SN Computer Science (2024) 5:446
https://doi.org/10.1007/s42979-024-02651-5

ORIGINAL RESEARCH

Deep Learning for Stock Market Prediction Using Sentiment and Technical Analysis

Georgios-Markos Chatziloizos[1] · Dimitrios Gunopulos[2] · Konstantinos Konstantinou[2]

Received: 11 March 2022 / Accepted: 24 January 2024

Abstract

Machine learning and deep learning techniques are applied by researchers with a background in both economics and computer science, to predict stock prices and trends. These techniques are particularly attractive as an alternative to existing models and methodologies because of their ability to extract abstract features from data. Most existing research approaches are based on using either numerical/economical data or textual/sentimental data. In this article, we use cutting-edge deep learning/machine learning approaches on both numerical/economical data and textual/sentimental data in order not only to predict stock market prices and trends based on combined data but also to understand how a stock's Technical Analysis can be strengthened by using Sentiment Analysis. Using the four tickers AAPL, GOOG, NVDA and S&P 500 Information Technology, we collected historical financial data and historical textual data and we used each type of data individually and in unison, to display in which case the results were more accurate and more profitable. We describe in detail how we analyzed each type of data, and how we used it to come up with our results.

Keywords Technical analysis · Sentiment analysis · Machine learning · Stock market · Deep learning

Introduction

Forecasting the prices and the trends of the stock market is one of the most challenging and competitive domains for scientists and financial experts. Many people have lost their savings trying to time the market and make a fortune off of it. The most dominant advice given by financial advisors and the most traditional investors is to just invest a part of your income into the S&P 500 and just wait, in any other case you will probably lose your money. Also, according to the Efficient Market Hypothesis (EMH), every stock price is trading at a fair value, and an investor cannot continuously beat the market. On the other hand, many investors, hedge funds, and scientists, often deny EMH since they can predict stock market trends using fundamental, technical, and sentiment analysis.

Predicting the price of the stock market is a multivariate equation that many researchers and financial experts have tried to find its components. At the beginning of 2021, we observed a massive movement by private investors against hedge funds and their private dinner parties. What they did was to buy the stock shares of companies that hedge funds shorted and thus leading them to a short squeeze and resulting in heavy damages for very famous and large funds. This massive movement was organized mainly through Reddit and Twitter. As we can imagine with the help of real-time crawlers at Twitter and Reddit, data can be obtained and then fed into a fast and highly effective real-time NLP system. Therefore, this could be potentially beneficial and profitable in the stock market prediction.

So, to understand the stock market, researchers need to retrieve two different types of information, hard and soft. The closing price of a stock, the revenue and the number of sales made by a company are considered hard data as they can be represented in numeric values. On the other hand, news, articles and tweets are considered soft data, as these are more abstract information that can also be represented

This article is part of the topical collection "Data Science, Technology and Applications" guest edited by Slimane Hammoudi and Christoph Quix.

✉ Georgios-Markos Chatziloizos
g.chatziloizos@gmail.com

[1] PSL Research University, Paris, France

[2] Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece

Published online: 18 April 2024

numerically, but there is a loss of information. Because of their intrinsic differences, these two types of data are relevant to different tasks as they can produce better outcomes on different issues than each other. In [1], it is detailed in depth what each type of information is and how they differ in the banking and financial world.

While many people have studied the field of forecasting the prices and the trend of the stock market, most have tried to apply only one of the two types of data, hard [2–7] or soft [11, 12]. When they choose the latter, they mostly raise the question of whether a correlation between the sentiment of textual data and the stock market exists without applying it in an actual decision-making tool. The amount of noise in this kind of data and the difficulty to implement it into a viable prediction model for the stock market, have deterred researchers from fully trying to realize the potential of textual analysis in making such a tool. While this stands true, in recent years we have seen many occasions where the collective of internet users change, with their synchronous movement, the course of a stock price. As more and more people join these collectives and online communities in their effort to gain an advantage in their decision-making in regard to finance and the stock market, this analysis of sentiment can become more powerful and accurate. Ideas, personal opinions and inside information, all make their appearance in internet posts thus possibly providing an insight into how the stock trend is going to fluctuate. All of these kinds of posts portray the feelings of people towards companies and their stock tickers, and using sentiment analysis, it is possible to assess how positive or negative these feelings are. What this means is that the correlation of internet corpora of text and the stock trends becomes more apparent, and these decision-making tools based on sentiment analysis can actually become more efficient and trustworthy. Sentiment analysis analyzes people's feelings and moods toward an entity, such as a stock ticker, using textual data to determine how negative or positive their thoughts are [9, 10].

