******

Enhancing Stock Return Forecasts

Using Sentiment Analysis

**Thesis**

Submitted in partial fulfillment of the degree of

**Bachelor of Science**

**in International Business Information Systems**

at the Faculty Business Information Systems

of the Furtwangen University of Applied Sciences

submitted by the author

**Aly Elhadad**

First supervisor: Prof. Dr. Peter Schanbacher

Second supervisor: Prof. Dr. Holger Ziekow

Summer semester 2024

Submitted on: 28.07.2024

Affirmation in lieu of an oath

I assure that I have written the above work independently and have not used any other aids for this purpose than those indicated. All passages of the work that have been taken literally or analogously from external sources are marked as such.

The paper has not been submitted in the same or similar form as an examination paper in any other course of study or published elsewhere.

I am aware that a false declaration can have legal consequences.

Place, Date Signature

Furtwangen, den 28.07.2024 Name of the author: **Aly Elhadad**

# Abstract

# Table of contents

Abstract 4

Table of contents V

List of figures VII

List of tables IX

1 Introduction 1

1.1 Background 1

1.2 Problem statement and task definition 2

1.3 Significance of the Study 4

2 Literature Review 6

3 Methodology 9

3.1 Data Collection 9

3.1.1 Historical Stock Price Data 9

3.1.2 Sentiment Data from Trustpilot 9

4 Conclusion and outlook 15

5 Bibliography and references 16

6 Glossary 17

7 Annex 17

# List of figures

# List of tables

Table 1, List of scraped companies.

.

# Introduction

## Background

In most countries, the stock market is the biggest player in the economy. It is the strongest indicator of how an economy is performing, and over the years it has allowed us – as both retail investors and venture investors – to put investments where they are most needed and where it provides the most value. This in turn has boosted technological advances over the years and solidified the stock market’s position as the free spirit of the economy.

This, of course, also means that those who pay close attention to the stock market and monitor its movements over a period of time have the chance to generate personal profit, by investing in stocks and seeing their investments skyrocket in value when times are good and sell right before the stock price starts to plumet down. However, this is an optimal situation; one that is not always easy to replicate, as it requires not only faith in the stock market and the specific stock in mind but also a lot of technical analysis that considers the many variables that influences the price of a stock.

Luckily, over the years a lot of technological advancements have been made in this field, making it easier to access the historical price figures of a certain stock and automatically assessing all the possibilities for that stock and their probabilities as well. One should be very careful nevertheless, as those are just predictions and are made to help an investor make the decision and not make the decision for them.

One area where this thesis is interested in exploring is customer sentiment, and whether it has an influence on stock prices and turnover. It’s not wild or far fetched to think that even big companies and cooperations are, at the end of the day, not invulnerable to being affected by bad sentiment. After all, customer is always right and their needs should always be accommodated.

## Problem statement and task definition

Stock price prediction always deals with many variables, such as the economic condition, industry trends, and the specific financial performance of companies. It is very clear that the very complexity of the market is what makes the accurate prediction of stock price movements so tricky. Traditional models, most based on historical prices and technical indicators, often don't come close to encompassing the full range of factors that affect market dynamics. Models can use these along with sentiment analysis models to give an insight into the market sentiment that might be directed at the different stocks, all sourced from textual data and various other platforms like social media and review sites.

In this thesis, we aim to improve a pre-trained stock price prediction model –prophet- by incorporating sentiment data sourced from Trustpilot. Trustpilot is an excellent source of user-generated reviews and ratings that reflect consumer sentiment toward companies and their products. We hypothesize that integrating this sentiment data will enhance predictive accuracy for the model and provide an enriched view of factors driving stock price movements. The enriched model will be tested through rigorous estimation by sentiment data contribution to predictive performance improvement.

