Exploring the Use of Sentiment Data in machine learning stock Market Predictions

WARD DE LANGE, University of Twente, The Netherlands

ABSTRACT

This research paper investigates the utilization of sentiment data from social media (specifically Twitter) and financial news articles in machine learning models for predicting stock market movements. Various machine learning algorithms, including Neural Networks (NN), Support Vector Machines (SVM), Naïve Bayes (NB), Long Short-Time Memory (LSTM), and Random Forest (RF), are examined for their effectiveness in sentiment analysis and stock market prediction. The study finds that sentiment analysis of social media and news articles provides valuable insights into stock market sentiment, with LSTM and SVM showing high accuracy in predicting stock movements. The results highlight the potential benefits of incorporating sentiment data alongside historical price data for stock price predictions, but also show that variations exist in effectiveness of sentiment data for predictions of different stocks. This research contributes to the existing literature, guiding researchers in developing more robust stock market forecasting models, ultimately improving investment strategies.

Additional Key Words and Phrases: stock market prediction, machine learning, sentiment analysis

1 INTRODUCTION

For multiple centuries, the stock market has been a central part of the economy. The place where companies and investors come together to come to agreements over the value of a company stock. In the past, this interaction happened in centralized places like a stock exchange. However, since the coming of the internet, the stock market has changed a lot. Stocks can be traded from anywhere in the world by anyone who wants to, and has the means to invest. Where in the past only the very rich and knowledgeable would invest, now anyone can do so with the click of a button. With this change, the reasoning behind these financial decisions has also changed drastically. Where in the past most investors would do thorough research into the business situation of the company before investing, nowadays, a company like Tesla can get funding for years by having a good public opinion while it didn't make any profit until recently [6]. With the change in the stock market, there has also been a change in the reasoning of investors. These days, public opinion and general sentiment toward a company can greatly impact the value of a company on the stock market. More recent development in academic research and the financial market might bring along a new change in the stock market: The use of Artificial Intelligence and Machine Learning to try to predict the direction of a stock. Predicting stock movements accurately has been a challenging task for researchers. Traditional approaches using historical stock price data [4, 20, 21, 39, 41] often fail to capture the complexity of the stock market. To address this issue, researchers have turned to alternative data sources such as sentiment data [8, 22, 25, 28] to improve prediction accuracy. This paper focuses on the use of sentiment data, such as social media and news articles, in machine learning models for stock market predictions. Sentiment analysis can be used to determine the opinion of the wider public on a certain stock or the market as a whole. By incorporating this sentiment data into machine learning models, researchers hope to improve their prediction models and provide more accurate insights into stock movements to investors. This paper begins by discussing the two primary sources of sentiment data: social media and news articles. This section also explores the importance of sentiment analysis to understand public opinion and discusses the improvements in methods for extracting sentiment from textual sources. Secondly, this work evaluates the various machine-learning techniques used for stock price prediction models. The strengths and applications to sentiment analysis and stock price prediction of each machine learning algorithm is discussed. Finally, the results and conclusions of various research papers that incorporated sentiment data into their prediction models are examined. That included papers that solely used sentimental data, as well as studies that used models with historical data paired with sentiment data. With the aim to identify the impact of sentiment data on the prediction accuracy of machine learning algorithms. By investigating the existing research, this paper aims to provide insights into the use of sentiment data in machine-learning stock price predictions. The findings will help researchers better understand the relevance of sentiment analysis and identify the most suitable machine-learning methods for incorporating sentiment data into prediction models. Overall, this research contributes to the ongoing efforts to improve stock market predictions and offers valuable insights for investors seeking more accurate and informed prediction tools.

2 METHODOLOGY

This paper will mainly review previously published literature. This will be done with the systematic literature review(SLR) method. This method consists of systematic steps to review published papers that make use of machine learning to predict the financial market. The first step of SRL is establishing a good research question, with a few sub-research questions to help answer the main question. For this paper, the following research question was chosen:

Main question: *How can the addition of sentiment data sources such as social media sentiment and news articles improve the accuracy of financial market predictions using machine learning models?*

To help answer this question, the following sub-questions will be answered first:

SRQ1: *Which data sources have effectively been used for sentiment analysis to assess public opinion on stocks?*

TScIT 39, July 7, 2023, Enschede, The Netherlands

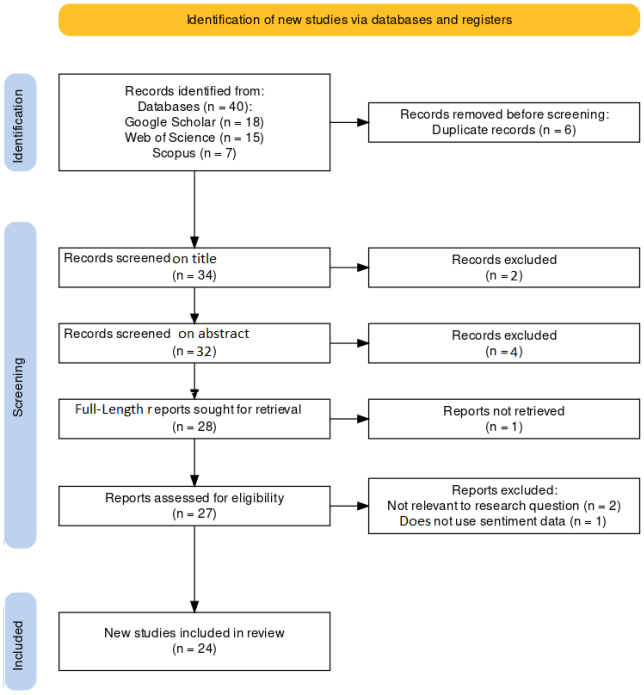

Fig. 1. PRISMA flow chart of literature search

SRQ2: *What machine learning methods are best suited for stock market predictions with the public sentiment?*

