Forecasting News Sentiment-Oriented Stock Market based on Sequential Transfer Learning Approach

Bea Khean Thye

MASTER OF PHILOSOPHY

FACULTY OF BUSINESS AND FINANCE
UNIVERSITI TUNKU ABDUL RAHMAN
JULY 2023

Forecasting News Sentiment-Oriented Stock Market based on Sequential Transfer Learning Approach

By

Bea Khean Thye

A dissertation submitted to the Department of Finance,
Faculty of Business and Finance,
Universiti Tunku Abdul Rahman,
in partial fulfillment of the requirements for the degree of
Master of Philosophy
July 2023

ABSTRACT

Forecasting News Sentiment Oriented Stock Market based on Sequential Transfer Learning Approach

Bea Khean Thye

Forecasting stock movements is a significant and challenging task due to the dynamic and high variability of market attributes. The massively available online text information holds the key to reveal the unexplained variability and facilitate the forecasting accuracy. However, text data is largely unstructured and exists in the form of natural language, necessitating an alternative way to process and capture the insight. This study tapped into the potential of cutting-edge Neural Networks and Natural Language Processing (NLP) techniques, with a specific emphasis on the sentiment-based approach for financial forecasting. Sequential transfer learning is a recent advancement in neural networks applied to Natural Language Processing (NLP). It adopts the "pre-train then fine-tune" paradigm to leverage existing knowledge to enhance the performance of different downstream NLP tasks and achieve state-of-the-art results. Our objective is to assess the performance of various pre-trained models such as BERT, FinBERT, and SKEP, in the context of sentiment-based financial forecasting. This study is novel in its application of these models to the scenario of algorithmic trading, offering a fresh perspective in this research area. Specifically, these pre-trained models served as the sentiment analyzers to transform the financial news titles into sentiment features. These are then combined with a range of technical indicators and used as input to the (MRM-LSTM) predictive model. The KLCI market index is primary prediction focus in this study to demonstrate the effectiveness of the proposed approach. The findings indicate that the inclusion of sentiment features extracted from financial news via these pre-trained models results in a substantial improvement in the accuracy of KLCI index price movement predictions. These results are statistically significant in achieving the highest average point return of 336.10 with F1 score of 53.59, underlining the validity of our approach. In summary, this study finds that sequential transfer learning is effective in extracting sentiment features from financial news and provides superior financial predictive performances when use in conjunction with other market data.

ACKNOWLEDGEMENT

First, I would like to express my gratitude to my academic supervisor, Dr Cheng Wai Khuen for his support and guidance throughout the study and always being available with his critical suggestion to enhance the quality of work. Besides, I would like to thank my co-supervisor Mr Jireh Chan Yi-Le for giving me valuable advice and motivation.

Second, I would like to appreciate Dr Chen Yen-Lin to provide the publication support with funding opportunity. I also like to acknowledge Mr Steven Leow Mun Hong, my fellow research partner for his assistance in research and creating a friendly environment at work. Moreover, I would like to extend my thanks to the Ministry of Higher Education of Malaysia and Ministry of Science and Technology in Taiwan to provide funding throughout my research.

Last, I would like to convey my gratitude to my parents who always contributing themselves to me over the years. In addition, the completion of this study is an aggregated efforts of all the parties, thus, I hereby would like to express my deepest gratitude for their tolerance, patience, effort, time etc.

APPROVAL SHEET

This dissertation/thesis entitled **"<u>Forecasting News Sentiment-Oriented Stock Market based on Sequential Transfer Learning Approach</u>"** was prepared by BEA KHEAN THYE and submitted as partial fulfillment of the requirements for the degree of Master of Philosophy at Universiti Tunku Abdul Rahman.

Approved by:

____

(TS. Dr.CHENG WAI KHUEN)
Date:…10 July 2023….

Assistant Professor/Supervisor

Department of Computer Science

Faculty of Information and Communication Technology

Universiti Tunku Abdul Rahman

____

(MR JIREH CHAN YI-LE)

Date:…10 July 2023..

Professor/Co-supervisor

Department of Finance

Faculty of Business and Finance

Universiti Tunku Abdul Rahman

FACULTY OF BUSINESS AND FINANCE

UNIVERSITI TUNKU ABDUL RAHMAN

Date: 6 September 2023

SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS

It is hereby certified that Bea Khean Thye (ID No: 20ABM04790) has completed this final year project/ dissertation/ thesis* entitled "Forecasting News Sentiment-Oriented Stock Market based on Sequential Transfer Learning Approach" under the supervision of Cheng Wai Khuen (Supervisor) from the Department of Computer Science, Faculty of Information and Communication Technology, and Jireh Chan Yi-Le (Co-Supervisor)* from the Department of Finance, Faculty of Business and Finance.

I understand that University will upload softcopy of my dissertation in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

Thye

(*Bea Khean Thye*)

*Delete whichever not applicable

DECLARATION

I hereby declare that the dissertation is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTAR or other institutions.

Name Bea Khean Thye

Date 10 July 2023

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Adaptive Market Hypothesis	AMH
Aspect-sentiment Pairs	AP
Attentional Encoder Network	AEN
Autoencoder	AE
Autoregressive	AR
Bag of-words	BOW
Bidirectional Encoder Representation from Transformer	BERT
Bidirectional Gated Recurrent Unit	Bi-GRU
Chinese Emotion Word Ontology	CEWO
Context Vector	CoVe
Continuous Bag of Words	CBOW
Convolutional Neural Network	CNN
Cumulative Moving Average	CMA
Deep Neural Generative	DNG
Deep Random Subspace Ensembles	DRSE
Dilated causal convolution networks with attention	Att-DCNN
Dow Jones Industrial Average	DJIA
Efficient Market Hypothesis	EMH
Embeddings from Language Models	ELMo
Event Attention Network	EAN
Event Embedding Convolutional Neural Network	EB-CNN
Event-Neural Network	E-NN
Exponential Moving Average	EMA
Financial domain of BERT	FinBERT
Gated Recurrent Unit	GRU
Generative Pretraining Model	GPT
Google Profile of Mood States	GPOMS
Graph Convolutional Network	GCN
Hierarchical Complementary Attention Network	HCAN
Hybrid Attention network	HAN
K-nearest neighbors	KNN
Knowledge Graph Event Embedding Convolutional Neural Network	KGEB-CNN
Kuala Lumpur Composite Index	KLCI
Kuala Lumpur Stock Exchange	KLSE
Language Modelling	LM
Latent Dirichlet Allocation	LDA
Long Short-Term Memory	LSTM
Machine Reading Comprehension	MRC
Machine Translation	MT
Malayan Stock Exchange	MSE
Mask Language Modelling	MLM
Multicollinearity Reduction Module	MRM
Multi-task Structured Stock Prediction model	MSSPM
Named-Entity Recognition	NER

Nasdaq 100	NDQ
Natural Language Inferring	NLI
Natural Language Processing	NLP
Neural Network Language Model	NNLM
Next Sentence Prediction	NSP
Numerical-based Attention	NBA
Out-of-vocabulary	OOV
Part-of-speech Tagging	POS tagging
Question Answering	QA
Recurrent Convolutional Neural Network	RCNN
Recurrent Neural Network	RNN
Restricted Boltzmann Machine	RBM
Robust enhanced Bidirectional Encoder Representation from Transformer	RoBERTa
Russell 2000	RSL
Self-Organizing Fuzzy Neural Network	SOFNN
Sentiment Analysis	SA
Sentiment Knowledge-enhanced Pretraining	SKEP
Sentiment Masking	SM
Sentiment Word prediction	SW
Sentiment-Oriented Word Embedding	SOWE
Sentiment-Specific Word Embedding	SSWE
Simple Moving Average	SMA
Singular Value Decomposition	SVD
Standard & Poor’s 500	S&P 500
Stock Exchange of Malaysia	SEM
Structured Stock Prediction model	SSPM
Support Vector Machine	SVM
Term Frequency-Inverse Document Frequency	TF-IDF
Universal Language Model Fine-tuning	ULMFit
Word Polarity prediction	WP

CHAPTER 1

INTRODUCTION

1.1. Background of the Study

Stock is a financial instrument representing the business ownership and its price is usually determined by the market demand and supply. The stock prices usually reflect the investor's expectation toward the future value of a particular business. In other words, the stock trading activities are information-driven where the investors regularly react to the new information and revise their expectations toward the future business status.

According to the theory of Efficient Market Hypothesis (EMH) (Fama, 1970), the rational investors should fully respond to all available information instantly when trading under a perfectly efficient market. Under such conditions, the price deviation is eliminated by the arbitrage action and the market prices should be fair and neither overvalued nor undervalued. Based on this, the price movement should obey the Random Walk Hypotheses (Malkiel, 1973) and negating the forecasting attempt to achieve more than 50% accuracy. However, it remained a question, "Is the perfect market condition exists in practice?".

In past literature, extensive studies such as Bollen, Mao and Zheng (2011), Hu et al. (2018), Zhang et al. (2018), Chen et al. (2019), Wang, Wang and Li (2020) are continuously realized the excess return in the stock market. Those findings are contradicted with EMH and raise a fundamental question, “Where does the abnormal return come from?”. To this, the Adaptive Market Hypothesis (AMH) (Lo, 2004) is proposed to explain the market anomalies with the theories of behavioral finance. In practice, the ability of informational mining and investor’s behavior toward the information can be very different across the investors. The information asymmetry is the primary reason to distort the perfect market condition and imply the likelihood to realize the excess return in the stock market (Xing, Cambria & Wlsch, 2017).

Although excess return can be realized in practice, the nonlinear and dynamic attributes of financial markets are being the biggest challenges to be addressed in financial forecasting (Tang et al., 2022). Financial time series usually with a high degree of variability and extensive noise to disrupt the forecasting performance (Green and Zhao, 2022). The massive available text data in web could be used to reveal the unexplained variability in the financial market and improve the forecasting performance.

Yet, the natural language in the text is usually unstructured and implied with different inherent meanings. Those messages can be easily interpreted by humans but difficult to be understood by a machine. However, such a massive amount of text data is difficult to manually process by a human due to the constraint of time, ability, energy, etc. The invention of Natural Language

Processing (NLP) has provided a solution to develop computational models that enable the machine to understand human languages and provide an automatic solution to practical problems. Across different NLP applications, sentiment analysis is found to be naturally suited for financial forecasting since the behavioural information of investors can be captured to explain the market variability.

According to Liu (2015), sentiment analysis (SA) is defined as an automation tool to identify and analyze people's opinions, emotions, and attitudes toward a specific event, individual, or topic. According to Mao (2020), sequential transfer learning in NLP has recently succeeded in achieving state-of-the-art results across different sentiment-related tasks. The pre-trained model is a result of self-supervised training the model on a massive corpus to optimize the pretraining objective and learn the synergistic representation that can be transferred to benefit the downstream tasks.

According to Mishev et al. (2020), the pre-trained model showed an overwhelmed performance over the traditional sentiment analysis approaches across different financial sentiment evaluation datasets. However, different pre-train models might have different impact on an identical downstream task, since the pre-trained model is trained with different pre-training tasks, model architecture, and corpus sources (Qiu et al., 2020). Therefore, it remained a need to examine the potential improvement of different pre-trained model in sentiment-based financial forecasting. Among the list of pre-trained models, there are three pre-trained models chosen to conduct the comparison study which are BERT (Bidirectional Encoder Representation from Transformer), FinBERT (Financial domain of BERT) and SKEP (Sentiment Knowledge-Enhanced Pretraining) models.

Specifically, the BERT model is a generic pretrained model invented by Devlin et al. (2018) for self-supervised training the model in learning the bidirectional contextual representation of words from a massive unlabelled corpus. The training corpus of BERT was majorly denominated by the common English corpus such as English Wikipedia (2500M words) and Book corpus (800M words).

On the other hand, Yang, UY & Huang (2020) pointed out that the BERT model trained on general corpus might have the difficulty to interpret the meaning of the financial term due to the domain discrepancy. Based on this, the study introduced the FinBERT model which pretrained with the financial text sources such as Corporate Report 10-K & 10-Q, Analyst Reports and Earning Call Transcripts to learn additional financial domain knowledge.

Apart from the BERT and FinBERT models trained on different corpus, Tian et al. (2020) has introduced the sentiment knowledge-enhanced pretraining (SKEP) model to improve the performance of sentiment-oriented tasks. Different from general pretraining tasks, the model is pretrained with different sentiment-oriented objective to learn additional sentiment information.

In this study, the Malaysia stock market is the primary focus since the text-based financial forecasting studies available in Malaysia is limited. The Kuala Lumpur Composite Index (KLCI) remained the major market index that composed by the top 30 companies in Malaysia based on the market capitalization weighted method (Kwong et al., 2017) to evaluate the country economic growth. The selection of the companies is majorly based on their capitalization and influencing power across different sector in Malaysia.

From the perspective of economist, the performance of stock market index implied the action of fiscal or monetary policies to ensure the stability of country's economy. Therefore, the forecasting of stock market is being a significant research topic and valuable to be studied.

1.2. Problem Statement

Firstly, the use of sentiment features as input data in financial forecasting is relatively common in past literature (Bollen, Mao & Zheng, 2011; Nguyen, Shirai and Velcin, 2015; Chen, Cai & Lai, 2016; Ren, Wu, and Liu, 2018; Mohan et al., 2019; Maqsood et al., 2020; Gupta & Chen, 2020). However, those studies are solely relying on a single sentiment feature which is not sufficient to make the forecasting accurate. Although the technical indicators do provide additional market information that is useful to improve the forecasting performance, but the study to integrate the technical indicator with sentiment features in Malaysia for financial forecasting is limited. As up to current, existing local sentiment-based forecasting studies are merely drawn by Rahman et al. (2017), Shuhidan et al. (2018) and Kuan et al. (2019).

Secondly, the pre-trained model trained with different corpus sources might have different effects on an identical task. FinBERT, a financial domain pre-trained model proposed by (Yang, UY & Huang., 2020) is justified to outperform BERT (General pre-trained model) in various financial sentiment classification tasks. The result is not surprised since each specific domain is characterized by a unique vocabulary and the semantic expression is diverse across the domains.

For example, the specific financial term, "*long, short, put, call*" implied a different meaning from general understanding. When two domains are completely unrelated, the pre-trained knowledge might be insignificant to transfer and eventually degrade the performance of the downstream tasks. However, a contradict finding by Mishev et al. (2020) was neglected the superiority of FinBERT over BERT. As the evaluation study on FinBERT is limited, the actual impact of the pretrained models is remained uncertain in financial forecasting.

Lastly, the different in pretraining objectives might have different impact on the pretraining features. The pretraining objectives like *Language Modeling* (Sarzynska-Wawer et al., 2021), *Masked Language Modeling* (Delvin et al., 2019), *Permuted Language Modeling* (Yang et al., 2019), and *Next-Word Prediction* (Radford et al., 2018) are prevailing to capture the dependency between words and syntactic structures. However, the sentiment information is seldom to be encoded in the pre-trained general representation and limit the performance in sentiment analysis. To this, the SKEP model (Tian et al., 2020) is proposed to learn the unified sentiment representation via the three-novel sentiment pretraining objectives to improve the sentiment-related downstream tasks. However, the SKEP pre-trained model is yet to be evaluated for its performance in financial sentiment analysis.

1.3. Research Questions

1. What are the sentiment features compared to technical indicators to improve financial forecasting performance?
2. How does the financial domain pre-trained model (FinBERT) compare to the general pre-trained model (BERT) to improve sentiment-based financial forecasting?
3. How does the sentiment-oriented pre-trained model (SKEP) compare to the general pre-trained model (BERT) to improve sentiment-based financial forecasting?

1.4. Research Objectives

1. To examine the potential improvement of including sentiment features in technical indicator based financial forecasting.
2. To examine the potential improvement of FinBERT (Financial domain pre-trained model) over BERT (General pre-trained model) in sentiment-based financial forecasting.
3. To examine the potential improvement of SKEP (Sentiment-oriented pre-trained model) over BERT (General pre-trained model) in sentiment-based financial forecasting.

1.5. Significance of the Study

Firstly, this study expects to investigate the potential improvement of the application of different pre-trained models in sentiment-based financial forecasting. Ideally, the preciseness of sentiment measure may infer the actual investor behaviour to improve the forecasting result. Although the pre-trained models are proven to achieve state-of-the-art results across different sentiment-related tasks, none of the studies evaluate its sentiment outcome for practical financial forecasting. Therefore, the findings of this study are expected to serve as a milestone to push the pre-trained models on the ground of financial forecasting application.

