Predicting stock movement using public sentiment analysis of twitter feed

Application of machine learning models using sentiment indicators and financial lexicons

I. INTRODUCTION

Most psychological researches concur that emotion or sentiment plays a significant role in human-decision making [1,4,12]. This has been further supported by behavioural finance indicating that monetaristic choices are substantially determined by mood [6]. According to Bollen, Mao & Zeng [2], mining of public sentiment and its analysis through Twitter offers a robust and efficient approach rather than extraction through traditional means like surveys which prove to be comparatively high-priced and time-consuming. They further discussed that variations in the public temperament can definitely be traced from a wide-ranging Twitter feed which in turn corresponds with alterations in the Dow Jones Industrial Average (DJIA) values occurring a few days later.

During their investigation, Mäntylä, Graziotin, & Kuutila [7] discovered that most modern papers related to sentiment analysis (2014-2016) centred around stock market and human emotions amongst a few other topics. Taking this knowledge into consideration, the key objective of this report is to further explore current research done, particularly from the last four years, in the field of sentiment analysis of social media data to predict stock market fluctuations. This data is gathered from the microblogging platform Twitter. The rest of the report is organised as follows. Section 2 presents the search process for procuring the literature. Section 3 critically analyses the literature gathered. Section 4 discusses the methodology operated across the papers. Finally, inferences are drawn in Section 5.

II. SEARCH APPROACH

The Scopus database was used for the purposes of literature exploration and retrieval. This selection was done with consideration to its creator Elsevier, according to whom “Scopus is the largest abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings” [7].

To gain a better understanding of the knowledge space encompassing the subject matter, an audit of academic literature was conducted on the Scopus database. The keywords for the initial search query were “stock”, “twitter” and “sentiment”. The results were then limited to the subject area of computer science and to publications from 2016 and above.

Reading through the abstracts of the emerging 92 matches, 5 were selected for an exhaustive analysis based on their relevance to the topic, with

significant inclination towards publications with higher citations. The chosen publications deemed to epitomize the theme of discussion are listed as follows:

Publication	Year	Citations
The impact of micro blogging data for stock market prediction	2017	47
Stock market sentiment lexicon acquisition using microblogging data and statistical measures	2016	31
Sentiment analysis of Twitter data for predicting stock market movements	2017	19
Stock price prediction using linear regression based on sentiment analysis	2016	14
Discovering public sentiment in social media for predicting stock movement of publicly listed companies	2017	12

III. ANALYSIS OF LITERATURE

The first paper inspected described an approach for evaluating the significance of text-based sentiment indicators for predicting the stock market variables as well as their correlation to public mood surveys [10]. An assemblage of about 31 million tweets were gathered using Twitter's REST Application Programming Interface (API) from 22 December 2012 to 29 October 2015 encompassing stock data related to around 3800 US companies. A realistic rolling window of 300 days of for stock market data for training and 50 indices of survey indicators were extracted upon which 5 machine learning (ML) models were operated on the data namely multiple linear regression (MR), neural networks (NN), support vector machine learning (SVM), random forest (RF) and ensemble averaging (EA), under two classifications: an auto-regressive baseline and a microblog group. Furthermore, a recent lexicon fine-tuned for financial microblogs comprising of indicators in excess of 20000 was used to [9] produce sentiment indicators which were unearthed from 31 million tweets relevant to US stock market between December 2012 and October 2015. The authors further discussed the design of a Kalman Filter (KF) to merge different sentiment aggregators such as bullish ratio, variation, etc. with values of various periodicity: daily Twitter estimates, weekly survey indicators like American Association of Individual Investors (AAII) & Investors Intelligence (II). Lastly, a Diebold-Mariano (DM) test was conducted to examine the usefulness of predictions based on sentiment indicators compared to the baseline. The conclusion was that public sentiment on Twitter was beneficial in forecasting the proceeds of Standard & Poor's 500 (SP500) index, lower market capitalization portfolios and a few sectors like High Technology, Energy and Telecommunication. However, this might simply be due to the fac that there are generally more tweets on these sectors thus biasing the dataset. Moreover, public mood on Twitter was instructive in projecting negative values of AAII. The inadequacy for this approach is that it treats all tweets as equal and doesn't take into consideration the weightage of influential and informed social media users. Additionally, the authors argue that microblogging data, however objective, lags behind blogs for providing a competent judgement and so should also be incorporated in further studies.