2.1 Literature search

To systematically report the literature search process, the PRISMA 2020 statement [31] methodology was used to create a flowchart for clarity(see Figure 1). PRISMA is a guideline to clearly visualize the paper search and selection process in a flowchart for better reproducibility. To make sure papers were relevant, they had to adhere to a few requirements to be considered: published after 2015, written in English, and published in scientific journals or conference papers. The following quarries were used to find papers on Scopus, Google Scholar, and Web of Science:

- Predicting stock prices AND (machine learning OR artificial intelligence OR AI)
- (sentiment analysis OR social media) AND predicting stock prices AND (machine learning OR artificial intelligence OR AI)
- (Twitter data OR Financial news data) AND predicting stock prices AND (machine learning OR artificial intelligence OR AI)
- (social media OR sentiment OR sentiment analysis) AND predicting stock prices AND (machine learning OR artificial intelligence OR AI)

The resulting papers from these searches can be found in Table 1 sorted on the sentiment data source and ML algorithm.

3 SENTIMENT DATA SOURCES

An important step to being able to make use of public opinion to predict the stock market is the ability to assess the sentiment of the people at a specific time. Unlike historical data that is already readily usable, sentiment data consists of text in most cases, which has to be transformed into a usable data format before it can be used. Therefore, this section will focus on the different data sources and data sources that can be used for sentiment analysis of the stock market to answer the following sub-research question: *Which data sources have effectively been used for sentiment analysis to assess public opinion on stocks?*

3.1 Social media

The first data source that has been used for sentiment analysis of the stock market is Twitter data. This was first implemented for the purpose of stock predictions by Bollen et al. [2]in 2011 and has since been widely used in the field. The reason for the usage of Twitter data is mainly the easy access to the data by making use of Twitter APIs to download relevant tweets easily and the fact that tweets have a fairly low character limit which limits the data size. The tweets can be split up into individual words, which can then be used for sentiment analysis. Through the years, multiple different approaches have been tried to convert tweets from plain text to usable sentiment data. In the beginning, this involved manually making a dictionary, linking certain words to the appropriate emotion. At this point, the technique was solely used on individual stocks, but in 2015, Nguyen et al. [30] combined the sentiment analysis with topic recognition to be able to link sentiments to a specific stock and therefore make predictions of multiple stocks at the same time. In more recent studies, researchers have mostly gone to a more simple approach of sentiment analysis that just assigns positive, neutral, or negative sentiment to words instead of more complicated emotions. This move was made to simplify predictions because a negative or positive sentiment toward a stock is easier to process than the eight different emotions originally used by Bollen et al.[2] One of the most recent changes in the sentiment analysis of tweets to predict stocks was the additional usage of emoticons to assess sentiment by Pal et al. [33]. In previous studies, emoticons were discarded together with stop words and other information that was deemed irrelevant to stock sentiment. But Pal et al. included it in the analysis, which might help to better understand the sentiment of tweets.

3.2 News articles

The second widely used source of data for sentiment analysis in the field of stock market prediction with machine learning is Financial news articles. News is an important resource to be able to determine if a stock is doing well or not.[37] Because news articles can have such an effect on stock prices, predicting stock movements might be more accurate if financial news is considered as well as historical data. News articles are a way to learn new information about a certain company or event that has recently happened. This information is used by many investors to make investment choices. Which means it can be a good indication of the sentiment of investors. When a news article is written by a news outlet that is focused on the financial world, it can also include an analysis of the effect

ML algorithm	news articles	Twitter	news+Twitter	other
NN	[1, 10, 29, 40]		[38]	
SVM	[5, 7, 9, 10, 14, 17, 19, 36]	[22, 23]	[15, 24, 25]	[30]
NB	[14, 17, 19]	[16, 22]	[15, 24, 25]	
LSTM	[1, 8, 9, 27, 29, 35, 36]		[25]	
RF	[9, 14, 35, 36]	[16, 32]	[15, 24]	
Other	[9, 17, 19, 27]	[16, 23]	[15, 24, 33]	

Table 1. Papers ordered by sentiment source and ML algorithm

an event might have on the stock of a company. By analysing the words from the headline or even the entire article with sentiment analysis, a positive, neutral, or negative assessment can be made about the sentiment toward the stock or the market. In the stock prediction field news sentiment has been used in many variations with some [1, 8, 36] only using headlines for the analysis, while others [7, 9] make use of the entire text. Almost all of these studies seem to focus their efforts on news of a few stocks to test their models and predictions, but Gong et al. [9] made a prediction model with news sentiment data that included all individual stocks in the S&P500.

3.3 Combination of News and social media

Recent papers have explored using a combination of news articles and Twitter data for sentiment analysis (see Table 1). The reason for doing so is to combine the expert opinion and recent events expressed in news articles and the sentiment of the wider public found on Twitter. Though not that many studies have tried to combine the two yet, the combination might be used more in the future if it proves to give better results than the two data sources individually.