Secondly, the different pretraining corpus might have different effect on an identical task. The FinBERT (Financial domain pre-trained model) proposed by (Yang, UY & Huang, 2020) is proven to outperform BERT (General pre-trained model) in various financial sentiment classification tasks. However, a contradict finding by Mishev et al. (2020) was neglecting the superiority of FinBERT. As the evaluation of FinBERT is limited, the future study will be ambiguous to select the appropriate pre-trained models for financial application. Therefore, the finding of this study is expected to serve as a reference to evaluate the performance of FinBERT in financial forecasting applications.

Thirdly, the different pretraining tasks might significantly affect the properties of pretraining features. To this, the sentiment-oriented pre-trained model, SKEP (Tian et al., 2020) is the first study to encode additional sentiment knowledge into the pre-trained features throughout the three-novel sentiment pretraining objectives. However, SKEP is yet to apply in the financial sentiment-related task and this study will be the first to evaluate the performance of SKEP in the application of sentiment-based financial forecasting.

Lastly, the study aims to integrate the sentiment features with technical indicator to examine the potential improvement toward the forecasting performance. In practice, market movement will be affected by various factors. However, past study to investigate the integration of technical indicators with sentiment features for financial forecasting is very limited. Therefore, the finding is expected to serve as a reference for future studies to investigate the financial forecasting with multi-sources.

1.6 Conclusion

In this chapter 1, the study introduced the research background which claiming the difficulty of financial forecasting in addressing the dynamic and non-linear of stock market. With the advancement in technology, recent studies attempted to explore the text information to enhance the forecasting stock. Recently, the sequential transfer learning approaches has achieved the state-of-the-art performance across different NLP tasks. However, the pretrained model trained with different corpora, objective and model architectures might resulted of different properties and performance vary on a same task. Due to the limited studies in examining the potential impact of different pretrained model in market forecasting, this study attempted to examine the impact from three different pretrained models namely, BERT, FinBERT and SKEP in forecasting the Malaysia stock market, KLCI.

CHAPTER 2

LITERATURE REVIEW

2.1. Transfer Learning

Pan and Yang (2009) defined transfer learning as a methodology that utilizes knowledge gained from the source task to enhance the performance of a target task. To establish a taxonomy of transfer learning, we introduce the notations of domain (D) and task (T). Domain (D) comprises data features (X) and the distribution of those features (P(X)), denoted as D = X, P(X). On the other hand, task (T) encompasses label data (y) as well as the predictive function f(.), represented as T = y, f(.). Therefore, transfer learning can be formally described as an approach that aims to enhance the predictive function f(.) in the target domain Dt by leveraging knowledge acquired from a different source domain (Ds) and learning task (Ts), where Ds ≠ Dt and Ts ≠ Tt.

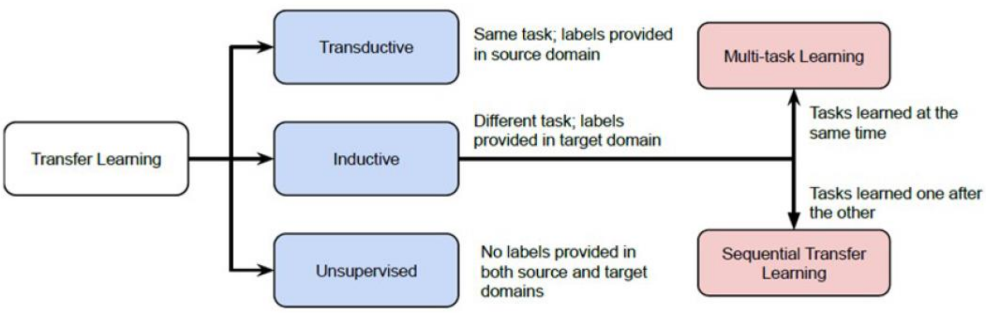

Figure 2.1 The taxonomy of transfer learning. Retrieved from (Chan et al, 2020c)

In accordance with Chan et al. (2022c), transfer learning can be classified into three distinct categories based on the tasks in the source and target domains and the availability of labels. These categories include inductive, transductive and unsupervised transfer learning. Based on the setting of transductive transfer learning, the tasks in the source and target domains are similar (Ts = Td), but only the source domain possesses labeled data. The domain adaptation is an example of transductive transfer learning, where a sentiment model trained on Amazon reviews is utilized to predict the financial news sentiment. The primary objective here is to tackle the issue of limited labeled data in a specific domain. However, the significant challenge in this study lies in bridging the gap between the data features and feature distributions of the source and target domains.

In the setting of inductive transfer learning, the tasks in the source and target domains differ, regardless of the similarity of the domains themselves (Ts ≠ Tt). In such setting, the labels are available in the target domain, but their presence in the source domain is optional. On the other hand, if neither the source nor the target domain has labels, it falls under the category of unsupervised transfer learning. Additionally, Mao (2020) fine-grained the categories of inductive transfer learning into two subcategories: (1) multitask learning and (2) sequential transfer learning, based on the availability of labels in the source domain.

Multitask learning involves having access to labeled data in both the source and target domains, with different tasks being performed. In this setup, the tasks Ts and Tt are learned simultaneously through the joint optimization of multiple objective functions. The primary objective is to utilize a larger pool of data to acquire task-invariant features that can enhance performance across various tasks. An additional advantage of multitasking learning is its potential to achieve strong generalization across diverse scenarios. However, one challenge in multitask learning arises when there is a considerable disparity between the tasks. This discrepancy can impede the model training process and lead to a decline in overall performance. It is crucial to address this challenge by employing appropriate techniques to effectively handle the differences between tasks within the multitask learning framework.

On the contrary, multitask learning relies on having labeled data available in both the source and target domains, which can pose a constraint when dealing with limited annotated data. In such cases, sequential transfer learning offers a solution to mitigate the cost of labeling. Sequential transfer learning involves first self-supervised learning of general knowledge in the source domain (Ds) and source task (Ts), followed by its application to enhance the predictive function fT(.) in the target domain (Dt), where the labels ys are unavailable and Ts ≠ Tt. According to Chan et al. (2022c), sequential transfer learning is also referred to as the "pretraining then fine-tuning" paradigm. During the pretraining phase, the model is trained using self-supervised learning techniques, leveraging labels derived from the data without manual labelling effort. The objective of pretraining is to optimize specific objectives like language modeling (LM), mask language modeling (MLM), or next sentence prediction (NSP). This process aims to learn a universal synergistic representation that can be transferred and applied to different downstream tasks.

The benefits of pretraining have been further elucidated by Hao et al. (2019), highlighting that pretraining can offer superior model initialization, effectively facilitating the error optimization process, enhancing model generalization, and mitigating overfitting issues, particularly when dealing with small datasets. In recent times, the advancement in computational power and the availability of vast amounts of unlabeled data have played a pivotal role in the development of sequential transfer learning approaches. These approaches involve pretraining a large language model to improve overall model generalization. Subsequently, the fine-tuning process is employed as a knowledge adaptation technique to directly train the pretrained model for a specific task. Given the rapid progress in sequential transfer learning, this study aims to explore its application in various sentiment-related tasks.

In the earlier stages, pretraining embedding was considered as a traditional technique for transferring features, widely utilized in various downstream natural language processing (NLP) tasks due to its simplicity and effectiveness. Specifically, the process involved pretraining word embedding using simpler model architectures like word2vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014), or FastText (Bojanowski et al., 2017). These pretrained embeddings were then used to initialize the embedding layer of a task-specific model framework, serving as the first layer of the pretrained model. Subsequently, the word embedding layer would be "frozen," meaning it remains unchanged, allowing each input text to be transformed into its respective word representation vector. This vector captured the semantic meaning of the text, providing a general and synergistic feature that could be shared across different downstream NLP tasks. It enabled the model to understand the underlying meaning of the words. Section 2.2 of the discussed material delves into the evolution of pretrained models, progressing from the initial pretrained word embedding technique to the contextual pretrained encoder.

2.2. Development of sequential transfer learning

2.2.1 The first phase: Pretrained word embedding.

According to Chan et al. (2022c), in the earlier research, words in a language were often quantified and represented using atomic symbols or one-hot representations. However, this approach had limitations in distinguishing the semantic similarity between words and treating each word independently. As a result, there was a need to explore semantic word representation.

The motivation to study semantic word representation was inspired by the distributional hypothesis (Harris, 1954), which suggests that words with similar meanings tend to appear in similar contexts or neighboring words. Building on this idea, Schütze (1992) introduced the "word space" approach, which aimed to measure the semantics of words based on the statistical co-occurrence of 4-gram letters.

Nevertheless, as the vocabulary size increases, the dimension of the co-occurrence matrix also grows, leading to a challenge known as the "curse of dimensionality." This issue arises when the data distribution is spread thin across high-dimensional spaces, which can weaken the robustness of the model. To address this problem, the singular value decomposition (SVD) technique is commonly employed. Additionally, the SVD approach is inflexible when it comes to incorporating new words into the existing word representation framework.

Afterward, Bengio et al. (2000) made significant advancements by introducing a neural network language model (NNLM). The NNLM aimed to model the joint probability function of word sequences and learn distributed word representations through a product-based approach. The key insight from Bengio et al.'s work was that the proposed NNLM model had the ability to generate coherent sentences, even when encountering unknown words. This was achieved by leveraging the nearby representations of known words to construct sentences containing previously unseen or unknown words.

In contrast to Bengio et al. (2000), Collobert et al. (2011) took a different approach to learn distributed word representations using a neural network. They proposed a method where word embeddings were pretrained objectively using a large unlabeled dataset. The pretrained embeddings were then transferred to downstream natural language processing (NLP) tasks such as semantic role labeling, named-entity recognition (NER), part-of-speech tagging (POS tagging), and chunking.

The popularity of word embeddings can be largely attributed to the work of Mikolov et al. (2013b), who introduced a simpler model architecture known as word2vec. They proposed two models, namely continuous bag of words (CBOW) and skip-gram, for pretraining word embeddings using a large unlabeled corpus. Their study aimed to improve the computational efficiency of the previously pretrained neural network language model (NNLM) introduced by Collobert et al. (2011) by eliminating unnecessary hidden layers.

The CBOW model operates by predicting the current word based on its surrounding context words within a specified window. It aims to learn the relationships between the target word and its context words to generate meaningful embeddings. On the other hand, the skip-gram model takes the current word as input and predicts the surrounding context words occurring before and after it. This model also focuses on capturing the contextual information and word relationships.

It found that pretrained word vectors can exhibit semantic compositionality. For instance, by performing vector operations such as vec('France') - vec('Paris') + vec('Japan'), we can obtain a vector representation that is close to vec('Tokyo'). This ability to perform vector arithmetic and capture semantic relationships between words has been one of the significant advantages of pretrained word embeddings. The similarity between word vectors is often measured using cosine similarity, which quantifies the angle between two vectors. In subsequent work, Mikolov et al. (2013a) made several advancements to improve the training efficiency of both the CBOW and skip-gram models. They introduced two techniques: negative sampling and subsampling.

In addition to the word2vec model, Pennington et al. (2014) and Bojanowski et al. (2017) introduced alternative approaches for word embeddings known as GloVe and FastText, respectively. These approaches aimed to further enhance the quality of word vectors. Pennington et al. (2014) proposed the Global Vectors for Word Representation (GloVe) model. They argued that while skip-gram models, like those used in word2vec, focus on local context windows for training word embeddings, they may overlook important global statistical information. To address this limitation, GloVe incorporates additional word co-occurrence statistics into the training process. By considering the overall co-occurrence patterns of words in the corpus, GloVe aims to capture both local and global semantic information, resulting in improved word representations.

Furthermore, Bojanowski et al. (2017) recognized a limitation in previous word embedding approaches, which assigned a separate vector to each word. This approach posed challenges when dealing with unknown words that were not present in the training corpus, often referred to as the out-of-vocabulary (OOV) issue. To address this limitation, they introduced the FastText model, which enriches word embeddings with subword information. The FastText model addresses the OOV issue by incorporating subword information into the word representation process. Instead of representing each word as a single vector, FastText learns representations for character n-grams (subword units) individually. These n-gram vectors capture information about the character sequences present within each word.

Indeed, traditional word embeddings have limitations when it comes to encoding sentiment information. In these embeddings, words with opposite sentiment polarities, like "good" and "bad," are often mapped to similar embeddings, which can negatively impact sentiment analysis performance (Tang et al., 2014). To address this issue, Tang et al. (2014) introduced sentiment-specific word embedding (SSWE) to align the semantic meaning of words with their sentiment polarities in a sentiment embedding space.

Word embeddings can face limitations when applied to specific domains like finance. For example, words with sentiment implications may have different interpretations in financial contexts. For instance, the word "underestimate" carries a negative sentiment but can actually indicate a positive investment opportunity. To address this challenge, Li and Shah (2017) introduced sentiment-oriented word embedding (SOWE) specifically designed to capture domain-specific sentiment in the financial domain. They accomplished this by training embeddings using a sentiment lexicon tailored for the stock market. In this subsection 2.2.1, the study walks although the development of word embedding, and the discussions are summarized in Table 2.1.

Table 2.1 Summary table of Pretrained word embedding.

Publication	Technique	Contribution
-	One-hot-representation	Representing word with zero-one vector
Schutze (1993)	Word Space	Measure the word semantic based on the lexical co-occurrence statistic
Bengio et al. (2003)	NNLM	Apply language modelling to learn the distributed word representation
Collobert et al. (2011)	NNLM	Pretrain the word embedding on unlabeled corpus and applied to downstream NLP tasks
Mikolov et al. (2013b) Mikolov et al. (2013a)	Skip-gram	Remove the hidden layer to effectively pretrain the word embedding
Pennington et al. (2014)	GloVe	Pretrain embedding with additional statistical word co-occurrence information
Bojanowski et al. (2017)	FastText	Pretrain embedding with sub-word information
Tang et al. (2014)	SSWE	Align the word semantic with sentiment polarity in pretraining the word embedding
Li and Shah (2017)	SOWE	Pretrain the word embedding based on the sentiment lexicon in financial domain

2.2.2 The second phase: contextual pretrained encoder

During the initial stages, pretrained word embeddings were often used to kick-start the architecture for task-specific networks. However, this approach, which relied heavily on a single pretrained embedding layer, had difficulty managing issues such as polysemy, syntactic structures, semantic roles, and anaphora. Consequently, later research efforts focused on creating a more robust neural encoder capable of learning higher-level representations, like the varying meanings of a word in different contexts (Chan et al., 2022c).

McCann et al. (2017) proposed a model known as Context Vector (CoVe), which used a deep two-layer LSTM to learn context-dependent word representations. Unlike prior single-layer transfers, this approach also transferred the static embedding layer along with the LSTM encoder's contextual state, enabling a more potent representation. This context-dependent state enabled dynamic word representations based on context.

However, the limitations of cross-lingual data potentially impeded the pretraining process for contextual representations using machine translation tasks (McCann et al., 2017). To address this, Sarzynska-Wawer et al. (2021) introduced "Embeddings from Language Models" (ELMo), a method to pretrain deep context-dependent word representations using a large unlabeled corpus and a bidirectional language model. The aim was to enhance pretrained feature quality and tackle out-of-vocabulary issues by advancing from a unidirectional LSTM (McCann et al., 2017) to a two-layer bidirectional LSTM.

Howard and Ruder (2018) developed a framework known as ULMFit for fine-tuning universal language models based on LSTM. This framework aimed to improve classification performance through a three-stage fine-tuning process. Initially, a language model is pretrained on a vast general domain corpus. Then, this pretrained model is further fine-tuned on a task-specific corpus, and finally, a task-oriented classifier is further fine-tuned with the target dataset. Additionally, novel fine-tuning techniques like "discriminative fine-tuning" and "slanted triangular learning rates" were introduced.

Following this, Radford et al. (2018) first applied the transformer neural network to enhance the Bi-LSTM structure in ELMo (Sarzynska-Wawer et al., 2021), introducing the Generative Pretraining Transformer (GPT). This model, further improved to GPT-2 by Radford et al. (2019), used a multi-layer, left-to-right transformer to manage long-range text dependencies and was trained on a larger dataset with a multitask objective. Unlike previous models, transformers allow the pretrained model to be directly fine-tuned without needing other task-specific architectures.

