*TRAN DUC ANH DANG | 103995439*

*Swinburne University of Technology*

Leverage sentiment analysis for predictive modelling in stock market dynamics

Research Report

# Abstract

This study aims to explore the relationship between public sentiment, as expressed through digital media, and stock market dynamics. Utilizing sentiment analysis and machine learning algorithms, this research will examine how different sentiments correlate with stock market movements, potentially unveiling predictive patterns. The findings of this study could provide investors and policy makers with valuable insights into how public sentiment influences market behaviour, thus contributing to a more comprehensive understanding of stock market dynamics.

Table of Contents

[Abstract 1](#_Toc149489666)

[Introduction 3](#_Toc149489667)

[Background 3](#_Toc149489668)

[Objective 3](#_Toc149489669)

[Scope and Limitations 3](#_Toc149489670)

[Scope 3](#_Toc149489671)

[Limitations 3](#_Toc149489672)

[Literature Review 4](#_Toc149489673)

[Previous Work on Sentiment Analysis 4](#_Toc149489674)

[Sentiment Analysis in Finance 4](#_Toc149489675)

[Gap Identification 4](#_Toc149489676)

[Research 5](#_Toc149489677)

[Data Collection 5](#_Toc149489678)

[Sentiment Analysis Framework 5](#_Toc149489679)

[Predictive Modelling 5](#_Toc149489680)

[Evaluation Metrics 6](#_Toc149489681)

[Summary 7](#_Toc149489682)

[Training Process 7](#_Toc149489683)

[Predict Process 7](#_Toc149489684)

[Implementation 7](#_Toc149489685)

[Practical Implications 7](#_Toc149489686)

[Theoretical Contributions 7](#_Toc149489687)

[References 7](#_Toc149489688)

# Introduction

## Background

The financial market is a complex ecosystem where myriad factors interact to dictate the price movement of a securities. One such factor, which has recently garnered attention is the sentiment prevalent among investors and the general populace. Sentiment analysis often associated with the field of Natural Language Processing (NLP), provides a mechanism to gauge public sentiment through the analysis of textual data available on various platforms like social media, news, financial forums. In recent years, the explosion of digital data has provided a fertile ground for the application of sentiment analysis in understanding and predicting stock market dynamics. Previous studies have demonstrated a correlation between public sentiment and stock market movements, indicating the potential of sentiment analysis as a tool for predictive modelling in financial markets.

## Objective

The primary objective of this research is to explore the potential of sentiment analysis in predicting stock market trends, which include:

* Developing a sentiment analysis framework capable of processing and analysing large volumes of textual data related to stock market.
* Investigating the correlation between public sentiment, as gauged through the sentiment analysis framework, and stock price movement across different time frames.
* Creating predictive models to leveraging sentiment analysis to forecast stock market trends and assessing the accuracy and reliability of these models in comparison to traditional financial models.
* Identifying the practical implications and potential benefits for investors and financial analysis.

## Scope and Limitations

### Scope

* The study will focus on major U.S stock market indices over a period of five years.
* Data for sentiment analysis will be sourced from publicly available platforms including social media, financial news, etc.
* Various machine learning and NLP techniques will be employed to develop the sentiment analysis framework and predictive models.

### Limitations

* The accuracy and reliability of sentiment analysis and predictive modelling are contingent on the quality and volume of data that available.
* Sentiment analysis may be influenced by external factors such as political events or global crises which are outside the scope of this study.
* The rapidly evolving nature of financial markets are trading technologies may introduce unforeseen challenges in the application and effectiveness of sentiment based predictive models.

# Literature Review

## Previous Work on Sentiment Analysis

Sentiment analysis, situated at the intersection of Natural Language Processing (NLP) and machine learning, has evolved significantly over the years. Its applications span across various domains ranging from consumer product reviews to political discourse analysis. A systematic review on sentiment analysis revealed the employment of hybrid tools and techniques to enhance the accuracy and efficiency of sentiment analysis.

In another dimension, sentiment analysis is often juxtaposed with emotion detection, where the former primarily concerns polarity (positive, negative, neutral) while the latter delves into the emotional or psychological state or mood of the text. This distinction highlights the subjective nature of sentiment analysis as compared to the more objective and precise nature of emotion detection.