3.4 Challenges of using sentiment data sources

When making use of textual data and sentiment analysis, there are some challenges that have to be addressed to make sure the resulting data is useful and an accurate representation of the sentiment of the source text. The first issue that is common with any data source is the chance of the data source being biased. Twitter data has the problem that a small group of people that have a very strong opinion on a topic will be very vocal, while the people that are not as passionate, stay quiet. This means that the data from Twitter might not accurately reflect the opinions of the actual public. However, this vocal group still influences the population with their loud opinions on social media, which might convince more people to adopt that standpoint and in turn influence stock prices. In that sense, it might not always be an accurate representation of public opinions, but at least an indication of shifting opinions, which is still useful for stock predictions. The same can be said about any potential biases in newspaper articles. Even though the article itself may be biased, it can still have an influence on the opinions of the people that read the article. Therefore, the news article still remains an accurate indicator of shifts in public opinion. The second challenge that a researcher might face when using sentiment analysis is accurately representing the sentiment of the text. The difficulty in this lies mainly in sarcasm, local nuances, context, and more recently emojis.[26] To be able to accurately determine the sentiment of some text, all of these aspects need to be considered. Some of the mentioned papers try to take this into account by making their own dictionary with sentiment associations of specific words and phrases, and Pal et al. [33] even included the sentiment of emojis. But, the broader challenge of accurate sentiment analysis of textual data, is yet to be fully solved.

4 MACHINE LEARNING TECHNIQUES

This section will focus on the different machine learning algorithms used by the researched papers for stock market predictions. Though most algorithms will probably be suitable for the purpose, the right algorithm might improve the accuracy of stock predictions. Therefore, this section will answer the following research question: What machine learning methods are best suited for stock market predictions with public sentiment?

4.1 Neural Network

Some of the studies made use of neural networks for the stock prediction models. These models make use of connected nodes to process input data to influence internal states and the subsequent output of each node, the output is then compared to the expected output. When the output does not match the expected outcome, the weights of different factors are readjusted for the next iteration until an optimal situation is reached. This connected system of nodes is supposed to imitate the way a human brain functions, with many connected brain cells that are capable of learning and eventually coming to the right conclusions. For the prediction of stock prices with sentiment analysis, a few different variations of neural networks were used to be more compatible with unstructured data types. The first of these variations is the convolutional neural network(CNN). This variation is mainly known for its ability to analyse images, which can be used with stock price graphs as input instead of traditional numerical data.[1] Though CNN is mainly used for 2D data input, it seems to also function well with textual data.[18] The second variation is the Recurrent Neural Network (RNN). This variation adds internal states and connections so that the output of a node can be used as its own input. This feature is very useful for stock prediction because it allows for the reuse of information over multiple iterations. This is crucial for future predictions, as some events do not only affect stock prices for one day. Though, this feature falls short when the context is longer term, which is often the case with stock prediction. In this situation, LSTM is better suited than RNN.

4.2 Support Vector Machine

Another commonly used machine learning model is a support vector machine. It is most well known for its ability in classification of data that is not linear. This is likely also the reason it was preferred by many as a model to use for stock price predictions. SVMs allocate every data point to a point in a vector space and use that to classify them into certain groups.[13] This allocation of points to vectors in a space is likely also the reason for its naming. The points that are spaced together, belong to the same class, and the different classes are separated by gaps between points. Another attribute that SVM is known for is its ability to classify text, which seems like a perfect fit for sentiment analysis data because that is in most cases text. Therefore, the use of SVM seems to be well suited for stock predictions that make use of sentiment analysis because it can efficiently classify non-linear and textual data.

4.3 Naïve Bayes

Another often-used machine learning algorithm is the naive Bayes (NB) classifier. This algorithm is particularly known for its assumption of class conditional independence[9], which means that each value of a feature that corresponds to a class is seen as an independent value and is not linked to any other feature of said class. This means that each feature is considered independently for classification. Though it seems to be simple in principle, its simplicity is also one of its strengths because it means it doesn't take too long for large datasets with many features. Its ability for text classification combined with its simplicity and independence assumption seems to make it a popular algorithm for sentiment analysis.

4.4 Long Short-Time Memory

One of the most used machine learning methodologies in more recent research papers is Long Short-Time Memory(LSTM). LSTM is a more advanced version of the recurrent neural network. LSTM makes use of a network of memory cells instead of the nodes used in RNN mentioned earlier. [12] The output of the network is influenced by the state of each cell the data passed through. The memory cells are able to hold information that is deemed important, while forgetting unnecessary information. This combination means a cell can hold important information for a long time, while less important information is only held for a short time. This is an important feature when predicting the stock market because some information will have long-term effects, while other influences only last a few days. a memory cell is made up of three gates that influence the state of a memory cell. The first is the input gate, which regulates the amount of new information passed into the cell. The forget gate determines what parts of that information are important and should be remembered, and which information should be forgotten. The last gate is the output gate, which regulates what information is passed to the next cell of the network. The model has forward pass as well as backward pass capabilities. What makes this model especially well suited for stock prediction is its capability of remembering strong influences from the past and its ability to use textual data as input.

4.5 Random Forest

The last machine learning technique that seems to be popular for stock prediction with sentiment analysis is Random forest. This algorithm makes use of multiple decision trees with randomized features for each decision node.[3] To ensure that the different trees are not correlated to each other, a feature bagging method is used that generates a separate dataset from the test data with replacement for each individual tree.[11] Each tree is then given the same input, which all the different trees will then use to generate an output. The output of the entire random forest algorithm is then determined by the number of times the individual decision trees come to the same classification output or the average value in the case of regression. This way of using decision trees reduces overfitting that normal decision trees are prone to. The main idea is that using many of these unconnected trees will outperform individual models. Random forest is particularly suited for sentiment analysis and stock prediction because it can be used for both regression, which is used for price prediction, and classification for sentiment analysis.