However, Devlin et al. (2018) argued that both ELMo and GPT are unidirectional, meaning they sequentially read text input from one direction only, thus limiting the model's context during pretraining. To resolve this, BERT was introduced, which used an autoencoder approach to pretrain a deep bidirectional contextual representation based on "masked language model" and "next sentence prediction" tasks. For instance, in the case of GPT, the model processes the input text sequentially from left to right. However, its self-attention mechanism only considers the preceding tokens and does not consider the contextual information from the tokens on the right. This limitation poses challenges for tasks like question answering (QA) that require understanding and utilization of contextual information in both directions.

To address this issue, the "bidirectional encoder representation from transformer (BERT)" was introduced as a self-supervised pretraining method. BERT learns deep bidirectional contextual representations by training on a large unlabeled corpus using two novel unsupervised tasks: "masked language model (MLM)" and "next sentence prediction (NSP)". In the MLM task, tokens are randomly masked, and the transformer encoder processes the entire text sequence bidirectionally to predict the masked tokens based on contextual information. The NSP task involves predicting the next sentence based on the surrounding sentences to capture the relationship between sentences during pretraining.

In addition, the BERT model is typically pretrained using general corpora like BookCorpus and English Wikipedia, which may limit its performance in specific domains. Therefore, subsequent research efforts have aimed to extend the applicability of BERT to various specialized domains. For instance, BioBERT was pretrained on a vast biomedical corpus, SciBERT on a multidomain corpus of scientific publications, and FinBERT on a financial corpus comprising corporate reports, earnings call transcripts, and analyst reports. In tasks involving financial sentiment classification, fine-tuning the FinBERT model has shown improved results compared to the original BERT model.

Furthermore, there have been several extensions of pretrained BERT models, including RoBERTa, ALBERT, and DistillBERT. RoBERTa enhances the BERT model by utilizing additional training techniques, expanding the dataset, removing the next sentence prediction (NSP) objective, and introducing dynamic masking during pretraining. On the other hand, ALBERT and DistillBERT were developed to reduce computational requirements and accelerate the pretraining process. ALBERT achieves this by sharing parameters across layers and decomposing the embedding matrix to reduce the model size. It also replaces the NSP objective with sentence order prediction (SOP) to enhance the coherence between sentences. DistillBERT employs knowledge distillation, a compression technique, to transfer knowledge from a well-pretrained teacher model to a smaller student model through joint optimization.

Yang et al. (2019c) identified a limitation in BERT, where it is considered an autoencoder (AE) pretraining model that learns contextual representations through reconstructing corrupted input data. This approach resulted in a discrepancy between pretraining and fine-tuning, as the artificial symbols [MASK] used during pretraining were absent during fine-tuning. Unlike autoregressive (AR) models like ELMo and GPT that model the probability distribution of a text sequence, BERT is constrained to learning joint probability due to its assumption of independence for the predicted token [MASK]. This limitation hinders its ability to handle higher-order or longer-range dependencies in natural language. To address these issues, Yang et al. (2019c) introduced a generalized autoregressive pretraining model called XLNet, which learns bidirectional contextual representation by maximizing the expected likelihood over all permutations of the factorization order.

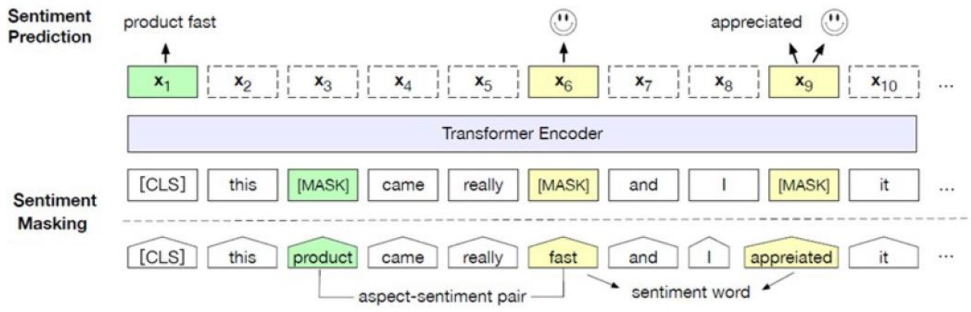

Figure 2.2 General Framework of SKEP. Retrieved from Tian et al. (2020)

Recently, Tian et al. (2020) introduced the sentiment knowledge-enhanced pretraining (SKEP) framework, aiming to learn a comprehensive sentiment representation that can be effectively applied to various sentiment-related tasks. The SKEP framework begins by automatically extracting sentiment knowledge and subsequently utilizing sentiment masking (SM) to identify and conceal sentiment information using the [MASK] token. Following this, SKEP employs three sentiment pretraining objectives: sentiment word prediction (SW), word polarity prediction (WP), and aspect-sentiment pairs (AP). These objectives facilitate the encoding of sentiment word, polarity, and aspect information into a unified sentiment representation. The SKEP model's transformer encoder is initialized with RoBERTa and pretrained using an unlabelled sentiment corpus. The fine-tuned SKEP model demonstrates promising results across different sentiment tasks. Figure 2.2 shows the general framework of SKEP.

Recently, Brown et al. (2020) improved the GPT-2 (Radford et al., 2019) by proposing GPT-3 that incorporating meta-learning and in-context learning to enhance the model performance in the scenarios with limited examples or no examples. Additionally, GPT-3 has a parameter scale 100 times larger than GPT-2, surpassing a landmark of 100 billion parameters and been trained with a size of 45TB pretraining data compared to only 40GB pretraining data for GPT-2.

Afterward, the pilot version of GPT3.5 regarded as the ChatGPT (InstructGPT) are proposed to adapt the reinforcement learning with human feedback (RLHF) to incrementally train the GPT-3 model (OpenAI, 2023). This approach enables the model to better understand and align with the user's intent. Furthermore, GPT-4 is a large multimodal model that accepts both image and text inputs, producing text outputs. In short, ChatGPT/GPT-4 demonstrates human-level performance on various professional and academic benchmarks.

In summary, subsection 3.2 reviews various types of pretraining models that can be directly fine-tuned for the sentiment analysis task, and the discussion is summarized in table 2.2. The following subsection 2.3 will be discusses the considerations in fine tuning the pretrained model.

Table 2.2 Summary table of pretrained contextual encoders

Publication	Framework	Pre-trained objective	Model
McCann et al. (2017)	CoVe	Machine Translation	LSTM
Sarzynska-Wawer et al. (2021)	ELMo	LM	Bi-LSTM
Howard & Ruder (2018)	ULMFit	LM	LSTM
Radford et al. (2018)	GPT	LM	Transformer
Radford et al. (2019)	GPT 2.0	LM with multi-task learning	Transformer
Delvin et al. (2018)	BERT	Masked LM Next sentence prediction (NSP)	Transformer
Liu et al. (2019)	RoBERTa	Masked LM	Transformer
Lan et al. (2019)	ALBERT	Marked LM Sentence order prediction (SOP)	Transformer
Yang et al. (2019)	XL-Net	Permuted Language Modelling (PLM)	Transformer-XL
Tian et al. (2020)	SKEP	Sentiment word prediction (SW), Word polarity prediction (WP), Aspect sentiment pairs (Ap)	Transformer

2.3 Knowledge adaptation: The fine-tuning approaches

According to Qiu et al. (2020), there are several important factors been emphasized in the fine-tuning the pretrained models for downstream tasks. The initial consideration is the selection of the pretrained model itself, as variations in the pretraining task, model architecture, and pretraining corpus can have a notable impact on the quality of the pretrained features. For instance, when it comes to sentiment analysis, the pretrained features derived from SKEP (Tian et al., 2020) are more suitable than those from BERT

(Devlin et al., 2018) due to SKEP's sentiment-oriented pretraining tasks and incorporation of sentiment knowledge.

Another important consideration, as highlighted by Qiu et al. (2020), is the selection of layers to be transferred during fine-tuning. For instance, in the case of word2vec, only the pretrained embedding layer is transferred to train the task-specific framework from the beginning. However, adapting the knowledge of a large pretrained model is more intricate. Take the BERT encoder as an example, which comprises 12 transformer blocks (layers), 12 self-attention heads, and a hidden size of 768 (Devlin et al., 2018). Each layer captures distinct types of information, making the decision of which layers to transfer a crucial one.

In Section 2.4, the study discusses various text-based financial forecasting approaches, including sentiment-based, semantic-based, event-extraction based, and hybrid approaches.

2.4 Text-based financial forecasting

This study classifies recent text-based stock forecasting approaches into three categories: sentiment-based, semantic-based, and event extraction-based approaches, based on the principles of NLP techniques applied in stock forecasting. Xing et al. (2018) divided available financial text into six main groups, including corporate disclosure, financial reports, professional periodicals, aggregated news, message boards, and social media, considering factors such as length, subjectivity, and frequency of updates.

In the scenario of investment decision-making, investors can be influenced passively or actively by text information. For instance, social media platforms often contain user comments with strong opinions and stances, which can actively influence investors. In such cases, sentiment-based approaches are preferred to analyze the emotional state for forecasting stock movements. On the other hand, sources like news articles and financial reports provide objective information that passively influences investors. Investors are then required to process this information to determine its impact on relevant stocks. Based on this, the following section 2.4.1, 2.4.2 and 2.4.3 attempted to discuss and summarize the approach in text-based stock forecasting studies such as sentiment-based, semantic-based and event extraction-based approaches.

2.4.1 Sentiment-based stock forecasting approach

Sentiment analysis (SA) is a branch of NLP that focuses on studying people's opinions, emotions, and attitudes towards specific events, individuals, or topics (Liu, 2015). In the financial domain, sentiment analysis plays a crucial role in transforming unstructured financial text into sentiment signals that reflect investors' inner thoughts. Tetlock (2007) observed a downward trend in market prices following reports with high pessimism scores, supporting the idea that emotional text information is key to revealing investor expectations in response to different textual information. As a result, sentiment analysis has become a popular research direction in the financial field.

In previous literature, sentiment-based financial forecasting typically involves a two-step procedure. First, a sentiment analyzer is used to extract sentiment features from textual sources, which are then adapted to a prediction model for stock forecasting. According to Yadav and Vishwakarma (2020), existing sentiment classification techniques can be categorized into three main types: (1) lexicon-based approaches that use sentiment dictionaries to determine the sentiment polarity of each word, (2) machine learning approaches that train sentiment classifiers based on statistical semantic features like n-grams, term frequency-inverse document frequency (TF-IDF), bag-of-words, etc., and (3) deep learning approaches that automatically extract feature representations using neural networks for sentiment analysis.

The lexicon-based approach is the most popular in sentiment-based financial forecasting due to its simplicity and efficiency. Once a sentiment wordlist is compiled, researchers can easily measure the sentiment of corresponding texts. Approaches to compiling sentiment wordlists can be divided into two categories: dictionary-based and corpus-based approaches. The former involves constructing sentiment wordlists based on synonyms and antonyms of predetermined sentiment words found in general dictionaries such as WordNet (Fellbaum, 1998), ConceptNet (Liu and Singh, 2004), SentiWordNet (Baccianella et al., 2010), Harvard General Inquirer, Henry Wordlist (Henry, 2008), Opinion Finder (Wilson et al., 2005), etc. On the other hand, the latter exploits the syntactic patterns of co-occurring words in a corpus to compile sentiment wordlists. Additionally, Yekrangi and Abdolvand (2021) attempted to combine both approaches by considering both statistical and semantic aspects to develop a specialized financial lexicon.

The general sentiment dictionary has been criticized for its weak performance in domain-specific sentiment analysis. For example, Loughran and McDonald (2011) discovered that approximately 73.8% of negative words in the Harvard General Inquirer dictionary were not considered negative in the financial domain. As a result, Loughran and McDonald developed a specialized financial lexicon, known as the Loughran and McDonald's wordlist (Financial lexicon), using a corpus-based approach that exploited financial sentiment words from U.S. Securities and Exchange Commission reports. Subsequent studies, such as Picasso et al. (2019), utilized the McDonald dictionary and AffectiveSpace2 to extract sentiment embeddings from summarized financial news articles related to the top twenty companies listed in the NASDAQ 100 index. These sentiment embeddings were then combined with technical indicators for market analysis.

Li et al. (2014) mapped words from financial news articles onto the emotional spaces of two different dictionaries, the Harvard IV-4 Dictionary and the Loughran-McDonald Financial Dictionary. The study found that sentiment features based on dictionaries yielded better results than the bag-of-words model. Seifollahi and Shajari (2019) preprocessed news headlines using approaches such as Relevant Gloss Retrieval, Similarity Threshold, and Verb Nominalization before applying SentiWordNet to measure sentiment scores. Their study demonstrated that identifying the appropriate sense of significant words in news headlines improves the sentiment-based market forecasting.

In addition to the aforementioned open-source dictionaries, several studies have taken the initiative to develop their own financial lexicons in order to enhance the quality of sentiment analysis. For instance, Shah et al. (2018) manually constructed a domain-specific sentiment wordlist in the pharmaceutical industry to determine the sentiment polarity of individual words. The sentiment score was computed based on the count of sentiment words and directly utilized for trading decisions. Oliveira et al. (2016) created an alternative financial lexicon using data from financial microblogging. This lexicon includes around 7,000 unigrams and 13,000 bigrams, each with their respective sentiment scores, and takes into account affirmative and negated contexts.

Subsequently, Oliveira et al. (2016) applied their proposed lexicon to measure the sentiment of each Twitter message, and the aggregated results were used to compute various Daily Twitter Sentiment Indicators, including the Bullish Ratio, Bullishness Index (BI), Agreement (AG), and Variation (VA). The study employed the Kalman Filter (KF) procedure to combine these Daily Twitter Sentiment Indicators with weekly values from the American Association of Individual Investors (AAII), Investors Intelligence (II), monthly values from the University of Michigan Surveys of Consumers (UMSC), and Sentix values. The aggregated sentiment index was then used to predict daily returns, trading volume, and volatility of various indices such as the Standard & Poor's 500 (SP500), Russell 2000 (RSL), Dow Jones Industrial Average (DJIA), Nasdaq 100 (NDQ), as well as portfolios based on size and industries.

Day and Lee (2016) combined sentiment features from four different lexicons (NTUSD, HowNet-VSA, NTgUFSD, and iMFinanceSD) with statistical word information to examine the relationship between financial news and stock price trends. The study emphasized that the accuracy of stock forecasting is influenced by the sources of financial news. Yu et al. (2013) proposed a sentiment classification approach based on the existence and intensity of emotion words in financial news. They employed the context entropy model to measure semantic similarity between words by comparing their contextual distributions using entropy.

Similarly, Bollen et al. (2011) utilized two mood tracking methods, Opinion Finder and Google Profile of Mood States (GPOMS), to analyze daily content on Twitter. They found that psychological features derived from mood dimensions improved predictive performance for the Dow Jones Industrial Average (DJIA) index. Specifically, the mood dimension "Calm" was identified as a statistically significant predictor for the daily price changes of the DJIA.

In the context of Chinese microblogging data, Chen et al. (2016) filtered Weibo posts related to three influential topics and applied the Chinese

Emotion Word Ontology (CEWO) to measure sentiment scores in seven categories (Happiness, Sadness, Surprise, Fear, Disgust, Anger, and Good). The discrete sentiment scores were aggregated to construct daily emotional time series for each category. The study found that public mood states of "Happiness" and "Disgust" were granger-causal factors influencing stock price changes in China. Wang et al. (2018) expanded the Tsinghua Sentiment Dictionary by incorporating specific financial terms and adjectives to analyze sentiment in Sina Weibo posts. The sentiment scores, along with technical indicators, were used as input for a novel "Deep Random Subspace Ensembles" (DRSE) model for market forecasting.

Wei et al. (2017) developed the "aggregate news sentiment index" (ANSI) using term frequencies of optimism and pessimism characteristic terms in Chinese financial news to investigate the relationship between financial news and the Taiwan stock market. Their findings indicated that the sentiment level of financial news significantly influences the construction of financial portfolios. Qian et al. (2020) utilized the Word2Vec model to represent user comments from the website www.eastmoney.com, and then applied CNN to measure the bullish or bearish tendencies of users. They emphasized that bearish tendencies among users are associated with higher market volatility and increased market returns.

