FINANCIAL NUMERIC AND TEXTUAL DATA BASED STOCK PREDICTION USING MACHINE LEARNING TECHNIQUES

BY

MOHAMMAD RABIUL ISLAM

A thesis submitted in fulfilment of the requirement for the degree of Doctor of Philosophy in Computer Science

Kulliyyah of Information and Communication Technology
International Islamic University Malaysia

December 2020

ABSTRACT

Source of web-text and numerical data analysis for stock prediction is the challenging tasks in today's stock market engineering. Day traders in the stock market face the common issues of decision making, which is mostly dependent on daily or weekly bases data analysis. To overcome this problem, web text mining and data mining analysis techniques applied on stock market closing values, which brings the most technical approach. In terms of stock-market textual data classification, the applicable soft-computing technique holds two classifications as binary and multinomial with clustering algorithms used to apply for analysis. New prediction models overcome the drawback of previous research and indicate the necessity of classification by creating prediction algorithm with a token or a polar based financial text weighting scheme of intensive scale value (ISV) system. Binary classification helps to improve the sense of positivity and negativity with intensive value to evaluate the vast amount of financial textual data for trading decision. This research improves with the technical correlation that addressed the problem of categorical financial textual and numerical data throughout various soft-computing techniques. Targeted numerical and textual data rely on subsequently neural network, binary and multinomial classification to improve the prediction techniques by feature engineering. In terms of textual data, the novel financial data dictionary is prepared based on Harvard reference weighting schemed valued that defined as a likely result in new Stock Prediction Model. Financial news-based text analysis techniques improve the classification scenario with Naïve Bayes binary classification through financial data dictionary. Beside the text analysis, the feed-forward neural network architectural model also improved based on backpropagation neural network structural that approached by defining the correlation between the actual and prediction trend of a daily basis. Daily stock price prediction is the main objective of this research and very essential to generate accurate prediction through online daily basis financial news data. The new architectural neural network model performs with sequential data as hidden with the dataset which applied by the multi-objective optimization algorithm. Throughout feature engineering, setting by scaling value determines the weight factors of developing a neural network that used to define more precious trend within this model. This model enabled to calculate the highest frequent value that occurred on a large dataset, and clustering indicate the stock trend of the prediction. Based on the numerical financial data, new Stock Prediction Model (SPM) have developed for analyzing market movement from two benchmarks numerical stock market dataset those are S&P 500index and OHLCV dataset. Developing integrated classification techniques conducting with prediction analysis based on its classification accuracy as defined in this research 82% which is obvious and better than previous research. The performance with feature engineering in text classification also gain 93%, whereas multilevel and binary classification have performed as combined to gain the best accuracy level. Performance of the proposed approach was estimated by evaluating various parameter as part of the information retrieval field, as seen in this experimental result. However, developing model impacts on academical research philosophy in terms of financial data classification but not highly recommend using in real trading analysis.

