

# **SENTIMENT ANALYSIS FOR FINANCIAL MARKET PREDICTION**

T. H. H Methmal

168246D

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

April 2020

# **SENTIMENT ANALYSIS FOR FINANCIAL MARKET PREDICTION**

Thommaya Hewa Hasanga Methmal

168246D

Dissertation submitted in partial fulfillment of the requirements for the degree Master of  
Science in Computer Science specializing in Data Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

April 2020

## DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: .....

Date: .....

Name: T. H .H Methmal

I certify that the declaration above by the candidate is true to the best of my knowledge and that this report is acceptable for evaluation for the CS6997 - MSc Research Project. The above candidate has carried out research for the Masters Dissertation under my supervision.

Signature: .....

Date: .....

Name: Dr. Surangika Ranathunga

## ABSTRACT

Today's world is highly dependent on financial markets. Financial markets are very dynamic, making it difficult to predict prices. If we can make good predictions, we can make financial gains without any risks. There are two main categories of data that we can use to predict the market price; historic market data and textual data. Most importantly, textual analysis on sources like news, social media and reports is more popular today among researchers. This process can be introduced as sentiment analysis. The whole idea behind sentiment analysis is checking the opinion behind the text; whether it is a positive, negative, or neutral polarity.

This research focuses on sentiment-analysis-based financial market prediction using deep learning. Market prediction using sentiment analysis is a very challenging task. There are complex linguistic issues to solve and using a microblog dataset like Twitter for the prediction task makes it even more difficult. However, the current prediction approach rarely exceeds the seventy percent accuracy mark. This research is based on the SEMEVAL 2017 fifth task and will use the same dataset shared by the SEMEVAL team.

This thesis presents an improved version for above mention reported baseline. We experiment with different techniques in both machine learning and deep learning domains. Lexicon based dictionaries are heavily used here in each model since this is a small dataset with train set (1693) and test set (793). We had to enlarge the dataset as much as possible to achieve good accuracy. We created mainly four models which are based on machine learning and deep learning techniques. Support vector regression algorithm is used for the machine learning model. Also we used convolutional neural network (CNN) , long short term memory (LSTM ) and gated recurrent unit (GRU ) as deep learning architectures which are performed better than any of the baseline models on this dataset. Our deep learning models are achieved maximum similarity scores than any of the single system.

Finally, we experiment three main ensemble techniques which are Averaging Linear Regression and multilayer perceptron. We achieved best results from averaging ensemble model.

**Keywords:** Financial Sentiment Analysis; Opinion Mining; Reviews; Text Analysis; Deep learning; Sentiment Analysis Challenges; Semeval.

## **ACKNOWLEDGEMENTS**

I would like to express my sincere gratitude to my supervisor, Dr Surangika Ranathunga for the continuous support from the first day of my M.Sc. study, for her patience, motivation, enthusiasm, and immense knowledge. Besides my supervisor, I would like to thank the rest of my teachers throughout the journey for guiding me towards the right path in life. Last but not least, I would like to thank my mother, without whom I would not be here, and for her support in guiding me spiritually throughout my life.

## TABLE OF CONTENTS

DECLARATION .....	i
ABSTRACT .....	ii
ACKNOWLEDGEMENTS .....	iii
TABLE OF CONTENTS .....	iv
LIST OF FIGURES .....	vii
LIST OF TABLES .....	viii
LIST OF ABBREVIATIONS .....	ix
Chapter 1 : Introduction .....	1
1.1Background .....	1
1.2Sentiment Analysis .....	1
1.3Motivation .....	2
1.4Objective .....	2
1.5 Thesis Contribution.....	2
1.6Thesis Organization .....	3
Chapter 2 : Literature Review .....	4
2.1Input dataset .....	4
2.1.1Textual Data .....	4
2.1.2Market Data.....	5
2.2Pre-processing .....	5
2.2.1Feature Selection.....	5
2.2.2Feature Reduction .....	8
2.2.3Feature representation .....	9
2.3Regular Machine Learning Approach.....	11
2.3.1Support Vector Machines (SVM) .....	11
2.3.2Regression Algorithms.....	13
2.3.3Naïve Bayes .....	14
2.3.4Decision Rules and Trees.....	15
2.3.5Combinatory Algorithms .....	16
2.4 Lexicon – Based Approach .....	17