As technology evolved and machine and deep learning algorithms became more sophisticated and successful hard data have been used to predict stock market prices or trends [2–7, 18–20]. The reader is also referred to [8] for a recent survey by Ferreira et al. Also, there are several papers in which authors implemented textual data for predicting stock market movements. The authors of [11] explore the link between stock market movements and twitter sentiments using sentiment analysis and supervised machine learning approaches. The authors of [12] used sentiment data and stock price market data to create an SVM model for predicting stock movements the next day. The approach proposed in [13] initially labels a stock market-related tweet dataset, then compares various deep learning models, and ultimately introduces an LSTM model that outperforms all other models.

Taking all of this into account, in this paper we propose our methods to add to the direction of analyzing and implementing machine/deep learning techniques to correctly anticipate stock prices and trends using both numerical/economical data and textual/sentimental data. We use deep learning/machine learning approaches on both types of data with the purpose of not only predicting stock market trends but also understanding how a stock's Technical Analysis may be strengthened by using Sentiment Analysis. This article builds upon our previous work presented in [14]. In [14] we have employed three deep/machine learning methods [15] i.e., Long Short-Term Memory (LSTM), k-nearest neighbors (KNN) and Decision Trees. In this paper, we used three more deep/machine learning approaches that are, Convolutional Neural Networks (CNN), Support Vector Classification(SVC) and Multilayer Perceptron (MLP) and we applied them to the following three different sets of historical data (a) numerical/economical data such as stock closing prices, technical analysis indicators, labels, etc. (b) sentimental data e.g. scores computed using lexical methodologies on textual data collected from Twitter and labels (c) combined data that include all the above data in sets (a) and (b).

In our tests, we used data from four stock tickers: AAPL, GOOG, NVDA, and S&P 500 Information Technology. The data consists of numerical/economic data collected over a 20-year period and textual data (about 29,000 tweets for each of the above tickers) collected over an 8-year period. Out of all six algorithms we compared the two best ones i.e. LSTM and CNN, also on extended new datasets. Specifically, the “financial_phrasebank” dataset [21] was additionally used which includes 5000 labeled sentences from financial news articles about Finnish Banks.

The results demonstrate that the extended datasets improve our profits in most cases and the most profits for the extended datasets came from the CNN on Numerical data. However, the most profits among all datasets and methods came from the LSTM method on Numerical data from the original dataset, which was presented in [14]. Sentiment analysis also proved to have future promise, as it was profitable and in most cases a better option than a passive investment. Sentiment analysis appears to produce better findings when additional high-quality data is included, such as news titles and articles, and the number of tweets collected is increased.

This paper is structured as follows: "Sentiment Analysis" describes the soft information we used, as well as how we analyzed and applied it. The technical analysis indicators used are shown in "Technical Analysis". "Application" describes the data used in the deep/machine learning methods, as well as the application's remaining parameters. "Results" summarizes the findings, and "Conclusions" presents the conclusions.

Sentiment Analysis

In the last few years, the stock market has been greatly influenced by the power of words and mass transactions that are stimulated and coordinated by social media users. Naturally, many researchers have begun trying to understand the sentiment behind such users and their posts, so that the patterns can be identified, to make it possible to predict a stock trend. It is evident that not only well-thought and well-written news articles by huge news networks are influencing this market, but a simple internet post with rashly put together text can have the same impact especially if the latter comes in huge quantities. To emphasize this, the now historic stock market incident, which can simply be referred to as the "GameStop short squeeze", was a landmark display of how internet users can collectively change the course of a stock.

As mentioned above, there have been several studies on the sentiment analysis of internet posts in order to make future predictions on stock trends or price. In our work these internet posts are referring to as Twitter posts, or as they will be named from this point on, "tweets". Twitter has been a powerful ally to researchers and a trustworthy prediction tool. Ussama et al. [16] researched the power of predictability that Twitter possesses in a very important and serious matter which is the US 2016 elections. Pagolu et al. [11] and Rao and Srivastava [17] tried to prove that there is a correlation between the stock market and the sentiment in tweets and to further analyze this relationship. Researchers have proven again and again how Twitter, and social media platforms like it, Reddit, 4chan and so on, can play a huge role in shaping or predicting trends.

