SENTIMENT ANALYSIS OF FINANCIAL STOCK MARKET NEWS USING PRE-TRAINED LANGUAGE MODELS

Widana Arachchi Sachith Kaushalya

(209343D)

Thesis/Dissertation submitted in fulfillment of the requirements for the degree Master of Science in Computer Science

Department of Computer Science

University of Moratuwa

Sri Lanka

July 2022

DECLARATION

I declare that this is my own work and this thesis/dissertation2 does not incorporate without acknowledgement any material previously submitted for a 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 thesis/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:

The above candidate has carried out research for the Master's Dissertation under my supervision.

Name of the Supervisor: Dr. Surangika Ranathunga

Signature of the Supervisor: Date:

ABSTRACT

Sentiment analysis helps data analysts to find public opinion, actual meaning of the given text (positive meaning, neutral meaning or negative meaning) conduct market research, monitor brand and product reputation, and understand customer experiences of newly introduced items or service.

Stock news sentiment analysis is a useful task in the financial domain. However, this is different from the customer feedback for a product or brand, movie review and customer support reviews. This huge difference is because of the domain specific language in stock markets and lack of labeled data. This research implements a stock news sentiment analysis system using the latest transformer-based pre-trained language models in NLP. I could get higher sentiment classification results for the transformer-based pre-trained language models than the traditional classifications models in this research. Also I could reduce classification result bias for the particular stock market specific words, because of the transfer learning method. And I could introduce correlation between stock news sentiment and stock price change percentage value. This proposed model can predict the percentage change value of the stock when received a news.

Additional key words and phrases: Sentiment analysis, Deep learning, Language transformer models, Transfer learning

ACKNOWLEDGMENTS

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank my supervisor, Dr Surangika Ranathunga, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would like to extend my thanks towards DirectFN for providing me with all-important data to work on my research. These results and insights would not be possible without the valuable data you provided free of charge.

TABLE OF CONTENTS

DECLARATION I

ABSTRACT II

ACKNOWLEDGMENTS III

TABLE OF CONTENTS IV

LIST OF FIGURES VII

LIST OF TABLES VIII

LIST OF ABBREVIATIONS X

1. INTRODUCTION 1

1.1. Background 1

1.1.1. Stock Market 1

1.1.2. Stock Market News 1

1.1.3. Stock Price 4

1.1.4. Previous Day Close Price 4

1.1.5. Change Percentage 4

1.2. Research Problem 5

1.3. Research objectives 5

1.4. Thesis Structure 5

2. RELATED WORK 6

2.1. Overview 6

2.2. Dictionary based methods 6

2.3. Deep Learning models 7

2.4. Transformer-based pre-trained Language models 7

2.5. Zero shot text classification 7

2.6. Summary 8
3. PROPOSED METHODOLOGY 9
3.1. Dataset 9
3.2. Proposed method to select correct polarity 9
3.3. Evaluation models 10
3.4. Implementation 11
3.4.1. Zig Zag Indicator 11
3.4.2. Find best k value and stock news sentiment. 12
4. RESULT AND ANALYSIS 18
4.1. ALBERT-base model optimization and evaluation 18
4.1.1. Training parameter optimization 18
4.1.2. Model evaluation 19
4.2. ALBERT-large model optimization and evaluation 19
4.2.1. Training parameter optimization 19
4.2.2. Model evaluation 20
4.3. BART-base model training, optimization and evaluation 20
4.3.1. Training parameter optimization 20
4.3.2. Model evaluation 21
4.4. BART-large model training, optimization and evaluation 21
4.4.1. Training parameter optimization 21
4.4.2. Model evaluation 22
4.5. BERT-base model training, optimization and evaluation 22
4.5.1. Training parameter optimization 22
4.5.2. Model evaluation 23
4.6. BERT-large model training, optimization and evaluation 23

4.6.1. Training parameter optimization 23
4.6.2. Model evaluation 24
4.7. FinBERT model training, optimization and evaluation 24
4.7.1. Training parameter optimization 24
4.7.2. Model evaluation 25
4.8. RoBERTa-base model training, optimization and evaluation 25
4.8.1. Training parameter optimization 25
4.8.2. Model evaluation 26
4.9. RoBERTa-large model training, optimization and evaluation 26
4.9.1. Training parameter optimization 26
4.9.2. Model evaluation 27
4.10. XLNet-base model training, optimization and evaluation 27
4.10.1. Training parameter optimization 27
4.10.2. Model evaluation 28
4.11. XLNet-large model training, optimization and evaluation 28
4.11.1. Training parameter optimization 28
4.11.2. Model evaluation 29
5. CONCLUSION 32
6. CONTRIBUTION 33
7. FUTURE WORK 34
8. REFERENCES 35