4.6 Analysis

Though, all the aforementioned machine learning techniques have been used for stock prediction with sentiment analysis in the past and seem to be suited for the task. Some of them perform better than others in specific situations. Though many of the researched papers only make use of a single algorithm, the studies that made use of multiple techniques with the same data set can be used to get an idea of which algorithm performs best in the field. In the earlier papers, SVM was a very popular ML algorithm among the researched papers[5, 7, 22, 34], its compatibility with sentiment data and stock prediction was validated by [19] that compared the accuracy of NB, KNN, and SVM on their news sentiment data set and concluded that SVM is most accurate of the three with an average accuracy of 75.45% followed by NB with 72.64% while KNN lagged far behind with 47.99%. This result was further reinforced by [24] who tested their own novel model as well as SVM, RF, and NB on the multiple sentiment data sets. Though, their testing measure was F-measure instead of accuracy. While their novel algorithm outperformed the traditional models with an average F-measure of 73.6, SVM performed the best of the remaining three with 67.8 followed by RF with 61.3 and NB with 60.25. [17] seems to come to different conclusions as they compared the accuracy of SVM, KNN, and NB as well, where NB averaged an accuracy of 80.6% while KNN scored 72.7% and SVM 64.8%. These differences in result could be caused by a difference in the dataset because this paper focussed on the NB algorithm and only used the other two as an additional test. while [19] used different algorithms to test their dataset. After 2019 the preference for SVM seems to shift to LSTM [8, 27, 35] as the standard algorithm, This shift seems to be justified by the performance of LSTM when compared with other ML algorithms. This was done by [25], who compared the accuracy performance of LSTM, SVM, and MB on a Twitter and news sentiment data set. Where LSTM scored highest with an accuracy of 92.45% followed by SVM with 89.46% and NB with 86.72 %. [1] seems to reaffirm this conclusion when comparing the abilities of LSTM and CNN to predict stock closing prices with news sentiment data. This paper

concluded that LSTM performs better on this data set than CNN, but also state that a hybrid model might increase performance. which was researched by [9] which came to the conclusion that a model that combines multiple machine learning and deep learning algorithms can achieve better prediction accuracy than the individual algorithms on their own. So, to answer the question posed at the start of this section: all the mentioned ML models are well suited for the task of stock market predictions with sentiment data. LSTM and CVM seem to perform the best, but choosing the best-suited ML model also depends on the types of data used in the dataset. Lastly, a combination of multiple models into an ensemble model can improve the prediction accuracy even further.

5 RESULTS

To determine the significance of sentiment data for stock predictions with machine learning, the various results and conclusions of the papers that made use of it in the past can be analysed.(see Appendix A for a table with all the results and datasets of the researched papers) The results of these papers seem to fall into a few different categories. The first group consists of papers that use sentiment data and historical price data for their prediction model, but do not test if the accuracy changes when the sentiment data is not used. The second group tests the significance of sentiment data by predicting the market with sentiment data only. The last group tests their model with a dataset that includes sentiment and historical data and a dataset with only historical price data. This last group might be the most useful because different papers have their own dataset with different timeframes and data sources, which makes it impossible to directly compare model accuracy with each other to see if a model that used sentiment data is performing better than a model that doesn't.

5.1 Predictions with historical and sentiment data

Even though the papers in this group did not compare if their model performed better or worse without sentiment data, their results and conclusions can still be useful to consider. The first paper in this group is [14] which concluded that stock trends can be predicted using news articles and stock price history with a prediction accuracy of up to 92%. This prediction accuracy was echoed by [25] which achieved a similar accuracy of 92.45%, which led to the conclusion that there is a strong relationship between news article sentiment and stock movements. [35] took the same prediction approach with historical data combined with news headline sentiment and tested the model with stock trend predictions on TSLA, AMZ, and GOOG stock with prediction accuracies ranging from 90% to 92.3%. While all these papers didn't check prediction accuracy without sentiment data, their high performance might indicate that historical data combined with sentiment data can be really effective for stock prediction. Although another explanation might be that these models were simply overfitted on the datasets used, but this can't be proven because the models were not tested on a new data set.

5.2 Predictions with only sentiment data

This group of papers that used only sentiment data for stock price prediction starts with [7] which tried to predict the Vietnamese stock index(VN30) with only news sentiment data. The model managed to reach an accuracy of 80% for the VN30 and 60-90% when tested on 5 individual companies. [19] tested a news sentiment prediction model by predicting the yearly upward or downward movement of a stock from 2005-2014 with an accuracy of 70%. These results were reinforced by [32] which shows results that opinions and emotions expressed on Twitter about a company can be used to predict the rise and fall of the stock value of that company with a model that achieved 69% accuracy. [17] agrees with the other papers from this group and concludes that a strong relationship between stock prices and financial news articles exists. This conclusion follows from a prediction model that makes use of solely newspaper paper sentiment to predict stock price movements with an accuracy of 59.18% and when historical data was added to the model it reached an accuracy of 89.8%, which comes close to the accuracies found in the first group. [33] modified an initial Twitter and news dataset to consider follower and retweet amount to filter out less impactful data, which helped them reach a prediction accuracy of 73% with solely Twitter data and 81% when both Twitter and news sentiment were taken into consideration. The last paper of the second group is [36] which used news headline sentiment to predict the stock trend of the next day with an accuracy of 84.92%. The results of the studies in the second group show that the use of solely sentiment data can predict stock movements with reasonable accuracy without making use of historical stock data.