Given all of the above, it is safe to assume that there can be a significant percentage of accurate predictions that use textual data to track a stock's trends. While in our work we used both numerical and textual data, in the rest of this Section we will explain how we used the latter.

Data Collection

In our previous work [14], we collected approximately 29.000 English tweets for each of the three different tech giants, Google, Nvidia and Apple, using their respective stock ticker GOOG, NVDA and AAPL. We used their tweets to make predictions for each respective corporation and for the S&P Information Technology Sector as well. The tweets ranged from the 1st of January 2012 to the 31st of December, 2019. For each ticker, we collected 10 tweets for each day and therefore, we have gathered about 29,000 tweets. These same tweets from our previous work were used again so that there is a valid benchmark to make the necessary comparisons with the new findings of our work.

To expand our research, we looked for a labeled dataset that could provide the right amount of data well focused on the topic of stock market and finance. The "financial_phrasebank" dataset which was developed and used in the work of Malo et al. [21] met the needs of our research. This dataset included 5000 labeled sentences from financial news articles about Finnish Banks. The sentiment behind each sentence was identified by people with sufficient knowledge of the financial world. These sentences were appropriate for our sentiment analysis research, as they included news on corporate finances as well as news unrelated to corporate internal affairs, focusing on external sentiments and assessments. Using this data we were able to expand and strengthen our research, especially when it was used in combination with the datasets mentioned above.

With these two types of data, the unlabeled and the labeled, we were able to move on. In the following, we discuss how all this data was used and in what ways.

Stemming and Cleaning

To use each of the datasets, both the unlabeled and the labeled one, we implemented a word removal process followed by a stemming process. Unnecessary and redundant words and characters, such as stopwords and punctuation, were eliminated throughout the removal process. Links and user mentions were also removed because they were no longer useful for our current purpose. Only the words that could be valuable to us remained in place. After that, the stemming phase can start.

Stemming refers to the process of disassembling a word, so that it can be returned to its original form. For example, word derivatives such as the word "weakening" and "weakness" can be represented after a stemming process as "weak". This is done so that the number of different words can be limited to boost computational times, which can be crucial at a later stage of the research. It is also used to be able to match certain words, with a word in a lexicon or a vector representation model, so that we can have a better overall analysis of the sentiment of the data. The sentiment of each word and combination of words is mostly unaffected by this process.

Therefore, after cleaning each tweet, we tokenized the words and stemmed them, using the Snowball Stemmer. With this stemmer, we stripped each word from suffixes and we kept it at its basic form. The order of the words was not changed, so the word combinations that could substantially modify the meaning and sentiment of a sentence, were not changed.

Evaluation

After the sentences have been cleaned and filtered, the evaluation process can begin. The evaluation process involves labeling (categorizing) the tweets of the unlabeled dataset as either positive, negative, or neutral based on their sentiment. To categorize the data, we implemented two methods: the Lexicon and the Machine Learning methods. The Lexicon method involves using three pre-defined sentiment lexicons to assign a sentiment score to each tweet. For this method, we get a score for each lexicon that we later use for the stock prediction. However, the Machine Learning method involves training machine learning models on the labeled dataset to predict the sentiment of new tweets of the unlabeled dataset. For this method, we use the accuracy as a metric for the evaluation.

Lexical Methodology

The Lexicon method or the lexical methodology uses a lexicon to identify and evaluate with a numerical score each word of a sentence. These scores can then be aggregated to demonstrate the overall score of the sentence, and hence the sentiment behind a tweet. How or how well the score reflects the sentiment depends to a large extent on the quality of the lexicon. We used this technique for three different lexicons: the VADER, the Loughran-McDonald and a generic lexicon.

- The VADER Lexicon [22] is a lexicon that is mainly used for social media analysis, as it contains words and their respective sentiment score, focused around social media posts.
- The Loughran-McDonald Lexicon [23] although limited in the number of words it contains, the reason for its development, that is the understanding of words around finance, as well as its ability to use words in unison, make it a very powerful lexicon for our research.
- A Generic Lexicon is a simple lexicon that does not use a certain viewpoint (finance, social media, etc.) to determine the sentiment score of a word. The number of words in such a lexicon could fill in the gaps when the other two could not provide a score.