خلاصة البحث

تعد المصادر الإلكترونية للنصوص والتحاليل الرقمية للبيانات للتنبؤ بالأسهم مهمة صعبة في هندسة سوق الأوراق المالية في وقتنا هذا. حيث يواجه المتداولون اليوميون في سوق الأسهم قضايا مشتركة لاتخاذ القرارات والتي تعتمد في الغالب على تحليل البيانات اليومية أو الأسبوعية لسوق الأوراق المالية. وللتغلب على هذه المشكلة، يتم تطبيق تقنيات التنقيب عن البيانات وتحليلها على قيم إغلاق هذه الاسواق، وهو ما يحقق أفضل نهج تقني. وفيما يتعلق بتصنيف البيانات النصية لسوق الأسهم، فإن الأساليب الحاسوبية المعمول بها تحمل صنفين أحدهما ثنائي والآخر متعدد الحدود مع خوارزميات التجميع المستخدمة للتحليل. وتتغلب نماذج التنبؤ الجديدة على عيوب البحوث السابقة كما تشير إلى ضرورة التصنيف عن طريق إنشاء خوارزمية التنبؤ مع رمز أو نظام ترجيح للنص المالي القائم على نظام قيمة المقياس المكثف (ISV). اما التصنيف الثنائي يساعد على تحسين المعنى الإيجابي والسلبي مع أهمية مكثفة لتقييم الكم الهائل من البيانات النصية المالية لاتخاذ قرار التداول. يتم تحسين هذه الدراسة من خلال الارتباط الفني الذي تناول مشكلة البيانات المالية الرقمية والنصية عبر مختلف التقنيات الحاسوبية. حيث تعتمد البيانات الرقمية والنصية المستهدفة على الشبكات العصبية والتصنيف الثنائي والمتعدد الحدود لتحسين تقنيات التنبؤ من خلال هندسة الميزات. ومن حيث البيانات النصية، تم إعداد قاموس البيانات المالية الجديد استناداً إلى مخطط ترجيح المرجع في جامعة هارفارد، والذي تم تحديده باعتباره نتيجة محتملة لنموذج التنبؤ بالأسهم الجديد. تعمل تحليل بيانات النصوص المالية المستندة على الأخبار المالية على تحسين سيناريو التصنيف باستخدام تصنيف Naïve Bayes الثنائي من خلال قاموس البيانات المالية. وإلى جانب تحليل النص، تم أيضًا تحسين بنية الشبكة العصبية للتغذية المسبقة استناداً إلى بنية الشبكة العصبية العكسية والتي تمت من خلال تحديد العلاقة المتبادلة بين الاتجاه الفعلي والتوقع للأساس اليومي. كما يعد التنبؤ بأسعار الأسهم اليومية الهدف الرئيسي لهذا البحث وهو ضروري للغاية لتقديم تنبؤ دقيق من خلال بيانات الأخبار المالية اليومية على الإنترنت. حيث يعمل نموذج الشبكة العصبية الجديد ببيانات متسلسلة مخفية مع مجموعة البيانات التي يتم تطبيقها بواسطة خوارزمية التحسين متعددة الأهداف. وخلال هندسة الميزات، يتم تحديد الإعدادات حسب قياس قيمة عوامل الوزن لتطوير شبكة عصبية تستخدم لتحديد اتجاه أكثر قيمة ضمن هذا النموذج. تم تمكين هذا النموذج لحساب أعلى قيمة متكررة حدثت في مجموعة البيانات، ويشير التجميع إلى اتجاه الأسهم من التنبؤ. واستنادًا إلى البيانات المالية الرقمية، تم تطوير نموذج جديد للتنبؤ بالأسهم (SPM) لتحليل حركة السوق من خلال مجموعة بيانات رقمية لسوق الأوراق المالية هما S&P 500index وOHLCV. تطوير تقنيات التصنيف المتكاملة التي تجري مع تحليل التنبؤ بناءً على دقة التصنيف حددت بنسبة 82٪ في هذا البحث وهو أمر واضح انه أفضل من دقة التصنيف في البحوث السابقة. كما تحصل الأداء مع هندسة الميزات في تصنيف النصوص على 93%، في حين أن اداء كل من التصنيف الثنائي والمتعدد المستويات فقد تم دمجهما معا للحصول على أفضل مستوى من الدقة. تم تقدير أداء المنهج المقترح من خلال تقييم مختلف المتغيرات كجزء من مجال استرجاع المعلومات، كما تم توضيحه في النتائج التجريبية. ومع ذلك، فإننا لا نوصي باستخدام تطوير تأثيرات النموذج المبني على فلسفة البحوث الأكاديمية من حيث تصنيف البيانات المالية في تحليل التداول الحقيقي.