5.3 Predictions tested with and without sentiment data

To be able to make conclusions about the added value of sentiment data for stock price prediction, the third group is the most insightful, because these papers also tested the predictions without sentiment data. This group starts with [30] which used their model to predict stock directional movements for five US stocks. The model could predict directional stock movements with an average accuracy of 56.4%. The main conclusion of this paper was that the model that added sentiment data outperformed the model without with an average of 6.07% accuracy, though they also concluded that some stock prices seemed to be more influenced by the sentiment data than others. [24] came to a similar conclusion after testing a multitude of datasets. The model they created was used to test a dataset with only stock price data, two separate datasets that added either Twitter sentiment or news sentiment data to the stock price data, and a dataset with all these data sources combined. Their results revealed that the data set with news performed best with an accuracy of 73.9%, followed by the combined dataset of all features with 73.6%, with a tied last place for the stock price dataset and Twitter data with 67.6% accuracy. These results indicate that news sentiment can increase the efficiency of price predictions, while Twitter sentiment does not really add anything extra and might even be detrimental. The paper also concluded that certain stock prices seemed to be more influenced by sentiment data than others. Moving on to, [27] which concludes that a strong connection between stock prices and financial news articles exists, by testing a model with and without sentiment data. As a testing measure, Mean Absolute Percentage Error was used. The data set with sentiment data scored 2.03%, while the dataset that only used historical data got a MAPE value of 2.13%.

[38] went a different route and tested if a model with sentiment data from Twitter and news articles could better predict stock price directions than a model without sentiment data. This resulted in the model with sentiment data outperforming the model without 77.78% of the time. The paper then tried to determine if this outcome is statistically significant with multiple probabilistic calculations, but they could not prove a significance level of 5%. The paper did conclude, however, that evidence supports sentiment can contribute to stock price prediction. In contrast to the other papers in this group, [16] concluded that the impact of sentiment data on a specific stock price was minimal. The paper offered two possible reasons for their results: Sentiment data has no influence, or the sentiment analysis techniques used are not good enough to provide enough evidence for a connection between sentiment and a stock price movement. The results of the paper show that the added accuracy of sentiment data ranged from 0 to 3%. [29] Agrees with that conclusion because the used LSTM model only showed an increase in accuracy of 0.6% when sentiment data was taken into consideration. The paper did conclude that the addition of recurrence and attention to the ML model showed more promise.

5.4 Analysis of results

The one conclusion that all these papers seem to agree on is that sentiment data does have an effect on the prediction accuracy of stock price prediction with machine learning models. This is shown by the many results where sentiment data alone can predict stock movements to some degree, and many papers show an increase in the accuracy of prediction when sentiment data is added. Though, the level of effect is debated. Another conclusion that might explain the discrepancies in the results is that some stocks are more affected by sentiment data than others. Expanding on that, some stocks seem to be more affected by Twitter data, while others seem to be more influenced by news. So, to answer the main research question of this literature review: There certainly is an effect on accuracy when sentiment data is used in addition to historical price data. The degree of the effect depends on the stock and the sentiment source used.

6 DISCUSSION

6.1 Limitations

This research paper had a number of limitations because of the scope of a literature review. Which limited the research to interpreting the results of previously published literature. One of the consequences of this limitation is that there is a lot of focus on the accuracy of the machine learning models because that was the main testing measure that the different papers used to assess their models. To more accurately determine the performance of these models, other financial indicators such as return on investment and risk-adjusted returns might be able to provide more insight. Another consequence of solely looking into previously published literature in the field is that the recommendations and findings of this paper will reflect on what is popular in the field. Therefore, these findings do not take into account improvements and practices in other fields that have not been used for stock price predictions with sentiment analysis. Investigating developments that have not been tried for this purpose before might impact prediction accuracies better than those that are already established in the field.

6.2 Future work

This paper lays a good groundwork for further study in the field of stock price predictions with sentiment data. So, the next step would be to put the insights of this paper into practice. This would mean making a machine-learning model for stock predictions with historical and sentiment data. Which uses Twitter and news articles for sentiment analysis. The model should be made with LSTM, SVM, or a combination of multiple algorithms into an ensemble model. To further improve this model, more research can be done into advanced sentiment analysis methods that make use of natural language processing algorithms and contextual embedding. Furthermore, the differences in company characteristics should be investigated to determine what influences the level of effect sentiment data has on the stock price predictions of individual companies. Lastly, to be able to really test the performance of such a prediction model, it has to be tested on the real stock market. This can be accomplished by devising a trading strategy for selling and buying stocks according to the model's predictions. With such a strategy in place, the relevancy of the model can be shown with economic performance measures such as return on investment and risk-adjusted returns. Which will be more relevant to investors and other stakeholders than the model accuracy that has been the main testing measure of the researched literature.

7 CONCLUSION

In conclusion, this literature review explores the use of sentiment data in stock price predictions with machine learning. The paper reviewed various previous literature on the subject to provide some insight into the most effective approaches for incorporating sentiment data in stock price predictions. This work identifies social media, Twitter in particular, and news articles as primary sources of sentiment data for assessing public opinion on stocks. These sources provide valuable insights into the sentiment of investors and the impact of events on stock prices. Moreover, a combination of Twitter data and news articles has shown some potential in combining expert analysis with the sentiment of the wider public. Different machine-learning algorithms were explored, which offer the ability to efficiently process unstructured data and find complex patterns for stock price predictions. Most of the researched algorithms are compatible with the purpose of stock price prediction with sentiment data, but SVM and LSTM really stand out as the preferred options in recent literature. finally, the results of all the previous literature indicate at least some correlation between sentiment data and stock price movements. This effect is different for each stock, where some are heavily influenced while others seem to barely be affected. By leveraging sentiment analysis in machine learning models, researchers and investors can certainly benefit and gain a deeper understanding of market dynamics and make more informed decisions in the ever-changing stock market.