Table 1 The accuracy of the models

Method	Bag-of-words accuracy	TF-IDF accuracy
Naive Bayes	0.50412	0.50515
DT	0.71030	0.67525
KNN	0.62061	0.64020
SVC	0.70309	0.71237
MLP	0.63195	0.65360

Using these three lexicons, we created a score for each tweet. Then we calculated the average score of the 10 tweets of each day for each respective lexicon. This means that for each day we had 3 average scores, produced by the three lexicons.

Machine Learning Method

For the Machine Learning Method, the labeled dataset is used to train models so that they can learn patterns between the features (that we introduce in the following paragraph) in the tweets and their corresponding sentiment labels (positive, negative, or neutral). Once the models have been trained on the labeled dataset, they can be used to predict the sentiment of new tweets from the unlabeled dataset. This is accomplished by feeding the new-unlabeled tweets into the model and letting it make predictions based on what it has learned from the labeled data.

Creating vector representation models We used vectors to represent each sentence from the labeled dataset and each tweet from the unlabeled datasets. These vectors consist of different numbers based on the words of the tweets and sentences, which are used so that patterns can be identified to make predictions. Through the stemming process described above, we prepared the "financial_phrasebank" dataset to be used in vector representation models to train our models. We used two of the most basic such vector representation techniques: the Bag of Words (BOW), and the Term Frequency–Inverse Document Frequency (TF-IDF).

- The BOW model is a very common vector representation model used in Machine Learning. When we enter a dataset, it keeps every unique word encountered in it, and the number of occurrences of the word in the entire dataset. Because of its simplicity, it is used as a benchmark vector representation model.
- The TF-IDF model is also a common vector representation model used in Machine Learning. This model uses the frequency of a word in a given dataset and the frequency of the same word in a given sentence to produce a weight. This weight basically describes the rarity and therefore the significance of the word, making sentences with such rare words leaning more to those words' sentiment.

We gave as input to these models the clean and filtered "financial_phrasebank" dataset. When they were finalized, we split them into training and testing vectors, 80% of the dataset as training and 20% as testing, to be used later by our Machine Learning Algorithms.

Training the models We used the BOW and TF-IDF models to train five different Machine Learning Algorithms so

that we can later use them in labeling the tweets. These algorithms were:

- *Naive Bayes* The Naive Bayes classifiers are a collection of supervised learning algorithms based on the Bayes' theorem. The classifier we used, is a very commonly used classifier applied on many Machine Learning projects and researches, as well as on real-world problems, because of its efficiency relative to its simplicity.
- *Decision Trees* Decision Trees (DT) is a non-parametric supervised learning method. This model creates nodes and splits into new ones depending on the features and their different variations derived from the training data given as input. These nodes create a structure resembling a tree, hence the name.
- *K-Nearest Neighbors* K-Nearest Neighbors (KNN) is a non-parametric supervised learning method. The KNN model uses "k" number of points from the training dataset to make an assumption about a new data point taken from the testing dataset. These k points are chosen based on how close they are locally to the data point in question, in an n-dimensional space made from the n features, so that we can identify to which "neighborhood" of data this new instance belongs, or to put it differently, to which class.
- *SVC* Support Vector Classification or SVC is a supervised learning method that splits the data depending on their features into two classes, therefore solving a binary problem. While it is used for binary classification, if this process is used multiple times to solve sub-classification problems within the dataset, it can produce a multi-classification result.
- *MLP* Multilayer Perceptron or MLP is a deep learning method. MLP is an Artificial Neural Network (ANN), which means that it creates different layers within it, using the training input.

For each algorithm, two different inputs were given, the BOW and the TF-IDF, with a total of ten different models.

Testing the models When the training process was finished, the testing input was fed to the ten models. Here we can get the idea of how well the models could produce results, and act accordingly (Table 1). The accuracy for each model was:

As we can see from the above table, DT and SVC are the most accurate ones with KNN and MLP following closely. Naive Bayes seems to be subpar to the rest with just over 50%.

Applying the models on the Tweets Once the models were trained and ready, we used the dataset of tweets we collected, to make decisions on each tweet using each method. As we did with the lexical methodology, we produced an average sentiment for each day using its 10 tweets. But in this case, instead of creating an average number score, which was created in the lexical methodology using the average of the weights of each word, we ended up with just a signal for the sentiment that is, − 1, 0 or 1 for negative, neutral, and positive, respectively. These signals were then exported into a dataset, in which for every day we had 10 different columns, each corresponding to a model.