APPROVAL PAGE

The thesis of Mohammad Rabiul Islam has been approved by the following:

_____
Imad Fakhri Al-Shaikhli
Supervisor

_____
Rizal Mohd Nor
Co-Supervisor

_____
Afidalina Tumian
Co-Supervisor

_____
Asadullah Shah
Internal Examiner

_____
Shamala Subramaniam
External Examiner

_____
Hilal Mohammad Yousif Albayatti
External Examiner

_____
Mohamed Elwathig Saeed Mirghani
Chairman

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Mohammad Rabiul Islam

Signature ..... Date .....

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

ONLINE FINANCIAL NUMERIC AND TEXTUAL DATA ANALYSIS FOR STOCK PREDICTION WITH COMPREHENSIVE DATA ANALYSIS THROUGH MACHINE LEARNING AND SOFT-COMPUTING TECHNIQUES

I declare that the copyright holders of this thesis are jointly owned by the student and IIUM.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Mohammad Rabiul Islam

……..…….. ……..

Signature Date

ACKNOWLEDGEMENTS

Thank you from the bottom of my heart to the Most of all, the Almighty “ALLAH”, the beneficent and the merciful to mankind. Secondly my beloved Professor Dr. Imad Fakhri Taha Al-Shaikhli and rest of the special acquaintance in my life.

Of course, it’s my utmost pleasure to dedicate this work to my dear parents and my family, who granted me the gift of their unwavering belief in my ability to accomplish this goal: thank you for your support and patience.

I wish to express my appreciation and thanks to those who provided their time, effort and support for this research. To the members of my dissertation committee, thank you for sticking with me.

Once again, the special thanks to Professor Dr. Imad Fakhri Al-Shaikhli and co-supervisor Dr. Rizal Mohd Nor and Asst. Prof Dr. Afidalina Tumian for their continuous support, encouragement and leadership, and for that, I will be forever grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
ARMA	Auto-Regressive-Moving-Average
BPNN	Back Propagation Neural Network
CNBC	Consumer News and Business Channel
CR	Challenge of Research
CBR	Case Based Reasoning
EMH	Efficient Market Hypothesis
ES	Existing Solution
FN	False Negative
FP	False Positive
FS	Feasible Solution
FFNN	Feed Forward Neural Network
GM	Genetic Algorithm
HTML	Hyper Text Markup Language
ISV	Intensive Scale Value
IR	Information Retrieval
IFR	Financial Information Retrieval
LOCF	Last Observation Carried Forward
LR	Linear Regression
ML-Algo	Machine Learning Algorithm
MLPs	Multilayer Perceptron
MSE	Mean Squared Error
NDTV	New Delhi Television Limited
NMT	News Mining Techniques
NB	Naïve Bayes
NLP	Natural Language Processing
OHLCV	Open, High, Low, Close price and Volume
PTM	Pattern Taxonomy Method
PNN	Probabilistic Neural Networks
PSNR	Pick Signal-to-Noise Ratio
PCA	Principle Component Analysis
PBM	Phrase Based Method
PSP	Potential Solution for Problem
RNN	Recurrent Neural Nets
ReLU	Rectified Linear Unit
RG	Research Gap
SPM	Stock Prediction Model
SVM	Support Vector Machine
S&P 500	Standard and Poor 500 Index
TDFDD	Term Document of Financial Data Dictionary
TBM	Term Based Method
TP	Truth Positive
TN	Truth Negative
WWW	World Wide Web

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF THE RESEARCH

Stock market prediction is a very active, optimistic and challenging field as to explore in the field of soft computing techniques. The area of financial text mining application is still in its infancy but the potentiality of its future under evolution (Khadjeh Nassirtoussi *et al.*, 2014). A vast amount of big data as collections of the stock market that usually called from the primary source of the World Wide Web (www). In this emerging source of information where a large amount of data able to submerge for the big data system. Having estimated every 20 months, the source of information found double in world wide web (Frawley and Piatetsky-Shapiro, 1992).