Task Definition: The primary goal of this study is to find out the impacts of sentiment analysis extracted from Trustpilot reviews on a pre-trained model of stock price prediction. Thus, the study will pursue the following tasks:

Data Collection: Historical stock prices for listed companies will be collected from reliable financial databases -Yahoo finance-. Sentiment data from Trustpilot, including reviews, dates and ratings, for these firms will also be collected.

Sentiment Analysis: Process the Trustpilot reviews with some sentiment analysis techniques to measure the sentiment being said in the reviews. This step is expected to leverage natural language processing and other tools in analyzing and classifying data on sentiment.

Model Integration: Integrate the quantified sentiment data with the existing pre-trained stock price prediction model and ensure that the model can make predictions with possibly included past price data and sentiment scores.

Training and Testing of the Model: Train the developed and refined model on a subset of collected data to ensure that training is robustly done. Test the trained model on a separate dataset for prediction performance evaluation.

Performance Evaluation: The review of the model is done using some main performance measures, such as root mean square error, mean absolute error, mean absolute percentage error, and R-squared. The performance of the model should be compared to the enhanced version of the baseline pre-trained model to demonstrate the impact of sentiment data.

Analysis and Discussion: Perform a deep analysis of how the model performs across different companies and industries. Provide your comments regarding possible benefits and limitations of incorporating sentiment analysis into stock price prediction. Also discuss any challenges faced during the research and outline possible directions for future studies.

## Significance of the Study

As previously mentioned, it has always been fairly challenging to accurately predict stock trends and price movements. After all, there is only so much that one can know beforehand and control. Traditional models attempt to predict stock prices and turnover using only historical price data. Some models attempt to refine the predictions by considering variables, such as industry trends and other technical indicators.

This study aims not to reinvent the wheel, but to refine it, or at least investigate the possibility of refining it using user/customer sentiment, and it hopes to provide some insight as to whether it improves the process. Customer sentiment, especially from dedicated company review websites such as Trustpilot, offers a rich source of customer reviews. These reviews include initial sentiment represented in a 1–5-star rating and the review date, which is very helpful in determining trends over a certain period. By analyzing these reviews, we can potentially uncover patterns and correlations that were previously overlooked in traditional models.

Adding this sentiment to an already proven pretrained model, which in our case is Prophet from Facebook, should bring another dimension into play. This approach will enable us to incorporate qualitative data into our predictions, thus providing a more holistic view of the factors influencing stock prices. By integrating customer sentiment, we can capture the market's psychological aspects, which are often missed by models relying solely on quantitative data.

Prophet is an open-source model developed by Facebook, specifically designed to forecast a target variable by training from historical data. It is known for its ability to handle missing data and outliers, making it robust for various forecasting tasks. By combining the strengths of Prophet with customer sentiment data, we aim to enhance the accuracy of stock price predictions. This integration not only improves our understanding of market dynamics but also opens new avenues for research and application in financial forecasting.

The significance of this study lies in several key areas:

1. Improved Prediction Accuracy: The future prediction of stock prices will be made better by including customer sentiment. It will result in good decisions for investors, and strategies could be planned to manage the risks for the investor and financial analyst.
2. Market Analysis: This qualitative data will help to analyze the market and its dynamics better with a firm-level view. It could help pinpoint the underlying factors driving market movements, which go beyond what financial metrics dictate.
3. Innovative Methodology: This study contributed to the literature by designing a new approach to financial forecasting. Here, a mix of traditional quantitative models with qualitative sentiment analysis brings a new dimension to the modeling approach.
4. Practical Implication: In a practical sense, the results of the current study can be extended to the stakeholders, such as investors, financial institutions, and policymakers. Enhanced forecasting models will eventually facilitate improved decision-making and better market regulations.
5. Novelty Gap: The present research paper contributes toward closing the marked gap in the existing research because it tests the tendency of customer sentiment as an ex-ante variable. It would contribute to the literature on the validity of psychological and emotional factors that impact financial market outcomes.
6. Implications for Future Research: The results might encourage further studies in financial prediction. This study may motivate other future works using additional qualitative data sources and advanced modeling techniques.