REFERENCES

[1] Hetal Bhavsar, Anjali Jivani, Sameer Amesara, Smeet Shah, Prashant Gindani, and Sohamkumar Patel. 2023. Stock Price Prediction Using Sentiment Analysis on News Headlines. *Smart Innovation, Systems and Technologies* 311 (2023), 25–34. https://doi.org/10.1007/978-981-19-3571-8{_}4/COVER

[2] Johan Bollen and Huina Mao. 2011. Twitter Mood as a Stock Market Predictor. *Computer* 44, 10 (10 2011), 91–94. https://doi.org/10.1109/MC.2011.323

[3] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (10 2001), 5–32. https://doi.org/10.1023/A:1010933404324/METRICS

[4] Lin Chen, Zhilin Qiao, Minggang Wang, Chao Wang, Ruijin Du, and Harry Eugene Stanley. 2018. Which Artificial Intelligence Algorithm Better Predicts the Chinese Stock Market? *IEEE Access* 6 (7 2018), 48625–48633. https://doi.org/10.1109/ACCESS.2018.2859809

[5] Raymond Chiong, Marc T.P. Adam, Zongwen Fan, Bernhard Lutz, Zhongyi Hu, and Dirk Neumann. 2018. A sentiment analysis-based machine learning approach for financial market prediction via news disclosures. *GECCO 2018 Companion - Proceedings of the 2018 Genetic and Evolutionary Computation Conference Companion* (7 2018), 278–279. https://doi.org/10.1145/3205651.3205682

[6] Bradford Cornell. 2021. Making Sense of Tesla's Run-up. *SSRN Electronic Journal* (6 2021). https://doi.org/10.2139/SSRN.3857786

[7] Duc Duong, Toan Nguyen, and Minh Dang. 2016. Stock market prediction using financial news articles on Ho Chi Minh stock exchange. *ACM IMCOM 2016: Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication* (1 2016). https://doi.org/10.1145/2857546.2857619

[8] Shilpa Gite, Hrituja Khatavkar, Ketan Kotecha, Shilpi Srivastava, Priyam Maheshwari, and Neerav Pandey. 2021. Explainable stock prices prediction from financial news articles using sentiment analysis. *PeerJ Computer Science* 7 (1 2021), 1–21. https://doi.org/10.7717/PEERJ-CS.340/SUPP-1

[9] Jiaying Gong, Bradley Paye, Gregory Kadlec, and Hoda Eldardiry. 2021. Predicting Stock Price Movement Using Financial News Sentiment. (2021), 503–517. https://doi.org/10.1007/978-3-030-80568-5{_}41

[10] Petr Hajek and Aliaksandr Barushka. 2018. Integrating sentiment analysis and topic detection in financial news for stock movement prediction. *ACM International Conference Proceeding Series* (9 2018), 158–162. https://doi.org/10.1145/3278252.3278267

[11] Tin Kam Ho. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 8 (1998), 832–844. https://doi.org/10.1109/34.709601

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (11 1997), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

[13] Thorsten Joachims. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. (1998), 137–142. https://doi.org/10.1007/BFB0026683

[14] Joshi Kalyani, Prof. H. N. Bharathi, and Prof. Rao Jyothi. 2016. Stock trend prediction using news sentiment analysis. *International Journal of Computer Science and Information Technology* 8, 3 (7 2016), 67–76. https://doi.org/10.5121/ijcsit.2016.8306

[15] Wasiat Khan, Mustansar Ali Ghazanfar, Muhammad Awais Azam, Amin Karami, Khaled H. Alyoubi, and Ahmed S. Alfakeeh. 2022. Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing* 13, 7 (7 2022), 3433–3456. https://doi.org/10.1007/S12652-020-01839-W/FIGURES/20

[16] Wasiat Khan, Usman Malik, Mustansar Ali Ghazanfar, Muhammad Awais Azam, Khaled H. Alyoubi, and Ahmed S. Alfakeeh. 2020. Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis. *Soft Computing* 24, 15 (8 2020), 11019–11043. https://doi.org/10.1007/S00500-019-04347-Y

[17] Ayman E Khedr, S E Salama, and Nagwa Yaseen. 2017. Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. *MECSJ. Intelligent Systems and Applications* 7 (2017), 22–30. https://doi.org/10.5815/ijisa.2017.07.03

[18] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (8 2014), 1746–1751. https://doi.org/10.3115/v1/d14-1181

[19] Mr D K Kirange and Ratnadeep R Deshmukh. 2016. Sentiment Analysis of News Headlines for Stock Price Prediction. *An international journal of advanced computer technology* 5, 3 (2016). https://doi.org/10.13140/RG.2.1.4606.3765

[20] Yuming Li, Pin Ni, and Victor Chang. 2020. Application of deep reinforcement learning in stock trading strategies and stock forecasting. *Computing* 102, 6 (6 2020), 1305–1322. https://doi.org/10.1007/S00607-019-00773-W/TABLES/2

[21] Shu Liu, Bo Wang, Huaxiong Li, Chunlin Chen, and Zhi Wang. 2023. Continual portfolio selection in dynamic environments via incremental reinforcement learning. *International Journal of Machine Learning and Cybernetics* 14, 1 (1 2023), 269–279. https://doi.org/10.1007/S13042-022-01639-Y/TABLES/2

[22] Tejas Mankar, Tushar Hotchandani, Manish Madhwani, Akshay Chidrawar, and C. S. Lifna. 2018. Stock Market Prediction based on Social Sentiments using Machine Learning. *2018 International Conference on Smart City and Emerging Technology, ICSCET 2018* (11 2018). https://doi.org/10.1109/ICSCET.2018.8537242