Also, such type of information and volume of usable digital data as remained hidden in the web. Data mining and text mining techniques are being applied over the last few years to discover the knowledge from hidden data. (Fayyad, Piatetsky-Shapiro and Smyth, 1996) However, most academicians and practitioners believe that the prediction of the stock market is possible by merely analysing the past stock data. Based on this past numeric or text data, data mining techniques possible to improve by developing some algorithms and applying several strategies to get the prediction in a more effective and efficient way (Raj, 2017).

The technique of text mining enables to extract the high-quality, informative text that usually extracts from a set of textual documents. Analyzing text mining of any field, especially the types of firms that related to information and its application techniques that bring the priority of action-based attributes. These attributes produce the most logical approach instead of low priority (L. Song, 2015). Because the stock market

always deals with a large amount of numeric and non-numeric data that almost impossible to calculate manually in a short time for the traders. In the field of computer science, soft computing techniques utilized the mining techniques as used to predict the stock markets price movement (Oyland, 2015).

Accordingly, advance soft computing techniques such as fuzzy logic in data mining produce many options like operations and logical approaches. Last few decades, various soft computing techniques have been discovered for financial prediction. For example, complex keyword topples that generate the prediction of stock market movements (Wuthrich *et al.*, 1998). Human emotions engaged with hope, fear and worry such critical term come from news that increasing or decreasing amounts of the stock price influence on following day (Zhang, Fuehres and Gloor, 2011). So, this research exactly focus on stock prediction that manifested by soft computing methods.

In this case, "Corpus of Text" was a popular research technique from the last few years that usually contains the word of human emotion, sentiment for data and text mining result in the financial sector (Pang and Lee, 2006; Godbole, Srinivasaiah and Skiena, 2007; Leskovec, Backstrom and Kleinberg, 2009) and many more. Among those techniques, Neural Networks are the most widely used (Choudhry and Garg, 2008). For example, Time Delay Neural Networks for stock market trend prediction, Probabilistic Neural Networks (PNN) used for such model for classification problem. Recurrent Neural Nets (RNN) also used for predicting the next day stock price index with of Bayesian networks, evolutionary algorithms, classifier system, fuzzy algorithm as well as time series analysis (Chung and Shin, 2018) (Jeenanunta, Chaysiri and Thong, 2018) (Camara *et al.*, 2018) (Sharma, 2017) (Environment, 2010).

However, in the financial sector, the significant progress of (Support Vector Machine) SVMs from last several years also impact on financial time series forecasting

(Francis E H Tay & Cao 2001). This beneficial technique impact on technical analysis indicator as input to SVMs. The comparable result of previous BPN and SVM showed that SVM outperformed on BPN more often, but some specific market BPN has also been found better (Sheta, 2006). Instead of BPN and SVM, this research mainly focuses on Feed Forward Neural Network (FFNN) that applies to numerical financial data and Naïve Bayes (NB) for textual day trading data as discussed in subsequent chapters.

1.2 IMPORTANCE OF ANALYSIS FOR DAY-TRADING PREDICTION

Traders of daily basis are known as "day traders" used to active on one-day trading. Due to an intelligent portfolio management system, trading in a day is essential for financial decision making, that directly depend on feasible prediction (Lee, 2007). Moreover, the major stock market indices through the most important financial news that usually taken from daily new closing values as numerical result for decision making (Wuthrich, 1998). Indeed, daily based financial text news brings out the critical decision for day trading. The performance of numeric data or text mining applied on various predicting fields like a policy, engineering, economic, medical, meteorology and others. In this approach, traditional statistical methods also to discovered by hidden knowledge of raw data (Gharehchopogha and Khalifehlou, 2012; Soleimanian, 2012; Soleimanian Gharehchopogh and Reza Khaze, 2012). Throughout this field prediction of algorithm possible to develop in data mining path.