# Literature Review

**Literature Review**

**Introduction** Stock price prediction is one of the most exciting issues for researchers and financial analysts because of the very high potential profits that can be made from accurate predictions. The traditional method analyzes historical price data along with technical indicators. Still, the introduction of big data and machine learning has opened up new horizons on integrating alternative sources into this process, particularly sentiment analysis from social media and news articles. In other words, this literature review discusses how sentiment analysis can be integrated with machine learning models to improve the accuracy of stock price predictions.

**Stock Price Prediction Models** Traditional stock price prediction models are built on historical price data and technical indicators. However, such models often fail to capture minor factors responsible for market dynamics. Recent advances in machine learning have enabled complex, unstructured data, such as textual sentiment, to be included. For instance, Darapaneni et al. [36] developed a hybrid LSTM-RF model for predicting stock prices using social media sentiment data and other influencing factors such as gold, oil, and so on. Their study showed how sentiment analysis enhanced the traditional models in predictive power, and the metrics of RMSE also increased (Darapaneni et al., 2022).

**Sentiment Analysis in Stock Prediction** Sentiment analysis, which is the computational identification and categorization of opinions as expressed in text, has increasingly been applied to the stock market prediction domain. The hypothesis is that investor sentiment, as expressed in social media posts and news articles, can influence the behavior of investors and, by extension, influence stock prices. For instance, in a study by Ho and Huang (2021), they used a collaborative network of sentiment features extracted from social media and data from candlestick charts. They first used a one-dimensional CNN model for sentiment classification of the text part and then a two-dimensional CNN model for the technical part, i.e., to analyze candlestick charts. Their model did quite well against the traditional model, displaying remarkable accuracy compared to other methods for more extended prediction periods (Ho and Huang, 2021). The value of sentiment analysis combined with technical analysis was emphasized in the research of Ferreira et al. (2021). For example, with techniques related to deep learning and data on sentiment, the model indicated much better predictive precision than other models and also stated the worth of incorporating data from several sources (Ferreira, Gandomi, and Cardoso, 2021).

**Comparative Studies and Model Performance** It has also been shown that, in different comparative studies, the model's performance was superior due to the integration of sentiment analysis compared with models built upon historical price data. For instance, Sun, Lachanski, and Fabozzi (2016) and Pagolu et al. (2016) established that including sentiment data from Twitter in their models significantly improved the performance by a huge margin. (Sun, Lachanski, and Fabozzi, 2016; Pagolu et al., 2016). On the same note, Tabari et al. (2018) compared various approaches for analyzing the sentiment in stock market tweets. They concluded that neural-network-based models resulted in more accurate predictions than traditional methods. Their study attests to the proper application of advanced machine-learning techniques for the processing and extracting valuable insights from social media data.

**Limitation and Future Work** Although the empirical results seem to be very promising, there still exist a few limitations in applying sentiment analysis for stock price predictions. There is a big drawback to the sentiment scores due to noise and variability in social media data. Also, the models must be constantly updated and validated over new data since markets are dynamic. An area of future research interest may concentrate on the enhancement of algorithms for sentiment analysis to improve their accuracy and integration with other data sources, such as economic indicators and global news events. Another possibility would be the use of hybrid models integrating the diversity of machine learning techniques to gain further improvement in predictive performance.

**Conclusion** The integration of sentiment analysis within traditional stock price prediction models leads to strong potential with enhancement in accuracy and providing fresh insights into market dynamics. As computational capabilities and data availability expand, the potential for more sophisticated and reliable predictive models will likely follow, paving the way toward making investment decisions with more information and strategy.

# Methodology

## Data Collection

### Historical Stock Price Data

Thanks to Yahoo’s powerful API – yfinance - fetching historical financial data was not a difficult task, The data included daily closing prices, trading volumes and monthly turnover.