[23] Haider Maqsood, Irfan Mehmood, Muazzam Maqsood, Muhammad Yasir, Sitara Afzal, Farhan Aadil, Mahmoud Mohamed Selim, and Khan Muhammad. 2020. A local and global event sentiment based efficient stock exchange forecasting using deep learning. *International Journal of Information Management* 50 (2 2020), 432–451. https://doi.org/10.1016/J.IJINFOMGT.2019.07.011

[24] Manolis Maragoudakis and Dimitrios Serpanos. 2016. Exploiting Financial News and Social Media Opinions for Stock Market Analysis using MCMC Bayesian Inference. *Computational Economics* 47, 4 (4 2016), 589–622. https://doi.org/10.1007/S10614-015-9492-9/TABLES/14

[25] Pooja Mehta, Sharnil Pandya, and Ketan Kotecha. 2021. Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Computer Science* 7 (4 2021), 1–21. https://doi.org/10.7717/PEERJ-CS.476/SUPP-2

[26] Saif M Mohammad and S M Mohammad. 2017. Challenges in Sentiment Analysis. *Socio-Affective Computing* 5 (2017), 61–83. https://doi.org/10.1007/978-3-319-55394-8{_}4

[27] Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia, and David C. Anastasiu. 2019. Stock price prediction using news sentiment analysis. *Proceedings - 5th IEEE International Conference on Big Data Service and Applications, BigDataService 2019, Workshop on Big Data in Water Resources, Environment, and Hydraulic Engineering and Workshop on Medical, Healthcare, Using Big Data Technologies* (4 2019), 205–208. https://doi.org/10.1109/BIGDATASERVICE.2019.00035

[28] Sohrab Mokhtari, Kang K. Yen, and Jin Liu. 2021. Effectiveness of Artificial Intelligence in Stock Market Prediction based on Machine Learning. *International Journal of Computer Applications* 183, 7 (6 2021), 1–8. https://doi.org/10.5120/ijca2021921347

[29] Frederico G. Monteiro and Diogo R. Ferreira. 2021. How Much Does Stock Prediction Improve with Sentiment Analysis? *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12591 LNAI (2021), 16–31. https://doi.org/10.1007/978-3-030-66981-2{_}2/TABLES/3

[30] Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction. *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference* 1 (2015), 1354–1364. https://doi.org/10.3115/V1/P15-1131

[31] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLoS Medicine* 18, 3 (3 2021). https://doi.org/10.1371/JOURNAL.PMED.1003583

[32] Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. 2017. Sentiment analysis of Twitter data for predicting stock market movements. *International Conference on Signal Processing, Communication, Power and Embedded System, SCOPES 2016 - Proceedings* (6 2017), 1345–1350. https://doi.org/10.1109/SCOPES.2016.7955659

[33] Roshni Pal, Utkarsha Pawar, Karishma Zambare, and Varsha Hole. 2020. Predicting Stock Market Movement Based on Twitter Data and News Articles Using Sentiment Analysis and Fuzzy Logic. *Lecture Notes on Data Engineering and Communications Technologies* 44 (2020), 561–571. https://doi.org/10.1007/978-3-030-37051-0{_}63/FIGURES/8

[34] Alexander Porshnev, Ilya Redkin, and Alexey Shevchenko. 2013. Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis. *Proceedings - IEEE 13th International Conference on Data Mining Workshops, ICDMW 2013* (2013), 440–444. https://doi.org/10.1109/ICDMW.2013.111

[35] Marwa Sharaf, Ezz El Din Hemdan, Ayman El-Sayed, and Nirmeen A. El-Bahnasawy. 2022. An efficient hybrid stock trend prediction system during COVID-19 pandemic based on stacked-LSTM and news sentiment analysis. *Multimedia Tools and Applications* (11 2022), 1–33. https://doi.org/10.1007/S11042-022-14216-W/TABLES/9

[36] Sashank Sridhar and Sowmya Sanagavarapu. 2021. Analysis of the Effect of News Sentiment on Stock Market Prices through Event Embedding. *Proceedings of the 16th Conference on Computer Science and Intelligence Systems, FedCSIS 2021* (9 2021), 147–150. https://doi.org/10.15439/2021F79

[37] Paul C. Tetlock. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance* 62, 3 (6 2007), 1139–1168. https://doi.org/10.1111/J.1540-6261.2007.01232.X

[38] Bruce James Vanstone, Adrian Gepp, and Geoff Harris. 2019. Do news and sentiment play a role in stock price prediction? *Applied Intelligence* 49, 11 (11 2019), 3815–3820. https://doi.org/10.1007/S10489-019-01458-9/TABLES/6

[39] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, and Arun Kumar. 2020. Stock Closing Price Prediction using Machine Learning Techniques. *Procedia Computer Science* 167 (1 2020), 599–606. https://doi.org/10.1016/J.PROCS.2020.03.326

[40] Zhaoxia Wang, Seng Beng Ho, and Zhiping Lin. 2019. Stock market prediction analysis by incorporating social and news opinion and sentiment. *IEEE International Conference on Data Mining Workshops, ICDMW* 2018-November (2 2019), 1375–1380. https://doi.org/10.1109/ICDMW.2018.00195

[41] Xing Wu, Haolei Chen, Jianjia Wang, Luigi Troiano, Vincenzo Loia, and Hamido Fujita. 2020. Adaptive stock trading strategies with deep reinforcement learning methods. *Information Sciences* 538 (10 2020), 142–158. https://doi.org/10.1016/J.INS.2020.05.066