In contrast, Zhang et al. (2018) highlighted that measuring sentiment polarity alone is insufficient, as sentiment polarities can vary across different topics or domains. They employed Latent Dirichlet Allocation (LDA), a topic model, to calculate the distribution of topics over words. Zhao et al. (2016) also used LDA to filter unrelated topics in financial microblogs on Weibo and applied a financial lexicon to obtain sentiment polarities for predicting the market index. Similar studies by Si et al. (2013) and Nguyen et al. (2015) performed topic-based sentiment analysis to predict the stock market.

Besides, several studies attempted to assemble different type of machine learning models to improve the predictive performance in conjunction of sentiment features. For example, Deng et al. (2022) proposed LightGBM-NSGA-II-SW, which combines LightGBM, NSGA-II, and SW methods to predict and simulate trading in the Shanghai Stock Exchange (SSE) index using investor sentiment as features. Deng et al. (2023) proposes an explainable XGBoost-based method for stock index prediction and trading simulation in the Chinese security market, utilizing sentiment features of institutional, individual, and foreign investors as explanatory variables, achieving the best forecasting accuracy and identifying institutional investor sentiment as relatively more essential for index direction prediction.

Recently, Lopez-Lira and Tang (2023) explored the use of ChatGPT to predict stock market returns by analyzing sentiment in news headlines, finding a positive correlation between "ChatGPT scores" and subsequent daily stock market returns, outperforming traditional sentiment analysis methods, and suggesting that incorporating advanced language models can improve prediction accuracy and enhance quantitative trading strategies. In addition, Qiu, Song and Chen (2022) introduced a modified sentiment index that incorporates weighted sentiment analysis and the day-of-the week effect, leading to improve stock trend prediction accuracy across multiple machine learning model. Table 2.3 summarized the finding of sentiment-based financial forecasting studies.

Table 2.3 Summary table of sentiment-based stock forecasting approach

Author	Model	Method
Bollen et al. (2011)	Self-Organizing Fuzzy Neural Network	Applied two mood tracking methods (1) Opinion Finder and (2) Google Profile of Mood States (GPOMS) to analyse daily twitter content for prediction
Chen et al. (2016)	Neural Network	Applied Chinese Emotion Word Ontology (CEWO) to measure sentiment score in 7 categories and aggregated the score into a daily emotional time series
Zhao et al. (2016)	user-group model	Applied Latent Dirichlet Allocation (LDA) to filer irrelevant topic of financial microblogs from "Weibo" and then applied the financial lexicon to obtain the sentiment polarities to forecast the market index.
Picasso et al. (2019)	Feedforward Neural Network	Applied McDonald dictionary and AffectiveSpace2 to extract sentiment embedding
Wei et al. (2017)	VAR	Measure the market optimism and pessimism to construct the aggregate news sentiment index
Qian et al. (2020)	CNN	Applied CNN to measure the user's bullish-bearish tendencies for prediction

2.4.2 Semantic-based stock forecasting approach

Initially, text-based stock forecasting studies often relied on the bag-of-words (BoW) approach to represent financial texts. This approach treated each word independently, using word frequencies as features. For instance, Mittermayer and Knolmayer (2006) introduced the NewsCATS model, which applied BoW to represent press releases and used KNN and SVM models to analyze their impact on stock markets. Luss and d'Aspremont (2015) utilized BoW and tf-IDF weighting to represent press releases and employed multiple kernel learning for prediction.

In addition to BoW, researchers started incorporating different textual features such as noun phrases, named entities, and tf-IDF. Schumaker and Chen (2009) investigated the effect of breaking news on stock prices using SVM with BoW, named entities, and noun phrases as separate textual features. They found that the noun phrases model outperformed BoW and named entities models in terms of directional accuracy, simulated trading, and closeness. Dadgar et al. (2016) employed tf-IDF to represent news headlines from the BBC and 20 newsgroup datasets as input to an SVM for stock movement prediction.

However, the BoW approach struggled to capture the semantic relationships between words, treating each word independently. To address this, the n-gram feature was introduced, capturing contiguous sequences of n words from a text sequence and providing better syntactic information. Hagenau et al. (2013) evaluated the effectiveness of different n-gram features using ad hoc announcements data and found that bigram features performed better than trigram and unigram. However, Kiros et al. (2014) highlighted the issue of the "Curse of Dimensionality" with n-grams, as the co-occurrence matrix dimension grows with vocabulary size, leading to sparse data distribution and reduced model robustness. Furthermore, n-grams had limitations in addressing language-dependency.

To overcome these issues, word embedding techniques emerged, representing words as low-dimensional dense vectors. Each dimension of the vector represents a latent word feature, and linguistic patterns are encoded within the vector. Word vectors can be semantically composited, allowing for calculations like vec('France') - vec('Paris') + vec('Japan') ≈ vec('Tokyo'). Cosine similarity can measure the similarity between word vectors.

Pagolu et al. (2016) utilized n-grams and word2Vec to represent tweet text and employed random forest algorithms to measure the correlation between sentiment and stock price movement. Similarly, Garcia-Lopez et al. (2018) examined the impact of tweets on stock market trends using BoW and word embedding representations. While BoW vocabulary size improved, tweet vectors generated by Word2Vec with 300 dimensions outperformed BoW representation. This highlighted the strong semantic capability of word embeddings and led to their widespread adoption in subsequent studies, Vargas et al. (2017) utilized the Word2Vec model to embed Reuters' financial news titles into sentence embeddings and combined them with seven technical indicators to predict the movement of the S&P 500 index. Similarly, Huynh et al. (2017) applied Word2Vec embedding to a bidirectional gated recurrent unit (GRU) to predict the movement of the S&P 500. In a different context, Yun et al. (2019) employed the Skiagrams model to transform Korean sentences into vectors, which were then used as input to a CNN model for forecasting stock prices after five trading days.

Rather than assigning each word to a distinct vector using word embedding, dos Santos Pinheiro and Dras (2017) adopted character-level embedding to capture sub-word information in news for predicting S&P 500 movements. Character-level embedding has the advantage of preserving word morphology information and mitigating the issue of out-of-vocabulary (OOV) words missing from the training corpus. Additionally, some studies focused on embedding entire sentences or paragraphs for prediction. For example, Akita et al. (2016) applied paragraph vectors to embed newspaper articles, and Matsubara et al. (2018) employed paragraph vectors to embed online financial news, using a Deep Neural Generative (DNG) model for prediction.

The quality of articles can significantly impact the performance of text-based financial forecasting. However, online content often poses challenges in terms of its unstable quality, trustworthiness, and comprehensiveness. Hu et al. (2018) addressed these issues with their proposed Hybrid Attention Network

(HAN), which incorporates news-level attention to capture important information and temporal attention to account for time-varying effects, ensuring effective and efficient learning.

One issue associated with word embedding approaches is the long-range dependency problem. Compressing the entire information of a long text, such as a document, into a fixed vector without loss of information is challenging. Irrelevant text noise can also impact the quality of word vectors. To mitigate these challenges, several studies, such as Vargas et al. (2017), dos Santos Pinheiro and Dras (2017), and Matsubara et al. (2018), focused on analyzing the news titles instead of the content to reduce noise.

Duan et al. (2018) pointed out that news titles often provide a short description and ignore crucial information. To address this, they employed the news abstract as the target to weigh sentences in the news content and generate a target-specific abstract-guided representation of news documents, reducing noise while preserving important information.

Liu et al. (2018) proposed the Hierarchical Complementary Attention Network (HCAN) with two attention mechanisms: word-level attention to quantify the significance of words in news titles and content, and sentence-level attention to quantify the importance of each sentence. The title representation and content representation were concatenated to predict stock price movement. Liu and Wang (2018) aimed to enhance the interconnection between market and news data by introducing the Numerical-Based Attention

(NBA) method to align news embeddings with the stock vector. This approach assigns more weight to news impact on relevant stocks.

Recently, Xu, Cao and Li (2022) proposed the SELFGAN (Self-regulated Generative Adversarial Network) to predict the stock price movement. SELFGAN integrated the cooperative network and a generative adversarial network (GAN) to overcome the stochasticity and overfitting problems associated with stock price and financial text information. The cooperative network and GAN work together to reduce the stochasticity in both data sources and improve the model's generalization ability. The experimental results demonstrate that the SELFGAN achieves state-of-the-art performance in stock price movement prediction compared to existing methods. Table 2.4 summarized the finding of Semantic-based financial forecasting studies.

Table 2.4 Summary table of semantic-based stock forecasting approach

Author	Model	Method
Akita et al. (2016)	LSTM	Applied paragraph vector method to embed the newspaper articles
Vargas et al. (2017)	RCNN	Applied word2vec to generate word vectors of the news title and combine with technical indicator vector for prediction.
Hu et al. (2018)	HAN	Applied two attention modules (1) news level attention and (2) temporal attention to better capture asymmetric news and time dependency effect
Liu et al. (2018)	NBA	Proposed (NBA) method to improve the dual sources interaction between market and news data for stock market prediction.
Duan et al. (2018)	Document embedding model	Using news abstract as target model to weight the sentence in news content and generate a target specific abstract guided news documents representation
Liu et al. (2018)	HCAN	Proposed HCAN model with two attention mechanism (1) word-level attention and (2) sentence-level attention to retain important text information and mitigate the noise of irrelevant text.
Yun et al. (2019)	BiGRU	Applied word2vec to obtain the word vectors of the financial news before adapting (BiGRU) for prediction

2.4.3 Event extraction based financial forecasting approach

The event extraction approach is a specific application of NLP in financial forecasting that focuses on retrieving essential event information from text and representing it in a structured form. Unlike the semantic-based approach that processes entire sentences, paragraphs, or documents, event extraction aims to distill vital information and reduce irrelevant text noise. Xiang and Wang (2019) define an event as a specific occurrence happening at a particular time and place, involving one or more people, and resulting in a change of state. The task of event extraction involves detecting event-related sentences, identifying event triggers (keywords indicating a specific type of event), and, if present, determining the event type and its arguments.

In essence, event extraction summarizes unstructured natural language into a structured set of linked relations that can answer the "5W1H" questions ("who, when, what, where, why, and how") about a real-world event. This structural representation of events can be further utilized for logical reasoning and inference. Ding et al. (2014) argue that news events can impact investor sentiment, trigger trading actions, and influence stock movements. Event extraction finds applications in various areas of business and finance, including rapidly discovering market responses, providing trading signals, and conducting risk analysis (Yang et al., 2018).

In previous research, Nuij et al. (2013) used the ViewerPro system to extract company events from Reuters news articles related to the FTSE 50 stock index. The system filtered out irrelevant news and identified events through pattern matching in a domain-specific knowledge repository. Subsequently, Ding et al. (2014) represented structured news events as tuples of (Actor-Action-Object-Time) using the Open Information Extraction (Open-IE) approach to predict the movement of the S&P 500.

However, the predictive performance of models based on one-hot feature vectors representing structured event tuples is limited by the sparsity issues of discrete features in statistical models. This limitation motivated Ding et al. (2015) to enhance their study by representing structured events with dense vector event embeddings. Specifically, word embeddings for the "Agent," "Predicate," and "Object" components were extracted from raw text and combined to create the structured event embedding (A, P, O). The objective of event representation learning is to ensure that similar events are embedded close to each other in the vector space, while different events are further apart.

As a result, event embeddings led to better stock market prediction compared to the original discrete event-based approach proposed by Ding et al. (2014) because the embeddings capture structural relationships through semantic compositionality. Furthermore, they found that a CNN model performs better at capturing the long-term effects of events. Subsequently, the technique of structured event embedding was applied in the study by Nascimento and Cristo (2015) to forecast the movement of the S&P 500 index based on financial news from Reuters and Bloomberg. However, their study received criticism from Wang et al. (2019) for neglecting events related to small companies due to limited relevant reporting. Oncharoen and Vateekul (2018) combined the merits of event embedding (Ding et al., 2015) with news headlines from "Reuters," "Reddit," and "Intrinio," as well as the same set of technical indicators used by Vargas et al. (2017) and historical prices to improve the predictability of the movement of the S&P 500 and DJIA indices.

On the other hand, the limitations of the event embedding approach proposed by Ding et al. (2015) were addressed in their later study (Ding et al., 2018), where it was acknowledged that events with similar semantics and syntax might not have similar word embeddings. The approach was constrained by the assumption that events with similar word embeddings would have similar semantics. For instance, the event embeddings "Peter quits Apple" and "Steve Jobs leaves Apple" reflect a significant semantic difference as "Peter" is a customer while "Steve Jobs" is the CEO of Apple. However, the event embeddings do not capture these semantic differences. To enhance the quality of event embeddings, Ding et al. (2018) incorporated a knowledge graph into the training phase to encode background information.

Additionally, Wang et al. (2019) highlighted that previous studies overlooked certain event characteristics, which could significantly impact predictive performance. They identified four event properties: "imbalanced distribution of events," "inconsistent effect of events," "distinct importance of events," and "temporal effect of events." The imbalanced distribution of events refers to the tendency of financial news to report more on large enterprises than small enterprises, leading to sparse or dense event dictionaries for different stocks. The effect of events is inconsistent and diverse across industries, with the same news potentially having positive effects on one industry and negative effects on another. The magnitude of events' impact varies, highlighting the need to distinguish the significance of events. Finally, events have different long-lasting effects, indicating causal relationships or dependencies among events.

To address these event characteristics, Wang et al. (2019) proposed the event attention network (EAN) to leverage sentimental event embeddings. The EAN captures the simultaneous effects and sentiment properties of events to improve predictions for the stock trends of 20 different companies in the Hong Kong and Shenzhen markets. Event information is extracted from sources such as "Finet," "Tencent News," and "Sina News" and structurally embedded as (Time-Location-Name-Action). The attention mechanism distinguishes the importance of specific events, while a bidirectional LSTM with CNN layer captures sequential dependencies and extracts stock-driven feature representations. Sentiments are analyzed from social media platforms like "East Money," "Facebook," and "Twitter," classified into six dimensions (Happy, Vital, Kind, Sure, Calm, Alert), and their inclusion significantly improves predictive results.

Furthermore, Chen et al. (2019) criticized the coarse-grained event structures, such as (Subject, Predict, Object) proposed by Ding et al. (2014) and Zhang et al. (2018), for potentially omitting specific semantic information about different types of events. To address this, they introduced the Japanese financial event dictionary (TFED), which automatically extracts fine-grained events from financial news. TFED specifies the types of financial events, their corresponding trigger words, and event structures. For example, trigger words like "acquisition," "merge," and "acquire" detect M&A events, while "fund" and "funding" indicate funding events. The event details are further extracted in their corresponding structures (Firm-TimeMethod) and (Who-Action-Target-Time-Location). The Multi-task Structured Stock Prediction Model (MSSPM) was introduced to jointly learn event extraction and stock prediction, given the high correlation between these tasks.

Additionally, Daiya et al. (2020) introduced the Att-DCNN (Dilated Causal Convolution Networks with Attention) to generate event-knowledge embeddings, which capture direct and inverse relationships among events, and consider financial indicators for predicting the index based on the S&P 500. Their study outperformed previous baseline models with an accuracy of 72.23%.

Xu et al. (2021) introduces a framework called REST (Relational Event-driven Stock Trend Forecasting) for predicting stock trends. It addresses the limitations of existing methods by considering the stock-dependent influence of event information and the cross-stock influence between related stocks. REST incorporates a stock context model and a propagation layer on a stock graph to capture these influences. Table 2.5 summarizes the finding of the event extraction-based stock forecasting studies.

Table 2.5 Summary table of event extraction-based stock forecasting approach

Author	Model	Contribution
Ding et al. (2014)	E-NN	First attempt to represented news events in a structured tuple of (Actor-Action-Object-Time) based on Open-IE approach
Ding et al. (2015)	EB-CNN	Improve the discrete structure events into dense vector structured event embeddings (Agent, predicate, object)
Ding et al. (2018)	KGEB-CNN	Incorporated knowledge graph into training phase to encode background information to improve the representation of event embedding
Wang et al. (2019)	EAN	Incorporated sentiment information and adapted the attention mechanism to model the asymmetric event impact and time dependency effect to improve the prediction
Chen et al. (2019)	SSPM	Proposed a novel Japanese financial event dictionary (TFED) to extract fine-grained structured events information and fuse with news text to generate structure-aware text representation. SSPM model employed the structured events as distant supervised label to further training a multi-task framework for both event extraction and stock prediction
Daiya et al. (2020)	Att-DCNN	Proposed (Att-DCNN) model to better learning the direct and inverse relationship between events to improve the event-knowledge embedding for prediction

2.5. History of Malaysia' Stock Exchange

According to Arshad and Yahya (2016), the Malaysian stock exchange was established in the early 1930s and was initially known as the Malayan Stock Exchange (MSE). In the 1960s, the MSE changed its name to the Stock Exchange of Malaysia (SEM) and served as a common trading floor for Malaysia and Singapore. However, in 1973, the stock exchanges of both countries were separated due to the discontinuation of the exchange rate interchangeability between the Malaysian Ringgit and the Singapore Dollar. As a result, the Malaysian stock exchange was renamed the Kuala Lumpur Stock Exchange (KLSE), while the Singapore stock exchange became the Stock Exchange of Singapore. In 2004, the KLSE was rebranded as Bursa Malaysia to regulate its trading activities. As reported by Salim (2022), there were a total of 982 listed companies on Bursa Malaysia in 2022.