The Consumer Sentiment Index

To improve the prediction power of our final model, we used the Consumer Sentiment Index, a monthly index published by the University of Michigan. This index is a powerful tool and has been shown to accurately represent the sentiment of the public. The fears or ambitions of the public greatly influence the trend of the whole market system and therefore of the stock market. The index tries to capture these emotions and provide a score based on them. Our research, which focuses on a trend forecast rather than an accurate stock price forecast, could theoretically provide better outcomes if it was backed up by an index that tries to predict consumer sentiment. For this reason, we incorporated the Consumer Sentiment Index into our research using it along with the other sentiment data discussed above. Because the Consumer Sentiment Index is published monthly, we used the same value for each day of the month.

Final Products

The scores from the lexicons, and the signals from the Machine Learning models were put into their respective datasets-csvs and were then used by the final model. The Consumer Sentiment Index was put alongside each of these datasets to boost their results.

Technical Analysis

The technical analysis is applied to the raw numerical data of the stocks (opening, closing, high and low price of the stock ticker per day). Technical analysis techniques employ a number of technical indicators to forecast the stock trend/price. Common traders use at least two to three indicators in order to predict the trend/price of the market but the results usually are not good enough. On the other hand, trying to use too many indicators will also end up with inefficient results.

There are a number of indicators used in the stock market. The simplest technical indicators are the Simple Moving Average (SMA) and the Exponential Moving Average (EMA). The SMA is the unweighted mean of the previous n stock's prices. The number of days considered determines the value of n (e.g. 10 days for the short term, 80 days for the long term). The SMA considers all stock's prices equally

and is influenced disproportionally by old prices. This is addressed by the EMA which gives more weight to more recent prices while not completely ignoring older observations. Therefore, EMA has a stronger impact on recent price fluctuations. We have a rising tendency when a short-term moving average crosses over a longer-term moving average. We have a sliding tendency when a short-term moving average passes under a longer-term moving average. The information offered by moving averages has a time lag of many days, which is a disadvantage. To overcome this deficiency, more powerful indicators are used such as the following (https://www.investopedia.com):

(a) The MACD (Moving Average Convergence Divergence) is a trend-following momentum indicator that uses the relationship between two exponential moving averages. By subtracting the 26-day EMA from the 12-day EMA, the MACD line is calculated. The "signal line" which is a 9-day EMA of the MACD, can be used to trigger buy and sell signals. There is a purchase signal if MACD is above the signal line; otherwise, there is a sell signal.
(b) The RSI (Relative Strength Index) is a momentum indicator that evaluates the value conditions in a stock's price by measuring the magnitude of recent price fluctuations. The RSI is represented as an oscillator with a range of [0,100]. When the RSI is above 70, it shows that the stock is overvalued, indicating that we should sell, and when it is below 30, it suggests that the stock is oversold, indicating that we should purchase. If the RSI is between 30 and 70, we hold.
(c) The Stochastic oscillator is a momentum indicator that attempts to predict price turning points by comparing the closing price of a stock to its price range over a period of time. It is used to generate overbought and oversold indications, and it spans the [0,100] range. Typically, if the stochastic oscillator is over 80, it is overbought; if it is under 20, it is oversold; and when it is in the region of 20 to 80, it does not provide any additional information.
(d) The Bollinger Band (BB) is a technical analysis indicator defined by a set of lines plotted two standard deviations away from a simple moving average (upper and lower bands). The belief is that the closer prices get to the upper band, the more overbought the market becomes, and the closer prices get to the lower band, the more oversold the market becomes. The lower and higher bands are where the majority of the price action occurs. It's a rare occurrence when a breakthrough happens above or below these bands.

Fiol-Roig et al. [24] successfully use the indicators MACD, EMA(C) (Exponential Moving Average of the Closing Price), EMA(V) (Exponential Moving Average of the Volume), Stochastic oscillator and BB to generate a decision tree that classifies buying-selling orders.