### Sentiment Data from Trustpilot

As for the sentiment data, Trustpilot was chosen as the source for customer sentiment, as it hosts thousands of reviews for various companies over a long period of time ranging from 5-10 years.

A python script had to be developed, one that can easily browse through hundrades of pages and scrape the title, content as well as the date and star rating, all those reviews are then stored in a csv formatted file for later processing.

The following companies have been chosen for this study:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Ticker | Industry | | PSX | Financial services | | JNJ | Healthcare | | GE | Industrials | | LOW | Retail | | F | Transportation | | BAC | Financial services | | KHC | Consumer goods | | CMCSA | Telecommunications | | AVGO | Technology | | V | Financial services | | KO | Consumer goods | | ORCL | Technology | | TSLA | Transportation | | MMM | Industrials | | CAT | Industrials | | HD | Retail | | PEP | Consumer goods | | TXN | Technology | | PYPL | Financial services | | CAG | Consumer staples | | MDT | Healthcare | | TMO | Telecommunications | | DIS | Entertainment | | SLB | Energy | | HSY | Consumer goods | | BA | Industrials | | MDLZ | Consumer goods | | |  |  | | --- | --- | | Ticker | Industry | | JPM | Financial services | | MPC | Energy | | DUK | Utilities | | KMB | Consumer goods | | BLK | Investment management | | SBUX | Consumer discretionary | | MA | Financial services | | GM | Transportation | | GIS | Consumer goods | | MRK | Pharmaceuticals | | UNH | Healthcare | | MS | Technology | | CB | Financial services | | CPB | Consumer goods | | D | Utilities | | ABT | Healthcare | | PG | Consumer goods | | HON | Industrials | | CL | Consumer goods | | EL | Chemicals | | SCHW | Financial services | | IBM | Technology | | BYDDY | Transportation | | LLY | Pharmaceuticals | | GS | Financial services | | CSCO | Technology | | INTC | Technology | | CVX | Energy | |

Table 1, List of scraped companies.

# Conclusion and outlook

# Bibliography and references

* Darapaneni, N., Paduri, A. R., Sharma, H., Manjrekar, M., Hindlekar, N., Bhagat, P., Aiyer, U., and Agarwal, Y. (2022) 'Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets', *arXiv*. Available at: <https://arxiv.org/abs/2204.05783> (Accessed: 28 June 2024).
* Ho, T. T., and Huang, Y. (2021) 'Stock Price Movement Prediction Using Sentiment Analysis and CandleStick Chart Representation', *Sensors*, 21(23), 7957. Available at: <https://doi.org/10.3390/s21237957> (Accessed: 28 June 2024).
* Ferreira, F., Gandomi, A., and Cardoso, R. (2021) 'Deep Learning for Stock Market Prediction Using Sentiment and Technical Analysis', *SN Computer Science*. Available at: <https://link.springer.com/article/10.1007/s42979-021-00777-7> (Accessed: 28 June 2024).
* Sun, A., Lachanski, M., and Fabozzi, F. (2016) 'Trade the tweet: social media text mining and sparse matrix factorization for stock market prediction', *International Review of Financial Analysis*, 48, pp. 272–281. Available at: https://doi.org/10.1016/j.irfa.2016.10.009 (Accessed: 28 June 2024).
* Pagolu, V. S., Reddy, K. N., Panda, G., and 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), pp. 1345–1350. Available at: <https://doi.org/10.1109/SCOPES.2016.7955659> (Accessed: 28 June 2024).
* Tabari, N., Seyeditabari, A., Peddi, T., Hadzikadic, M., and Zadrozny, W. (2018) 'A comparison of neural network methods for accurate sentiment analysis of Stock Market Tweets', *ECML PKDD 2018 Workshops. MIDAS 2018, PAP 2018*, LNCS, vol 11054, Springer. Available at: https://doi.org/10.1007/978-3-030-10997-4\_16 (Accessed: 28 June 2024).

# Glossary

API:

# Annex