Furthermore, the top 30 listed companies in Malaysia contribute to the composition of the Kuala Lumpur Composite Index (KLCI), which is calculated using the market capitalization weighted method (Kwong et al., 2017). The KLCI serves not only as an investment indicator but also as an economic indicator reflecting the stability of the Malaysian economy (Alzaid, 2016).

Given the significance of market forecasting for investors and policymakers, it is crucial to conduct research in this area. In this study, the primary focus is on the KLCI index due to the limited number of studies on text-based financial forecasting in the Malaysian market.

Among the few sentiment-based financial forecasting studies in Malaysia, Rahman et al. (2017) extracted handcrafted features (n-grams and TF-IDF) from Malaysian news articles (The Edge) and used an SVM model for binary classification of stock movements. The average prediction accuracy of the study was found to be 56%. Shuhidan et al. (2018) proposed a hybrid lexicon-Naïve Bayes method to classify the sentiment polarity of financial news titles using lexicon-based sentiment word counts. Similarly, Kuan et al. (2019) classified the sentiment polarity of "relevant news" for the Malaysian company "Genting" using a Naïve Bayes model. The study achieved a sentiment classification accuracy of 68.75% and found a correlation of 58.41% between the historical stock price of Genting and the relevant news.

On the other hand, Lopez-Lira and Tang (2023) point that the quality of sentiment measure is the key factor to affect the performance of sentiment-based financial forecasting. Previous studies such as Rahman et al. (2017), Shuhidan et al. (2018), Kuan et al. (2019) reled on handcrafted features and lexicon-based sentiment analysis may limit the accuracy and adaptability of the models. The use of pre-defined sentiment lexicons might not adequately capture the subtleties and evolving nature of sentiment expressions in financial news. It is important to leverage more advanced natural language processing techniques, such as contextualized word representations, to enhance sentiment analysis performance. Besides, the studies primarily focused on binary classification of stock movements and sentiment polarity, which oversimplifies the complexity of financial markets. It is suggested to adapt more nuanced approaches, such as regression-based models, to capture the degree of sentiment impact on stock prices.

2.6 Conclusion

In this chapter 2, the study first introduced the framework of transfer learning before discussing the development of sequential transfer learning. The study illustrated the evolutional phrase of sequential transfer learning from earlier pretrained word embedding to recent contextual pretrained encoder. Afterward, the knowledge adaptation approach of fine-tuning method is discussed to adapt the pretrained model to different downstream NLP task. For example, the aspect-based sentiment analysis based on sequential transfer learning.

Apart from the methodology related review, the study discussed the recent text-based financial forecasting and classified the approaches into sentiment-based, semantic-based, event extraction-based and hybrid approaches. Lastly, the history of Malaysia' Stock Exchange is discussed to ensure the understanding on the relative forecasting target.

CHAPTER 3

METHODOLOGY

3.1. The Overall Framework and Data Collection

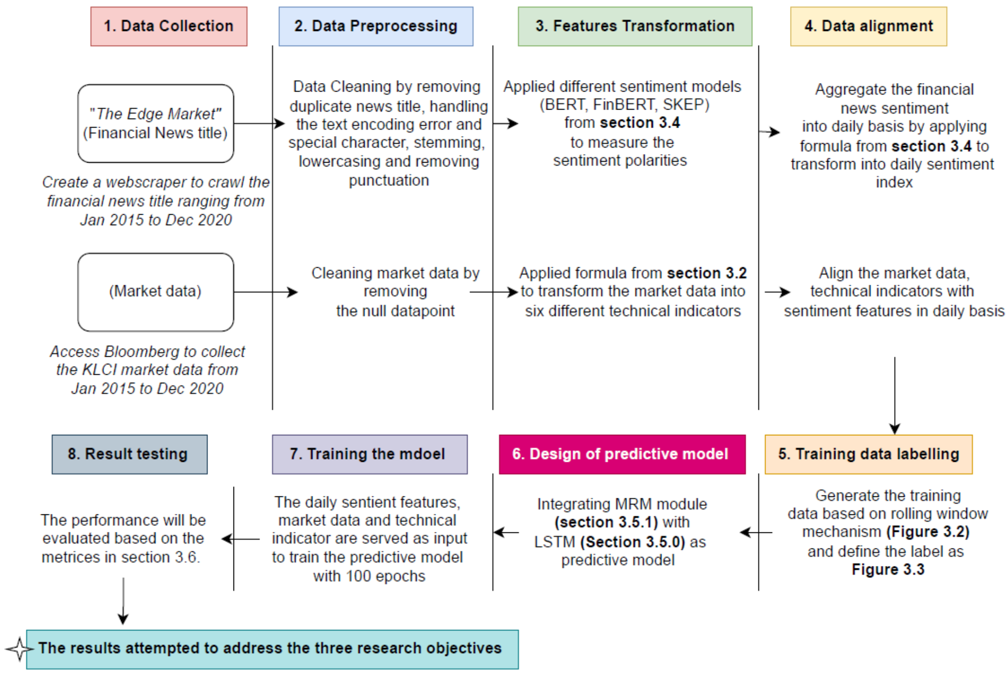

Figure 3.1 Proposed Sentiment-based financial forecasting framework

This study applies different pre-trained models to extract the financial news sentiment before integrating it with different technical indicators to forecast the KLCI index movement. As compared to traditional sentiment approaches, the pretrained models can capture the additional contextual information for semantic understanding. Besides, the insufficient of training data might degrade the performance of traditional machine learning model.

Therefore, this study deploys the pretrained model for better initialization to ensure the model generalization.

In Figure 3.1, the overall sentiment-based financial forecasting framework are exhibited to enhance the understanding of the methodological flow. First, the study collected the financial news text data and the market data correspondingly before processing them into different features correspondingly. Afterward, the study transformed the market data into technical indicators (Prices features) based on the formulas reported in section 3.2. For news text data, the study preprocessed the financial news title to reduce noises. The preprocessing methods include of removing repeating and irrelevant news, cleaning unnecessary punctuations and URLS, lowering the case and tokenizing the news title.

Afterward, the study fine-tuned the pretrained models such as BERT, FinBERT and SKEP with financial phrase bank dataset (Malo et al., 2014) before inputting the preprocessed financial news title data to perform the sentiment classification task. The outputted sentiment polarities are standardized and transformed into daily sentiment features which showed in section 3.4. Besides, the study is required to label and generate the training data based on methodology showed in section 3.3. Finally, the market data, prices feature, and sentiment features are serving as inputs to the predictive model (section 3.5) for final stock movement prediction.

Afterward, the prediction results are evaluated with the metrices shown in section 3.6. The sentiment-features based model (BERT, FinBERT and SKEP) is compared to the baseline MRM-LSTM (Chan et al. 2022) in order to examine the effectiveness of sentiment features to improve the forecasting performance. Then, it followed with two different pair of comparison between sentiment-features-based model (FinBERT- BERT) and (SKEP-BERT) to examine the different between different sentiment models.

After introducing the overall research framework, the study would like to discuss the data collection process. First, the study constructs a web scraper to automatically extract the *news title*, *reporting time,* and *categories tags* from the Malaysian financial news websites, “The Edge market”. The local news source, “*The Edge Market*” are recommended by Rahman et al., (2017) and Shuhidan et al. (2018) due to its creditability as one of the long-standing local financial news in delivering reliable information to investor. It retained the track record of news and provide a relatively depth coverage across local and international markets. However, the news content may consist of irrelevancies to disrupt the accuracy of sentiment measure, the study merely focuses on using the news title to evaluate the sentiment features.

On the other hand, the market data used in this study is the KLCI index which was retrieved from Bloomberg. The market data included the open, low, high, closed and volume. This study employed a span of 6 years from 1 January 2015 to 31 December 2020. The news and market data are collected in daily interval. Afterward, the market data is further employed to compute the six technical indicators shown in section 3.2. According to Chan et al. (2022a), technical indicator may suffer from the multicollinearity problem, thereby the multicollinearity reduction module (Chan et al., 2022b) is proposed to ensure the forecasting performance.

3.2 Technical indicators

This section attempted to discuss the market data preprocessing method in transforming the market data into technical indicators (price features). There are a total of six different technical indicators namely Simple Moving Average (SMA), Cumulative Moving Average (CMA), Exponential Moving Average (EMA), Stochastic Oscillator, Williams %R, and Bollinger Bands are adopted in accordance with Shynkevich et al. (2017). These suggested technical indicators enable to reflect the insight of the market trends such as the potential entry and exit points, momentum, overbought/oversold conditions, and volatility.

The moving averages provide a smooth representation of the overall trend and aid in identifying potential support/resistance levels and trend reversals (Chan et al. 2022b). The Stochastic Oscillator and Williams %R offer insights into overbought and oversold conditions, helping traders anticipate potential reversals. Bollinger Bands provide information on volatility, indicating periods of contraction and potential breakouts. The subsection below provided the explanation of each feature together with their corresponding formula.

3.2.1 Simple Moving Average (SMA)

The SMA is an unweighted moving average of the closing price of the preceding n days. The investor can determine the n period based on the desired trend period, such as short-term, medium-term, or long-term trends. Also, the SMA line can be referred as the stock's support or resistance level which often used in conjunction with SMA lines of other period to identify whether the stock is in an uptrend or downturn (Chan et al. 2022b). For example, a stock price above the SMA signaled an uptrend and, as a result, a buy signal.

$$SMA = \frac{p_1 + p_2 + p_3 + \cdots p_n}{n}, \qquad (1)$$

where p_n is the price at period n and n is the total number of periods in a fixed window.

3.2.2 Cumulative Moving Average (CMA)

CMA is a moving average where current average computed based on the cumulated value of all data until the current data point (Chan et al., 2022b).

$$CMA_n = \frac{p_1 + p_2 + p_3 + \cdots p_n}{n}, \qquad (2)$$

where p_n is the price at n period, and n is the total number of periods.

3.2.3 Exponential Moving Average (EMA)

The EMA is one of the weighted moving averages. It enables the latest data with a highest weight, while exponentially lowering the weight for older data (Chan et al., 2022b). The formula of EMA is exhibited as below,

$$S_t = \alpha \times Y_t + (1 + \alpha) \times S_{t-1} , \qquad (3)$$

Where,

α : Parameter of weight

Y_t : Observation at time t.

3.2.4 Stochastic Oscillator

A stochastic oscillator is a momentum indicator that uses the price range over a period. The current price is a percentage of the recent highest and lowest price. The indicator is to capture the signal of prices tend to approach the extremes of the range before turning. The following is the formula of the oscillator (Shynkevich et al., 2017).

$$\%K = 100 \frac{closing\ price - L}{H - L} , \qquad (4)$$

$$\%D = 3\ period\ moving\ average\ of\ \%K ,$$

3.2.5 William %R

The William %R referred to an oscillator ranging from −100 to 0. The indicator tells whether the stock price is trading near the highest or lowest in past recent period. When the value comes to −100, it implied the closing price is approaching the lowest in past n days (Shynkevich et al., 2017).

$$\%R = \frac{(high)\ - close}{-min\ (low)} * \ -100 \tag{5}$$

3.2.6 Bollinger Band

The Bollinger Band is a price-based indicator implied of a peaks and valleys in relative to price. It usually consists of an upper and lower band. When the price approximating either band, it indicates a potential reversal in trend (Shynkevich et al., 2017).

$$Bollinger\ High = SMA_n + 2 * \sigma_n\ , \tag{6}$$

$$Bollinger\ Low = SMA_n - 2 * \sigma_n\ ,$$

Where,

SMA : Simple moving average

σ : Standard deviation

n : Number of days in the smoothing period.

3.3. Data Generation

After introducing the transformation method of market data, this subsection 3.3 attempted to discuss the labelling or generating training data for the forecasting model. In this study, the major reason of applying rolling window mechanism is to construct the training samples for training the model since the stock price is majorly affected by the lag information. Figure 3.2 illustrated an example of rolling window mechanism in a time series of 10-time steps. Based on the window size setting of 3-time steps, a total of eight rows of data will be generated in which the first row of data is from t =1 to t=3, second row of data is from t =2 to t =4 and so on. This study defaulted the window size setting of 15 to employ a lag of 15 trading days to label and generate training sample. The forecasting problem is a quaternion classification task where the training sample is labelled in 4 different categories of price level as training target (Figure 3.3). Based on figure 3.3, the labeling methodology of four-classification outcome are defined as below:

Label 0: Predicted price is lower than the pre-defined bottom threshold price level. (Significant loss)

Label 1: Predicted price is higher than the pre-defined upper threshold price level. (Significant profit)

Label 2: Predicted price is lower than current price but within the pre-defined threshold price level. (loss)

Label 3: Predicted price is higher than current price but within the pre-defined threshold price level. (profit)

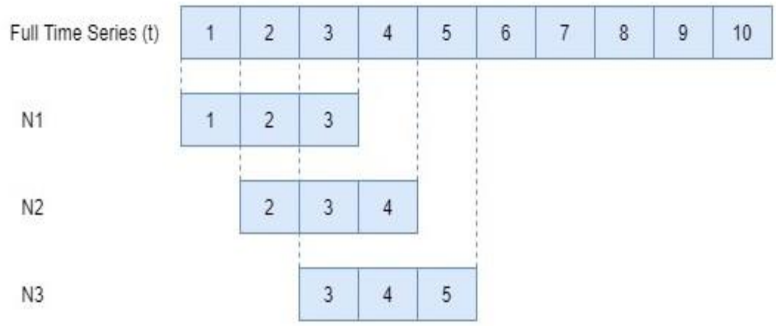

Figure 3.2 Illustration of the rolling window mechanism.

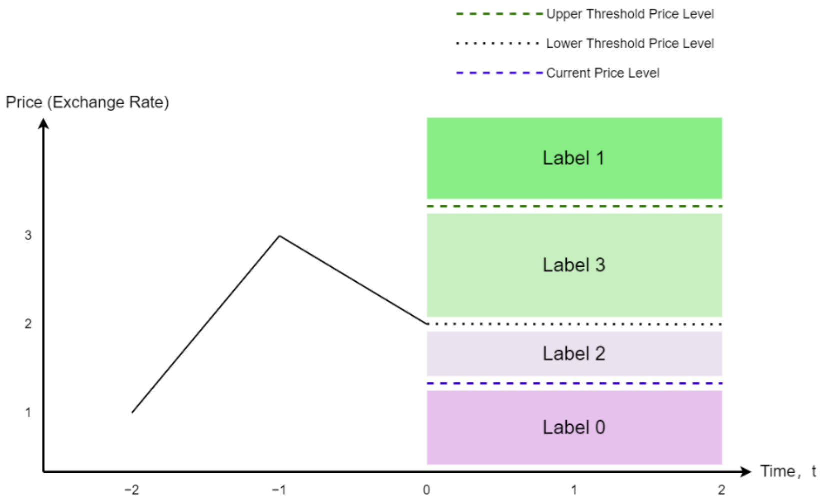

Figure 3.3 Illustration of dataset labeling methodology.

3.4 Sentiment Analyzer Module

For this subsection, the study attempted to introduce different pretrained model known as BERT, FinBERT and SKEP which applied as sentiment model in this study. The selection of the pretrained models is based on their unique properties in specific financial domain-oriented and sentiment-oriented. While the BERT-base model with (12 encoder layers, 768 hidden size, 12 self-attention head and 110M parameter) is employed as a baseline to examine the potential improvement of the specific pretrained models. First, the study required to fine-tune each of the pretrained models with financial phrase bank dataset (Malo et al., 2014).

