is influenced by many factors: Technical, fundamental, sentimental and normalization factors. The stock prices indicators are generally dynamic and noisy. Although, the majority of existing approaches incorporate data normalization as well as sentiment analysis in their prediction model, they don't pay a great deal of attention to the context and the area of interest of noisy words by making a strong assumption that taking the first canonical form present in the correction dictionaries is the best one or they focus on spelling errors and orthographic noise. Which is not realistic owing to the fact that one noisy word can be restored to different standard forms on account of the context, the type of interest and the time period of the microblogging text and also because social media noisy data that affect predictive studies and sentiment analysis go beyond simple orthographic errors or misspelling ones. The sentiment analysis features as well as technical and fundamental features aren't enough to give an accurate predicted price movements. The accuracy of stock price directional movements can be significantly improved by the incorporation of an additional feature set constructed from normalization from noisy data. Early stages of tests indicate that there is a strong impact of normalization features on stock market price change. Hence, a well-studied and accurate data normalization model will have a strong impact on any stock prediction model. Under those purposes, we proposed to combine the stock prediction model with our context aware and type of interest, time period sensitive data normalization model (Mezhar *et al*., 2016). Second, to add a normalization feature constructed from areas of interest inferred for noisy data to the MKL-SVM learner. The proposed model is expected to lead to a better results than traditional stock price movements' prediction existing approaches.

REFERENCES

Amelia, T.N., 2016. Causality test between exchange rate, inflation rate and stock price index in Southeast Asia. J. Administrative Bus. Stud., 2: 101-106.

Aramaki, E., S. Maskawa and M. Morita, 2011. Twitter catches the flu: Detecting influenza epidemics using twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing, July 27-31, 2011, ACM, Stroudsburg, Pennsylvania, ISBN:978-1-937284-11-4, pp: 1568-1576.

Ayodele, A.A., A.K. Charles, A.O. M arion and O.O. Sunday, 2012. Stock price prediction using neural network with hybridized market indicators. J. Emerg. Trends. Comput. Inform. Sci., 3: 1-9..

Bing, L., K.C. Chan and C. Ou, 2014. Public sentiment analysis in twitter data for prediction of a company's stock price movements. Proceeding of the IEEE 11th International Conference on E-Business Engineering (ICEBE), November 5-7, 2014, IEEE, New York, USA., ISBN:978-1-4799-6563-2, pp: 232-239.

Bollen, J., H. Mao and X. Zeng, 2011. Twitter mood predicts the stock market. J. Comput. Sci., 2: 1-8.

Clark, E. and K. Araki, 2011. Text normalization in social media: Progress, problems and applications for a pre-processing system of casual English. Procedia Soc. Behav. Sci., 27: 2-11.

Cohen, A.M., 2005. Unsupervised gene protein named entity normalization using automatically extracted dictionaries. Proceedings of the Acl-Ismb Workshop on Linking Biological Literature, Ontologies and Databases: Mining biological semantics, June 24-24, 2005, ACM, Stroudsburg, Pennsylvania, pp: 17-24.

Devi, K.N. and V.M. Bhaskaran, 2015. Semantic enhanced social media sentiments for stock market prediction. Int. J. Soc. Behav. Educ. Econ. Manage. Eng., 9: 1-5.

Dickinson, B. and W. Hu, 2015. Sentiment analysis of investor opinions on Twitter. Soc. Networking, 4: 62-71.

Ding, T., V. Fang and D. Zuo, 2013. Stock market prediction based on time series data and market sentiment. Ph.D Thesis, Northwestern University, Evanston, Illinois. https://pdfs.semanticscholar.org/2c91/447c35fe2d4426b6661b8c8c97f439f3172e.pdf.

Fama, E.F., 1965. The behavior of stock-market prices. J. Bus., 38: 34-105.

Haim, B.R., E. Dinur, R. Feldman, M. Fresko and G. Goldstein, 2011. Identifying and following expert investors in stock microblogs. Proceedings of the Conference on Empirical Methods in Natural Language Processing, July 27-31, 2011, ACM, Stroudsburg, Pennsylvania, ISBN:978-1-937284-11-4, pp: 1310-1319.

Hung, L.C., 2011. The presidential election and the stock market in Taiwan. J. Bus. Policy Res., 6: 36-48.

Katsios, G., S. Vakulenko, A. Krithara and G. Paliouras, 2015. Towards Open Domain Event Extraction from Twitter: REVEALing Entity Relations. MODUL University Vienna, Vienna, Austria, pp: 35-46.

Layyinaturrobaniyah, D.M. and G. Sekartadjie, 2016. Fundamental and technical analyses for stock investment decision making. J. Administrative Bus. Stud., 2: 1-7.

Liew, C. and T.N. Chou, 2016. The prediction of stock returns with regression approaches and feature extraction. J. Administrative Bus. Stud., 2: 107-112.

Linoff, G.S. and M.J.A. Berry, 2011. Data Mining Techniques: For Marketing, Sales and Customer Relationship Management. 3rd Edn., John Wiley and Sons, New York, ISBN: 9781118087459, Pages: 888.

Maizir, H., R. Suryanita and H. Jingga, 2016. Estimation of pile bearing capacity of single driven pile in sandy soil using finite element and artificial neural network methods. Int. J. Appl. Phys. Sci., 2: 45-50.

Mezhar, A., M. Ramdani and E.A. Mzabi, 2016. A novel weakly supervised approach for casual english normalization. Proceeding of the 2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA), October 19-20, 2016, Mohammedia, Morocco, ISBN: 978-1-5090-5781-8, pp: 1-6.

Schumaker, R. and H. Chen, 2006. Textual analysis of stock market prediction using financial news articles. Proceedings of the 12th Americas Conference on Information Systems, August 4-6, 2006, AMCIS, Acapulco, Mexico, pp: 1432-1440.

Si, J., A. Mukherjee, B. Liu, S.J. Pan and Q. Li *et al.*, 2014. Exploiting social relations and sentiment for stock prediction. Proceedings of the 2014 International Conference on Empirical Methods in Natural Language Processing (EMNLP), Octobre 25-29, 2014, ACM, Doha, Qatar, pp: 1139-1145.

Wanjawa, B.W. and L. Muchemi, 2014. ANN Model to Predict Stock Prices at Stock Exchange Markets. Cornell University, Ithaca, New York, USA.

Zhang, X., H. Fuehres and P.A. Gloor, 2011. Predicting stock market indicators through twitter i hope it is not as bad as i fear. Procedia Soc. Behav. Sci., 26: 55-62.