Forecasting Stock Market Movement Based on Political and Financial News Using Machine Learning: Evidence from Developed and Developing Countries

Ngoc Vo

A thesis report

requirements for the award of the degree of

MASTER OF DATA SCIENCE

CHARLES DARWIN UNIVERSITYCOLLEGE OF ENGINEERING,
INFORMATION TECHNOLOGY AND ENVIRONMENT

November 2021

DECLARATION

I hereby declare that the work herein, now submitted as an interim report for the degree of Master of Data Science at Charles Darwin University, is the result of my own investigations, and all references to ideas and work of other researchers have been specifically acknowledged. I hereby certify that the work embodied in this thesis report has not already been accepted in substance for any degree and is not being currently submitted in candidature for any other degree.

Signature: Ngoc Vo

Date: November 2021

ABSTRACT

Keywords: *News sentiment, Extreme Gradient Boosting, Sentiment analysis, Developed countries, Developing countries*

Market sentiment plays a crucial role in the predictability of stock market. Over the past ten years, researchers have consistently searched for different news types and their impacts on stock prediction. Furthermore, existing machine learning literature with regards to stock movement prediction has mostly concentrated on one single market, whose results may mask the true degree of conclusion to other markets. This research aims to fill these gaps by examining the predictive value of daily political news and firm-specific financial news on stock market movement prediction; as well as performing a cross-border study to investigate the impact of news sentiment in three countries- the US, Australia, and Vietnam from developed and developing economies. Sentiment analysis is integrated into an extreme gradient boosting algorithm (XGB) to predict stock trends for the next period. The prediction results are employed for portfolio selection using a mean-variance model. Empirical results confirm the predictive value of news sentiment is consistent across countries. In addition, the proposed portfolios are superior to benchmarks in terms of risk-adjusted returns. The research findings are valuable for stock market participants, and future researchers to expand the study sample based on these two types of news.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my supervisors, Dr Bharanidharan Shanmugam and Dr Fen Nee Chong for their brilliance, patience, and kindness, for without the encouragement and guidance, as well as sympathy at every point during the project, my thesis report would not have come to fruition. I am also immensely grateful to all my professors, faculties and teaching assistants who constantly supported me during my time of writing this thesis.

I would also like to thank my family for their unconditional love and support, especially throughout the pandemic. In addition, my student years would not have been fulfilled without my close friends Nguyet Nguyen and Hieu Nguyen, and my special friend who shall be anonymous, my classmates, and many other beloved friends that I made during the journey in Australia.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Forecasting Stock Market Movement Based on Political and Financial News using Machine Learning: Evidence from Developed and Developing Countries

Abstract— **Market sentiment plays a crucial role in the predictability of stock market. Over the past ten years, researchers have consistently searched for different news types and their impacts on stock prediction. Furthermore, existing machine learning literature with regards to stock movement prediction has mostly concentrated on one single market, whose results may mask the true degree of conclusion to other markets. This research aims to fill these gaps by examining the predictive value of daily political news and firm-specific financial news on stock market movement prediction; as well as performing a cross-border study to investigate the impact of news sentiment in three countries- the US, Australia, and Vietnam from developed and developing economies. Sentiment analysis is integrated into an extreme gradient boosting algorithm (XGB) to predict stock trends for the next period. The prediction results are employed for portfolio selection using a mean-variance model. Empirical results confirm the predictive value of news sentiment is consistent across countries. In addition, the proposed portfolios are superior to benchmarks in terms of risk-adjusted returns. The research findings are valuable for stock market participants, and future researchers to expand the study sample based on these two types of news.**

Keywords*— *News sentiment, Extreme Gradient Boosting, Sentiment analysis, Developed countries, Developing countries

I. Introduction

Background

For decades, the global financial industry has operated based on the foundation of Efficient Market Hypothesis (EMH) (Fama, 1970), where investors cannot make a prediction or beat the market as all currently available information is fully reflected in today's stock price, leaving no room to make profits by investing in the assets that are already and accurately priced. However, in the years following, researchers and practitioners have provided substantial works to confront Fama's view in both theoretical and empirical sides (Andrew, 1997; Dhar & Zhu, 2006; French, 1980; Keim, 1983; Malkiel, 2003). The main arguments are based on the grounds that the current capital market is dominated by uninformed and irrational players: not all market agents have access to all available information in a timely manner and the majority of players make investment decisions based on their own psychological biases (Malkiel, 2003**.** Therefore, investor irrationality and market anomalies could cause pricing irregularities and predicted patterns for short periods, when market participants are able to make arbitrage opportunities before asset price correction that "fully reflect all available information" (Malkiel, 2003). This leads to continuous efforts searching for the ultimate prediction methods under fundamental, technical, and technological approaches to maximize excess returns. The paper is concerned with the technological approach using machine learning modelling predictive movement in the stock market.

In order to achieve higher prediction accuracy, it is crucial to determine the appropriate features for modelling techniques since different features have different predictive purposes and impacts. The paper utilizes different methods of sentiment analysis, and machine learning models to study the predictive value of news sentiment. In addition, this research investigates the predictive value of daily political news, which have been rarely examined in prior work. To fully capture the modelling prediction, the full data set consists of historical stock price, technical indicators, and the combination of daily political news and firm-specific financial news. The robustness of the proposed model is further tested on two types of markets: developed countries (the US and Australia) and developing countries (Vietnam) to reveal the true degree of informational efficiency and prediction performance across multiple countries.

Sentiment analysis will be used to extract market sentiment in news data. This method provides the text's polarity by weighing up its positive and negative components (Feuerriegel & Prendinger, 2016). Among the different approaches for sentiment extraction, a dictionary-based approach is utilized in this study. Specifically, a financial lexicon dictionary created by (Loughran & McDonald, 2011) and Vader dictionary is used to produce more accurate sentiment classification in financial texts. The features are used as inputs to the machine learning model using extreme gradient boosting algorithm. In order to demonstrate the persuasiveness of empirical results, this paper incorporates prediction results into the portfolio selection process and adopts mean-variance optimization. The proposed portfolios have been tested to for a consistent performance overtime in both developed and developing markets.

Finally, the empirical results from prediction models and portfolio allocation are further discussed in light of the EMH. This supports machine learning literature to understand the reasons and feasibility of prediction results.

Aim of Research

In summary, the research project presents the following contributions:

- Exploring the predictive values of daily pollical news which have not been studied in prior work, to verify the impact on predicting models.
- Forecasting stock market movement using disparate data inputs including political news, business news and technical indicators.
- Incorporating news sentiment and prediction results for portfolio selection
- Conducting a cross-border study to investigate the consistency of the proposed methods over a period of time and across countries
- Providing evidence of the EMH from empirical evidence in the machine learning context.

Structure of Paper

The remainder of this paper is organized as follows. In Section II, market setting, and problem formulation are given through literature review. Section III describes the methodology in detail. In Section IV, data collection is described along with the experiment design. Section V presents empirical results. In Section VI, conclusions are drawn along with future work.

II. Literature Review

The section provides insights into several important fields of financial theories such as efficient market hypothesis, behavioural finance as well as stock prediction, machine learning and sentiment applications to address the research objectives. Further, the most significant literature papers over the past ten years have also been reviewed to highlight the research gap.

A. Efficient Market Hypothesis (EMH)

For decades, the global financial industry has operated based on the foundation of Efficient Market Hypothesis (EMH) (Fama, 1970), where investors cannot make a prediction or beat the market as all currently available information is fully reflected in today's stock price, leaving no room to make profits by investing in the assets that are already and accurately priced. The EMH presented by (Fama, 1995) implies that it is impossible to make such a prediction based on the random walk hypothesis. In specific, the price of a financial instrument is based on a coin flip (random walk), resulting in no overall trend of stock; therefore, it is impractical to predict a stochastic direction, let alone a magnitude changes (Fama, 1970; Malkiel, 1999). Asset prices are economically believed to adjust or change direction in response to news (Cutler et al., 1988) but market prices already reflect all current and available information (Fama, 1970), which altogether support the view of EMH that it is unnecessary and unfeasible for an attempt to forecast asset prices. (Fama, 1970) investigated from his empirical work with a proposal of three types of market efficiencies in terms of weak-form, semi-strong form, and strong-form, which are categorised by the information availability reflected on asset price. The notion behind this experiment is that practitioners are only able to yield profit i.e., excess risk-adjusted returns by making an investment decision based on available information that is not priced in the market (Fama, 1970). The first form studied information contained historical price, which is all reflected in the market prices; therefore, the future price cannot be predicted by past price, and traders (technicians) cannot earn any abnormal return. The second one implies that market price is already reflected on historical price and any publicly available information so that investors cannot use either of them to gain higher returns in the market. The strong form states that stock price fully reflects both public and private information so that traders (esp. insiders) cannot gain any profit from their trading (Fama, 1970).

Despite Fama's theory, academic researchers and practitioners have made numerous attempts to dispute such hypotheses. Several opponents perceive weak form and semi-strong efficient markets as the milder versions of the EMH relaxing the assumption of prices incorporating all relevant information (Malkiel, 2003). Specifically, critics of weak-form efficiency theory argue that investors can utilise publicly available information such as financial statements and corporate announcements to determine the intrinsic value of potential stocks. Likewise, investors who gain access to inside information- information that is not readily available to the public, can increase their chances of making higher-than-average profits (Malkiel, 2003). Behavioural finance also argues with the imperfections in financial markets, showing that the capital market is also driven by cognitive biases of market participants, such as overaction, overconfidence, and various other human errors that change asset's movements significantly (Fromlet, 2001; Kahneman & Tversky, 2013; Slovic, 1992). Market anomalies and investor irrationality has been further proved with persistent academic evidence, such as the weekend effect (French, 1980), and January effect (Keim, 1983), and disposition effect (Dhar & Zhu, 2006; Grinblatt & Keloharju, 2001), which cause mispricing for asset prices. According to (Andrew, 1997), a financial market exists "trend" so that it is predictable. In this line of reasoning, market participants are believed to earn an excess return by exploiting these anomalies.

B. Behavioral Finance

Behavioural finance proposes that market participants are not always rational and have cognitive biases such as information bias, loss aversion, overreaction, overconfidence, representative bias, and other biases in reasoning and information processing (Friesen & Weller, 2006). This is because investors' interpretation and behaviours towards information vary greatly, which leads to a significant impact on market dynamics and stock returns when new information is released (Bollen & Huina, 2011) (Robertson et al., 2006,). Therefore, the concept of behavioural finance provides strong evidence to substantial market anomalies like market corrections and financial recessions Wisniewski and Lambe, 2013.

C. Adaptive Market Hypothesis (AMH)

The applicability of market efficiency to different countries and stock markets has still been questionable and ongoing research topic with conflicting results for decades (Majumder, 2013). Therefore, Lo (2004) produced an evolved theory – the adaptive market hypothesis (AMH) to reconcile EMH and behavioural finance. The AMH states that markets evolve and adapt following economic cycles and important events at certain times. Hence, the degree of market efficiency varies at different times and countries (Lo, 2005) Charles et al. (2012), Hiremath and Narayan (2016). In this line of reasoning, the behaviour of stock returns and their predictability vary from time to time depending on the market conditions and other economic factors (e.g., unemployment rate, interest rate, and GDP). Recent researchers have provided several empirical evidence on the AMH in stock markets across the globe, including the studies of the US (Ito and Sugiyama, 2009) and (Kim et al.,2011), the UK and Japan Urquhart and Hudson (2013), and emerging markets Lim et al., 2006. Their empirical results investigated that the studied countries went through different levels of efficiency in different periods. In addition, (Lim et al., 2006) shows evidence of AMH in developed and emerging countries. Therefore, as the nature of AMH, this section implies that the results obtained from the stock prediction in different markets cannot be concluded solely based on the overall classification of the markets but based on the classification in a particular time for each market. This leads to inconsistent results in the study of market efficiency, especially in the studied countries, including the US, Australia, and Vietnam.

Research has shown that there is no consensus on the type of market efficiency of the US stock market. For the US stock market, empirical tests are generally concentrated on the "semi-strong" and the "strong" form (Malkiel, 2003; Chong and Lam 2010; Kim et al. 2011; Ito et al., 2012; Sabbaghi & Sabbaghi, 2018). Furthermore, (Urquhart and Hudson, 2013) provide evidence in support of AMH for the US.

The Australian stock market follows the same trends, whose evidence of efficient markets remains inconclusive. (A. C. Worthington & Higgs, 2009) confirmed the weak form inefficiency for Australian stock returns from 1975 to 2006. (Tong et al., 2014) investigated that the Australian market is weak-form efficient with little or no evidence for short-term return predictability in the period 2000-2012. Meanwhile, (Hasanov, 2009) rejected the weak-form efficiency in Australia are not weak-form efficient from 1973 to 2006. (A. C. Worthington & Higgs, 2009) also concluded the Australian equity market is not weak-form efficient when examining the period from 1987 to 2003. Brown et al. (1983), as well as Gaunt and Gray (2003), report similar evidence against weak-form efficiency in the Australian market. The Australian economy is a particularly relevant context in adaptive market hypotheses, such as the possibility of 'switches' between efficiency and inefficiency, and the potential for levels of observed efficiency to diverge in the context of different sectoral performance (Deo et al., 2017).

In Vietnam, the majority of the literature also reached the same inconclusive outcome of efficiency testing. The majority of authors have concluded that the market did not achieve the weak form of efficiency, indicating the low level of information transparency (Hoai & Khuyen, 2010), (Dong Loc, 2010), (Long, Huyen, 2017), (Nghia & Blokhina, 2020). The Vietnamese stock market is also found in line with AMH from 2006 to 2008 and 2011 (Phan Tran Trung & Pham Quang, 2019). However, there is no study to validate market efficiency the recent years, i.e., the period from 2019 to 2021.

D. The Predictability of Stock Markets

Stock market prediction is the act of determining the future direction and movement of an index or value of a financial instrument. As mentioned earlier, the EMH presented by (Fama, 1995) implies that it is impossible to make such a prediction as market prices already reflect all current and available information (i.e., intrinsic value). Nevertheless, the theory of behaviour finance together with the adaptive market hypothesis provides strong evidence to the predictability of stock markets. In this line of reasoning, the search for different methods and models still receives numerous attention from academic researchers. The prediction approaches can be classified into traditional methods including fundamental analysis and technical analysis, as well as modern technological methods (Cavalcante et al., 2016). This paper is concerned with the latter method, prediction using machine learning approach to be specific, which is believed to address non-linearity and randomness issues of financial market and outperform the traditional methods *(*Hsu et al., 2016).

E. Sentiment Analysis

Sentiment analysis or opinion mining identifies emotional sentiment in textual data by using natural language processing, computational linguistics, and text analysis (Turner et al., 2021). Sentiment analysis has been utilised in market prediction through an analysis of news data to retrieve investor and market sentiment. Over the past ten years, investor sentiment has been verified of showing a strong dynamic link with asset returns, especially when it comes to the stock market. From an economical perspective, it is believed that when investor sentiment is significant, it can drive prices away from their fundamental values, which can be explained by the appearance of "noise trader", who has random beliefs about future dividends, and "rational arbitrageurs", who maintain Bayesian beliefs (Black, 1986; De Long et al., 1990). Specifically, when it comes to negative news, noise traders might react irrationally and tend to sell their positions to arbitrageurs, resulting in the temporary downward pressure on asset prices (Shleifer & Summers, 1990). Furthermore, there are growing research papers regarding a promising effect of the news on stock returns (Alfarano et al., 2011; Andersen et al., 2007; Chan, 2003; Lavrenko et al., 2000; Nartea et al., 2009; Tetlock et al., 2008). Therefore, researchers have gradually integrated their stock prediction models with investor sentiment and news data such as financial news, corporate announcements, blogs, and messages from social media-tweets. Recent technological advances support academic researchers and practitioners to directly extract sentiment value from textual data, which can be collected from three main sources: firms-specific announcement or corporate disclosures and fillings, media articles, and internet messages, whose sentiment is retrieved from comments and opinions about firms, institutions, and markets (Kearney & Liu, 2014). According to (Kearney & Liu, 2014), corporate disclosure such as interim or annual reports, earnings press releases and earnings conference calls can reveal the fundamental value between sentiment and future asset returns. Meanwhile, studies of media articles such as firm-specific news, general news, financial news and internet posting such as blogs, messages, and comments from social media, concentrate on the short-term effects of sentiment on various market variables such as stock prices, returns, trading volumes and volatility. Each information sources have its unique advantages and drawbacks that attracts numerous researchers and investors continue to explore its predictivity power.