After the fine-tuning process, this study applied an output layer with the softmax activation function on the pretrained encoder to classify the sentiment of news title into a trinary class (positive, neutral, and negative). The outputted sentiment polarities are further computed into different sentiment features such as daily simple sentiment index, log sentiment index and moving average of sentiment index before combining with technical indicator as inputs to the predictive model. In short, the input data will be used are the sentiment polarities, simple sentiment index, log sentiment index, moving average of simple and log sentiment index. The formula of simple sentiment index and log sentiment index are as below.

$$Simple\ Sentiment\ Index = \frac{n^{postive} - n^{negative}}{n^{postive} + n^{negative}} \quad (7)$$

$$Log\ Sentiment\ Index = \log\frac{1 + n^{postive}}{1 + n^{negative}}$$

3.5 Predictive Model Framework

In past literature, recurrent neural network (RNN) has achieved many great results in different real real-world applications that involve of sequential data. Based on Young et al. (2018), the term “recurrent” explained the general architecture idea where an identical function is applying on each element of the sequence while the output of the previous element will be aggregately retained over the internal memory” of RNN until the end of sequence. Based on this, RNN enables to compress the information and produce a fixed-size vector to represent a particular sequence. The recurrence operation of RNN is advantageous in time series data since the temporal information of a sequential can be effectively captured. RNN is also flexible to model a variable length of a sequence to capture unbounded contextual information. However, RNN might be suffered from the issues of vanishing and exploding gradients.

Therefore, LSTM has proposed by Schmidhuber and Hochreiter (1997) to address the corresponding problem in standard RNN by using the gate mechanism to selectively preserve or forget the previous information. The recurrent structure of LSTM is more complicated than RNN where each memory cell of LSTM consists of the output, input, and forget gates to control the read, write, and reset operations respectively over its internal state. The gate operation may ensure longer preservation for the state of the memory cell to mitigate the long-range dependencies issue of RNN. The diagram of LSTM unit cell is exhibited in figure 3.4 and the formula of LSTM cell unit is as below.

$$f_t = \sigma\left(W_f[h_{t-1}, x_t]\right), \tag{8}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t]),$$

$$o_t = \sigma(W_o[h_{t-1}, x_t]),$$

$$Q_t = f_t \odot Q_{t-1} + i_t \odot tanh(W_C[h_{t-1}, x_t]),$$

$$h_t = o_t \odot tanh(Q_t).$$

Where, $f_t, i_t, o_t \in R$ are the sigmoid gate unit for forget gate, input gate and output gate respectively at time t. At time t, three different gates unit processed on previous hidden state h_{t-1} and current input, x_t before applying the sigmoid (σ) function to output a value between 0 to 1. When sigmoid output of zero, it implies the forget gate to ignore previous information. For input gate and output gate, it decides which value to be update. $Q_t \in R^m$ is the cell state at time t, and $W_f, W_i, W_o \in R^{m \times (m+V)}$ are the parameters.

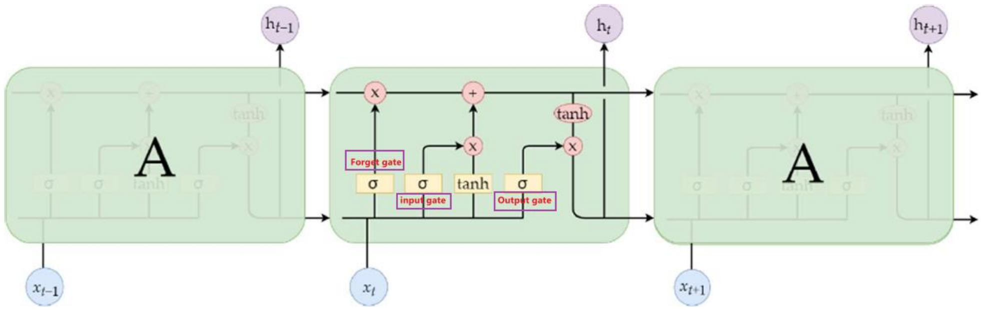

Figure 3.4 The diagram of LSTM cell

However, Bahdanau et al. (2014) criticized that the recurrent-based model may be problematic in handling the long-range dependencies in data due to the memory compression issue in which the neural network is difficult to compress all the necessary information from a long sequence input into a fixed-length vector. In order words, the fixed-length vector is difficult to represent the entire input sequence without any information loss. Despite the help of the gated activation function, the forgetting issues of RNN-based model is getting serious as the length of input sequence is growing. Based on this, the attention mechanism is proposed to deal with the long-range dependencies issue by enabling the model to focus on the relevant part of input sequence when predicting a certain part of the output sequence.

According to Zhang et al. (2018), the idea of attention mechanism is first stimulated by the human visual attention where human usually adjust their focal point over time to perceive a “high resolution” when focusing a particular region of an image but perceive a “low resolution” for the surrounding image. Thus, attention mechanism enables the model to learn to assign different weight in accordance with the input contribution to capture the asymmetric influence among inputs. In this study, the LSTM model is adopted as the predictive model and the attention mechanism is also applied to capture the asymmetric effect from different features.

3.5.1 Multicollinearity Reduction Module (MRM)

In addition to LSTM, this study proposed to enhance the model capability based on the Multicollinearity Reduction Module (MRM) developed from attention mechanism. The MRM framework has been applied in our published works (Chan et al., 2022b) in mitigating the multicollinearity problem and achieve a great result in forecasting the forex. The attention mechanism enables the model to learn the relevance of each feature in contributing to final prediction. The attention module composed by a linear layer, dropout layer, sigmoid function and followed with another linear layer and SoftMax layer. The SoftMax function is to ensure all the weightings sum equal to one.

Afterward, the study adopted the correlation embedding to measure the relative relationship between variables. Correlation measures the strength of relationship between two variables. The correlation value ranging from -1.0 to 1.0, where 0 indicate the variables are having no relationship. The correlation matrix is then processed by a single neural network layer and output the correlation embedding where the size is equal to input. The formula of correlation is exhibited as below.

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} * C_{jj}}}, \tag{9}$$

Where

R_{ij} : Correlation coefficient of X_i and X_j

C_{ij} : Covariance matrix of X_i and X_j

C_{ii} : Variance of X_i

C_{jj} : Variance of X_j

The general framework of predictive model (LSTM-MRM) is exhibited in figure 3.5. Specifically, the study first input the data, X_t with the shape of (batch, sequence length, input size) to the input attention module, which output the weight of each feature. The weight is then applied on the input data to generate the weighted input data. The weighted input is then multiplied with the correlation embedding (*Cr*) before passing into the LSTM layer. Based on the input attention module with correlation embedding, the model enables to learn the relevance on each feature. Afterward, the LSTM output is then processed by two more linear layers to reduce the hidden size before output the prediction of classes.

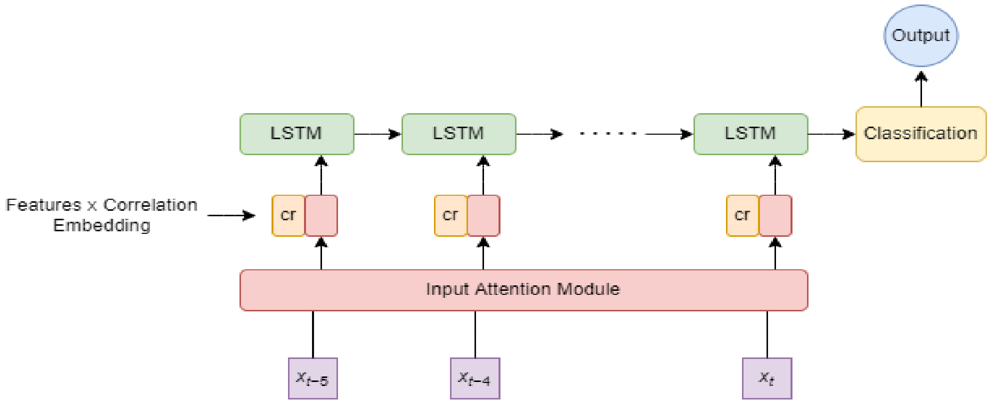

Figure 3.5 Framework of Predictive Model (LSTM-MRM)

3.6 Performance evaluation metrices

3.6.1 Accuracy.

Accuracy refers to the ratio of correct prediction made by the model over the total number of predictions. This metric is relatively effective when the classes in the dataset are balanced with an equal number of instances for each class (Goutte et al. 2015). However, accuracy can be misleading in imbalanced datasets as it assumes equal cost for both false positives and false negatives, which often leads to overestimating the performance of a model that largely predicts the majority class while ignoring the minority class. The formula of accuracy as below.

$$\text{Accuracy} = \frac{True\ positives + True\ negatives}{True\ positives + True\ negatives + False\ negatives + False\ positives} \tag{10}$$

3.6.2 Precision

Precision measures the proportion of correctly predicted positive observations out of the total predicted positive observations. It measures the relevancy of the classified results and usually implied the classifier's exactness. However, it is highlighted that the result with high precision is possible to be associate to a poor recall, as precision is indifferent to false negatives (Goutte et al. 2015). Therefore, it is a better way to evaluate the precision in conjunction with recall or F1 score to obtain a comprehensive view of performance evaluation. The formula of precision as below.

$$\text{Precision} = \frac{True\ positives}{True\ positives + \ False\ positives} \tag{11}$$

3.6.3 Recall

Recall can be known as sensitivity measure or true positive rate to measure the proportion of predicted positive observation that are correctly identified as such. In other words, it quantifies the ability of a model to find all the relevant cases within a dataset. However, recall does not consider of false positives where high recall rate does not necessarily indicate good performance in the cases that consisted of high false positives. Therefore, it is better way to evaluate the recall in conjunction with F1-score to get a comprehensive performance evaluation. (Goutte et al. 2015). The formula of recall as below.

$$\text{Recall} = \frac{True\ positives}{True\ positives + \ False\ negatives} \tag{12}$$

3.6.4 F1-score

F1-score is the harmonic mean of precision and recall which being a comprehensive metric to evaluate the model performance. F1-score is particularly useful in dealing with imbalanced datasets where the score is ranging from 0 to 1. The higher F1 score indicates the model with both high precision and high recall to avoid the type 1 and 2 errors (Goutte et al. 2015). The formula of F1-score as below.

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{13}$$

and hence, tends to favor models that have similar precision and recall metrics. This is a particularly useful feature when dealing with imbalanced datasets. The F1 score reaches its best value at 1 (which represents perfect precision and recall) and worst at 0.

3.6.5 Trading return

Trading return refer to the gain or loss made from a trade which is a key metric to evaluate the performance of a trading system. Positive trading return indicate a profitable effective trading strategy, vise versa. The formula of trading returns as below.

$$Return = \frac{Selling\ price - Purchase\ price}{Purchase\ price} \tag{14}$$

3.6.6 T-test

According to Jones, Schlomer & Wiggs (2014), the t-test, also known as Student's T-Test, is a statistical procedure used for hypothesis testing. It involves comparing values on continuous variables between two groups or samples. Researchers collect raw data and apply the T-test to determine its significance level, indicated by the p-value. The p-value reflects the likelihood of obtaining the observed outcomes by chance. If the p-value falls below certain thresholds such as 0.01 (1%), 0.05 (5%), or 0.1 (10%), the researchers reject the null hypothesis (H0) and conclude a significant relationship between the explanatory and responding variables.

The hypothesis test performed as below:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Decision rule: Reject H_0, if the p-value of T-test is less than or equal to the alpha value else do not reject H_0.

3.7 Conclusion

In short, this chapter illustrated the overall methodology of this research. The study collected the data in terms of news textual data and numerical market data. Both data are preprocessed and transformed into different features such as sentiment index and technical indicator before integrating them as input to the predictive model for forecasting. Besides, the LSTM-MRM is being selected as the predictive model in this study. The model enabled to concentrate on important features in capturing the asymmetric impact between the features.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Data Description

News data	BERT	FinBERT	SKEP
Number of Raw News	97554	97554	97554
Number of Processed News	82056	82056	82056
Number of Positive News	22229	25683	29518
Number of Negative News	15726	21006	32280
Number of Neutral News	44101	35367	20258
Label	**Number of observations**		
0 (Significant loss)	812		
1 (Significant profit)	535		
2 (Loss)	10		
3 (Profit)	36		
Total processed market data observations	1393		

Table 4.1 Summary of data description

Based on Table 4.1, there is a total of 97554 financial news been crawled from the website of "*The Edge Market*". After pre-processing the news data by removing the duplicate news, encoding error news and the undesirable length of news. Each of the three models (BERT, FinBERT, and SKEP) has 82,056 articles available for analysis across 6 years data span. These articles have been labelled based on their sentiment, which can be positive, negative, or neutral.

The sentiment analysis results show that BERT identified 22,229 positive articles, 15,726 negative articles, and 44,101 neutral articles. FinBERT identified 25,683 positive articles, 21,006 negative articles, and 35,367 neutral articles. SKEP had the highest number of positive articles, with 29,518, along with 32,280 negative articles and 20,258 neutral articles.

On the other hand, the market data been labelled with four different categories based on the rolling window and labelling method in section 3.3. The labels represent significant loss, significant profit, loss, and profit. There were 812 observations of significant loss, 535 observations of significant profit, 10 observations of loss, and 36 observations of profit, totalling 1,393 processed market data observations. Figure 4.1 showed the example of the processed news and market data.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	date	open	high	low	close	volume	negative	positive	neutral	SMA_5	SMA_10	SMA_20	CMA	EMA	%K	%D	%R	Bollinger High	Bollinger Low	BI_Simple	BI	BI_MA	BI_Simple_MA	profit/loss	target
2	1/30/2015	-3.23	-5.43	2.09	-0.92	1.61E+08	4	7	10	-4.364	3.769	1.4245	1.188	-2.478	65.63	73.1	-44.74	2.45822668	0.390773324	0.272727	0.47	0.0624	0.036363636	15	1
3	2/4/2015	26.36	42.02	19.83	21.76	2.06E+08	5	7	8	1.316	4.971	3.32	2.069	5.601	71.34	68.77	-33.1	5.76046891	0.879531088	0.166667	0.288	0.0288	0.016666667	-10.93	0
4	2/5/2015	-10.93	-27.16	-4.68	0.07	1.42E+08	6	7	10	-0.016	5.298	4.3255	1.892	3.758	71.41	69.46	-35.42	3.30570138	5.345298621	0.076923	0.134	0.105	0.057692308	-9	0
5	2/6/2015	2.52	10.05	9.66	6.35	1.45E+08	2	7	13	2.712	3.935	5.013	1.999	4.622	77.82	73.52	-32.44	1.91076247	8.115237529	0.555556	0.981	0.1338	0.075555556	-9	0
6	2/9/2015	9.31	4.01	-0.91	2.14	1.18E+08	1	6	13	5.88	2.983	4.176	1.925	3.795	76.92	75.38	-37.96	3.34601886	5.005981144	0.714286	1.253	0.056	0.033928571	-9	0
7	2/10/2015	-2.88	-1.69	3.53	-0.46	1.44E+08	10	4	9	5.972	0.804	3.934	1.759	2.376	76.34	77.03	-38.84	2.50825816	5.359741835	-0.42857	-0.79	-0.189	-0.102857143	-9	0
8	2/11/2015	-1.76	-4.54	-17.9	-12.2	1.38E+08	5	6	13	-0.814	0.251	3.1935	1.178	-2.472	59.4	70.89	-62.14	-0.13328972	6.520289718	0.090909	0.154	0.0377	0.023376623	-20.29	0
9	2/12/2015	-20.29	-21.08	-10.6	-9.88	1.75E+08	4	3	11	-2.804	-1.41	2.0085	0.741	-4.942	37.49	57.74	-81.05	-0.30833113	4.32533113	-0.14286	-0.22	-0.101	-0.057142857	15	1
10	2/13/2015	0.99	9.95	10.1	11.88	1.45E+08	7	7	9	-1.698	0.507	2.947	1.1	0.666	41.69	46.19	-58.31	-0.66166686	6.555666858	0	0	-0.018	-0.011111111	15	1
11	2/16/2015	13.26	8.76	9.93	7.94	1.36E+08	2	10	11	-0.538	2.671	3.1945	1.291	3.09	56.89	45.36	-43.98	-0.29668319	6.68568319	0.666667	1.299	0.1076	0.054166667	15	1
12	2/17/2015	3.47	1.19	4.28	1.2	81090400	5	5	13	-0.206	2.883	3.326	1.246	2.46	59.19	52.59	-41.64	-1.92033606	8.572336062	0	0	-0.047	-0.027272727	15	1
13	2/18/2015	-1.87	5.05	1.65	-2.22	62964300	2	3	2	1.784	0.485	2.728	1.099	0.9	54.94	57.01	-27.4	-1.95316938	7.409169378	0.2	0.288	0	0.003333333	15	1
14	2/23/2015	4.47	-3.2	-1.22	1.52	1.22E+08	0	7	4	4.064	0.63	2.964	1.079	1.107	57.85	57.33	-23.41	-4.91087345	10.83887345	1	2.079	0.1946	0.092307692	15	1
15	2/24/2015	1	6.49	7.06	9.29	1.42E+08	4	3	4	3.546	0.924	2.4295	1.288	3.835	75.14	62.64	-1.435	-0.13404175	4.993041748	-0.14286	-0.22	-0.12	-0.06984127	-9	0
16	2/25/2015	6.21	-0.89	-3.36	-2.82	1.4E+08	9	4	13	1.394	0.428	1.7055	1.134	1.616	69.63	67.54	-8.66	0.55526897	2.855731031	-0.38462	-0.69	-0.195	-0.10989011	-9	0
17	2/26/2015	-4.34	3.32	4.25	5.01	1.17E+08	4	5	12	2.156	0.975	0.8895	1.21	2.748	98.07	80.95	-1.93	2.40692664	-0.62792664	0.111111	0.182	0.0971	0.053968254	-9	0
18	2/27/2015	8.39	3.98	-1.16	0.34	1.85E+08	10	4	13	2.668	2.226	1.2385	1.154	1.945	90.23	85.98	-9.771	2.24700004	0.229999963	-0.42857	-0.79	-0.094	-0.051948052	-9	0
19	3/2/2015	0.64	-0.32	0.61	-4.08	1.35E+08	3	6	12	1.548	2.806	0.698	0.986	-0.063	81.25	89.85	-24.11	1.29202967	0.103970326	0.333333	0.56	0.0783	0.047619048	-9	0
20	3/3/2015	-2.48	0.46	5.59	4.12	1.5E+08	2	8	12	0.514	2.03	1.2685	1.041	1.331	90.04	87.17	-17.77	1.74368428	0.793315716	0.6	1.099	0.1099	0.06	-9	0
21	3/4/2015	5.23	0.26	1.24	4.29	1.42E+08	1	9	15	1.936	1.665	2.168	1.096	2.317	98.89	90.06	-2.369	0.45633023	3.879669775	0.8	1.609	0.031	0.013333333	-9	0
22	3/5/2015	-6	-6.38	-13.4	-19.5	2.01E+08	8	8	8	-2.956	-0.4	1.2415	0.568	-4.938	56.46	81.79	-94.6	-2.98093901	5.463939015	0	0	0	0	-14.8	0