## Sentiment Analysis in Finance

The application of sentiment analysis in the financial domain aims at understanding market dynamics by analysing investors' sentiments through various textual data sources like news articles, social media, or financial forums. A critical review of sentiment analysis in predicting financial markets highlighted the examination of 24 papers focusing on this domain, indicating a growing interest in leveraging sentiment analysis for financial market predictions.

Social media, being a rich source of raw data reflecting public sentiment, has been especially utilized for sentiment analysis in the financial domain. The methods explored include various machine learning algorithms to process the textual data from social media platforms and derive insights regarding market sentiment.

## Gap Identification

Despite the promising strides, several gaps persist in the current literature:

* **Granularity of Analysis**: The granularity with which sentiments are analysed varies across studies, with many adopting a broad categorization into positive, negative, or neutral sentiments. A more nuanced approach capturing a wider spectrum of emotions could potentially offer more precise correlations with stock market movements.
* **Temporal Dynamics**: The temporal aspect of the influence of sentiment on stock market dynamics has not been exhaustively explored. Understanding how sentiment impact evolves over different time frames could be pivotal for developing accurate predictive models.
* **Predictive Modelling**: The transition from identifying correlations to developing reliable predictive models leveraging sentiment analysis is still in the nascent stages. The rigorous testing and validation of sentiment-based predictive models across different market conditions and time frames are requisite for advancing the field.
* **Integration with Traditional Models**: The amalgamation of sentiment analysis with traditional financial models and indicators has been less explored. Investigating how sentiment analysis can complement existing financial models could provide a more robust framework for predicting stock market dynamics.
* **Big Data and Sentiment Analysis**: The advent of big data has provided an avenue for classifying diverse sentiments and deriving commercial insights from text-oriented content. However, the systematic integration and analysis of big data in sentiment analysis frameworks in the context of financial markets require further exploration.

# Research

## Data Collection

Data collection for sentiment analysis in finance can be accomplished through various platforms where financial discussions are held. A notable example is the utilization of microblogging platforms like StockTwits, where a dataset of one million messages was employed to evaluate pre-processing methods and machine learning algorithms for sentiment analysis in finance. Other sources include financial news data, as depicted in a study where time-series data analysis was conducted on the closing price data of Infosys Company from 2014 to 2018, showing a strong relation between finance news data and stock market prices​​. Furthermore, a unified solution for data collection, analysis, and visualization in real-time stock market prediction was proposed, retrieving, and processing relevant financial data from news articles, social media, and company technical information​.

## Sentiment Analysis Framework

Various frameworks have been proposed for sentiment analysis in financial markets. A framework introduced in a study combined text mining methods to filter relevant and meaningful information from the textual content, addressing the challenge of information overload as financial markets become faster and more complex​. Other frameworks have employed lexicon-based sentiment analysis models for opinion mining and deep sentiment analysis to understand how stock markets respond to various news categories in the short, medium, and long term​​. More advanced frameworks like the knowledge graph-based sentiment analysis system have also been developed for the stock investment market, consisting of three main modules: knowledge graph module, embedding layer module, and recurrent convolutional neural network (RCNN) module​.

## Predictive Modelling

Predictive modelling in this context aims to harness the potential of sentiment analysis for forecasting stock market dynamics. Utilizing sentiment scores extracted from digital media as crucial input features can significantly enhance the model's ability to forecast stock market trends.

* Sentiment Analysis Integration
  + The heart of predictive modelling in this study revolves around the integration of sentiment analysis. By employing advanced NLP techniques to extract sentiment scores from a vast array of textual data including social media, news articles, and financial forums, a robust sentiment analysis framework can be established. This framework can serve as a vital foundation for extracting sentiment features which will be instrumental in feeding the predictive models.
* Sentiment-driven Feature Engineering
  + Feature engineering is a cornerstone for building powerful predictive models. In this study, sentiment scores, along with other derived sentiment features such as sentiment volatility, sentiment momentum, and aggregated sentiment indices, can be engineered to serve as input features for the predictive models.
* Model Development
  + Model such as LSTM, GRU, ARIMA, and SARIMAX can be leveraged to develop predictive models. LSTM and GRU can be particularly adept at capturing temporal dependencies in sentiment scores and stock prices over time. Meanwhile, ARIMA and SARIMAX can be employed for univariate and multivariate time-series forecasting respectively, with sentiment scores serving as exogenous inputs in SARIMAX.
* Hybrid Sentiment-Time Series Models
  + Developing hybrid models that marry the strengths of time-series analysis and sentiment analysis can potentially lead to more accurate and insightful forecasts. For instance, a hybrid model could be designed to incorporate sentiment scores and other sentiment-derived features alongside historical stock prices and macroeconomic indicators to forecast stock market dynamics.