With respect to media-expressed sentiment literature, this paper also reviews recent studies in the past ten years to emphasize the contribution of such a new type (Table I). It should be noted that prior research has yet to officially categorise media articles such as general stock market news, sector-related news, specific stock related news, and macroeconomic news including politics, economics, government policies, and terrorism, which prevent researchers from effective and consistent findings (Usmani & Shamsi, 2021).

Table I shows rare attempts to assess the effect of macroeconomic news on stock market return despite the proven impact of macroeconomic news on stock returns. Long before, macroeconomic news has been found to have a strong correlation with stock price (Caporale et al., 2016, 2017, 2018), which was explained by the theory of capital asset pricing model, claiming that factors affecting macro series such as interest rate and unemployment should also affect asset prices (Merton, 1973). In addition, the empirical results of (Cutler et al., 1988) indicated that macroeconomic news has more predictive power than financial news. Furthermore, (Fan et al., 2020) showed the strong impact of public sentiment on firm-level volatility (stock price and trading volume) around the political event (i.e., US Presidential election) through social media. It should be noted that the study of political information impact on the financial market is not new, but what remained unanswered so far is the evaluation of predictive power and the use of such information when modelling forecasts of stock market movement. Therefore, this study expands the existing literature by not only exploring the impact of daily political news on stock movement, but also regard to how such news can be further exploited to improve the performance of prediction model. Furthermore, another finding of table 5 is that papers with financial news have been tremendously applied recently; therefore, the use of this news type together with the studied political news can make a promising combination to enhance model predictivity performance.

F. Cross-border Studies

(Bustos & Pomares-Quimbaya, 2020) ,(Henrique et al., 2019), (Ozbayoglu et al., 2020) provided a novel review of the state-of-art machine learning applications in stock market prediction. However, previous studies do not attempt to compare their proposed modelling techniques with multiple stock markets or discuss the implication of empirical results on cross-border effects. This is significantly important as results obtained from one stock market may not generalise to other markets (Hsu et al., 2016), (Agrawal et al., 2020). Recall that the purpose of any predictive model is to improve its prediction quality, (Bleyer et al., 2016) pointed out the fact that a model built to predict one group might not be as accurate for predicting another group, and suggested that the dataset should be stratified for an equitable algorithm. In addition, (Hsu et al., 2016) also emphasises conclusions drawn from findings derived from a single market, as is undertaken in the majority of existing machine learning studies, cannot be generalized to a different market, may mask the true degree of informational efficiency in financial markets. From an investor sentiment point of view, investors sentiments differ across different firms/sectors/countries with respect to their demand for stocks (Baker & Wurgler, 2006; Curatola et al., 2016). Furthermore, (Bustos & Pomares-Quimbaya, 2020), based on their systematic review, discovered that stock market prediction from the frontier and emerging markets achieves better modelling performance results than developed countries due to the randomness and complexity of developed countries challenging the predictive quality. However, empirical results from (Hsu et al., 2016) found that the model-based approach has higher predictive performance in the high-income financial market when compared to medium and low-income stock markets. This inconsistency in research findings motivates an investigation to perform a cross-borders prediction utilising consistent machine learning models on different stock markets, which is believed to offer a clearer picture of the generalizability of empirical studies. In addition, Table 5 also shows that empirical results with only one research country have been obtained sparsely in previous studies.

Table I Reviewed Studies of Stock Market Prediction via Sentiment Analysis and Machine Learning

Reference	News Data	Stock Markets	Technical Indicators	Applications
(Schumaker & Chen, 2009)	Financial news	US	No	Trading Strategy
(Huang et al., 2010)	Financial News	Taiwan	No	Trading Strategy
(Schumaker et al., 2012)	Financial news	US	No	Trading Strategy
(Hagenau et al., 2013)	Financial News and corporate announcements	Germany	Yes	Trading Strategy
(Kalyanaraman et al., 2014)	Financial news	India	No	N/A
(Nguyen et al., 2015)	Text in message board	US	No	N/A
(Van de Kauter et al., 2015	Financial News	Belgium	No	N/A
Ding et al. (2016)	Financial News	US	No	N/A
(Luo et al., 2016)	Stock message	Shanghai	No	N/A
(Song et al., 2017)	Financial News	US	No	Trading Strategy
(Chan and Chong, 2017)	Financial News, Blogs	China	No	N/A
(Gálvez & Gravano, 2017)	Stock Message Board Posts	Argentina	Yes	N/A
Vargas, De Lima & Evsukoff (2017)	Financial News	US	Yes	N/A
(Weng et al., 2018)	Financial News, Google Trends, Wikipedia Hits	US	Yes	Trading Strategy
(Malandri et al., 2018)	Twitter comments	US	No	Portfolio Allocation
(Khan et al., 2019)	Political events and stock messages from social media	Pakistan	No	N/A
(Chen et al., 2019)	Financial news	Tokyo	Yes	N/A
(Deng et al., 2019)	Financial news	US	No	N/A
Jin, Yang & Liu (2019)	Twitter comments	US	No	N/A
Li, Wu & Wang (2020)	Firm-specific financial news	Hongkong	Yes	N/A
(Maqsood et al., 2020)	Global and Local Events	US, Hongkong, Turkey, Pakistan	No	Trading Strategy
(Khan et al., 2020)	Political Events, Twitter News Feeds	Pakistan	No	N/A

G. Modern Portfolio Theory

Another research gap that can be deducted from table 5 is a financial application from stock predictions results. In particular, the majority of existing papers have focused on the adoption of trading strategies for the evaluation of prediction outcomes (Schumaker & Chen, 2009; Huang et al., 2010, Huang et al., 2010; Schumaker et al., 2012; Hagenau et al., 2013) while there only one study utilising in the portfolio allocation from sentiment prediction (Malandri et al., 2018). This motivates the author to apply the application of portfolio management to fill the gap of existing research. The below section introduces the Modern Portfolio Theory, which served as the base theory for this research portfolio application.

The Modern Portfolio Theory uses the mean-variance approach to solve portfolio selection issues by optimizing the asset allocation in modern portfolios about an objective function. In specific, investors can construct a portfolio aiming to maximize the portfolio returns (i.e., the expected returns) while maintaining the risks (i.e., variances), or minimise the portfolio risks and maintain the unchanged returns. In recent years, from financial journals, existing papers have incorporated investor sentiment in the process of portfolio formation through financial indicators such as the relative strength index and the volatility index (Li and Xu, 2013; Li et al., 2015; W. Chen et al., 2021). However, there have been few research incorporating news sentiment and prediction stage before applying mean-variance optimization (Chen et al., 2021). In this regard, this study continues to narrow the existing gap by introducing the prediction stage with news sentiment to the portfolio optimization method.

H. Extreme Gradient Boosting in Stock Market Prediction

Extreme gradient boosting (XGB) has been widely adopted in recent years by its proven high speed, performance, and capability to solve non-linear time-series data (rf). The algorithm is based on distributed gradient boosting framework with the approach of parallel processing and tree-pruning to handle missing values and regularization to bias and overfitting (Tianqi Chen, 2019; Deng et al., 2019) uses XGB to predict the weekly movement direction of the S&P 500 Index and show its superiority to other classification methods. Further, there has been increasing research studying the integration of XGB with several classification methods to maximize prediction results. (Chen et al., 2019) utilizes XGB to explore the NIKKEI 225 Index, achieving the average prediction performance of 81 per cent. (Chen et al., 2021) adopted XGB with the firefly algorithm to enhance stock price prediction, proven by the positive cumulative returns of the proposed methods when compared with other algorithms such as Long short-term memory (LSTM) and Support Vector Regression (SVR). (Wu et al., 2019) integrates XGB with sentiment analysis to predict the stock trends of three indexes, namely DJIA Index, New York Stock Exchange Index, and S&P 500 Index, showing that the proposed method can effectively predict the index movement in volatility periods. As proven by the mentioned key papers, this research paper utilises the combination of sentiment analysis and XGB to predict stock market movement to forecast stock market movement for multiple markets such as the US, Australia, and Vietnam.

III. Methodology

This research aims to explore the predictive value of news sentiment and the performance of the proposed portfolio in the US, Australia, and Vietnam. In addition, the study also intends to shed light on the connection between findings from machine learning literature and market efficiency in finance. Therefore, the methodology consists of two stages: stock forecast and portfolio selection. The first stage describes the process of predicting stock market direction using news sentiment. Extreme gradient boosting (XGB) is adopted as the main algorithm for the prediction model. All the predicted results will be sorted in descending order and the top stocks will move into the next phase. The second stage utilised prediction results for stock selection and presents the mean-variance optimization problem. To summarise, the stock trend in the next following days will be predicted in the first stage, and stocks with a higher return in the future will be chosen for the second stage. The prediction results and portfolio performance are evaluated based on a rolling window approach for each market. The overview of the proposed system is illustrated in Fig.1.

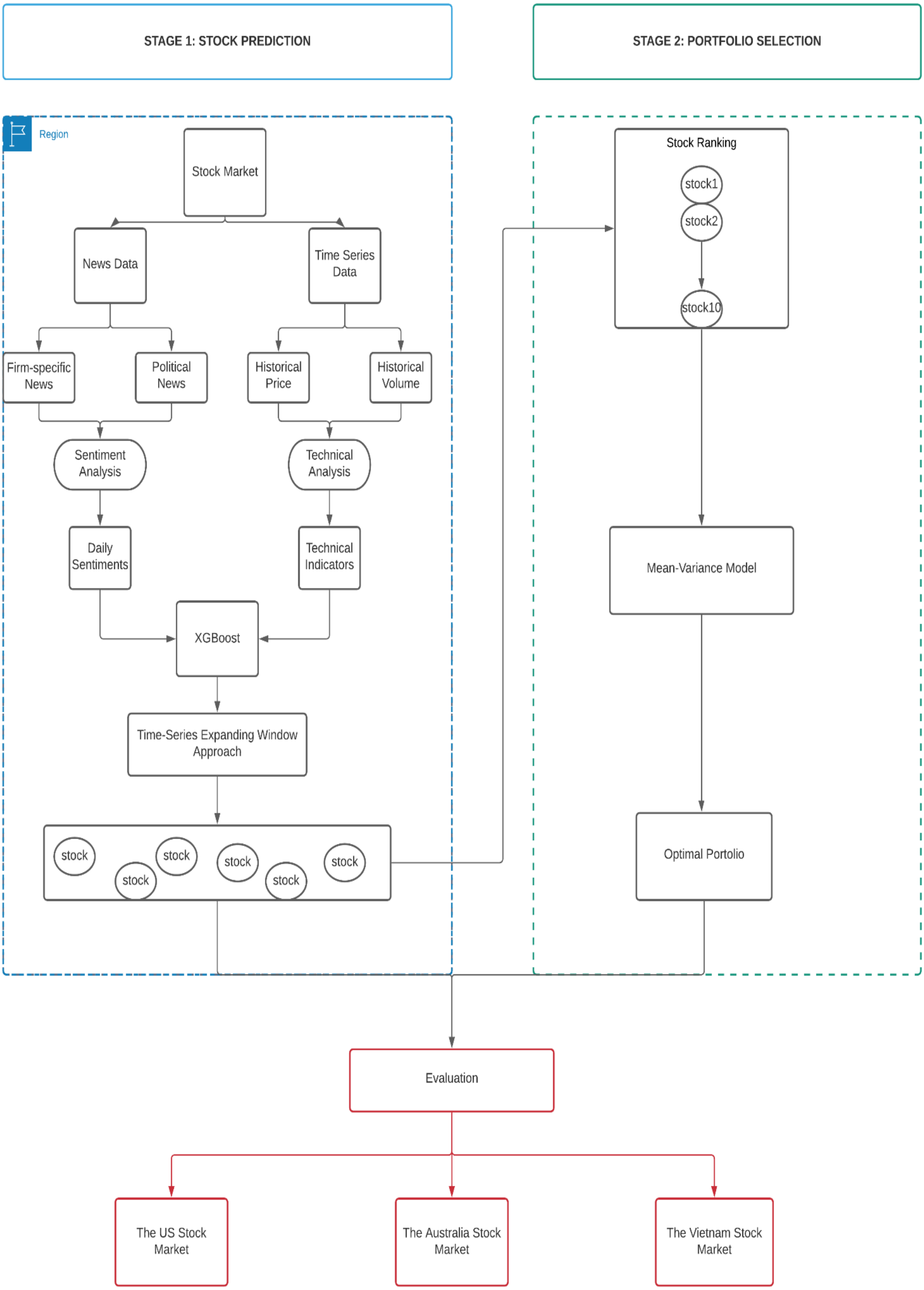


Fig. 1. The Proposed Model

A. Sentiment Analysis

Section 2. E has introduced the concept and function of sentiment analysis for investor sentiment extraction. Recall that sentiment analysis is utilized to retrieve opinions from a given text, which adds valuable information about the stock's sentiment perceived by the public so that investors can make a well-informed trading decision. This paper takes a dictionary-based approach (lexicon-based approach) for sentiment analysis due to the nature of the empirical dataset. Given the fact that most textual data's characteristics are unstructured and noisy, it is recommended to conduct firstly with a dictionary-based approach as it is more efficient, as well as less time-consuming and computationally expensive in terms of memory and CPU usage when compared to machine learning technique, where the textual dataset has to be labelled in advance (Drus & Khalid, 2019). Additionally, it also suits the initial research purpose of examining the predictive power of investor sentiment from the proposed news source; if the prediction is strong, it is expected that all other sentiment analysis methods can achieve relatively correct results for models despite the variation in sentiment extraction quality.

1. Sentiment Dictionaries

In the experiment, two sentiment dictionaries are used to calculate the sentiment vectors for the news articles, which are Vader and Loughran McDonald Financial Dictionary 2018. The choice of this combination is based on their domain functions, including general-purpose sentiment extraction (i.e., VADER) that can be used in any domain and domain-specific for the finance industry, which significantly improves sentiment analysis performance for financial text data and boosts stock directional accuracy (Li et al., 2020).

VADER is a well-known lexicon and rule-based dictionary to detect sentiments expressed in social media (Abraham et al., 2018; Li et al., 2020; Nemes & Kiss; Sohangir et al., 2018). It is commonly preferred by researchers as it contains a group of well-established sentiment lexicons incorporating thousands of lexical features such as sentiment expressions, emotions, sentiment related acronyms and initialisms to optimize sentiment-extracted results (Hutto & Gilbert, 2014). The choice of using VADER for sentiment analysis is based on empirical results from (Sohangir et al., 2018), showing that it has outperformed other dictionaries such as TextBlob as well as the machine learning approach using Support Vector Machine and Logistic Regression in the process of extracting sentiment from financial media. VADER classifies sentiment into four dimensions including positive, negative, neural, and compound scores. These scores are a proportional ratio of text that fall in three categories that are calculated and normalized by the compound score. It should be noted that the paper is only concerned with the compound score for the experiment despite positive, negative, and neutral scores for standardization purposes.

Compound score: a composite score of sentiment polarity, being normalized in the range between -1 (most extreme negative) and +1 (most extreme positive).