LIST OF FIGURES

Figure 1 Crude oil price change and news 1[13] 2
Figure 2 Crude oil price change and news 2 [13] 2
Figure 3 Crude oil price change and news 3 [13] 3
Figure 4 Twitter price change and stock news [13] 3
Figure 5 Tesla price change and stock news [13] 4
Figure 6 Dictionary-based news sentiment model 6
Figure 7 Supervised NLI System [14] 8
Figure 8 Complete proposed architecture 9
Figure 9 Zig-Zag Indicator chart [16] 11
Figure 10 Mechanism of the stock sentiment calculation with stock price data 12
Figure 11 Number of data points in train data against different k values. 14
Figure 12 Number of data points in validation data against different k values. 15
Figure 13 Number of data points in test data against different k values. 15
Figure 14 zero shot result and proposed method for result comparison. 16
Figure 15 Zero shot classification result against different k values 17

LIST OF TABLES

Table 1: Data points count and their percentage when K = 5 for final output classes..... 13
Table 2: Data points count and their percentage when K = 6 for final output classes..... 13
Table 3: Data points count and their percentage when K = 7 for final output classes..... 13
Table 4: Data points count and their percentage when K = 8 for final output classes..... 13
Table 5: Data points count and their percentage when K = 9 for final output classes..... 14
Table 6: Result of zero shot classification for different k values..... 16
Table 7 Accuracy values against learning rate for ALBERT-base..... 18
Table 8 Accuracy values against seed for ALBERT-base..... 18
Table 9 Accuracy values against evaluation steps for ALBERT-base..... 18
Table 10 Accuracy values against batch size for ALBERT-base..... 18
Table 11 Accuracy values against learning rate for ALBERT-large..... 19
Table 12 Accuracy values against seed for ALBERT-large..... 19
Table 13 Accuracy values against evaluation steps for ALBERT-large..... 19
Table 14 Accuracy values against batch size for ALBERT-large..... 19
Table 15 Accuracy values against learning rate for BART-base..... 20
Table 16 Accuracy values against seed for BART-base..... 20
Table 17 Accuracy values against evaluation steps for BART-base..... 20
Table 18 Accuracy values against batch size for BART-base..... 20
Table 19 Accuracy values against learning rate for BART-large..... 21
Table 20 Accuracy values against seed for BART-base..... 21
Table 21 Accuracy values against evaluation steps for BART-large..... 21
Table 22 Accuracy values against batch size for BART-large..... 21
Table 23 Accuracy values against learning rate for BERT-base..... 22
Table 24 Accuracy values against seed for BERT-base..... 22
Table 25 Accuracy values against evaluation steps for BERT-base..... 22
Table 26 Accuracy values against batch size for BERT-base..... 23
Table 27 Accuracy values against learning rate for BERT-large..... 23
Table 28 Accuracy values against seed for BERT-large..... 23
Table 29 Accuracy values against evaluation steps for BERT-large..... 23

Table 30 Accuracy values against batch size for BERT-large.....24
Table 31 Accuracy values against learning rate for FinBERT 24
Table 32 Accuracy values against seed for FinBERT 24
Table 33 Accuracy values against evaluation steps for FinBERT.....24
Table 34 Accuracy values against batch size for FinBERT.....25
Table 35 Accuracy values against learning rate for RoBERTa-base.....25
Table 36 Accuracy values against seed for RoBERTa-base.....25
Table 37 Accuracy values against evaluation steps for RoBERTa-base 25
Table 38 Accuracy values against batch size for RoBERTa-base.....26
Table 39 Accuracy values against learning rate for RoBERTa-large.....26
Table 40 Accuracy values against seed for RoBERTa-large.....26
Table 41 Accuracy values against evaluation steps for RoBERTa-large 26
Table 42 Accuracy values against batch size for RoBERTa-large 27
Table 43 Accuracy values against learning rate for XLNet-base 27
Table 44 Accuracy values against seed for XLNet-base 27
Table 45 Accuracy values against evaluation steps for XLNet-base.....28
Table 46 Accuracy values against batch size for XLNet-base.....28
Table 47 Accuracy values against learning rate for XLNet-large 28
Table 48 Accuracy values against seed for XLNet-large 28
Table 49 Accuracy values against evaluation steps for XLNet-large.....29
Table 50 Accuracy values against batch size for XLNet-large.....29
Table 51 Final model evaluation accuracy values 29
Table 52 Final evaluation results for financial phrase bank data trained data.....30

LIST OF ABBREVIATIONS

NLP = Natural Language Processing

BERT = Bidirectional Encoder Representations from Transformers

FinBERT = BERT model pre-trained on financial communication text

BART = Denoising Sequence-to-Sequence Pre-training for Natural Language model

RoBERTa = Robustly optimized Bidirectional Encoder Representations from Transformers

ALEART = A Light BERT for Supervised Learning

RNN = Recurrent Neural Network

CNN = Convolutional Neural Networks