## Evaluation Metrics

The effectiveness and precision of the predictive models constructed in this research are pivotal, as they directly influence the reliability of the findings. To gauge the performance of these models, a range of evaluation metrics will be employed. Here's a breakdown of these metrics for further understanding:

1. Mean Absolute Error (MAE):

* This metric calculates the average of absolute differences between the predicted and actual values, offering a straightforward measure of the model's accuracy.

1. Root Mean Square Error (RMSE):

* RMSE is the square root of the average squared differences between the predicted and actual values, which gives more weight to larger errors.

1. Mean Absolute Percentage (MAPE):

* This metric expresses forecast errors as a percentage, making it a scalable and easy-to-interpret measure of accuracy.

1. Confusion Matrix and Derived Metrics:

* These metrics are vital for assessing the model's classification performance, especially in binary or multiclass classification problems.

1. Economic Performance Metrics:

* Metrics such as cumulative returns, maximum drawdown, and Sharpe ratio are essential for evaluating the economic performance of the trading strategies derived from the models.

# Summary

## Training Process

### Data Collection

### Models

### Models Performance

## Predict Process

# Implementation

## Practical Implications

## Theoretical Contributions

# References

Ashtiani, M. N. and Raahemi, B. (2023) “News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review,” Expert systems with applications, 217(119509), p. 119509. doi: 10.1016/j.eswa.2023.119509.

Chai, T. and Draxler, R. R. (2014) “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature,” Geoscientific model development, 7(3), pp. 1247–1250. doi: 10.5194/gmd-7-1247-2014.

Drus, Z. and Khalid, H. (2019) “Sentiment analysis in social media and its application: Systematic literature review,” Procedia computer science, 161, pp. 707–714. doi: 10.1016/j.procs.2019.11.174.

Guo, Y. (2022) “Financial market sentiment prediction technology and application based on deep learning model,” Computational intelligence and neuroscience, 2022, pp. 1–10. doi: 10.1155/2022/1988396.

Hajiali, M. (2020) “Big data and sentiment analysis: A comprehensive and systematic literature review,” Concurrency and computation: practice & experience, 32(14). doi: 10.1002/cpe.5671.

Ligthart, A., Catal, C. and Tekinerdogan, B. (2021) “Systematic reviews in sentiment analysis: a tertiary study,” Artificial intelligence review, 54(7), pp. 4997–5053. doi: 10.1007/s10462-021-09973-3.

Mäntylä, M. V., Graziotin, D. and Kuutila, M. (2018) “The evolution of sentiment analysis—A review of research topics, venues, and top cited papers,” Computer science review, 27, pp. 16–32. doi: 10.1016/j.cosrev.2017.10.002.

Nandwani, P. and Verma, R. (2021) “A review on sentiment analysis and emotion detection from text,” Social network analysis and mining, 11(1). doi: 10.1007/s13278-021-00776-6.

Powers, D. M. W. (2020) “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” arXiv [cs.LG]. Available at: http://arxiv.org/abs/2010.16061 (Accessed: October 29, 2023).

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), pp. 1–13. doi: 10.1007/s42521-019-00014-x.

Tuarob, S. et al. (2021) “DAViS: a unified solution for data collection, analyzation, and visualization in real-time stock market prediction,” Financial innovation, 7(1). doi: 10.1186/s40854-021-00269-7.

Willmott, C. J. and Matsuura, K. (2005) “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” Climate research, 30(1), pp. 79–82. Available at: http://www.jstor.org/stable/24869236.

(No date a) Nih.gov. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8053016/#:~:text=Figure%2011%20depicts%20the%20time,SM%20prices%20using%20LSTM%20methodology (Accessed: October 29, 2023).

(No date b) Ieee.org. Available at: https://ieeexplore.ieee.org/document/6974028 (Accessed: October 29, 2023).

(No date c) Researchgate.net. Available at: https://www.researchgate.net/publication/298725275\_A\_new\_metric\_of\_absolute\_percentage\_error\_for\_intermittent\_demand\_forecasts (Accessed: October 29, 2023).

(No date d) Ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=2308659 (Accessed: October 29, 2023).