- Positive (pos): compound score >= 0.05
- Negative (neg): compound score < 0.05
- Neutral (neu): compound score within (-0.05; 0.05)

(Hutto & Gilbert, 2014)

Loughran McDonald is considered as the state-of-art for the financial market because of its financially-manually made to the financial text, proven by a strong correlation with financial text (Loughran & McDonald, 2011). There are seven dimensions for sentiment analysis, including positive, negative, uncertainty, litigious, weak model, strong modal, and constraining. Following other papers' methods such as (Bollen et al., 2011; Li et al., 2018; Liu & Zhang, 2012; Sprenger et al., 2014), if textual data does not belong in those dimensions, their frequencies are taken as neutral sentiment in this experiment

2. Daily Sentiment Calculation

Sentiment analysis through a lexicon-based approach is conducted through a sentiment processor and analyzer by using the sentiment dictionaries. A sentence-based with word embeddings is adopted to further improve the analyzer performance (Rudkowsky et al., 2018). In specific, each sentence is interpreted as a unit to interpret and be assigned a sentiment score. The process is as the following: each news headline is tokenized and converted into separate words. The word combinations are then projected into sentiment space to be processed. The analyzer uses the given sentiment values from sentiment dictionaries, assign a sentiment score to each sentence. As the daily news headlines for each company is usually more than one, daily sentiment score S_t is also calculated by averaging all daily sentiment scores obtained from each company's news data.

$$S_t = \begin{cases} 2{M_t}^{\mathrm{bull}}/({M_t}^{\mathrm{bull}} + {M_t}^{\mathrm{bear}}) - 1, & {M_t}^{\mathrm{bull}} > {M_t}^{\mathrm{bear}} \\ 0, & {M_t}^{\mathrm{bull}} = {M_t}^{\mathrm{bear}} \\ 1 - 2{M_t}^{\mathrm{bear}}/({M_t}^{\mathrm{bull}} + {M_t}^{\mathrm{bear}}), & {M_t}^{\mathrm{bull}} < {M_t}^{\mathrm{bear}} \end{cases} \tag{1}$$

where M_t^{bull} denotes the number of positive headlines, M_t^{bear} denotes the number of negative headlines in day t. The value of S_t is also defined as. Its value ranges from -1 to 0, where 0 means market sentiment holds a neutral view to this stock for day t. S_t is larger than 0 when market sentiment has positive, and S_t is less than 0 if market sentiment takes a negative view (Ren et al., 2019).

B. Technical Analysis

Technical analysis is also incorporated with the machine learning model to enhance the prediction performance. This technique utilises historical data to forecast future price movements. It assumes that past prices have patterns, from which investors can study and predict future trends

(Chen & Guestrin, 2016). Therefore, this study also considers technical analysis to obtain market signals, enriching the prediction model. Time series data offers two dimensions of data: prices and trading volume, where adjusted close price and volume are employed to retrieve technical indicators:

$$P = \left\{ P_i \middle| P_i = \left[Close_i, Volume_i \right], i = 1, ..., n \right\} \quad (2)$$

where i denotes a trading day.

It should be noted that different technical indicators are not consolidated with the same timestamp due to different purposes and parameters (e.g., time lengths). Therefore, each day preserves available technical indicators while others with missing values are neglected (Dash & Dash, 2016)

1. Technical Indicator Calculation

Predictive-modelling researchers have started to incorporate technical indicators as features for their dataset (i.e., model inputs) to predict future price trends, resulting in an improvement in prediction performance (Bustos et al., 2017; Ghanavati et al., 2016; Li et al., 2020; Liu et al., 2018; Yang et al., 2016). Following the works of (Li et al., 2020) that facilitate technical indicators with news sentiment, this paper adopts a similar approach in order to enhance the predictive performance of the machine learning model. Taken all technical indicators affecting stock returns in the studies of (Hsu et al., 2016; Menkhoff & Taylor, 2007; Taylor & Allen, 1992) The choice of these technical indicators are motivated based on the findings of (Bjerring et al., 2017) pointing out that recent machine learning literature has not considered cautiously technical indicators by either using a lengthy list or a commonly-used list of the indicators, resulting in computational cost and contradiction with real-world trading decision making. For instance, the combination of Stochastic D% and Stochastic K% indicators are commonly used as technical inputs in recent papers without considering the mathematical operations of each one. Specifically, the first one is simply the arithmetic mean of the latter (Dash & Dash, 2016). Therefore, this paper considers technical indicators based on their high impact and proven predictive value in the studies of (Hsu et al., 2016; Menkhoff & Taylor, 2007; Taylor & Allen, 1992).

Ten indicators are selected based on different statistical functions, the choice of ten indicators is proven by (Alsubaie et al., 2019) mentioned that 5-10 technical inputs support maximizing investment return and minimizing misclassification costs without compromising the achievement of highest-accuracy model performance. In addition, considering the work of (Yıldırım et al., 2021), this also establishes the choice of a technical timeframe as this study shares a similar short-term investing objective. These technical indicators are categorized into:

- Trend indicators are used to measure the direction of a trend from stocks or markets.
- Momentum indicators determine the speed and duration of a stock’s trend.
- Volatility indicators assess the rate of price movement as well as its acceleration and deceleration.
- Relative strength indicators measure the velocity and magnitude of price movements, implying a trend reversal or market correction
- Volume indicators identify the strengths of stocks’ trends and recognize the force behind the stock movement.

(Wilder, 1978)

Table II Technical Indicators

Symbol	Name	Types
ROC	Rate of Change	Momentum
MFI	Money Flow Index	Momentum
ADX	Average Directional Movement Index	Trend
EMA	Exponential Moving Average	Trend
BB	Bollinger Band	Volatility
ATR	Average True Range	Volatility
STOC	Stochastic	Relative Strength
RSI	Relative Strength Indicator	Relative Strength
OBV	On-Balance Volume	Volume
ADL	Accumulation Distribution Indicator	Volume

C. Extreme Gradient Boosting

Extreme gradient boosting (XGB or XGBoost) was introduced by (Chen and Guestrin, 2016), is a gradient descent algorithm, advancing in minimising the loss by adding new models and correcting the errors of existing models. Therefore, it has several characteristics such as high classification performance, low computational complexity, and fast running speed.

Boosting technique is used for a classification problem by sequentially combining the results of weak learners (models) to make a strong aggregated prediction. The essence of this technique is the combined accuracy results from multiple models to deduce a final and more accurate prediction result, instead of using outcome from a single machine learning model. Gradient boosting is an implementation of boosting technique using the principle of gradient descent (Friedman, 2000). The advancement of gradient boosting is based on the principle of gradient descent that when combined with all the weak models along with their initial prediction, the best possible next model is generated by setting target outcomes tagged into it to minimize the prediction error. In specific, the target outcomes are set based on the gradient of prediction error so that the next model follows this direction aiming to minimize the target prediction error. XGB is an improved version of gradient boosting, designed to minimize the time taken from the above process, aiming for efficacy and computational speed without compromising model performance (Chen & Guestrin, 2016). The process of how XGB works can be explained through the following steps:

(1) The tree space is the process in which the input x$_i$ maps to a leaf node. f$_t$(x$_i$) relates to tree structure q and leaf node weight ω

(2) The predicted output :

$$\hat{y}_i = \sum_{t=1}^{m} f_t(x_i)$$

Where m denotes the number of trees, ft(xi) represents the function decision of each tree. The predicted value of XGB is the total values of leaf nodes related to each tree.

(3) The objective function represents the complexity of t trees:

$$\mathrm{Obj} = \sum_{i=1}^{n} \mathrm{l}(\mathrm{y}, \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t})}) + \sum_{i=1}^{t} \Omega(\mathrm{f_i}).\ \Omega(f_i)$$

(4) Boosting approach is used to optimize the parameters:

$$\mathrm{Obj}^{(\mathrm{t})} = \sum_{i=1}^{n} \mathrm{l}(\mathrm{y}, \hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t})}) + \sum_{i=1}^{t} \Omega(\mathrm{f_i})$$

(5) The second-order Taylor expansion:

$$Obj^{(t)} = \sum_{i=1}^{n} l(y, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + 0.5 h_i f_t^2(x_i) + \Omega(f_t) + constant$$

(6) The optimal value:

$$Obj^* = -0.5 \sum_{i=1}^{T} \frac{G_i^2}{H_j+\lambda} + \gamma T$$

It is worth noting that XGB generates a better structure of tree if the value of the objective function is relatively small.

D. Mean-variance Optimisation

Mean-variance optimization (MVO) was introduced by Markowitz (Markowitz, 1952) to optimize the portfolio selection process, which initiates the foundation of Modern Portfolio Theory (MPT). In specific, portfolio selection depends on the acceptance level of risks and expected returns from investors' point of view. For rational investors, either the lower risk portfolio with constant expected returns or the portfolios of higher expected return in the considerate risk level is preferred. To balance the process, the principal of MVO is to create an optimal portfolio that not only maximized expected return but also minimized risks, which are quantified by the portfolio's expected return and variance, respectively.

The mean-variance optimization is then formulated

$$\underset{w_i,\dots,w_n}{Min} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \delta_{ij}$$

$$\underset{w_i,\dots,w_n}{Max} \quad \sum_{i=1}^{n} w_i \mu_i \qquad (3)$$

Subject to: $\begin{cases} \sum_{i=1}^{n} w_i = 1 \\ 0 \le w_i \le 1, \forall i = 1,\dots,n \end{cases}$

where w_i *and* w_e denote the allocation (i.e., weight) invested stock *i* and stock j. μ_i is expected return on stock *i*. Following (Chen et al., 2019), a variable λ called risk aversion coefficient is included to represent investors' behaviour corresponding to the risk investment choices. The value of λ ranges from 0 to 1 to illustrate the acceptance level of risk and returns. The λ = 0 represents the portfolio whose objective function is to maximize expected return without considering risk; whereas, λ= 1 demonstrate a portfolio of risk minimization without considering a return. MVO determines the efficient point to balance between the two extreme values. This leads to a mono-objective formulation to derive an optimal portfolio:

$$\underset{w_i,\dots,w_n}{Min} \quad \lambda\left[\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \delta_{ij}\right] - (1-\lambda)\left[\sum_{i=1}^{n} w_i \mu_i\right]$$

Subject to: $\begin{cases} \sum_{i=1}^{n} w_i = 1 \\ 0 \le w_i \le 1, \forall i = 1,\dots,n \end{cases}$

(4)

IV. EXPERIMENT

A. Data Description

Based on the literature review, the primary data used in this study are of listed companies from three major market indices representing the stock exchange of the US, Australia, and Vietnam. This includes S&P 100 (OEX) Index, the S&P/ASX 100 (XTO) Index, and the VN30 (VNI30) Index, respectively (Bloomberg, 2021).

For comparative purposes, the indices have similar attributes of market-capitalisation weighted and float-adjusted stock listed on their domestic stock exchange. The markets are selected as they contain the most liquidity and the largest market-cap companies including large-cap and mid-cap companies across all sectors, constituting above 70 per cent of share market capitalisation. This is beneficial for the study due to the availability of news for long-term analysis. In addition, all the indices are designed to replicate the performance of the national share markets or benchmark against a representative portfolio of tradable stocks. This smoothens the process of portfolio selection by the provision of high-quality stocks, positive exposure to certain market risks, and sufficient diversification. Furthermore, these indices play a crucial role in domestic economic development at both macro and micro levels; therefore, the correct prediction of the key players is beneficial to all stakeholders to form profitable portfolios, create relevant economic policies, prevent financial risks, and ensure the dynamic funds flow in the capital market.

The dataset includes stock price data and news corpus (i.e., financial news and political news) from January 1, 2010, to July 31, 2021, covering 11 years. The hold-out period 2016-2021 is used for the out-of-sample test as it contains market dynamics of the sample countries. This period is considered since the continuity of financial stock data, the longer the sample data is involved, the more likely it is to capture history information memory (F. E. Harrell, 2001). In addition, the choice of the dataset is to examine the performance of proposed methods in normal and extreme markets. The initial training data between 2010 and 2015 simply represents the bull markets with several market corrections, excluding the famous Global Financial Crisis 2007-08. Meanwhile, the out-of-sample test period encompasses important market cycles such as the bull market and the bear market as well as major events such as the 2020 stock market crash and the Black Swan event (i.e., The COVID-19 pandemic) (Bartholomeusz, 2020).

The following sections present the datasets and functions for the experiments.

1.Time Series Data

Time-series data includes historical stock prices sector indices and of all companies. The historical stock prices are used for both stages of the experiments. In the first stage, the price data is employed to generate technical indicators through technical analysis and used as separate features for prediction models. In the second stage, historical data is employed to compute portfolio returns and index returns.

Data are requested directly from Yahoo Finance, a reliable source of data. It should be noted that this paper does not employ Yahoo Finance API due to the closure in 2017 in order to ensure a stable dataset with reliable price adjustments (Yahoo Finance API, n.d.). The daily historical data includes the trading date and historical series of the adjusted opening price, closing price, highest and lowest prices for the day, and trading volume. The utilization of adjusted price to obscure the impact of key nominal price in the short term and obtain an accurate record of the stock performance (Bacidore et al., 2002; Rechenthin et al., 2013). The return is defined as follows:

$$Rd,t = \frac{Pt - Pt - d}{Pt - d} \qquad (5)$$

where P_t is the adjusted closing price at time t and d is the daily returns (d=10)

In the absence of the total return index during the study period, the adjusted monthly closing price is utilized to calculate index returns, which are commonly used when the index return data is not readily available (Brooks, 2008). The risk-free rates are obtained from national central banks.

2.News Data

News data is mainly used for the prediction stage, from which sentiment is extracted from sentiment analysis. The news corpus is composed of political news, company-specific, and market-related news. In specific, political news includes all political coverages such as elections, political polls, diplomatic issues, and the political impact of critical economic and business development in each country. In addition, financial news that is specifically related to the company regarding operational, economic, social, and financial activities, is also considered in this article. All news was written in English and automatically crawled from major news vendors such as Bloomberg, The Guardian, investing.com, and cafef.vn. Each piece of news is tagged with a timestamp showing the time the news is released, which helps classify news by dates. Companies that are related to the news are listed at the end of the news using their stock symbols, which helps establish the mapping from the news articles to stocks and vice versa. The headlines are adopted to obtain only important words and minimize noise. Reference (Wang et al., 2015) carried out a set of experiments and showed that news titles are more useful to forecast than news contents. In addition, previous studies have proved its sentiment capability for machine learning prediction models (Deng et al., 2019; Deveikyte et al., 2020; Ding et al., 2015).

The total number of news headlines collected are 281238 after denoising, which is relatively comparable to other news sentiment studies' text corpus size ranging from around 100,000 to 500,000 news text in the average range from 5 to 20 years (Deng et al., 2019; Ding et al., 2015; Li et al., 2020; Turner et al., 2021). Considering the size of the dataset, a web scraper is developed to collect data, textual information, and stock prices from data sources. This data is then fed into a pipeline, which processes data and sends it to the machine learning engine to detect sentiment in the provided texts.

B. Data Processing

Data cleaning and pre-processing are the required first steps before applying machine learning models since the quality of inputs has a significant impact on the classification effectiveness (Rahm & Do, 2000), (Symeonidis et al., 2018). The raw historical stock price is processed through data cleaning, integration, and transformation to ensure the larger value inputs (e.g. large-cap stock price) does not overwhelm smaller value inputs, reducing prediction error. (e.g., medium and small-cap stock price) (Dash & Dash, 2016; Ji et al., 2019; Kumar & Thenmozhi, 2006).

News data is pre-processed through tokenization and token normalization processes (Symeonidis et al., 2018). News data are first tokenized; "noise" from text data such as stop words- function words appearing frequently across the data such as "a", "the", "is", white spaces and punctuation are also removed. All news headlines are then enabled with the loss of capitalization to reduce the dimensionality problem (Dos Santos & Gatti, 2014; Zhang et al., 2015). Finally, text lemmatization is used to reduce inflectional words to their root form (Guzman & Maalej, 2014).

For comparative purpose, the experiment collects local news written in Vietnamese and translate them into English. This experimental design was employed because the majority of news articles, especially high impact news on economic indicators is written in the national language. Also, prior literature has proven the effectiveness of a unified syntax for a comparative and experimental purpose that can be used for sentence-level sentiment analysis (Araújo et al., 2020; Kaity & Balakrishnan, 2020; Rudkowsky et al., 2018). In addition, a recent study by (Araújo et al., 2020) confirmed the field of machine translation systems has achieved the level of maturity to acquire a competitive prediction performance.

As choosing machine translators can affect the empirical results, the study employs Yandex API, a novel hybrid translation model based on the statistical and deep learning approach (Savenkov et al., 2011). This neural machine translation achieves the average mean of 0.73 performance with a standard deviation of 0.12, whose similar performance to Google Translator (Araújo et al., 2020). To the author's knowledge, no related study in the Vietnam market has adopted the method, which makes it one of the main contributions for this paper and makes the neural machine translation area particularly appealing for future research.