The second paper discusses an efficient methodology for generating lexicons relevant to stock market [9]. This is derived through statistical procedures employed over a set of messages from StockTwits. A holdout split method was operated, using 75% StockTwits messages as training set and the remaining 25% as test set. 3 prevalent statistical

measures namely, Term frequency–inverse document frequency (TF–IDF), Information Gain (IG) and Pointwise Mutual Information (PMI) as well as 2 new statistics (Pdays(l) and Massoc(l)) were employed to calculate sentiment scores and a total of 12 lexicon sets were created. Sentiment analysis (SA) was performed on the correctness of all sets and evaluated against 6 general lexicons (Harvard General Inquirer (GI), opinion lexicon (OL), Macquarie Semantic Orientation Lexicon (MSOL), MPQA subjectivity lexicon (MPQA), SentiWordNet (SWN) 3.0 and financial sentiment dictionaries (FIN)). Additionally, the author described using Twitter based indicators as a comparable proxy for traditional stock market lexicons like as it showed significant Pearson's correlation with AAII and II indices. The results of the paper showed a statistically noteworthy enhancement for SA using the generated lexicons compared to the other 6 showing greater correlation values against Twitter extracted sentiment indicators. A major drawback of this approach is the requirement of labelled stock market documentation for efficient creation of the lexicons as such comprehensive documentation is hard to come by.

The third paper focused on the correlation of a company's stock value variations (e.g. Microsoft) to the public opinion of the said company over Twitter [11]. A collection of 250,000 tweets from 31 August 2015 to 25 August 2016 were collected and categorised using keywords such as $MSFT, #Windows, #Microsoft, etc. Stock values for the same period were obtained from Yahoo! Finance. Furthermore, the data gathered was then cleaned through tokenization, stop-word removal and Regex matching. The authors deliberated on employing n-grams and Word2vec text depiction techniques for classification of tweets as positive, negative or neutral. Of these, Word2vec proved more sustainable showing better performance and was then fed through a RF model with a split of 90% showing a precision of 70.2%. The deductions from this thesis revealed that a convincing association exists between a company's stock prices and its public social media opinion. This restraint for this paper is its reliance of only Twitter data as a huge portion of the stock trading population doesn't use Twitter but rather StockTwits, a specialized financial platform built specifically for trading among other things. Moreover, data can further be procured from other sources like news outlets to provide a comprehensive prediction model.

The fourth paper analysed the use of regression models for predicting stock prices fluctuations in the Indonesian market [4]. The dataset was gathered for a period two weeks, from 14 April 2015 to 30 April 2015. Stock prices for the same period were obtained from Yahoo Finance CSV API. SA operated 5 ML models; Naïve Bayes (NB), Decision Tree (DT), RF, NN and SVM for classifying the dataset into positive, negative or neutral. Of these RF showed the greatest precision with 60.39% and NB coming in second at 56.30%. After classification, the percentage of positive tweets were calculated, and models generated using the dominant tweet percentage of the prior days as dependent variable for regression. The results showcased the price fluctuation models to have NB as most accurate at 67.37% followed by RF at 66.34%. Potential shortcomings here may arrive from the minute dataset of just 2 weeks. A larger dataset would provide far better analysis. Additionally, utilization of an improved lexicon relevant to stock market [9] would result in a superior classification.

The fifth and final paper discussed an approach called SMeDA-SA (Social Media Data Analyzer – Sentiment Analysis) that explores public sentiment towards a catalogue of 30 companies listed in the New York Stock Exchange and NASDAQ to

forecast their stock movement [8]. The data set used here comprised of around 200 million tweets from October 2011 to March 2012. SMeDA-SA's approach involves utilization of Natural Language Processing (NLP) techniques for exploring twitter data for sentiment extraction and then incorporating association algorithms to construct concept/context maps for mined keywords. The outcome of the research propose that SMeDA-SA has a significant forecast implementation for certain sectors like Media and IT (at 76.12%). Additionally, this approach has the greatest prediction for a span of current (T) + 3 days. Further scope of research involves considering the opening, highest and lowest measure of rather than just the closing as in this paper. Moreover, the study only involved trading days data; the authors are of the opinion that public sentiment may hoard across non-trading days and thus should also be considered for better prediction models.

IV. METHOD ASSESSMENT

A. Data collection

A twitter post, also known as a Tweet, comprises of a large quantity of meta-data such as users and their geo-location as well as other entities like URLs hashtags, user mentions, etc. Across all paper, various APIs were used to amass a collection of tweets pertaining to a pre-decided timeframe. These included Twitter's own REST API [4,10], StockTwits [9], Twitter4J [3]. Furthermore, the stock data relevant to their respective markets was collected from mainly well reputed indexes such as Yahoo Finance CVS [4,11]. However, Oliveira, Cortez & Areal [10] went above and beyond with their across-the-board set of portfolios like SP500, Russell 2000 (RSL), Dow Jones Industrial Average (DJIA), Nasdaq 100 (NDQ), Excess return on the market (RMRF), etc thereby enabling them to analyse the influence of sentiment across stocks of various sizes as well as momentum.

The data was contained in MongoDB (https://www.mongodb.org), as the extracted files from the above-mentioned APIs are generally JavaScript Object Notation formatted (JSON) so MongoDBs NoSQL structure would benefit its storage. [4,10,11]. R language was utilised for all statistical procedures [9,10,11] except Pagolu, Reddy, Panda, & Majhi's [11] paper which uses the Weka tool running on Java Virtual Machine (JVM).