Figure 4.1 Example of processed market and news sentiment data.

4.2. Performance Results

The study proposed to simulate the experiment multiple times and selecting the best-performing models for comparison is a recommended approach to mitigate the impact of random initialization in training a neural network from scratch. Due to the presence of multiple local minima in the optimization landscape, different initializations can lead to significantly different solutions (Du et al., 2019). By conducting multiple simulations, we increase the chances of finding an average solution closer to the global minimum. Selecting the best ten models for comparison allows us to assess the stability and generalization abilities of the network while reducing the influence of individual random initialization, providing more reliable insights into the network's performance.

Thus, this study trained each model with 100 epochs and conducted the experiment with thirdly simulation. Afterward, the best ten simulations from each model are retrieved to evaluate the performance with different metrics such as accuracy, precision, recall, F1-score, and return. Every simulation contributed a vary result, the high deviation between the results might cause of bias. Based on this, this study proposed an idea to conduct a large scale of stimulations and select their best ten for comparison. The intention was to ensure the results robustness in making a fair comparison between different models.

Specifically, each of the model was trained on the prior 80% of data while the remaining 20% of data was used for test set. The reason of not allocating the validation dataset is due to the limit amount of data where 80% of data accounting 1115 data points have been used for training. It is highlighted that the lack of training data is insufficient for the model to learn the pattern. Table 4.2 showed the simulation results of each model form test set.

Table 4.2 Accuracy, F1 and Returns for each model.

Mean	Baseline without Sentiment LSTM-MRM (Chan et al. 2022)	Sentiment-based models		
		BERT	FinBERT	SKEP
Accuracy	55.80%	55.64%	54.50%	57.54%*
Precision	54.11%	54.00%	51.60%	57.11%
Recall	52.00%	56.12%	57.36%**	56.86*
F1	49.51%	51.20%	52.42%**	53.59%**
Return	207.75	219.82	247.47	336.10***

Notes: *, **, and *** denote as the significance level of 10%, 5%, and 1% respectively for the t-test. The null hypothesis indicates the improvement of metric is insignificant.

Based on Table 4.2, the study to attempted to discuss the result in two different perspectives to address our research objective and question, which are (1) analysis of sentiment features, and (2) analysis of sentiment model. There are a total five metrics have showed in table 4.2 while the most significant three metrices will be majorly discussed which are return, accuracy and f1-score in evaluating the impact of sentiment features in trading performance and the impact of different sentiment models.

4.3 Result Analysis

4.3.1 Analysis of sentiment features

Based on Figure 4.2 and Table 4.2, it is observed that the average trading return for the baseline model which is LSTM-MRM without sentiment features is 207.75, while the average trading return for LSTM-MRM with different sentiment-based features are 219.82, 247.47, and 336. 10, respectively. This implied that the profits are having a respective 5%, 19.29%, and 61.78% improvement over five years. Although the inclusion of sentiment features yielded an overall return improvement, that is not a case that only one of the results is statistically significant. Nevertheless, the result significant at the 1% level implied the improvement reliability and supported the incremental effect of sentiment features in trading return.

Figure 4.2 Bar-chart of return in comparing to baseline

In terms of accuracy, the baseline LSTM-MRM without sentiment features model yielded the mean of 55.80%, while the sentiment-based models attained the average results of 55.64%, 54.50%, and 57.54%, respectively. Accuracy is the metric to evaluate the correct prediction of each label over the total number of predictions. Figure 4.3 illustrates the accuracy of each model in a box plot. Based on the result, the inclusion of sentiment features might not show a noticeable improvement in accuracy as compared to baseline, except for the SKEP model with a minor 1.74% improvement and statistically significant at 10%. However, a question remains where BERT and FinBERT models with lower accuracy than baseline but yield higher returns.

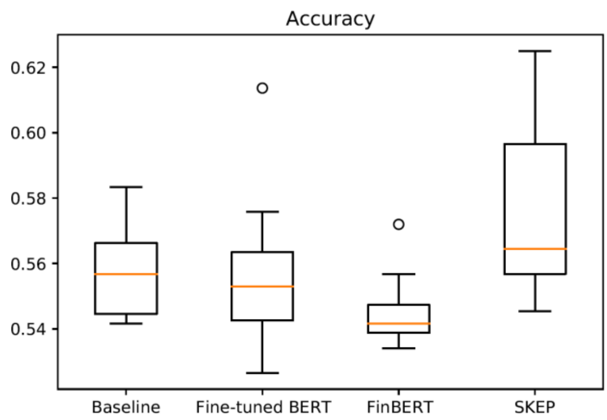

Figure 4.3 Plot-box of accuracy for each of the model

Based on the question above, it is essential to discuss another evaluation metric, such as the F1 score, to ensure the robustness of the result analysis. In classification tasks, there are two other vital metrics, namely precision, and recall. Precision refers to the number of true positives out of the total number of true positives and false positives. In other words, it told how much the classified results were relevant and usually implied the classifier's exactness.

In this experiment, the true positive is a correctly identified profitable signal (Classification of label 1), and the false positive is the case that the model incorrectly labels as a profitable signal, in fact, a non-profit or a worse case of loss signal. The high precision rate implied the model's capability to identify the exact profitable signal due to the number reduction of false positives. Based on this, there is a limitation of using accuracy to determine the model profitability because the model will only enter a trade when the model predict class 1 which is a profitable signal.

Another evaluation metric namely, “recall” which referred to the number of true positives out of the total number of true positives and false negatives. In other words, it was telling how much the total relevant results were correctly classified by the model and usually implied the classifier completeness. In this experiment, the true positive is correctly identified profitable signal, and the false negative would be the signal that the model labels as not profitable, in fact, a profitable signal. The high recall rate implied the model's capability to identify the profitable signal due to the number reduction of false negatives.

In short, precision and recall are relatively important as they will affect the total trading return. However, that is not the case to maximize both precision and recall at the same time. There is a trade-off between precision and recall where the increase in recall might result in a decrease in precision and vice versa. For example, the model can achieve recall near to 1 by completely identifying all the signals as profitable to reduce the number of false negatives near to 0, so the model does not miss any entry opportunity. On the other hand, it will result in extremely low precision due to the tremendous increase in false positives.

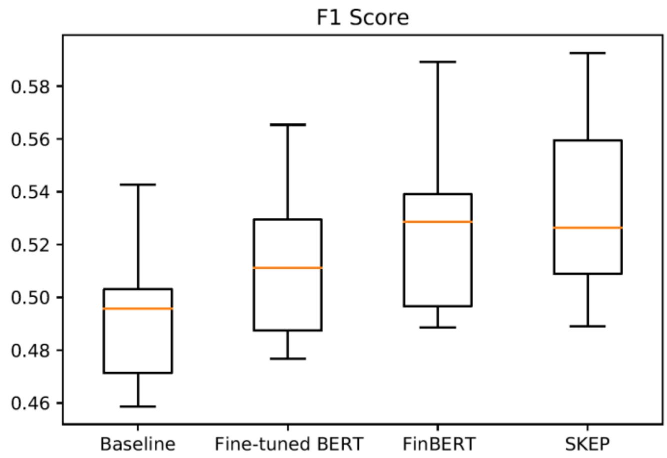

Figure 4.4 Plot-box of F1-score for each of the model

In this case, the F1 score is introduced to optimize the balance between recall and precision. F1 score is a measure of harmonic mean between recall and precision. In our experiment, the F1 score is more suitable for becoming the evaluation metric since precision and recall are crucial in contributing to the trading return profit. Figure 4.4 illustrates the F1-score of each model in a box plot. In terms of F1-score, the baseline model yielded the mean result of 49.51%, while the sentiment-based models attained the mean results of 51.20%, 52.42%, and 53.59%, respectively. It is obvious to show a general improvement in F1 score over baseline. The results of FinBERT and SKEP are found to be statistically significant at a 5% level. Based on this, the higher F1 score reasonably explains the lower accuracy but higher return generated. In short, it addressed the first research question where the sentiment features provided incremental improvement over the trading return.

4.3.2 Analysis of sentiment model

Based on the result shown in table 4.2, the study observed that different sentiment models influence the trading return differently. The result is not surprising since different sentiment models usually with different properties to generate distinct sentiment features. Therefore, the quality of sentiment feature has been regarded as one of the significant concerns in sentiment-based stock forecasting. In practice, sequential transfer learning has achieved state-of-the-art results in the task of sentiment analysis. However, the large number of pre-trained models available confuses the researcher in determining the sentiment model.

Therefore, this study examined the impact of three sentiment models, namely BERT, FinBERT, and SKEP, on trading returns. The study employed the accuracy, F1 score, and trading return as the evaluation metrics to discuss the following impact. In the first comparison pair between BERT and FinBERT, the result of FinBERT showed a minor improvement over the return and F1 score of BERT despite a slight fall in accuracy. In addition, the F1 score of FinBERT is statistically significant over the BERT model compared to baseline, which indicated the stability improvement on that metric.

Although the BERT model has been fine-tuned on the financial phrase bank dataset, it does not perform well as FinBERT. The possible reason might be the different pretraining text corpora since FinBERT was pre-trained on Reuter's financial texts while BERT was pre-trained on Wikipedia and BookCorpus. As a result, both models learned different extent of financial domain knowledge and processed the financial news differently. In short, it addressed the second research question where the pre-trained model with financial domain knowledge provided an incremental improvement over the general pre-trained model in sentiment-based financial forecasting.

In the second comparison pair between BERT and SKEP, it showed a noticeable improvement in the accuracy, F1 score, and return. All three metrics of SKEP are statistically significant over the BERT model compared to baseline, which indicated the improvement in the evaluation stability. The possible reason might be because of the difference in pretraining tasks. SKEP is mainly pre-trained on sentiment-oriented objectives such as (1) sentiment masking, (2) sentiment word prediction, (3) word polarity prediction, and (4) aspect-sentiment pair prediction. While BERT pre-trained with the general objective of (1) masked language model (MLM) and (2) next sentence prediction (NSP). In short, it addressed the third research question where the pre-trained model with sentiment-oriented knowledge provided an incremental improvement over the general pre-trained model in sentiment-based financial forecasting.

4.4 Conclusion

In short, this chapter reported the experimental results and provided the explanation of the results. The study trained the model based on the 80% of data and testing the result on the remaining 20% of data. Each of the model trained with 100 epochs and retrieved the best ten simulations out of a total of thirdly. Every simulation contributed different of results, the average result of the best ten simulations enabled the performance to be compared in par and avoid bias. The results addressed the research objectives and claiming an incremental improvement by sentiment features in forecasting the market movement. Besides, the result of SKEP and FinBERT are outperforming the generic BERT model which claiming the important of sentiment and financial domain knowledge.

CHAPTER 5

CONCLUSION

5. Conclusions

This study explores the potential of sequential-transfer learning and news-based sentimental information to improve the forecasting result toward the KLCI market index. The study first employed the web scraper to extract the daily news title from Malaysian financial news websites. Afterward, the study fine-tuned the pretrained models as sentiment classifier to analysis the news title. The consideration of news title is mainly due to the news content that might consist of irrelevancies to disrupt the accuracy of sentiment measure.

Afterward, the result of classifiers in sentiment polarities are further retrieved to compute the sentiment features such as daily sentiment index and moving average of sentiment index. The sentiment features are integrated with the technical indicators as inputs to the predictive model to classify the KLCI movement. The forecasting problem is formulated as a quaternion classification tasks which are (significant loss, loss, profit, significant profit).

The study evaluates the results based on the three most significant metrices which are accuracy, F1-score, and precision. In overall, the sentiment features showed a general improvement on the baseline model and the improvement are statistically significant. This provided evidence where sentiment features provide incremental improvement in forecasting the market.

In addition, this study explores the impact of different pretrained model in contributing the forecasting performance of KLCI index. There is a total of three pretrained models, namely BERT, FinBERT and SKEP models are compared. Our experiment shows a minor improvement of FinBERT over the fine-tuned BERT models in terms of return and F1. The possible reason might be the different of pretraining text corpora since FinBERT pretrained on financial text is likely to learn more on the financial domain knowledge.

The result of second comparison pair shows a noticeable improvement of SKEP over fine-tuned BERT in terms of accuracy, F1 score and return. All three metrics of SKEP are statistically significant over the baseline of BERT model. Both results suggested the essential criteria of involving financial domain knowledge and sentimental properties in determining of pretrained model for financial downstream application.

For the research contribution, the result of the study provided solid evidence to claim the effectiveness of pretrained model in measuring the sentiment features to improve the predictability of stock market movement. The method provided way to alleviate the problem of lacking label data. Unlike traditional machine learning method to train the model on scratch, researcher enabled to fine-tune the pretrained model with limited data and lesser of time to achieve a better performance in sentiment classification. In addition, the sentiment features are suggested to be included to improve the performance of stock forecasting.

Moreover, another finding from this research that provided evidence to support the useful of financial domain knowledge of FinBERT over the generic BERT model in financial forecasting despite of minimal improvement made. From the comparative result from SKEP, the finding support that the quality of sentiment features is significantly affecting the predictive performance. The finding also supports the effectiveness of SKEP in generating a better sentiment-word representation to improve the performance of sentiment-related task.

For the limitation of the study, the study only considers one news sources, "*The edge market"* out of a branch of news sources such as "The Star", "New Straits Times", "Bloomberg", etc. Based on this, the trustworthily and reliability of the news sources is relative important. This is because of the reporter preferences and their subjective opinion might affect the quality of news and cause of bias. Furthermore, the study only considers of the news text sources and omits of other text sources such as financial report, message boards, social media, professional periodicals, corporate disclosure, etc. Future studies are suggested to explore more on different text sources.

Furthermore, the study merely considers of the news title and omit the body content of the news. This is because of the limitation on the computing power as well as the significant challenges in addressing the noise of text. Although of the body content might carried of noise, it remained the potential opportunity to address additional information. Some of the news title is short where the important message is hiding in the body content and causing the news title to be meaningless. The study is forced to remove such the news title which cause of the information waste. Future work is likely to explore the text summarization approach to retrieve essential important message from body content to avoid noise.