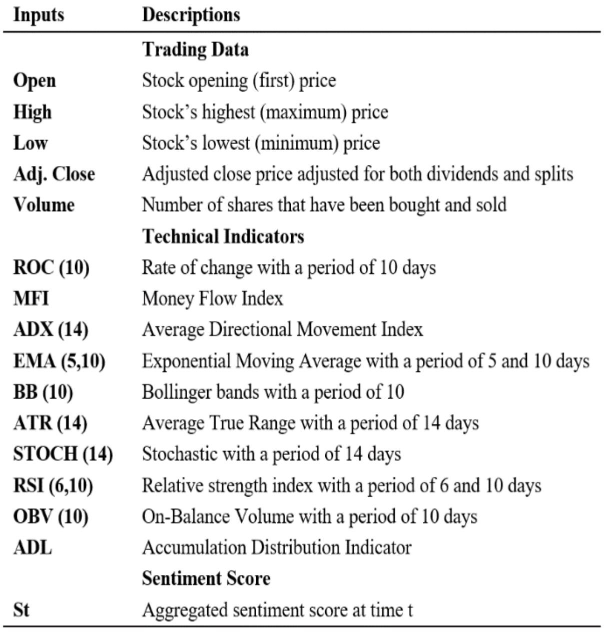
C. Stock Prediction

This section illustrates the process of stock movement prediction using the XGB model. Recall that the research experiment examines the impact of news sentiment and prediction models in different countries.

Model Inputs

Table III provides the list of variable inputs for the prediction model. It should be noted that indicators' time frame is usually customizable based on investors trading objectives; specifically, the longer the time frame, the less sensitive indicators respond to price changes (Taylor & Allen, 1992). Considering the work of (Yıldırım et al., 2021), this establishes the choice of a technical timeframe as this paper shares a similar short-term prediction objective.

Table III Input Variables



Inputs	Descriptions
	Trading Data
Open	Stock opening (first) price
High	Stock's highest (maximum) price
Low	Stock's lowest (minimum) price
Adj. Close	Adjusted close price adjusted for both dividends and splits
Volume	Number of shares that have been bought and sold
	Technical Indicators
ROC (10)	Rate of change with a period of 10 days
MFI	Money Flow Index
ADX (14)	Average Directional Movement Index
EMA (5,10)	Exponential Moving Average with a period of 5 and 10 days
BB (10)	Bollinger bands with a period of 10
ATR (14)	Average True Range with a period of 14 days
STOCH (14)	Stochastic with a period of 14 days
RSI (6,10)	Relative strength index with a period of 6 and 10 days
OBV (10)	On-Balance Volume with a period of 10 days
ADL	Accumulation Distribution Indicator
	Sentiment Score
St	Aggregated sentiment score at time t

Prediction Models

The empirical models are considered based on the choice of features.

Table IV Prediction Models

Model (M)	Features
M1	Historical Prices, Technical Indicators
M2	Historical Prices, Technical Indicators, Political News
M3	Historical Prices, Technical Indicators, Financial News
M4	Historical Prices, Technical Indicators, Political News, Financial News

Model 1 is considered as the benchmark to evaluate whether the integration of news inputs can outperform the traditional prediction method (technical analysis approach as mentioned in the literature review). Therefore, model 1 adopts historical stock prices together with technical indicators. From model 2 to model 4, the predictive values of different types of news are examined: model 2 considers political news and model 3 uses financial firm-specific news as input. Model 4 is aimed to achieve the optimal prediction output with the use of all features.

Label. A label variable is determined to identify the stock direction. Following the extreme method of (Li et al., 2020) a threshold of 1% stock price return is used for label determination. In particular, for each stock, if the return is equivalent or higher than 1%, the label is set as an uptrend (class 1). In contrast, if the return is less than 1%, the label is set as a downtrend (class 0).

Evaluation Metrics. In the experiment, F1-score is adopted for prediction evaluation. This serves the practical purpose of research objectives, which are to utilise prediction results for stock portfolio selection and perform comparison prediction-based portfolios in different countries. Therefore, the outcomes are concerned with true classes. The choice of F1-score is based on actual costs assigned to every error caused; in this study, the cost of misclassification for either a false positive (a mislabelled rising stock) and false negative (a mislabelled falling stock) is much higher when forming a portfolio (F. Harrell, 2017)**.** In addition, (Chicco & Jurman, 2020) has been found to significantly affects the performance of the optimized portfolio when compared to other metrics.

The metrics are defined as followed:

F1 Score - the harmonic mean of precision (i.e., the percentage of stock trend classified as up-trend that is up-trend) and recall (i.e., the percentage of up-trends that are classified as up-trend):

$$\mathrm{F1} = \frac{TP}{TP+(FP+FN)/2} \qquad (5)$$

where TP is truly positive, TN true negative, FP false positive, and FN false negative

Data Split. In this study, nested cross-validation is applied through the prediction model comparisons and evaluation approaches. As financial time series is non-stationary and complex, this approach is considered is prevent look-ahead bias and data leakage associated with the random sampling of the training and test data sample (Arlot et al., 2010) (Ratto et al., 2018). In addition, to evaluate the better performance of the prediction model and portfolio returns in each time point, a time window slicing cross-validation strategy is also adopted. In specific, the training and tests sets are moved across the timeline of a dataset. There are two parameters considered in this method: (i) Initial Window suggests the initial number of consecutive data points in each training and validation sample. The initial window only contains the data points occurring prior to data in the validation sample in order to prevent the invasion of future data points when training; (ii) Horizon dictates the size of a test sample. In this study, the initial window was set to be 80% of the data points and the Horizon parameter was set to be 20% of the observations. It is worth noting that the training set continues to expand when moving on to the next window. There is a total of 12 training and testing periods for each stock derived from the dataset (table V)

Table V Expanding Window for Prediction

	Initial Window (Train-Validate Sample)		**Horizon (Test Sample)**	
	Start	**End**	**Start**	**End**
Period 1	2010-01-01	2015-12-31	2016-01-01	2016-06-30
Period 2	2010-01-01	2016-06-30	2016-07-01	2016-12-31
Period 3	2010-01-01	2016-12-31	2017-01-01	2017-06-30
Period 4	2010-01-01	2017-06-30	2017-07-01	2017-12-31
Period 5	2010-01-01	2017-12-31	2018-01-01	2018-06-30
Period 6	2010-01-01	2018-06-30	2018-07-01	2018-12-31
Period 7	2010-01-01	2018-12-31	2019-01-01	2019-06-30
Period 8	2010-01-01	2019-06-30	2019-07-01	2019-12-31
Period 9	2010-01-01	2019-12-31	2020-01-01	2020-06-30
Period 10	2010-01-01	2020-06-30	2020-07-01	2020-12-31
Period 11	2010-01-01	2020-12-31	2021-01-01	2021-06-30
Period 12	2010-01-01	2021-06-30	2021-07-01	2021-12-31

D. Portfolio Selection

This section aims to explore whether the prediction results from news sentiment are of benefit to creating a successful portfolio for investment. Following the approach from (W. Chen et al., 2021) whose prediction model is based on technical indicator features, this paper also incorporates sentiment features and anchors on the premise of identifying patterns of behaviour in the historical series of asset prices.

Markowitz's Mean-Variance Optimisation method (MVO) will be used to obtain the capital allocation proportion for each stock that has been entered. This is to determine the proportion of capital allocated to each stock. To achieve the optimization and calculation of the minimum-variance portfolio, the measures of the classic Markowitz model are used: mean and covariance matrices. . To find a balance between return and risk, the Sharpe ratio is used to make better decisions, and available resources are allocated to the portfolio with the largest Sharpe ratio.

Since the purpose of this study is to explore the performance of the proposed models in various periods among countries, the portfolios were composed exclusively of risky assets, without taking into account investors’ risk preferences, risk-free assets. In addition, the portfolio excludes transaction costs and taxes as they are varied between countries and time periods. Further, the research is restricted to the buy-and-hold portfolio since Vietnam has only allowed a short position in the domestic stock market since February 2021 (ssc.gov, 2020). For consistent comparison, the portfolio consists of 10 stocks for all experiments. According to the recent literature, portfolio formation for individual investors is effective when focus on only fewer than 10 assets (Almahdi & Yang, 2017, Kocuk & Cornuéjols, 2018, Tanaka, Guoa, & Turksen, 2000), Paiva et al., (Paiva, Cardoso, Hanaoka, & Duarte, 2019) discover that the portfolio with ten assets performs better than others with different numbers of assets.

Metrics.

The portfolio performance is evaluated by Sharpe ratio by calculating the adjusted-risk return (Sharpe, 1994), which is defined as

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p} \tag{6}$$

Where R_p denotes the return of the portfolio, R_f denotes the return of risk-free asset (rf = 0) and σ_p denotes the standard deviation of the portfolio's excess return.

Benchmark.

The investment return is benchmarked with the market index for each country, namely the S&P 100 (OEX) Index, the S&P/ASX 100 (XTO) Index, and the VN30 (VNI30).

Table VI Benchmark Performance from 2016-2021

Countries	Avg. Annualised Return	Avg. Annualised Standard Deviation	Avg. Annualised Sharpe Ratio
The US (OEX)	0.20	0.14	0.79
Australia (XTO)	0.14	0.13	0.18
Vietnam (VNI30)	1.59	0.86	1.09

V. Analysis and Discussion of Results

A. Results of Stock Prediction

The section represents the results of stock prediction using the proposed news sentiment, including financial firm-specific news and daily political news. This is to investigate the responsiveness of prediction models to news across different countries.

Table VII reports the performance for all models in terms of average F1 score. The empirical results from Table VII confirm the predictive value of the news sentiment for stock movement prediction of all countries.

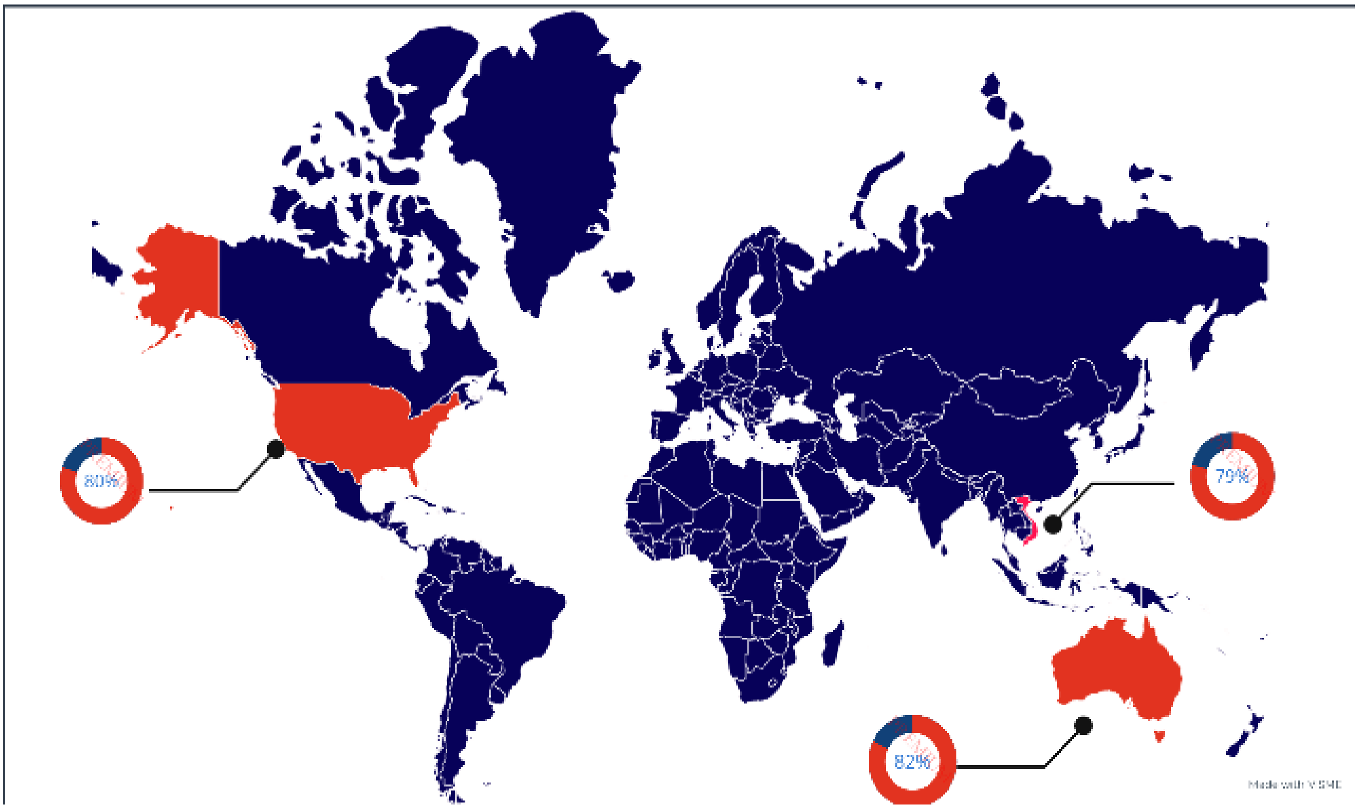
Table VII Results of Applying XGB using Political News, Finance News, or Both for Forecasting Stock Movement

Evaluation	Countries	Model 1	Model 2	Model 3	Model 4
Average F1	The US	0.790	0.794	0.792	0.802
	Australia	0.805	0.814	0.807	0.819
	Vietnam	0.754	0.761	0.780	0.791

It is evident from the results shown that news sentiment, irrespective of type, have an impact on the study sample. This result shows the performance improvement provided by exploiting news into the base model. The research outcomes are also parallel with the results in the range of 60-90 per cent gained from (Schumaker & Chen, 2009; Schumaker et al., 2012; Song et al., 2017; Nguyen et al., 2015; Weng et al., 2018; Malandri et al., 2018; Deng et al., 2019) when examining the impacts of news sentiment with prediction models.

The notable finding indicated the predictive value of market sentiment extracted from the proposed news is strong enough for the machine learning model to make directional predictions despite the variation in modelling performance. This can be proved by all of the performance results are in the range from 70 to 85.6 per cent. In addition, the findings confirm that the predictive value of news sentiment on a global scale, as illustrated in Fig.1.

In terms of prediction performance, the US-based model and the Australia-model outperformed Vietnam's model. In other words, with the higher average F1 score, the US and Australia were the most predictable market, followed by Vietnam, illustrated by results from Model 4 of 0.802, 0.819; and 0.791 respectively. In order to build a bridge between machine learning and financial economics literature, this research first attempts to evaluate the empirical results in the light of efficient market hypothesis. The empirical results mark a significant contribution by providing evidence in supporting the efficient market hypothesis. In specific, as the nature of semi to a strong form of market efficiency, all available information is fully reflected on the share price, therefore, the performance of the US and Australian models are better since the experiment is based on historical prices (Kim et al. 2011; Ito et al., 2012; Sabbaghi & Sabbaghi, 2018; Hasanov, 2009; Deo et al., 2017). The empirical finding from this study also provides clarification for machine learning literature, in which all predictive accuracy is significantly high in recent years, ranging from 60 to 80 over the decade and reaching 90 or higher in the past two years (Schumaker & Chen, 2009; Schumaker et al., 2012; Song et al., 2017; Nguyen et al., 2015; Weng et al., 2018; Malandri et al., 2018; Deng et al., 2019; Jin et al., 2019; Li et al. 2020; Maqsood et al., 2020; Khan et al., 2020). For prediction



models in developed-market papers, historical prices, which are used as model inputs, have all available information reflected in them; therefore, machine learning algorithms can easily find patterns and make a prediction based on historical patterns, resulting in significantly high prediction outcomes.