2.5 Word Embedding .....	19
2.6.1 Word2vec .....	19
2.6.2 Global Vectors (GloVe).....	22
2.6.3 FastText.....	23
2.6 Deep Learning Approach .....	24
2.7.1 Convolutional Neural Network (CNN).....	25
2.7.2 Recurrent Neural Networks (RNN) .....	26
2.7.3 Long Short Term Memory (LSTM).....	27
2.7.4 Gated Recurrent Units (GRU).....	27
2.7 Ensemble Approach .....	28
Chapter 3 : Methodology .....	31
3.1 Data .....	31
3.2 Data Preprocessing.....	32
3.3 Feature Engineering .....	33
3.3.1 Sentiment lexicon features. ....	33
3.3.2 Linguistic Features .....	35
3.3.3 Word embedding features .....	36
3.3.4 Domain – Specific Features .....	37
3.4 Evaluation Measure.....	38
3.5 Proposed Technique .....	39
3.5.1 Convolutional Neural Network model(CNN).....	39
3.5.2 Recurrent Neural Network (BI -LSTM) .....	39
3.5.3 Gated recurrent Units (GRU).....	39
3.5.4 Multilayer Perceptron (MLP).....	39
3.5.5 Regular machine Learning .....	40
3.5.6 Ensemble Model.....	40
Chapter 4 : Experiments.....	43
4.1 Regular Machine Learning Algorithms .....	43
4.1.1 Support Vector Regression (SVR).....	43
4.1.2 Other Machine Learning Algorithms .....	46
4.2 Bidirectional Long Short Term Memory Network (Bi-LSTM).....	47

4.3 Gated Recurrent Units (GRU).....	48
4.4 Multilayer Perceptron (MLP).....	48
4.5 Convolutional Neural Network (CNN).....	49
4.6 Ensemble Model.....	49
Chapter 5 : Discussion .....	51
5.1 Data preprocessing .....	52
5.2 Feature Engineering .....	52
5.3 Parameter Optimization .....	53
5.4 Comparison with Baseline .....	53
5.5 Future Works.....	54
Chapter 6 : Conclusion.....	55
REFERENCE.....	56



## LIST OF FIGURES

- Figure 2.1. System components diagram for machine learning models [9].
- Figure 2.2. Neural network architectures for CBOW single word context [70].
- Figure 2.3. Neural network architectures for CBOW group of word context [70].
- Figure 2.4. Neural network architectures for skip gram model [70]
- Figure 2.5. Overall accuracy on word analogy task between glove and word2vec. [73]
- Figure 2.6. Comparison between other deep learning methods and fastest on tag prediction[75].
- Figure 2.7. Comparison between other word representation methods and fastest on different languages [75].
- Figure 2.8. Ghosal et al [56] cosine similarity comparison between CNN and LSTM on validation set.
- Figure 2.9. simple RNN model. [76]
- Figure 2.10. Long Short-Term Memory [76].
- Figure 2.11. LSTM cell vs GRU cell [90]
- Figure 2.12. Ensemble MLP model [56].
- Figure 2.13. MLP based ensemble architecture [82]
- Figure 3.1. Proposed ensemble model
- Figure 3.2 Proposed method by Akhtar et al.

## **LIST OF TABLES**

Table 2.1. Semeval 2017 task 5 lexicons table.

Table 2.2. Usage of lexica by Semeval 2017 task 5 teams.

Table 3.1. Dataset Statistics

Table 3.2. Prediction scores of each model with sentiment scores of test set

Table 4.1. Cross validation score on individual feature sets

Table 4.2. Cross validation score on individual lexicon features

Table 4.3. Cosine similarity score on different features with SVR on test set.

Table 4.4. Cosine similarity score on different word vectors with SVR on test set.

Table 4.5. Cosine similarity score on different features with other machine learning algorithms.

Table 4.6. Cosine similarity score on different features with Bi-LSTM

Table 4.7. Cosine similarity score on different features with GRU

Table 4.8. Cosine similarity score on different features with MLP

Table 4.9. Cosine similarity score with word embeddings with CNN

Table 4.10. Experiment results with Ensemble models

Table 5.1. Selecting the data column

Table 5.2. Comparison between word embedding and hybrid features

Table 5.3. Summary of accuracy gain w.r.t parameter Optimization

Table 5.4. Comparison between baseline (Akhtar et al) and proposed model

## LIST OF ABBREVIATIONS

GDP	Gross Domestic Product
FOREX	Foreign Exchange Market
MSH	Morgan Stanley High-Tech Index
LDA	Latent Dirichlet Allocation
IG	Information Gain
CHI	Chi-square Statistics
DF	Document Frequency
TF	Term Frequency
IDF	Inverse Document Frequency
SVM	Support Vector Machine
CATS	Categorization And Trading System
SVC	Support Vector Classification
SVR	Support Vector Regression
RBF	Radial Basis Function Kernel
OLS	Ordinary Least Square
GI	The General Inquirer
LMFSD	Loughran-McDonald financial sentiment dictionary
CBOW	continuous bag-of-words
SG	skip-gram
GLOVE	Global Vectors
ABR	AdaBoost Regressor
BR	Bagging Regressor
RF	Random Forest
CNN	Convolutional Neural Network