Further categorisation of the data using keywords and cashtags was established. Oliveira, Cortez & Areal [9,10] describe using tokenization art of Speech (POS) tagging and lemmatization through applying Stanford CoreNLP. Pagolu, Reddy, Panda & Majhi [4] elaborated using tokenization, stop-words removal and regex matching for special characters removal. They also established cleaning of missing values for non-trading days by means of a concave function. Cakra & Distiawan Trisedya [4] utilized a simple words and it's weight approach with POS tagging.

B. Sentiment extraction

Indicators for classification and SA can be adapted from measures like AAII and II [10]. However, these are expensive and time-consuming to procure. So, twitter-based indicators can be incorporated as a proxy measure as they show strong correlation to the AAII and II index. Pagolu, Reddy, Panda, & Majhi [11] discuss using N-gram text representation and Word2vec representation, with preference given to Word2vec on account of having better performance. Cakra & Distiawan Trisedya [4] incorporated simple heuristic classifiers namely calm, alert, sure, vital, kind and happy. Oliveira, Cortez & Areal [9] further described a novel lexicon generation approach which is much more accurate that traditional baseline lexicons like GI, OL, MSOL, MPQA, SWN and FIN.

Additionally, a KF can be utilized to amalgamate measure of different periodicity for easier analysis [10].

C. Prediction models

A myriad of different ML algorithms, having their individual merits and demerits, were appointed across all papers for generating the prediction models. Among these were MR [4], RF [4,10,11], NN [4,10], SVM [4,10], EA [10] NB [4] and DT [4].

D. Prediction accuracy evaluation

The forecast precision of models by Oliveira, Cortez & Areal [10] were evaluated by a DM test operated on a comparison between a baseline dataset and the microblog biased dataset. A significant DM test would indicate a prediction inference for the microblog set.

Cakra & Distiawan Trisedya [4] evaluated their models based on the value of coefficient of determination (R^2).

Pagolu, Reddy, Panda, & Majhi [11] employed a relatively straightforward approach by transforming their correlation analysis into a classification problem.

V. CONCLUSION

This report emphasized the relevance of public sentiment analysis using a myriad of machine learning tools and approaches for accurately predicting the stock market movement. From the literature review it is evident that a significant correlation can be drawn between stock prices and public sentiment on social media. However, the relation can only be established for a select few sectors or companies. Further research is required to incorporate a greater degree of complexity like establishing a trust network of distinguished users.

Moreover, the literature also highlighted a methodology for gathering and pre-processing of microblogging data upon which a series of SA can be operated to harness its latent potential. Using a custom financial lexicon as well as Twitter indicators and uniting them using a KF to be utilized by a number of ML models, favourably RF across a periodicity of T + 3 days achieves the highest accuracy for predicting stock prices variations.

Word Count: 2301

REFERENCES

1. A.R. Damasio, Descartes' Error: Emotion Reason, and the Human Brain, Putnam, 1994
2. Bollen, J., Mao, H. & Zeng, X. 2011, "Twitter mood predicts the stock market", Journal of Computational Science, vol. 2, no. 1, pp. 1-8
3. Cakra, Y.E. & Distiawan Trisedya, B. 2016, "Stock price prediction using linear regression based on sentiment analysis", ICACSIS 2015 - 2015 International Conference on Advanced Computer Science and Information Systems, Proceedings, pp. 147.
4. D. Kahneman, A. Tversky, Prospect theory: an analysis of decision under risk, Econometrica 47 (2) (1979) 263–291.
5. Fersini, E., Liu, B., Messina, E. and Pozzi, F. (2016). Sentiment analysis in social networks. Elsevier Science.
6. J.R. Nofsinger, Social mood and financial economics, Journal of Behaviour Finance 6 (3) (2005) 144–160.
7. Mäntylä, M.V., Graziotin, D. & Kuutila, M. 2018, "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers", Computer Science Review, vol. 27, pp. 16-32.
8. Li, B., Chan, K.C.C., Ou, C. & Ruifeng, S. 2017, "Discovering public sentiment in social media for predicting stock movement of publicly listed companies", Information Systems, vol. 69, pp. 81-92.
9. Oliveira, N., Cortez, P. & Areal, N. 2016, "Stock market sentiment lexicon acquisition using microblogging data and statistical measures", Decision Support Systems, vol. 85, pp. 62-73.
10. Oliveira, N., Cortez, P. & Areal, N. 2017, "The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices", Expert Systems with Applications, vol. 73, pp. 125-144.
11. Pagolu, V.S., Reddy, K.N., Panda, G. & Majhi, B. 2017, "Sentiment analysis of Twitter data for predicting stock market movements", International Conference on Signal Processing, Communication, Power and Embedded System, SCOPES 2016 - Proceedings, pp. 1345.
12. R.J. Dolan, Emotion cognition, and behavior, Science 298 (5596) (2002) 1191–1194, http://www.sciencemag.org/cgi/content/abstract/298/5596/1191.