On the other hand, Vietnam as a weak-form efficiency causes a high level of randomness in past data, which makes it more challenging for the model to recognize the data's pattern (Hoai & Khuyen, 2010), (Dong Loc, 2010), (Long, Huyen, 2017), (Nghia & Blokhina, 2020). This can be observed by the lowest of 0.754 F1 scores in the base model M1, when compared to 0.790 and 0.805 for the US and Australia. In addition, the same trend is also applied when adding news sentiment, by which Vietnam has the lowest F1 score when compared to the others. The observations obtained in this paper is in line with (Hsu et al., 2016), whose results showed higher predictive performance in the high-income financial market when compared to medium and low-income stock markets. This finding contributes to the existing studies by highlighting the relative prediction performance between developed and developing markets. Regardless of any attempt to improve the prediction results, machine learning prediction models would achieve the same pattern when performed for each market type, which is higher predictive outcomes for developed markets and vice versa.

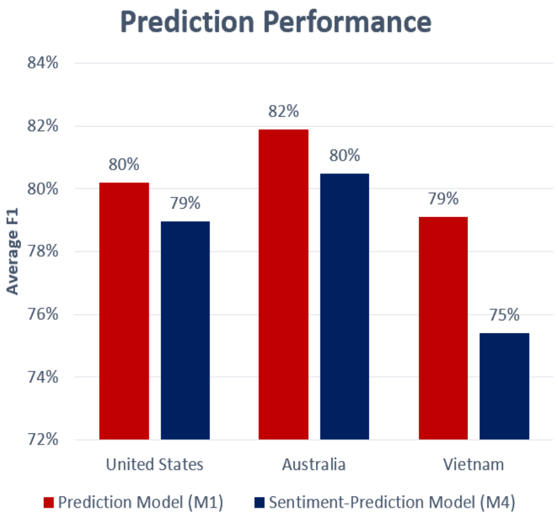


Fig. 2 Prediction Performance of Baseline Model and Sentiment-Prediction Model

For demonstrating the efficacy of news sentiment, the experiment has included the prediction results between 2016 and 2021. Table VIII shows the results from sentiment-prediction models (M4) of all countries in the study period. The prediction results have been very promising with the increasing performance of F1 scores for all countries from 2016 to 2021. In specific, M4 showed the improvement of F1 ranges from 78 to 83 per cent for the US and Australia, and 65 to 80 per cent for Vietnam.

Table VIII Prediction Results (F1-score) in 2016-2021

Year	The US	Australia	Vietnam
2016	0.793	0.783	0.655
2017	0.688	0.763	0.693
2018	0.811	0.781	0.725
2019	0.835	0.536	0.714
2020	0.856	0.841	0.812
2021	0.827	0.833	0.801

Table VIII shows the changes in prediction performance by year. It is evident that the predictive value of news sentiment contributes to the gradual improvement of forecasting results. However, in 2019, both Australia and Vietnam show poorer performance 2019, at the peak time of the COVID-19 pandemic. In particular, Australia suffered a significant decrement of F1 score, from 0.781 to 0.536, a drop of 32%; Vietnam has F1 score decreases from 0.725 to 0.714, by 1.5%. This result can be explained by the findings of (Steyn et al., 2020), investigating the impact of different types of sentiment (i.e., positive and negative sentiment) varied among stock markets. In specific, positive sentiment is highly correlated with prediction results in the US to get better results, whereas negative sentiment is found to be a significant predictor of stock movement in other countries in the sample such as India, Germany, and the UK. In this line of reasoning, the sentiment extracted from the news data might not have a sufficient "level of sentiment" to predict the stock movement for both countries when it comes to shock news. Another possible explanation is the fact that negative news results in a decrease in stock prices as investors keep selling stocks. Therefore, the stock direction in such time will likely be moving downwards. And applying F1 metrics to evaluate in the year with a great deal of negative news might not be suitable as the F1 metric focuses on the up-trend classification. Overall, the finding from table V confirmed the consistent prediction performance using sentiment data over time in the country sample. The exception was that the weaker F1 in 2019 for Australia and Vietnam.

To evaluate the impact of each type of news sentiment, the results are calculated following the percentage change formula for further clarification:

$$\boxed{\Delta_{news} = \frac{F_{base,sentiment} - F_{base}}{F_{base}}} \qquad (7)$$

Where $\Delta_{sentiment}$ denotes the relative improvement of F1 between the results of the base model and prediction models with news sentiment. $F_{base,\ sentiment}$ denotes the F1-score with news sentiment and F_{base} denotes the F1-score of the baseline model.

The improvements of different news types for each country are presented in Fig. 3. It can be observed that the sentiment from political news has a significant impact on the prediction models when compared to financial news for both the US and Australia. Australia is impacted by daily political news the most by a 1.2% increment, followed by 0.6% from the US. On the other hand, the Vietnam market is more sensitive to financial news rather than political news with the highest 3.45% change. The reason can be hypothesized by the strong level of correlation between political factors and economic factors with the countries. Therefore, it is suggested that an investigation for the root cause of these findings should be the subject of future study.

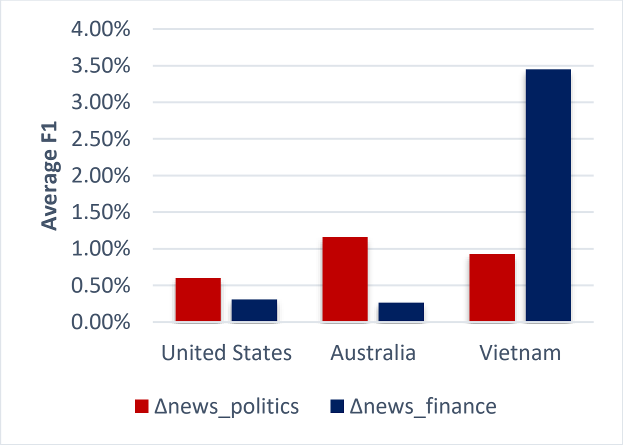


Fig. 3 The Δnews of each country, where the x-axis represents different countries

Table IX The Impact of Political News Sentiment

Evaluation	Countries	*Δnews_politics*	Δ*news*_finance	Δ*sentiment*
Average F1	The US	0.60%	0.31%	0.6%
	Australia	1.16%	0.27%	1.2%
	Vietnam	0.93%	3.45%	0.9%

The empirical results from table VII highlight the contribution of this research regarding the predictive value of daily political news for stock movement prediction. From table VII, it is evident that political news contributes to the prediction models by an increment in the range of 0.6-1.2% observed in all countries. It is worth mentioning that this paper is the first to apply daily political news as input to machine learning algorithms; therefore, the results cannot be used to compare with existing papers. However, (Khan et al., 2020) observed a 20% increase when the machine learning model was applied to political situation events (97 events) in Pakistan. As opposed to this research data which assess the impact of the daily intake of political news.

Table X The Impact of News Sentiment

Evaluation	Countries	Model 1	Model 4	$\Delta_{sentiment}$
Average F1	**The US**	0.790	0.802	1.6%
	Australia	0.805	0.819	1.8%
	Vietnam	0.754	0.791	<u>4.9%</u>

Overall, the empirical results are similar to findings on confirmation of the predictive value of the news sentiment as well as the positive relationship of sentiment on developed and developing countries (Brown & Cliff 2004).

Additional Sector Analysis

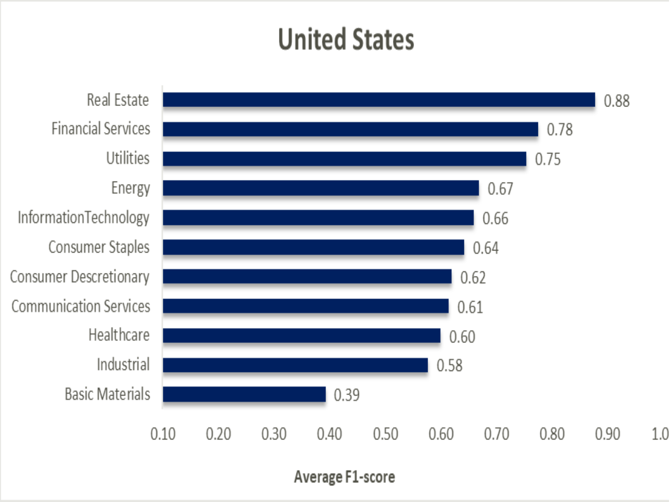


Fig. 5 Sector Analysis (The US)

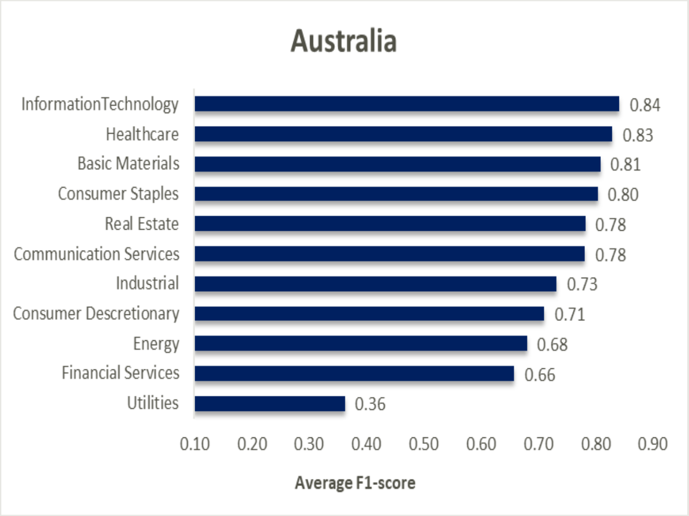


Fig. 6 Sector Analysis (Australia)

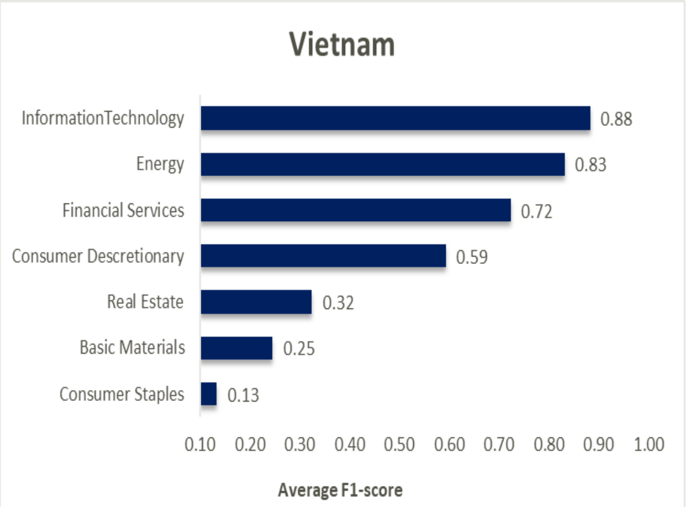


Fig. 4 Sector Analysis (Vietnam)

For the explanatory purpose, to investigate the effectiveness of the proposed news sentiment, the stability of the prediction model is checked by examining it from the industry level. Figures 4-6 evaluate the model's prediction performance of all sectors in the country sample. It is worth noting that contrasting with other papers, which only chose the representatives of sectors, the research considers all stocks in each index for better evaluation. By comparing the results sector by sector, in most cases, models with news perform better than the baseline with the average F1 ranging from 0.59 to 0.78. In the US, the prediction model has the highest value with real estate, financial services and utility sectors. In Australia, information technology, health care and basic materials have the highest results. In Vietnam, the proposed model has a better reaction with information technology, energy, and financial sectors. It is observed that the prediction model makes the most accurate prediction in most stocks of the technology sector for both Australia and Vietnam. Since all sectors in three countries achieved consistently better-than-random guess (F1 > 0.5), the proposed news sentiment can be concluded to have predictive power at the sector level.

B. Results of Portfolio Performance

Consistency of Proposed Models across Countries

After analysing prediction performance, this section validates if the sentiment-based prediction models can be applied in practice and how different investment returns perform in comparison. Fig 7 provides insights into the financial performance of the proposed portfolio when compared to market return indices. The figure represents the average annualised Sharpe ratio of different portfolios in the out-of-sample period (2016-2021).

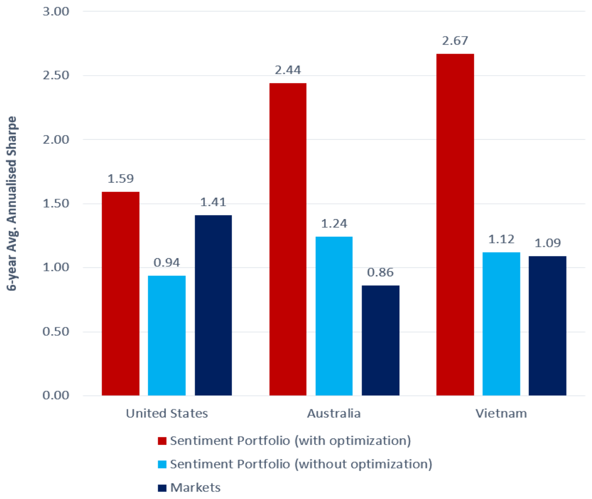


Fig. 7 Portfolio Results

It is evident that with the addition of news sentiment, mean-variance models showed consistent improvement across countries. Overall, the risk-adjusted returns of the sentiment portfolios have outperformed market indices of all countries. In other words, the proposed news sentiment is strong enough for stock selection to outperform the benchmarks for each market. The Sharpe ratio of the sentiment-based sector over the test period (2016-2021) ranges from 1.59 to 2.67 for the three countries.

Overall, the sentiment portfolio with mean-variance optimization performs better than the equal-weighted portfolio. This finding provides empirical evidence in the existing financial literature from machine learning analysis to support the superiority of the mean-variance optimization approach for portfolio construction. For the past decade, there has been a continuous battle regarding the risk-adjusted performance of mean-variance optimization and equal-weighted allocation strategy for portfolio selection (Plyakha et al., 2012; Schmidt, 2019).

Consistency of Proposed Models Overtime

Figures 8-10 presents portfolio risk-adjusted returns (Sharpe ratios) for the period 2016- 2021. It is important to note that all the portfolios are positive in returns for the testing period when compared to market index returns. This means that the proposed news sentiment and stock prediction provide a positive effect on portfolio construction over time. In terms of sentiment portfolio with mean-variance optimization, the highest Sharpe ratio observed in the US, Australia, and Vietnam was 3.44, 4.73, and 4.74, respectively. The highest Sharpe's ratio from sentiment-portfolio without optimization (equal-weighted portfolio) were 2.19, 2.89, and 3.3 in similar country's order. The Sharpe ratios were observed to gradually improve over time, especially from 2018 to 2021.

Important to note is 2018 where negative market indices for the US and Australia, which means no positive excess returns for the given level of portfolio risk in this year. Meanwhile, the proposed sentiment portfolios still outperformed the benchmarks. On the other hand, all the portfolios and market indexes in Vietnam fell to zero, and none of the studied portfolios outperformed the benchmark in 2018. The same trend applied in 2020 (i.e., the stock market crash 2020 and the peak time of COVID-19 pandemic) for Australia and Vietnam where market indices suffered a significant drop, but both of the experimental portfolios surpassed the markets (Schmidt, 2019). However, the US experienced a reverse trend, in which none of the portfolios performed better than the market index.

Although there have been fluctuations at different periods, the sentiment portfolios still achieve a reasonable number of times outperforming benchmarks, demonstrating the superiority of the proposed methods over a period of time.