A RESULTS AND DATASETS OF RESEARCHED PAPERS

Paper	Testing measure	results	dataset
[30]	Prediction Accuracy	56% average accuracy	Price data on five US stocks and message boards on same stocks from yahoo finance (July 2012- July 2013)
		Sentiment data improved accuracy by:XOM: 3.57%, EBAY: 3.58, IBM: 14.29%, KO: 12.5%.	
[24]	Fmeasure comparison of multiple data sets:	Fmeasure	historical stock data (analyzerxl.gr) on ALPHA, EUROB, OTE, AEGN. News articles (Neftemporiki, Capital), Twitter data (Twitter API all sources: jan 2013- jan 2014
	only technical data ,	67.6	
	technical + news data,	73.9	
	technical + news + twitter,	73.6	
	technical + twitter	67.6	
		AEGN and OTE more effected by sentiment	
[7]	Prediction accuracy	73%-78.9%	Financial news (vietstock.vn, hsx.vn, hsn.vn) ,daily stock prices (cophieu68.com) May 2014- April 2015
	Prediction accuracy of VN30 index	Only news sentiment data:80%	
	Prediction accuracy five individual stocks	60-90% accuracy	
[19]	prediction accuracy	KNN: 48%, SVM: 75.5%, NB: 72.6%	Financial news articles from press releases and regular financial news sources like wall street journal (2005-2014)
	Directional stock movement prediction per year(SVM)	70%	
[14]	Prediction accuracy	RF: 88%-92%	News data, stock price data for AAPL (yahoo, google news, reuters) Feb 2013-April 2016
		SVM: 86%	
		NB: 83%	
[32]	Prediction accuracy using sentiment data only	69.01%	Tweets on Microsoft stock (Twitter API) stock price data(yahoo finance) Aug 2015-Aug 2016
[17]	Prediction accuracy	sentiment: 59.18%	Price data on Microsoft, yahoo and facebook stock.News articles from multiple newspapers like Wall Street journal(No known time period)
		historical + sentiment: 89.80%	
[22]	paper did not use testing measures	SVM most efficient for prediction stock price movements with sentiment.	Twitter and stock price data (time period or stock used unknown)
[5]	Prediction accuracy	59.15% with sentiment data	news disclosures and price data. No time frame or sources mentioned
		54.5% without sentiment data	
[10]	Tested returns with buy and sell orders according to predictions	Greater return when making use of twitter and topic data than without.	News from reuters (2012-2017) on 15 US companies. Daily close price from yahoo finance
[40]	MSE(Mean square Error Graph of prediction price and actual data	Without sentiment: 3.72E-05	Yahoo finance data of DJIA (2007-2016), new york times articles from same period
		With sentiment: 3.57E-05	
[27]	MAPE(Mean absolute percentage error)	2.03 for price and sentiment	Closing stock prices and news articles of S&P 500 companies (feb 2013-mar 2017)
		2.13 for just price data	
[38]	Graph compares prediction with sentiment, historical and actual stock prices.	The model with Sentiment data performed better 77.78% of the time.	Twitter sentiment, newspaper sentiment and closing stock prices were all directly sources from Bloemberg(jan 2015 - juli 2018)
	Probability techniques to determine statistical significance of sentiment	Could not prove sentiment data influence at 5% level of statistical significance.	
[23]	RMSE,MAE (Root mean square deviation, mean absolute error)	Extensive tables of these measures for each coutry, ML algorithm and different companies.	Stock price data (yahoo) on 4 top companies from US, turkey, hong kong, and pakistan (jan 2000-oct 2018) Twitter data on major global and local events. (2012-2016)
[33]	Prediction accuracy	Twitter: 71%	Stock price data, twitter data, and news data(API's). time period of 45 days (specific stock and dates unknown)
		Twitter + news: 77%	
	Prediction accuracy after pre-processing.	Twitter: 73%	
		Twitter + news: 81%	
[16]	Prediction accuracy	68.56%	Stock price data (yahoo), Twitter(Twitter API). Juli 2016- June 2017
		Sentiment data 0-3% extra accuracy	
[8]	Prediction Accuracy	Only news data: 74.76%	news data, price data from same day and day after (Pulse), time period unknown
		Only stock price data: 88.73%	

Table 2. Results and datasets of researched papers

Paper	Testing measure	results	dataset
[25]	Prediction accuracy	LSTM: 92.45%	News data from multiple newspapers, twitter data, stock data(NSE),Oct 2014-Dec 2018
		NB:86.72%	
[29]	Prediction Accuracy	Sentiment data: max 0.6% increase	Market and news data from 2007-2016 on 3000 US listed companies
	With attention:	5.6%	
[9]	Directional price prediction accuracy	Price data:59.53%	News data from reuters, price data from yahoo finance.(feb 2013- feb 2018)
		Price+news sentiment:61.88%	
	Stock market profit	0.2%-1.7% profit	
		Avg. accuracy: 74.13%	
[36]	prediction accuracy	84.92%	Stock price data from DJIA and news headlines about DJIA(aug 2008-juli 2016)
[15]	Stock trend prediction accuracy with twitter model	historical: 75.16%	Twitter data, news articles from business insider, and stock data from yahoo(June 2016- juli 2018)
		added sentiment: 80.53%	
	Prediction accuracy with news model	historical: 69.79%	
		added sentiment: 75.16	
	Prediction accuracy both	Twitter + news: 79.86%	
[35]	Prediction accuracy with news sentiment and historical data.	TSLA: 90%	News data collected from Finviz, historical stock data from yahoo finance (dec 2019-aug 2020)
		AMZ: 91.6%	
		GOOG: 92.3%	
[1]	MAPE	Without sentiment: 36.75	Historical data (yahoo), news data (google news) April 2016- April 2021
		News sentiment: 47.99	

Table 3. continuation of: Results and datasets of researched papers