1.3 STATEMENT OF THE PROBLEM

Decision making in investment is very important for the stock market, whether the risk and the problem are in conditional states of a daily market phenomenon due to the vast amount of financial non-numerical data. Previous research found that stock market risk

dependent on daily return observation (Chevapatrakul, Xu and Yao, 2018). The stock price prediction always dependent on technical and fundamental analysis as the expected return comes from different analytical approach (Chaigusin, Chirathamjaree and Clayden, 2008). The technical approach bring on past price, and fundamental financial analysis depends on macroeconomic information, company sector information as well as the company itself (Thomsett, 2005). Both types of information always depend on the open-source sector basis. For example, daily newspaper, online forum, blog, social network, company analytical document that usually published daily, weekly, monthly or even yearly.

Risk is always a common issue as produced from several impacts for lacking the proper situation of the conditional market. A successful trader mostly depends on the contemporary accurate daily based trading price for further trading, but the fluctuation of the stock market is a normal phenomenon. That is why risk-free trading always expected and highly demanding platform for any kinds of financial investment (Prediction, 2016). However, the challenging task is to find the accurate and daily market price from a valuable large amount of textual information, which is usually released online. Most traders depend on daily or weekly price but taking proper decision for price movement gets complicated due to the trade volume in terms of market volatility. Analytical techniques of trade volume able to make a prediction but in some case, data complexity in trade analysis constrained with empirical research that need to extended for simplification (S.R. Dash, 2015) (Chen, 2018).

Nowadays, financial news sources (Numeric and Textual data) are available based on social media like Twitter, Facebook, Yahoo finance, CNN money and many more. Electronic news like textual-data is usually unstructured for testing of any model. This makes weak relationship between online news and one-day stock price movements

due to the noisy data (Noisy data mean other than financial data that cause in this term) experimented in the past (Oyland, 2015). However, this unstructured news is much related to noisy data so these irrelevant noise data need to eliminate from financial text documents that help to get the pure financial text data in terms of financial text-based polar form. The drawback of this research found that the result of accuracy for financial text analysis needs to be improved in terms of classification.

Besides, backpropagation neural network model for numerical data need to improve the architecture by building a new model for stock prediction with feature engineering of the scaling value which integrated principal component analysis (PCA) that enables to improve the accuracy level of stock data classification. The impact of risk defined by the current news and compare it with the actual value of numerical data. Daily basis online news always updated from time to time based on financial news and many researchers active on intelligence portfolio management system, which also recommends daily based stock trading (Lee *et al.*, 2007). The accuracy of prediction also depends on the fluctuation of the market price, which is analyzed by data mining through M-Learning algorithm.

This research points out three problem statements as found in previous betterment of stock-market data analysis.

1. There is a weak relationship between online news and stock volume price movement due to the noisy data and high possibility of variance in succeeding the component of the original data. So, it is inevitable to eliminate noisy data from the original news parameters as input to define more accurate prediction from the financial news (Oyland, 2015; I. Ibidapo, 2017)

2. Online textual news contains the most influential facts or terms for a stock market prediction that hasn't diverse on the dictionary in terms of intraday news-based stock price prediction over the one-day trade volume. This matter does not succeed due to the gap of late arrival news as an input of the text classification model of the same trading day as published (Oyland, 2015; Anbalagan and Maheswari, 2015; P. Gupta, 2017)
3. In the non-linear relationship between the methodology to topology having lack of architectural improvement in terms of weight Matrix value in Artificial Neural Network. That lacking of architectural configuration generate less accurate result with a minimum error of stock data classification (Saini, 2014)

1.4 RESEARCH QUESTIONS

Research questions bring out the crucial aspect and the analytical thought from unique research questions. As seen in the above research problems, three main questions have pointed out, as stated below, to forward for the research objectives.

1. How do the financial textual news documents overcome the noisy data by eliminating from the original financial textual news?
2. How do the technique and classification model apply to financial textual data to overcome on the intraday news-based stock price prediction over the one-day trade volume that suitable in the gap of late arrival financial news on the same trading day as published?
3. How does architectural neural network perform to minimise the error rate with the efficient algorithm that improves the accuracy level?