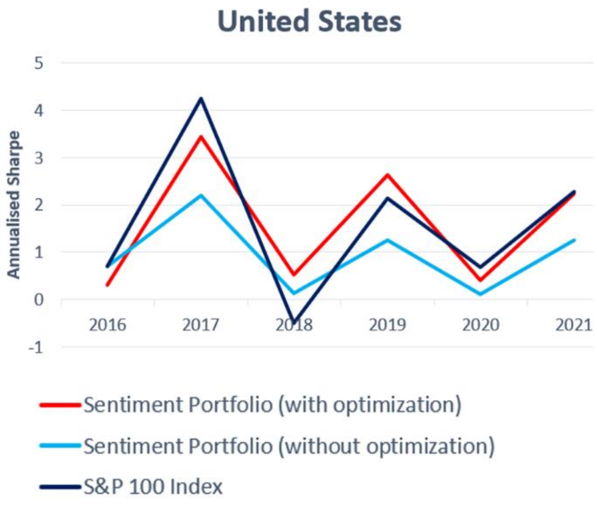


Fig. 8 Performance of Sentiment Portfolios in Testing Period (The US)

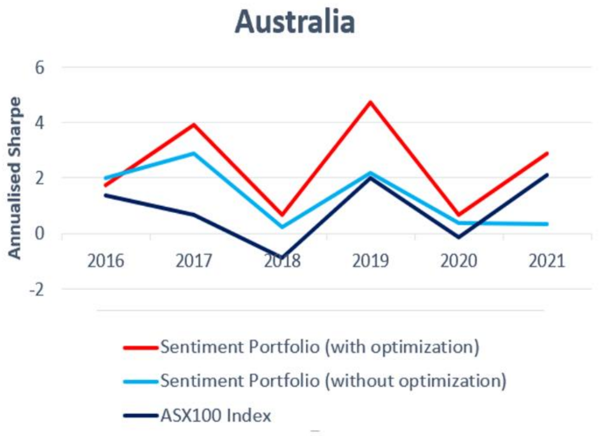


Fig. 9 Performance of Sentiment Portfolios in Testing Period (Australia)

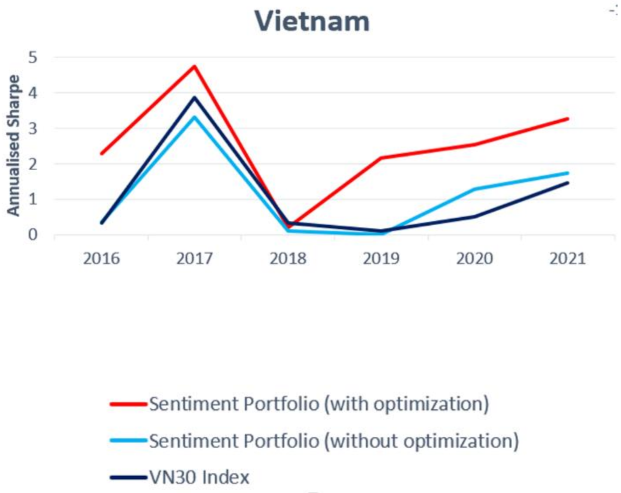


Fig. 10 Performance of Sentiment Portfolios in Testing Period (Vietnam)

The variation in portfolio results can additionally be explained by variation in volatility. Volatility (risk) has a significant impact on portfolio returns (Murphy, 1999). In particular, volatile stock allocation is more likely to achieve volatile returns. At different points of time in different countries, the choice of stocks to construct a portfolio is varied and depends on the outcomes from the prediction stage. In addition, from a financial perspective, stock allocation methods should depend on market conditions; for instance, equal-weighted stock allocation is suitable for a stable market as it is more vulnerable to high volatile markets (Murphy, 1999). Therefore, future research should consider including additional parameters for the first prediction stage as well as adding objectives and constraints during the stock construction process.

VI. Conclusion and Future Work

The main novelty of this paper is the investigation of predictive power for political news and financial news for the machine learning prediction model. Sentiment analysis is used to extract investor sentiment from textual data. Correspondingly, major financial websites of each country were selected to obtain the news corpus. Non-English news (i.e., Vietnamese news) was translated to English by the neural machine translation method. Then, a machine learning model based on the XGB algorithm is adopted to predict stock movement by implementing nested cross-validation and a realistic time-series expanding window approach. Overall, it can be confirmed that the proposed news sentiment has predictive value in the sample countries i.e., the F1-score can be as high as 85.4%. In order to demonstrate the persuasiveness of empirical results, this paper incorporates the news sentiment and prediction results into the portfolio selection process and adopts mean-variance optimization. The proposed portfolios have been tested for consistent performance over time in both developed and developing markets. The sentiment-based portfolio showed positive performance in terms of Sharpe ratio across countries from 2016 to 2021. Finally, the empirical results from prediction models and portfolio allocation are further discussed in light of the EMH. In specific, results obtained from prediction models and portfolio returns are significantly influenced by the maturity of the market. Further, the prediction performance from developed markets is higher than the developing market due to the use of historical prices. In specific, as the nature of semi to a strong form of market efficiency, all available information is fully reflected on the share price, therefore, the performance of the US and Australian models are better since the experiment is based on historical prices. On the other hand, Vietnam as a weak-form efficiency causes a high level of randomness in past data, which makes it more challenging for the model to recognize the data's pattern. The empirical results mark a significant contribution by providing evidence in supporting the efficient market hypothesis. Moreover, the impact of news sentiment for the Vietnam model is more significant when compared to other countries, highlighting the relative prediction performance between developed and developing markets. Regardless of any attempt to improve the prediction results, the machine learning prediction models would achieve the same pattern when performed in each market. Therefore, the observations can support machine learning literature to understand the reasons and feasibility of prediction results.

For future work, the proposed method can be further improved by the expansion of sample countries. In addition, an investigation of the proposed methods in a cross-sector study is also encouraged. Further, different choices of machine learning algorithms should be considered to test the robustness of prediction models and the validity of EMH conclusions.

APPENDIX A: DATA AND PROGRAMS USED FOR THIS REPORT

Source code can be founded in https://github.com/ocvo/Sentiment-based-Stock-Portfolio

APPENDIX B: FIGURES

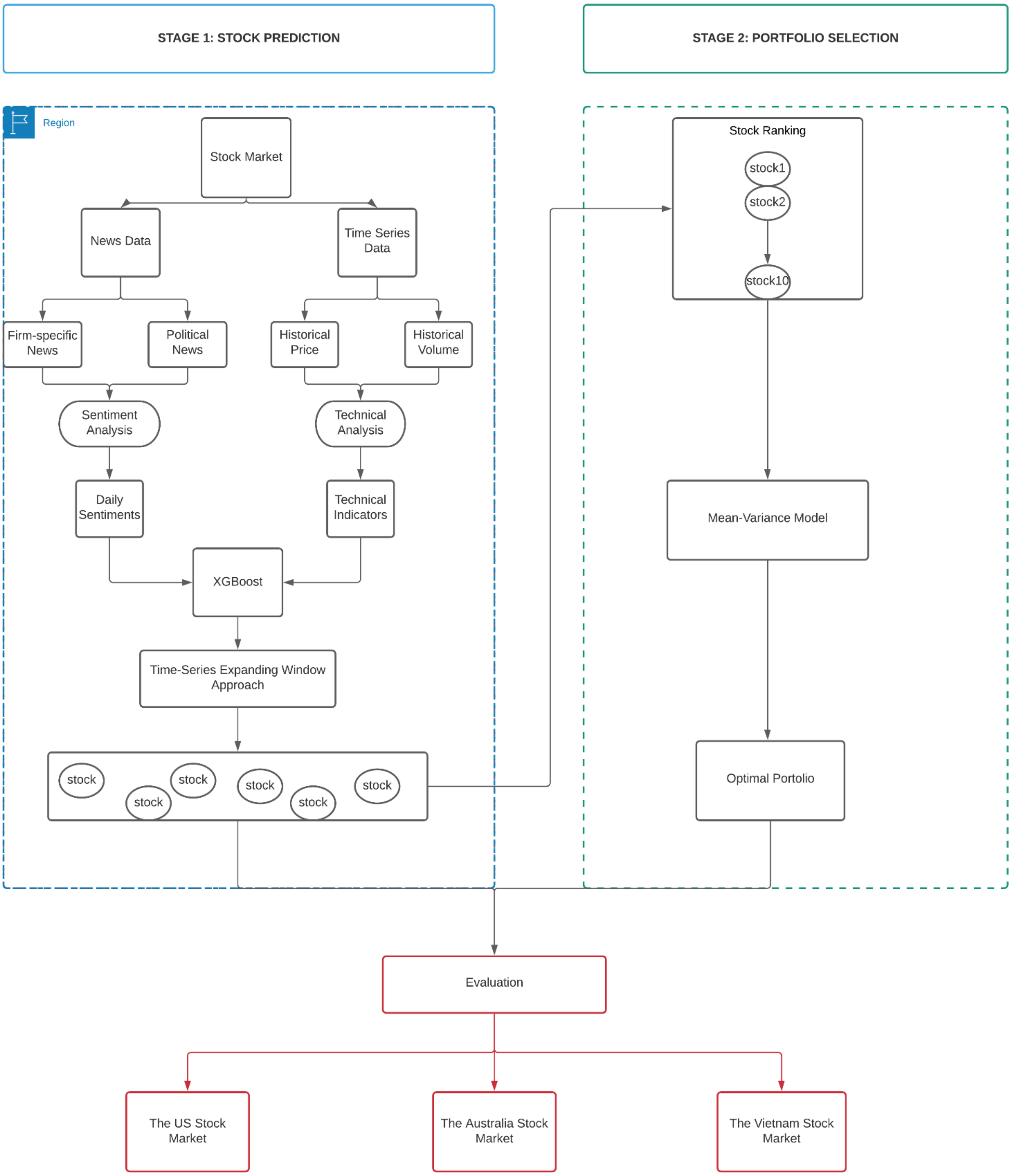


Fig. 1 The Proposed Model

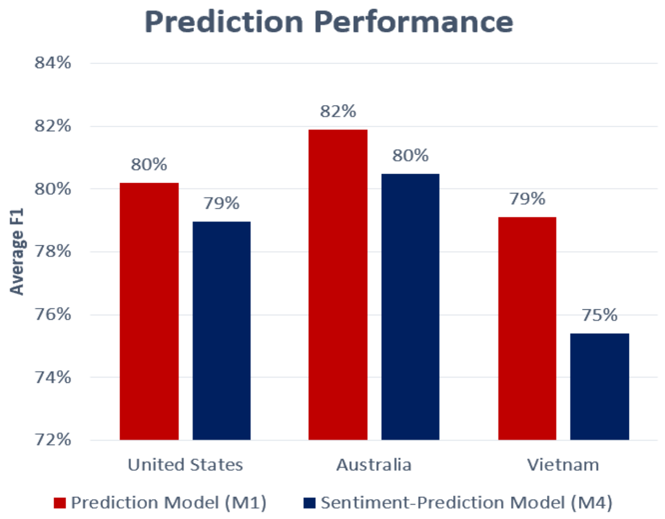


Fig. 2 Prediction Performance of Baseline Model and Sentiment-Prediction Model

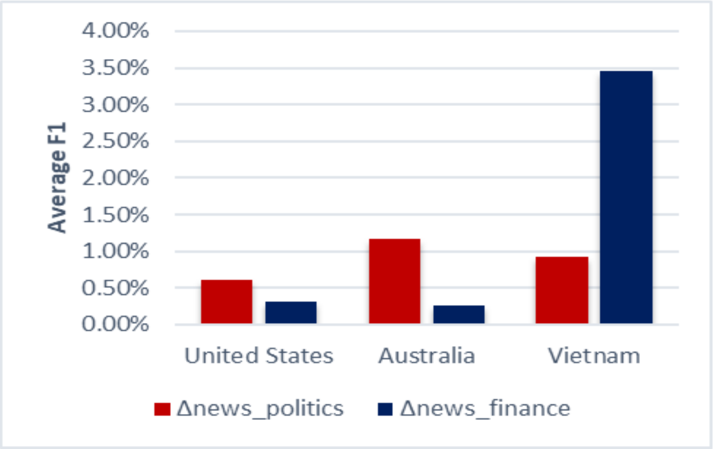


Fig. 3 The Δnews of each country, where the x-axis represents different countries

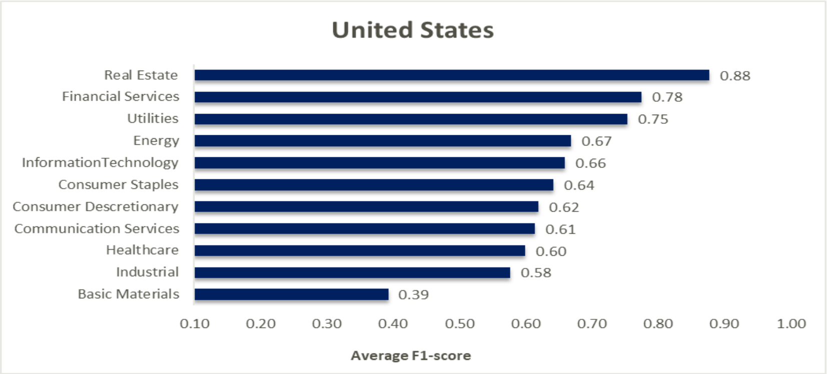


Fig. 4 Sector Analysis (US)

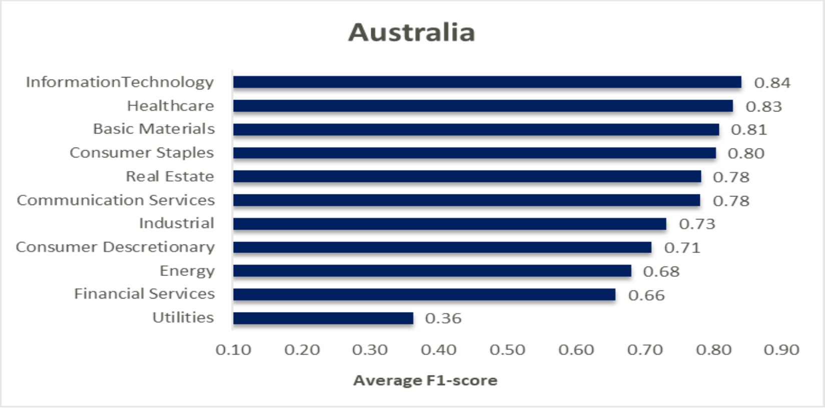


Fig. 5 Sector Analysis (Australia)

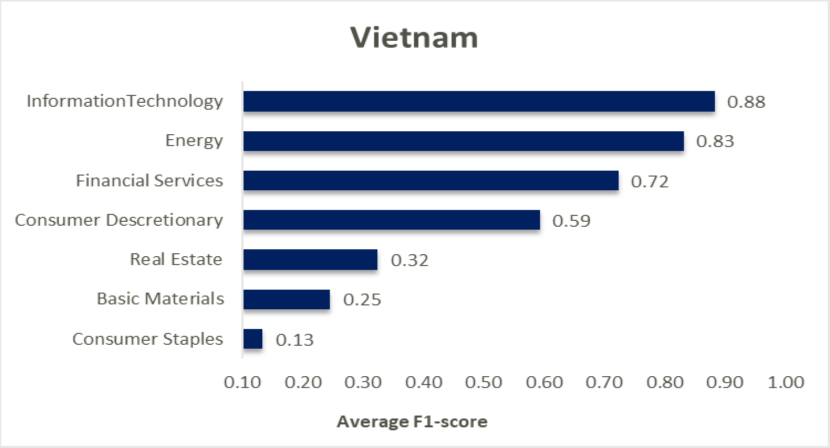


Fig. 6 Sector Analysis (Vietnam)

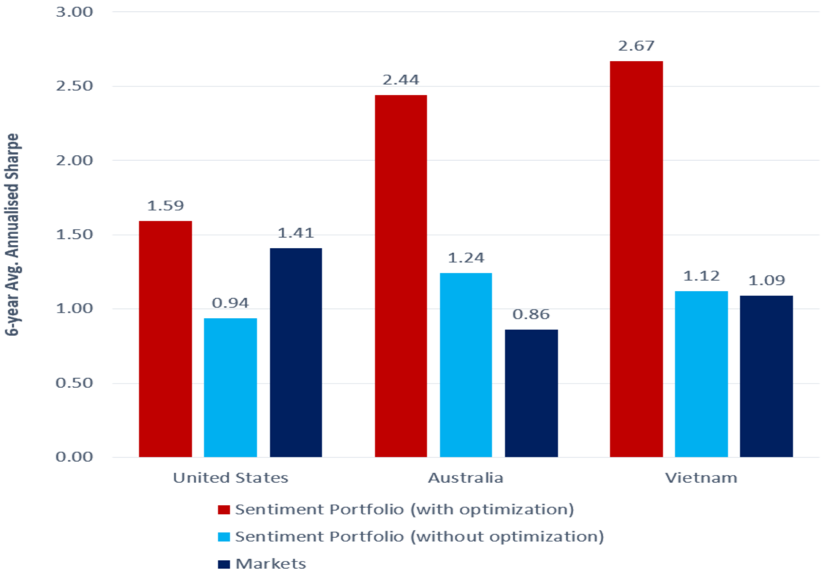


Fig. 7 Portfolio Results

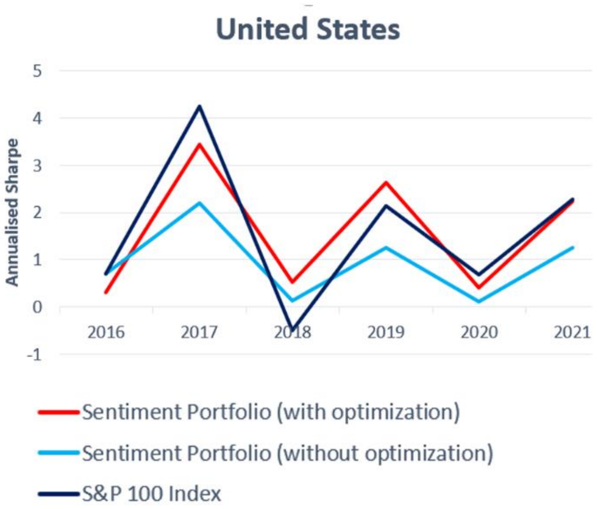


Fig. 8 Performance of Sentiment Portfolios in Testing Period (The US)