Future work of interest is to address the aspect-based sentiment stock forecasting. For example, opinion text is usually composed with the characteristic of diversity to express a varied opinion on different aspects. Due to computational efficiency, current sentiment-based stock forecasting studies mainly focus on measuring text sentiment at either document or sentence level. Although the emerged word embedding and deep learning approaches provide a way to globally analyse the semantical and syntactical text information, ignoring such fine-grained sentiment information is critical and might cause misleading. Such as “I love stock A but hate stock B.” The general sentiment model is limited in distinguishing the correct opinion on its corresponding aspect. Thereby, the general evaluation is a constraint and might not be accurately analysed. Based on this, a future sentiment-based forecasting study is suggested in looking forward to aspect-based sentiment analysis.

References

Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K. (2016, June). Deep learning for stock prediction using numerical and textual information. In *2016 IEEE/ACIS 15th* analysis and opinion mining. In *Proceedings of the Seventh International Conference International Conference on Computer and Information Science (ICIS)* (pp. 1-6). IEEE.

Alzaid, S. (2016). The Kuala Lumpur Stock Exchange Composite Index (KLSE CI) and economic forces. *South East Asia Journal of Contemporary Business, Economics and Law*, *10*(3), 53-64.

Arshad, M. N., & Yahya, M. H. (2016). Relationship Between Stock Market Returns and Exchangerates In Emerging Stock Markets. *Ikonomika: Jurnal Ekonomi dan Bisnis Islam*, *1*(2), 131-143.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). Sentiwordnet 3.0: An enhanced lexical resource for sentiment *on Language Resources and Evaluation (LREC'10)*.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137-1155.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1-8.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Chan, J. Y. L., Bea, K. T., Leow, S. M. H., Phoong, S. W., & Cheng, W. K. (2022c). State of the art: a review of sentiment analysis based on sequential transfer learning. *Artificial Intelligence Review*, 1-32.

Chan, J. Y. L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., & Chen, Y. L. (2022a). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics*, *10*(8), 1283.

Chan, J. Y. L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., ... & Chen, Y. L. (2022b). A Correlation-Embedded Attention Module to Mitigate Multicollinearity: An Algorithmic Trading Application. *Mathematics*, *10*(8), 1231.

Chen, D., Zou, Y., Harimoto, K., Bao, R., Ren, X., & Sun, X. (2019). Incorporating fine-grained events in stock movement prediction. *arXiv preprint arXiv:1910.05078*.

Chen, W., Cai, Y., Lai, K., & Xie, H. (2016, January). A topic-based sentiment analysis model to predict stock market price movement using Weibo mood. In *Web Intelligence* (Vol. 14, No. 4, pp. 287-300). IOS Press.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, *12*(ARTICLE), 2493-2537.
Dadgar, S. M. H., Araghi, M. S., & Farahani, M. M. (2016, March). A novel text mining approach based on TF-IDF and Support Vector Machine for news classification. In *2016 IEEE International Conference on Engineering and Technology (ICETECH)* (pp. 112-116). IEEE.
Daiya, D., Wu, M. S., & Lin, C. (2020, May). Stock movement prediction that integrates heterogeneous data sources using dilated causal convolution networks with attention. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8359-8363). IEEE.
Deng, S., Huang, X., Zhu, Y., Su, Z., Fu, Z., & Shimada, T. (2023). Stock index direction forecasting using an explainable eXtreme Gradient Boosting and investor sentiments. *The North American Journal of Economics and Finance*, *64*, 101848.
Deng, S., Xiao, C., Zhu, Y., Tian, Y., Liu, Z., & Yang, T. (2022). Dynamic forecasting of the Shanghai Stock Exchange index movement using multiple types of investor sentiment. *Applied Soft Computing*, *125*, 109132.
Day, M. Y., & Lee, C. C. (2016, August). Deep learning for financial sentiment analysis on finance news providers. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 1127-1134). IEEE.
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
Ding, B., Wang, Q., Wang, B., & Guo, L. (2018). Improving knowledge graph embedding using simple constraints. *arXiv preprint arXiv:1805.02408*.
Ding, X., Zhang, Y., Liu, T., & Duan, J. (2014, October). Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1415-1425).
Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015, June). Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.
dos Santos Pinheiro, L., & Dras, M. (2017, December). Stock market prediction with deep learning: A character-based neural language model for event-based trading. In *Proceedings of the Australasian Language Technology Association Workshop 2017* (pp. 6-15).
Du, S., Lee, J., Li, H., Wang, L., & Zhai, X. (2019, May). Gradient descent finds global minima of deep neural networks. In *International conference on machine learning* (pp. 1675-1685). PMLR.
Duan, J., Zhang, Y., Ding, X., Chang, C. Y., & Liu, T. (2018, August). Learning target-specific representations of financial news documents for cumulative abnormal return prediction. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2823-2833).
Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179-211.
Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, *25*(2), 383-417.

Fellbaum, C. (1998). A semantic network of English verbs. *WordNet: An electronic lexical database*, *3*, 153-178.
Garcia-Lopez, F. J., Batyrshin, I., & Gelbukh, A. (2018). Analysis of relationships between tweets and stock market trends. *Journal of Intelligent & Fuzzy Systems*, *34*(5), 3337-3347.
Green, J., & Zhao, W. (2022). Forecasting earnings and returns: A review of recent advancements. *The Journal of Finance and Data Science*.
Goutte, C., & Gaussier, E. (2005, March). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European conference on information retrieval* (pp. 345-359). Berlin, Heidelberg: Springer Berlin Heidelberg.
Gupta, R., & Chen, M. (2020, August). Sentiment Analysis for Stock Price Prediction. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* (pp. 213-218). IEEE.
Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, *55*(3), 685-697.
Hao, Y., Dong, L., Wei, F., & Xu, K. (2019). Visualizing and understanding the effectiveness of BERT. *arXiv preprint arXiv:1908.05620*.
Harris, Z. (1954). Distributional Hypothesis. *Word*, *10*(23), 146-162.
Henry, E. (2008). Are investors influenced by how earnings press releases are written?. *The Journal of Business Communication (1973)*, *45*(4), 363-407.
Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
Hu, Z., Liu, W., Bian, J., Liu, X., & Liu, T. Y. (2018, February). Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 261-269).
Huynh, H. D., Dang, L. M., & Duong, D. (2017, December). A new model for stock price movements prediction using deep neural network. In *Proceedings of the Eighth International Symposium on Information and Communication Technology* (pp. 57-62).
Jones, A., Schlomer, G., & Wiggs, C. B. (2014). T-Tests How To Guide. *University of Arizona*, 1–5.
Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.
Kuan, L. C., Ismail, M. A., Zayet, T. M., & Shuhidan, S. M. (2019, December). Prediction of Malaysian stock market movement using sentiment analysis. In *Journal of Physics: Conference Series* (Vol. 1339, No. 1, p. 012017). IOP Publishing.
Kwong, S. M., Tan, H. S., Tan, H. S., Tan, X. Y., & Tung, M. Y. (2017). Determinants of stock market performance in Malaysia (Undergraduate dissertation, Universiti Tunku Abdul Rahman). Retrieved from http://eprints.utar.edu.my/2449/1/FN-2017-1306066.pdf
Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Landauer, T., Laham, D., & Foltz, P. (1997). Learning human-like knowledge by singular value decomposition: A progress report. *Advances in neural information processing systems*, *10*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Li, Q., & Shah, S. (2017, August). Learning stock market sentiment lexicon and sentiment-oriented word vector from stocktwits. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (pp. 301-310).

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences*, *278*, 826-840.

Liang, B., Luo, W., Li, X., Gui, L., Yang, M., Yu, X., & Xu, R. (2021, October). Enhancing aspect-based sentiment analysis with supervised contrastive learning. In *Proceedings of the 30th ACM international conference on information & knowledge management* (pp. 3242-3247).

Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiment, and Emotions. *Cambridge: Cambridege University Press 2015*.

Liu, G., & Wang, X. (2018). A numerical-based attention method for stock market prediction with dual information. *Ieee Access*, *7*, 7357-7367.

Liu, H., & Singh, P. (2004). ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal*, *22*(4), 211-226.

Liu, Q., Cheng, X., Su, S., & Zhu, S. (2018, October). Hierarchical complementary attention network for predicting stock price movements with news. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 1603-1606).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lo, A. W. (2004). The adaptive markets hypothesis. *The Journal of Portfolio Management*, *30*(5), 15-29.

Lopez-Lira, A., & Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, *66*(1), 35-65.

Luss, R., & d'Aspremont, A. (2015). Predicting abnormal returns from news using text classification. *Quantitative Finance*, *15*(6), 999-1012.

Malkiel, B. G. (1973). *A Random Walk Down Wall Street [By] Burton G. Malkiel*. Norton.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782-796.

Mao, H. H. (2020). A survey on self-supervised pre-training for sequential transfer learning in neural networks. *arXiv preprint arXiv:2007.00800*.

Maqsood, H., Mehmood, I., Maqsood, M., Yasir, M., Afzal, S., Aadil, F., ... & Muhammad, K. (2020). A local and global event sentiment based efficient stock exchange forecasting using deep learning. *International Journal of Information Management*, *50*, 432-451.

Matsubara, T., Akita, R., & Uehara, K. (2018). Stock price prediction by deep neural generative model of news articles. *IEICE TRANSACTIONS on Information and Systems*, *101*(4), 901-908.

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems* (pp. 6294-6305).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: from lexicons to transformers. *IEEE Access*, *8*, 131662-131682.

Mittermayer, M. A., & Knolmayer, G. F. (2006, December). Newscats: A news categorization and trading system. In *Sixth International Conference on Data Mining (ICDM'06)* (pp. 1002-1007). Ieee.

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019, April). Stock Price Prediction Using News Sentiment Analysis. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 205-208). IEEE.

Nascimento, J. B., & Cristo, M. (2015, October). The impact of structured event embeddings on scalable stock forecasting models. In *Proceedings of the 21st Brazilian Symposium on Multimedia and the Web* (pp. 121-124).

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603-9611.

Nuij, W., Milea, V., Hogenboom, F., Frasincar, F., & Kaymak, U. (2013). An automated framework for incorporating news into stock trading strategies. *IEEE transactions on knowledge and data engineering*, *26*(4), 823-835.

Oliveira, N., Cortez, P., & Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, *85*, 62-73.

OpenAI. (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774v3*

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016, October). Sentiment analysis of Twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)* (pp. 1345-1350). IEEE.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345-1359.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Picasso, A., Merello, S., Ma, Y., Oneto, L., & Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems with Applications*, *135*, 60-70.
Qian, Y., Li, Z., & Yuan, H. (2020). On exploring the impact of users' bullish-bearish tendencies in online community on the stock market. *Information Processing & Management*, *57*(5), 102209.
Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872-1897.
Qiu, Y., Song, Z., & Chen, Z. (2022). Short-term stock trends prediction based on sentiment analysis and machine learning. *Soft Computing*, *26*(5), 2209-2224.
Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.
Rahman, A. S. A., Abdul-Rahman, S., & Mutalib, S. (2017, November). Mining textual terms for stock market prediction analysis using financial news. In *International Conference on Soft Computing in Data Science* (pp. 293-305). Springer, Singapore.
Ren, R., Wu, D. D., & Liu, T. (2018). Forecasting stock market movement direction using sentiment analysis and support vector machine. *IEEE Systems Journal*, *13*(1), 760-770.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533-536.
Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
Salim, S. (2022, July 19). Bursa Malaysia's total number of listed companies rises to 982, highest since 2008. *The Edge Markets*. Retrieved from https://www.theedgemarkets.com/article/bursa-malaysias-total-number-listed-companies-rises-982-highest-2008
Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, *304*, 114135.
Schmidhuber, J., & Hochreiter, S. (1997). Long short-term memory. *Neural Comput*, *9*(8), 1735-1780.
Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, *27*(2), 1-19.
Schütze, H. (1992). Word space. *Advances in neural information processing systems*, *5*.
Seifollahi, S., & Shajari, M. (2019). Word sense disambiguation application in sentiment analysis of news headlines: an applied approach to FOREX market prediction. *Journal of Intelligent Information Systems*, *52*(1), 57-83.
Shah, D., Isah, H., & Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. *International Journal of Financial Studies*, *7*(2), 26.
Shuhidan, S. M., Hamidi, S. R., Kazemian, S., Shuhidan, S. M., & Ismail, M. A. (2018, March). Sentiment analysis for financial news headlines using machine

learning algorithm. In *International Conference on Kansei Engineering & Emotion Research* (pp. 64-72). Springer, Singapore.

Shynkevich, Y., McGinnity, T. M., Coleman, S. A., Belatreche, A., & Li, Y. (2017). Forecasting price movements using technical indicators: Investigating the impact of varying input window length. *Neurocomputing*, *264*, 71-88.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X. (2013, August). Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 24-29).

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014, June). Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL (1)* (pp. 1555-1565).

Tang, Y., Song, Z., Zhu, Y., Yuan, H., Hou, M., Ji, J., ... & Li, J. (2022). A survey on machine learning models for financial time series forecasting. *Neurocomputing*, *512*, 363-380.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance, 62*(3), 1139-1168.

Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., ... & Wu, F. (2020). SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis. *arXiv preprint arXiv:2005.05635*.

Vargas, M. R., De Lima, B. S., & Evsukoff, A. G. (2017, June). Deep learning for stock market prediction from financial news articles. In *2017 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA)* (pp. 60-65). IEEE.

Wang, H., Wang, T., & Li, Y. (2020, April). Incorporating Expert-Based Investment Opinion Signals in Stock Prediction: A Deep Learning Framework. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 01, pp. 971-978).

Wang, K., Shen, W., Yang, Y., Quan, X., & Wang, R. (2020). Relational graph attention network for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.12362*.

Wang, Q., Xu, W., & Zheng, H. (2018). Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. *Neurocomputing*, *299*, 51-61.

Wang, Y., Li, Q., Huang, Z., & Li, J. (2019, June). EAN: Event attention network for stock price trend prediction based on sentimental embedding. In *Proceedings of the 10th ACM conference on web science* (pp. 311-320).

Wei, Y. C., Lu, Y. C., Chen, J. N., & Hsu, Y. J. (2017). Informativeness of the market news sentiment in the Taiwan stock market. *The North American Journal of Economics and Finance*, *39*, 158-181.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., ... & Patwardhan, S. (2005, October). OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations* (pp. 34-35).

Xiang, W., & Wang, B. (2019). A survey of event extraction from text. *IEEE Access*, *7*, 173111-173137.

Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, *50*(1), 49-73.

Xu, H., Cao, D., & Li, S. (2022). A self-regulated generative adversarial network for stock price movement prediction based on the historical price and tweets. *Knowledge-Based Systems*, *247*, 108712.

Xu, W., Liu, W., Xu, C., Bian, J., Yin, J., & Liu, T. Y. (2021, April). Rest: Relational event-driven stock trend forecasting. In *Proceedings of the Web Conference 2021* (pp. 1-10).

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, *53*(6), 4335-4385.

Yang, L., Xu, Y., Ng, J., & Dong, R. (2019, August). Leveraging BERT to improve the FEARS index for stock forecasting. In *The First Workshop on Financial Technology and Natural Language Processing, Macao, China, 12 August 2019*. ACL.

Yang, Y., UY, M. C. S., & Huang, A. (2020). FinBERT: A Pretrained Language Model for Financial Communications. *arXiv preprint arXiv:2006.08097*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems* (pp. 5753-5763).

Yekrangi, M. and Abdolvand, N. Financial markets sentiment analysis: Developing a specialized lexicon. *Journal of Intelligent Information Systems* **2021**, 57(1):127–146.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, *13*(3), 55-75.

Yu, L. C., Wu, J. L., Chang, P. C., & Chu, H. S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*, *41*, 89-97.

Yun, H., Sim, G., & Seok, J. (2019, February). Stock prices prediction using the title of newspaper articles with korean natural language processing. In *2019 international conference on artificial intelligence in information and communication (ICAIIC)* (pp. 019-021). IEEE.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1253.

Zhao, B., He, Y., Yuan, C., & Huang, Y. (2016, July). Stock market prediction exploiting microblog sentiment analysis. In *2016 International Joint Conference on Neural Networks (IJCNN)* (pp. 4482-4488). IEEE.