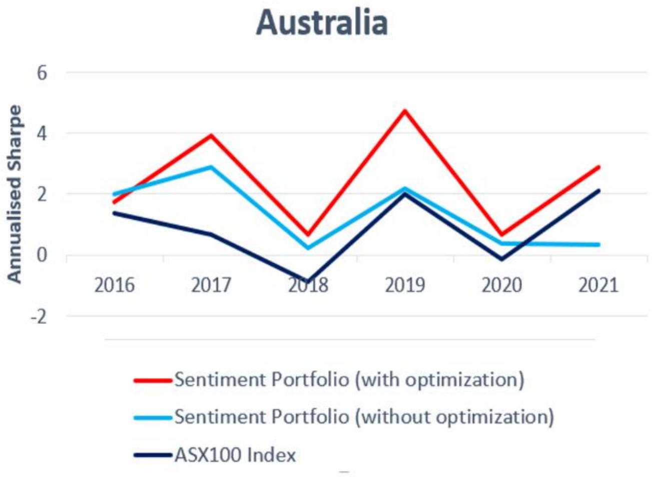


Fig. 9 Performance of Sentiment Portfolios in Testing Period (Australia)

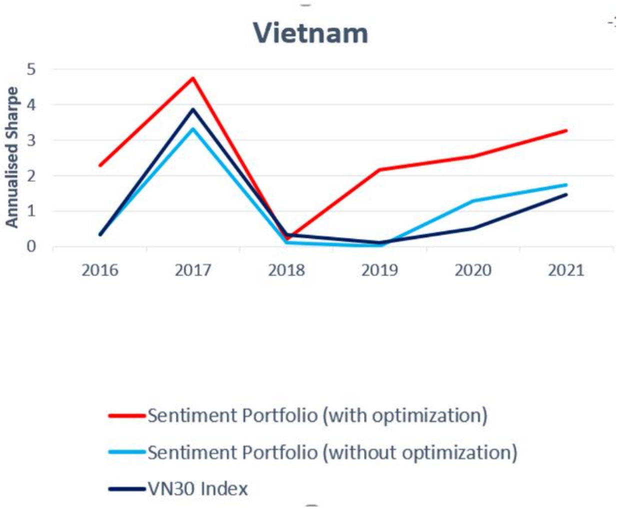


Fig. 10 Performance of Sentiment Portfolios in Testing Period (Vietnam

Appendix C: Tables

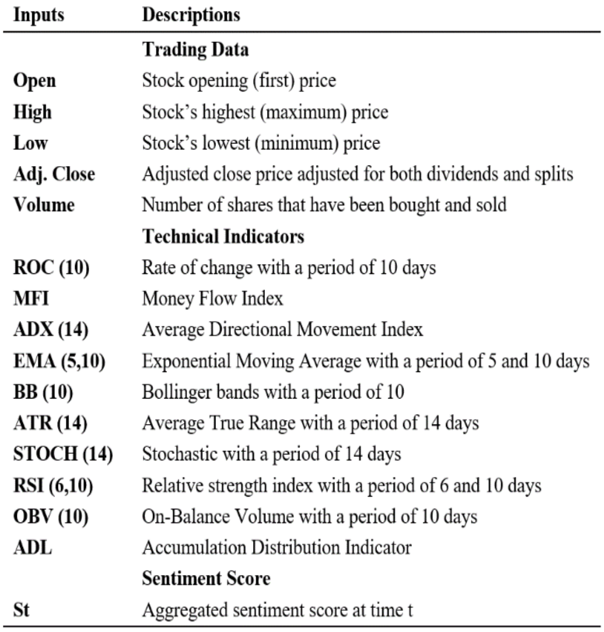
Table I Reviewed Studies of Stock Market Prediction via Sentiment Analysis and Machine Learning

Reference	**News Data**	**Stock Markets**	**Technical Indicators**	**Applications**
(Schumaker & Chen, 2009)	Financial news	US	No	Trading Strategy
(Huang et al., 2010)	Financial News	Taiwan	No	Trading Strategy
(Schumaker et al., 2012)	Financial news	US	No	Trading Strategy
(Hagenau et al., 2013)	Financial News and corporate announcements	Germany	Yes	Trading Strategy
(Kalyanaraman et al., 2014)	Financial news	India	No	N/A
(Nguyen et al., 2015)	Text in message board	US	No	N/A
(Van de Kauter et al., 2015	Financial News	Belgium	No	N/A
Ding et al. (2016)	Financial News	US	No	N/A
(Luo et al., 2016)	Stock message	Shanghai	No	N/A
(Song et al., 2017)	Financial News	US	No	Trading Strategy
Chan and Chong (2017)	Financial News, Blogs	China	No	N/A
(Gálvez & Gravano, 2017)	Stock Message Board Posts	Argentina	Yes	N/A
Vargas, De Lima & Evsukoff (2017)	Financial News	US	Yes	N/A
(Weng et al., 2018)	Financial News, Google Trends, Wikipedia Hits	US	Yes	Trading Strategy
(Malandri et al., 2018)	Twitter comments	US	No	Portfolio Allocation
(Khan et al., 2019)	Political events and stock messages from social media	Pakistan	No	N/A
(Chen et al., 2019)	Financial news	Tokyo	Yes	N/A
(Deng et al., 2019)	Financial news	US	No	N/A
Jin, Yang & Liu (2019)	Twitter comments	US	No	N/A
Li, Wu & Wang (2020)	Firm-specific financial news	Hongkong	Yes	N/A
(Khan et al., 2020)	Political Events, Twitter News Feeds	Pakistan	No	N/A

Table II Technical Indicators

Symbol	Name	Types
ROC	Rate of Change	Momentum
MFI	Money Flow Index	Momentum
ADX	Average Directional Movement Index	Trend
EMA	Exponential Moving Average	Trend
BB	Bollinger Band	Volatility
ATR	Average True Range	Volatility
STOC	Stochastic	Relative Strength
RSI	Relative Strength Indicator	Relative Strength
OBV	On-Balance Volume	Volume
ADL	Accumulation Distribution Indicator	Volume

Table III Input Variables



Inputs	Descriptions
	Trading Data
Open	Stock opening (first) price
High	Stock's highest (maximum) price
Low	Stock's lowest (minimum) price
Adj. Close	Adjusted close price adjusted for both dividends and splits
Volume	Number of shares that have been bought and sold
	Technical Indicators
ROC (10)	Rate of change with a period of 10 days
MFI	Money Flow Index
ADX (14)	Average Directional Movement Index
EMA (5,10)	Exponential Moving Average with a period of 5 and 10 days
BB (10)	Bollinger bands with a period of 10
ATR (14)	Average True Range with a period of 14 days
STOCH (14)	Stochastic with a period of 14 days
RSI (6,10)	Relative strength index with a period of 6 and 10 days
OBV (10)	On-Balance Volume with a period of 10 days
ADL	Accumulation Distribution Indicator
	Sentiment Score
St	Aggregated sentiment score at time t

Table IV Prediction Models

Model (M)	**Features**
M1	Historical Prices, Technical Indicators
M2	Historical Prices, Technical Indicators, Political News
M3	Historical Prices, Technical Indicators, Financial News
M4	Historical Prices, Technical Indicators, Political News, Financial News

Table V Expanding Window for Stock Prediction

	Initial Window (Train-Validate Sample)		Horizon (Test Sample)	
	Start	**End**	**Start**	**End**
Period 1	2010-01-01	2015-12-31	2016-01-01	2016-06-30
Period 2	2010-01-01	2016-06-30	2016-07-01	2016-12-31
Period 3	2010-01-01	2016-12-31	2017-01-01	2017-06-30
Period 4	2010-01-01	2017-06-30	2017-07-01	2017-12-31
Period 5	2010-01-01	2017-12-31	2018-01-01	2018-06-30
Period 6	2010-01-01	2018-06-30	2018-07-01	2018-12-31
Period 7	2010-01-01	2018-12-31	2019-01-01	2019-06-30
Period 8	2010-01-01	2019-06-30	2019-07-01	2019-12-31
Period 9	2010-01-01	2019-12-31	2020-01-01	2020-06-30
Period 10	2010-01-01	2020-06-30	2020-07-01	2020-12-31
Period 11	2010-01-01	2020-12-31	2021-01-01	2021-06-30
Period 12	2010-01-01	2021-06-30	2021-07-01	2021-12-31

Table VI Benchmark Performance from 2016-2021

Countries	Avg. Annualised Return	Avg. Annualised Standard Deviation	Avg. Annualised Sharpe Ratio
The US (OEX)	0.20	0.14	0.79
Australia (XTO)	0.14	0.13	0.18
Vietnam (VNI30)	1.59	0.86	1.09

Table VII Results of Applying XGB using Political News, Finance News, or Both for Forecasting Stock Movement

Evaluation	Countries	Model 1	Model 2	Model 3	Model 4
Average F1	**The US**	0.790	0.794	0.792	0.802
	Australia	0.805	0.814	0.807	0.819
	Vietnam	0.754	0.761	0.780	0.791

Table VIII Prediction Results (F1-score) in 2016-2021

Year	The US	Australia	Vietnam
2016	0.793	0.783	0.655
2017	0.688	0.763	0.693
2018	0.811	0.781	0.725
2019	0.835	0.536	0.714
2020	0.856	0.841	0.812
2021	0.827	0.833	0.801

Table IX The Impact of Political News Sentiment

Evaluation	Countries	Δ*news_politics*	Δ*news*_finance	$\Delta_{sentiment}$
Average F1	**The US**	0.60%	0.31%	0.6%
	Australia	1.16%	0.27%	1.2%
	Vietnam	0.93%	3.45%	0.9%

Table X The Impact of News Sentiment

Evaluation	Countries	Model 1	Model 4	$\Delta_{sentiment}$
Average F1	**The US**	0.790	0.802	1.6%
	Australia	0.805	0.819	1.8%
	Vietnam	0.754	0.791	4.9%

References

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review, 1*(3), 1.

Agrawal, A., Gans, J., & Goldfarb, A. (2020). How to win with machine learning. *Harvard Business Review*.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management, 39*(1), 45-65.

Alfarano, S., Camacho, E., & Morone, A. (2011). *The role of public and private information in a laboratory financial market*. IVIE.

Alkhatib, K., Najadat, H., Hmeidi, I., & Shatnawi, M. K. A. (2013). Stock price prediction using k-nearest neighbor (kNN) algorithm. *International Journal of Business, Humanities and Technology, 3*(3), 32-44.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Vega, C. (2007). Real-time price discovery in global stock, bond and foreign exchange markets. *Journal of international Economics, 73*(2), 251-277.

Andrew, L. (1997). A non-random walk down Wall Street. Proceedings of symposia in pure mathematics,

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance, 59*(3), 1259-1294.

Appel, G. (2005). *Technical analysis: power tools for active investors*. FT Press.

Araújo, M., Pereira, A., & Benevenuto, F. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences, 512*, 1078-1102.

Bacidore, J. M., Battalio, R. H., & Jennings, R. H. (2002). Depth improvement and adjusted price improvement on the New York Stock Exchange. *Journal of Financial Markets, 5*(2), 169-195.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance, 61*(4), 1645-1680.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of economic perspectives, 21*(2), 129-152.

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications, 42*(20), 7046-7056.

Basak, G. K., Das, P. K., Marjit, S., & Mukherjee, D. (2019). British Stock Market, BREXIT and Media Sentiments-A Big Data Analysis.

Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers. *The North American journal of economics and finance, 47*, 552-567.

Batra, R., & Daudpota, S. M. (2018). Integrating StockTwits with sentiment analysis for better prediction of stock price movement. 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET),

Bird, S. (2006). NLTK: the natural language toolkit. Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions,

Black, F. (1986). Noise. *The Journal of Finance, 41*(3), 528-543.

Bleyer, A., Baines, C., & Miller, A. B. (2016). Impact of screening mammography on breast cancer mortality. *Int J Cancer, 138*(8), 2003-2012. https://doi.org/10.1002/ijc.29925

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

Bollinger, J. (2002). *Bollinger on Bollinger bands*. McGraw-Hill New York.

Bustos, O., Pomares, A., & Gonzalez, E. (2017). A comparison between SVM and multilayer perceptron in predicting an emerging financial market: Colombian stock market. 2017 Congreso Internacional de Innovacion y Tendencias en Ingenieria (CONIITI),

Bustos, O., & Pomares-Quimbaya, A. (2020). Stock market movement forecast: A Systematic review. *Expert Systems with Applications, 156*, 113464. https://doi.org/10.1016/j.eswa.2020.113464

Caporale, G. M., Spagnolo, F., & Spagnolo, N. (2016). Macro news and stock returns in the Euro area: a VAR-GARCH-in-mean analysis. *International Review of Financial Analysis, 45*, 180-188.

Caporale, G. M., Spagnolo, F., & Spagnolo, N. (2017, 2017/05/01/). Macro news and exchange rates in the BRICS. *Finance Research Letters, 21*, 140-143. https://doi.org/https://doi.org/10.1016/j.frl.2016.12.002

Caporale, G. M., Spagnolo, F., & Spagnolo, N. (2018). Macro news and bond yield spreads in the euro area. *The European Journal of Finance, 24*(2), 114-134.

Cavalcante, R. C., Brasileiro, R. C., Souza, V. L., Nobrega, J. P., & Oliveira, A. L. (2016). Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications, 55*, 194-211.

Chan, W. S. (2003). Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics, 70*(2), 223-260.

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, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics, 21*(1), 1-13.

Curatola, G., Donadelli, M., Kizys, R., & Riedel, M. (2016). Investor sentiment and sectoral stock returns: Evidence from World Cup games. *Finance Research Letters, 17*, 267-274.

Cutler, D. M., Poterba, J. M., & Summers, L. H. (1988). *What moves stock prices?* (0898-2937).

Dash, R., & Dash, P. K. (2016, 2016/03/01/). A hybrid stock trading framework integrating technical analysis with machine learning techniques. *The Journal of Finance and Data Science, 2*(1), 42-57. https://doi.org/https://doi.org/10.1016/j.jfds.2016.03.002

De Bondt, W. F., & Thaler, R. (1985). Does the stock market overreact? *The Journal of Finance, 40*(3), 793-805.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of political Economy, 98*(4), 703-738.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Positive feedback investment strategies and destabilizing rational speculation. *The Journal of Finance, 45*(2), 379-395.

Deng, S., Zhang, N., Zhang, W., Chen, J., Pan, J. Z., & Chen, H. (2019). Knowledge-driven stock trend prediction and explanation via temporal convolutional network. Companion Proceedings of The 2019 World Wide Web Conference,

Deveikyte, J., Geman, H., Piccari, C., & Provetti, A. (2020). A Sentiment Analysis Approach to the Prediction of Market Volatility. *arXiv preprint arXiv:2012.05906*.

Dhar, R., & Zhu, N. (2006). Up close and personal: Investor sophistication and the disposition effect. *Management Science, 52*(5), 726-740.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. Twenty-fourth international joint conference on artificial intelligence,

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2016). Knowledge-driven event embedding for stock prediction. Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers,

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87.

Dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers,

Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: systematic literature review. *Procedia Computer Science, 161*, 707-714.

Emerson, S., Kennedy, R., O'Shea, L., & O'Brien, J. (2019). Trends and applications of machine learning in quantitative finance. 8th international conference on economics and finance research (ICEFR 2019),

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance, 25*(2), 383-417. https://doi.org/10.2307/2325486

Fama, E. F. (1995). Random walks in stock market prices. *Financial analysts journal, 51*(1), 75-80.

Fan, R., Talavera, O., & Tran, V. (2020, 2020/10/01). Social media, political uncertainty, and stock markets. *Review of Quantitative Finance and Accounting, 55*(3), 1137-1153. https://doi.org/10.1007/s11156-020-00870-4

Feuerriegel, S., & Prendinger, H. (2016). News-based trading strategies. *Decision Support Systems, 90*, 65-74. https://doi.org/10.1016/j.dss.2016.06.020

French, K. R. (1980). Stock returns and the weekend effect. *Journal of Financial Economics, 8*(1), 55-69.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. icml,

Fromlet, H. (2001). Behavioral Finance-Theory and Practical Application: SYSTEMATIC ANALYSIS OF DEPARTURES FROM THE HOMO OECONOMICUS PARADIGM ARE ESSENTIAL FOR REALISTIC FINANCIAL RESEARCH AND ANALYSIS. *Business Economics, 36*(3), 63-69. http://www.jstor.org/stable/23488166

Gálvez, R. H., & Gravano, A. (2017). Assessing the usefulness of online message board mining in automatic stock prediction systems. *Journal of computational science, 19*, 43-56.

Ghanavati, M., Wong, R. K., Chen, F., Wang, Y., & Fong, S. (2016). A generic service framework for stock market prediction. 2016 IEEE International Conference on Services Computing (SCC),

Grinblatt, M., & Keloharju, M. (2001). What makes investors trade? *The Journal of Finance, 56*(2), 589-616.

Guzman, E., & Maalej, W. (2014). How do users like this feature? a fine grained sentiment analysis of app reviews. 2014 IEEE 22nd international requirements engineering conference (RE),

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.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019, 2019/06/15/). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, *124*, 226-251. https://doi.org/https://doi.org/10.1016/j.eswa.2019.01.012

Hsu, M.-W., Lessmann, S., Sung, M.-C., Ma, T., & Johnson, J. E. V. (2016, 2016/11/01/). Bridging the divide in financial market forecasting: machine learners vs. financial economists. *Expert Systems with Applications*, *61*, 215-234. https://doi.org/https://doi.org/10.1016/j.eswa.2016.05.033

Huang, W., Nakamori, Y., & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & operations research, 32*(10), 2513-2522.

Hui, B., Zheng, X., & Jia-Hong, L. (2018). Investor sentiment extracted from internet stock message boards and its effect on Chinese stock market. *Journal of Management Sciences in China, 21*(4), 91-106.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media,

Ji, S., Kim, J., & Im, H. (2019). A comparative study of bitcoin price prediction using deep learning. *Mathematics, 7*(10), 898.

Jin, Z., Yang, Y., & Liu, Y. (2020). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing and Applications, 32*(13), 9713-9729.

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I* (pp. 99-127). World Scientific.

Kaity, M., & Balakrishnan, V. (2020). Sentiment lexicons and non-English languages: a survey. *Knowledge and information systems*, 1-36.

Kalyanaraman, V., Kazi, S., Tondulkar, R., & Oswal, S. (2014). Sentiment analysis on news articles for stocks. 2014 8th Asia Modelling Symposium,

Kara, M., Ulucan, A., & Atici, K. B. (2019). A hybrid approach for generating investor views in Black–Litterman model. *Expert Systems with Applications, 128*, 256-270. https://doi.org/10.1016/j.eswa.2019.03.041

Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications, 38*(5), 5311-5319.

Kearney, C., & Liu, S. (2014, 2014/05/01/). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis, 33*, 171-185. https://doi.org/https://doi.org/10.1016/j.irfa.2014.02.006

Keim, D. B. (1983). Size-related anomalies and stock return seasonality: Further empirical evidence. *Journal of Financial Economics, 12*(1), 13-32.

Khan, W., Malik, U., Ghazanfar, M. A., Azam, M. A., Alyoubi, K. H., & Alfakeeh, A. S. (2019). Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis. *Soft Computing*, 1-25.

Khan, W., Malik, U., Ghazanfar, M. A., Azam, M. A., Alyoubi, K. H., & Alfakeeh, A. S. (2020). Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis. *Soft computing (Berlin, Germany), 24*(15), 11019-11043. https://doi.org/10.1007/s00500-019-04347-y

Kowsari, Jafari, M., Heidarysafa, Mendu, Barnes, & Brown. (2019). Text Classification Algorithms: A Survey. *Information (Basel), 10*(4), 150. https://doi.org/10.3390/info10040150

Kumar, M., & Thenmozhi, M. (2006). Forecasting stock index movement: A comparison of support vector machines and random forest. Indian institute of capital markets 9th capital markets conference paper,

Lambert, D. R. (1983). Commodity channel index: Tool for trading cyclic trends. *Technical Analysis of Stocks & Commodities, 1*, 47.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., & Allan, J. (2000). Language models for financial news recommendation. Proceedings of the ninth international conference on Information and knowledge management,

Li, T., van Dalen, J., & van Rees, P. J. (2018). More than just noise? Examining the information content of stock microblogs on financial markets. *Journal of Information Technology, 33*(1), 50-69.

Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. *Information Processing & Management, 57*(5), 102212.

Li, X., Xie, H., Song, Y., Zhu, S., Li, Q., & Wang, F. L. (2015). Does summarization help stock prediction? A news impact analysis. *IEEE intelligent systems, 30*(3), 26-34.

Li, Y., Bu, H., Li, J., & Wu, J. (2020, 2020/10/01/). The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning. *International journal of forecasting, 36*(4), 1541-1562. https://doi.org/https://doi.org/10.1016/j.ijforecast.2020.05.001

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463). Springer.

Liu, Q., Cheng, X., Su, S., & Zhu, S. (2018). Hierarchical complementary attention network for predicting stock price movements with news. Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Liu, Y., Zeng, Q., Yang, H., & Carrio, A. (2018). Stock price movement prediction from financial news with deep learning and knowledge graph embedding. Pacific Rim Knowledge Acquisition Workshop,

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. https://doi.org/https://doi.org/10.1111/j.1540-6261.2010.01625.x

Luo, B., Zeng, J., & Duan, J. (2016). Emotion space model for classifying opinions in stock message board. *Expert Systems with Applications, 44*, 138-146.

Majhi, R., Panda, G., Sahoo, G., Panda, A., & Choubey, A. (2008). Prediction of S&P 500 and DJIA stock indices using particle swarm optimization technique. 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence),

Malandri, L., Xing, F. Z., Orsenigo, C., Vercellis, C., & Cambria, E. (2018). Public Mood–Driven Asset Allocation: the Importance of Financial Sentiment in Portfolio Management. *Cognitive computation, 10*(6), 1167-1176. https://doi.org/10.1007/s12559-018-9609-2

Malkiel, B. G. (1999). *A random walk down Wall Street: including a life-cycle guide to personal investing*. WW Norton & Company.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *Journal of economic perspectives, 17*(1), 59-82.

Manojlović, T., & Štajduhar, I. (2015). Predicting stock market trends using random forests: A sample of the Zagreb stock exchange. 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),

Maqsood, H., Mehmood, I., Maqsood, M., Yasir, M., Afzal, S., Aadil, F., Selim, M. M., & 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.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics, 5*(4), 115-133.

Menkhoff, L., & Taylor, M. P. (2007). The obstinate passion of foreign exchange professionals: technical analysis. *Journal of Economic Literature, 45*(4), 936-972.

Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica: Journal of the Econometric Society*, 867-887.

Mittermayer, M.-A. (2004). Forecasting intraday stock price trends with text mining techniques. 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the,

Moghaddam, A. H., Moghaddam, M. H., & Esfandyari, M. (2016). Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science, 21*(41), 89-93.

Murphy, J. J. (1999). *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin.

Naderi Semiromi, H., Lessmann, S., & Peters, W. (2020). News will tell: Forecasting foreign exchange rates based on news story events in the economy calendar. *The North American journal of economics and finance, 52*, 101181. https://doi.org/10.1016/j.najef.2020.101181

Nartea, G. V., Ward, B. D., & Djajadikerta, H. G. (2009). Size, BM, and momentum effects and the robustness of the Fama-French three-factor model: Evidence from New Zealand. *International Journal of Managerial Finance*.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications, 41*(16), 7653-7670.

Nayak, A., Pai, M. M., & Pai, R. M. (2016). Prediction models for Indian stock market. *Procedia Computer Science, 89*, 441-449.

Nayak, R. K., Mishra, D., & Rath, A. K. (2015). A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices. *Applied soft computing, 35*, 670-680.

Nemes, L., & Kiss, A. Prediction of stock values changes using sentiment analysis of stock news headlines. *Journal of information and telecommunication (Print), ahead-of-print*(ahead-of-print), 1-20. https://doi.org/10.1080/24751839.2021.1874252

Nenortaite, J., & Simutis, R. (2004). Stocks' trading system based on the particle swarm optimization algorithm. International Conference on Computational Science,

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.

Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020, 2020/08/01/). Deep learning for financial applications : A survey. *Applied soft computing, 93*, 106384. https://doi.org/https://doi.org/10.1016/j.asoc.2020.106384

Özorhan, M. O., Toroslu, İ. H., & Şehitoğlu, O. T. (2017). A strength-biased prediction model for forecasting exchange rates using support vector machines and genetic algorithms. *Soft Computing, 21*(22), 6653-6671.

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

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications, 42*(4), 2162-2172.

Plyakha, Y., Uppal, R., & Vilkov, G. (2012). Why does an equal-weighted portfolio outperform value-and price-weighted portfolios? *Available at SSRN 2724535*.

Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications, 181*(1), 25-29.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull., 23*(4), 3-13.

Rechenthin, M., Street, W. N., & Srinivasan, P. (2013). Stock chatter: Using stock sentiment to predict price direction. *Algorithmic Finance, 2*(3-4), 169-196.

Renault, T. (2020). Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages. *Digital Finance, 2*(1-2), 1-13. https://doi.org/10.1007/s42521-019-00014-x

Roy, S. S., Mittal, D., Basu, A., & Abraham, A. (2015). Stock market forecasting using LASSO linear regression model. Afro-European Conference for Industrial Advancement,

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures, 12*(2-3), 140-157.

S&P Dow Jones Indices. (2021). *S&P 500*. https://www.spglobal.com/spdji/en/indices/equity/sp-500.

Sable, S., Porwal, A., & Singh, U. (2017). Stock price prediction using genetic algorithms and evolution strategies. 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA),

Savenkov, D., Braslavski, P., & Lebedev, M. (2011). Search snippet evaluation at Yandex: lessons learned and future directions. International Conference of the Cross-Language Evaluation Forum for European Languages,

Schmidt, A. B. (2019). Managing portfolio diversity within the mean variance theory. *Annals of Operations Research, 282*(1), 315-329.

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.

Schumaker, R. P., Zhang, Y., Huang, C.-N., & Chen, H. (2012, 2012/06/01/). Evaluating sentiment in financial news articles. *Decision Support Systems, 53*(3), 458-464. https://doi.org/https://doi.org/10.1016/j.dss.2012.03.001

Shah, P., & Bhavsar, C. (2015). Predicting Stock Market using Regression Technique. *Res. J. Financ. Account, 6*(3), 27-34.

Sharpe, W. F. (1994). The sharpe ratio. *Journal of portfolio management, 21*(1), 49-58.

Shen, S., Jiang, H., & Zhang, T. (2012). Stock market forecasting using machine learning algorithms. *Department of Electrical Engineering, Stanford University, Stanford, CA*, 1-5.

Shleifer, A., & Summers, L. H. (1990). The noise trader approach to finance. *Journal of economic perspectives, 4*(2), 19-33.

Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance, 52*(1), 35-55.

Singh, S., Madan, T. K., Kumar, J., & Singh, A. K. (2019, 5-6 July 2019). Stock Market Forecasting using Machine Learning: Today and Tomorrow. 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT),

Slovic, P. (1992). Perception of risk: Reflections on the psychometric paradigm. In.

Sohangir, S., Petty, N., & Wang, D. (2018). Financial sentiment lexicon analysis. 2018 IEEE 12th International Conference on Semantic Computing (ICSC),

Song, Q., Liu, A., & Yang, S. Y. (2017). Stock portfolio selection using learning-to-rank algorithms with news sentiment. *Neurocomputing (Amsterdam), 264*, 20-28. https://doi.org/10.1016/j.neucom.2017.02.097

Sprenger, T. O., Sandner, P. G., Tumasjan, A., & Welpe, I. M. (2014). News or noise? Using Twitter to identify and understand company-specific news flow. *Journal of Business Finance & Accounting, 41*(7-8), 791-830.

Steyn, D. H., Greyling, T., Rossouw, S., & Mwamba, J. M. (2020). *Sentiment, emotions and stock market predictability in developed and emerging markets*.

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018, 2018/11/15/). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications, 110,* 298-310. https://doi.org/https://doi.org/10.1016/j.eswa.2018.06.022

Taylor, M. P., & Allen, H. (1992). The use of technical analysis in the foreign exchange market. *Journal of international Money and Finance, 11*(3), 304-314.

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.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance, 63*(3), 1437-1467.

Turchenko, V., Beraldi, P., De Simone, F., & Grandinetti, L. (2011). Short-term stock price prediction using MLP in moving simulation mode. Proceedings of the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems,

Turner, Z., Labille, K., & Gauch, S. (2021). Lexicon-based sentiment analysis for stock movement prediction. *Journal of Construction Materials, 2*, 3-5.

Usmani, S., & Shamsi, J. A. (2021). News sensitive stock market prediction: literature review and suggestions. *PeerJ Computer Science, 7*, e490.

Van de Kauter, M., Breesch, D., & Hoste, V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications, 42*(11), 4999-5010.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Vargas, M. R., De Lima, B. S., & Evsukoff, A. G. (2017). Deep learning for stock market prediction from financial news articles. 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA),

Vision, O. S. C. (2021). *Introduction to Support Vector Machines*. https://docs.opencv.org/3.4/d1/d73/tutorial_introduction_to_svm.html

Weng, B., Lu, L., Wang, X., Megahed, F. M., & Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. *Expert Systems with Applications, 112*, 258-273.

Wilder, J. W. (1978). *New concepts in technical trading systems*. Trend Research.

Wu, D. D., Zheng, L., & Olson, D. L. (2014). A decision support approach for online stock forum sentiment analysis. *IEEE transactions on systems, man, and cybernetics: systems, 44*(8), 1077-1087.

Yan, D., Zhou, G., Zhao, X., Tian, Y., & Yang, F. (2016). Predicting stock using microblog moods. *China Communications, 13*(8), 244-257.

Yang, J., Rao, R., Hong, P., & Ding, P. (2016). Ensemble model for stock price movement trend prediction on different investing periods. 2016 12th International Conference on Computational Intelligence and Security (CIS),

Yıldırım, D. C., Toroslu, I. H., & Fiore, U. (2021). Forecasting directional movement of Forex data using LSTM with technical and macroeconomic indicators. *Financial Innovation, 7*(1), 1-36.

Yu, L., Chen, H., Wang, S., & Lai, K. K. (2008). Evolving least squares support vector machines for stock market trend mining. *IEEE transactions on evolutionary computation, 13*(1), 87-102.

Zhang, X., Shi, J., Wang, D., & Fang, B. (2018). Exploiting investors social network for stock prediction in China's market. *Journal of computational science, 28*, 294-303.

Zhang, Z., Wu, G., & Lan, M. (2015). Ecnu: Multi-level sentiment analysis on twitter using traditional linguistic features and word embedding features. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)