In our previous work [14] we employed the indicators MACD, RSI, Stochastic oscillator and BB to apply on the raw data. In this paper, to improve our results for the LSTM and CNN models we use three more technical indicators. These indicators are:

- The Money Flow Index (MFI) is used to generate overbought and oversold signals, and it ranges in the interval [0,100]. If MFI is over 80, it is considered overbought, while if it is under 20 is oversold, while it does not provide more information when it is in the range of 20 to 80. To calculate the MFI we need the High, Low and Closing price and also the volume, the number of shares of stock traded that day, which is not information we used in our previous approach. The RSI and MFI are quite similar indicators but MFI's advantage is that it uses the volume of the stock.
- The Average True Range (ATR) is a technical indicator developed by Welles Wilder Jr. [25]. It was created for commodities such as gold, oil, beef, etc., but it is used in stocks and indices too. The ATR indicator is used mainly by traders to open and close positions, and it also helps to calculate the daily volatility of a stock with more precision. For this indicator, we did not implement any buy or sell signal as it is widely open for interpretation, but we just let the Machine and Deep learning algorithms reach their own conclusions.
- The Williams $\%R$ is a momentum indicator developed by Larry Williams [26] and it ranges in the interval $[-100,0]$. A stock is overvalued when the Williams $\%R$ is above -20, so it tells the trader to sell, while it is oversold when it is below -80, thus it indicates a buy signal. If Williams $\%R$ is between -20 and -80, then it does not provide any information.

Application

Data and Features

Firstly, we merge all the data that we discussed in "Sentiment Analysis" and "Technical Analysis" i.e.,

- the daily historical data of the stock ticker for (a) AAPL, GOOG, NVDA and Nasdaq Composite Index (because of the correlation between the stocks and the index) from yahoo finance and (b) SPIS (S&P 500 Information Technology Sector) from investing.com and

- the sentimental data and the Consumer Sentiment Index.

Afterwards, the technical analysis indicators are calculated from the historical data and we add them to our dataset.

The required features are:

(a) The Closing Price of the stock ticker
(b) The Closing Price of NASDAQ Composite Index
(c) The Volume of NASDAQ Composite Index
(d) MACD
(e) RSI
(f) Stochastic Oscillator
(g) Bollinger Bands
(h) Consumer Sentiment Index
(i) Score of the generic lexicon
(j) Score of VADER lexicon
(k) Score of Loughran-McDonald lexicon
(l) Labels

For our new Extended Datasets, where we test our two best models, LSTM and CNN, we decided to add to the existing dataset 3 more technical indicators and 10 more sentiment analysis ones, explained in "Sentiment Analysis" and "Technical Analysis" respectively. Namely, these are:

- Money Flow Index (MFI)
- Average True Range (ATR)
- Williams %R
- Naive Bayes Bag-Of-Words
- Naive Bayes TF-IDF
- Decision Trees Bag-Of-Words
- Decision Trees TF-IDF
- K-Nearest Neighbors Bag-Of-Words
- K-Nearest Neighbors TF-IDF
- SVC Bag-Of-Words
- SVC TF-IDF
- MLP Bag-Of-Words
- MLP TF-IDF

Creating the Labels and Scaling

The labels are used to indicate whether the stock market trend is positive or negative or it does not change significantly based on the price movements of a particular stock. To predict the stock ticker's trend 5 days later, we shift the closing price for 5 days and then compare the closing price to the closing price 5 days ahead to determine if the trend is bullish, bearish, or does not change significantly, so we merely hold. This is how our labels are created. If there is a bullish (positive) trend then we append number 2 for the specific day, number 1 for hold and number 0 for bearish (negative) trend.

The entire dataset is then modified, with the exception of the labels, using the MinMaxScaler offered by sklearn [27], so that each value in the dataset falls within the range [0,1].

The Datasets

From the original dataset, three datasets are formed at this step. The first comprises all of the features stated above and is called combined dataset (a − l). The second is made up entirely of numerical/economic data, such as closing prices, volume, technical analysis indications, and labels (a–g + l). The last one is the sentimental dataset, which includes the Consumer Sentiment Index, the scores from the three lexicons, and the labels (h–l).

For the Extended datasets, the combined Extended dataset consists of the features (a–l) and (1–13), the numerical Extended dataset includes the features (a–g + l) and (1–3), and finally the sentimental Extended dataset consists of the features (h–l) and (4–13).

Training/Testing Datasets

The datasets are split into training and testing datasets to train and test our models. The training dataset consists of the days between 01.01.2000 until 31.12.2017 and the

Table 2 AAPL Profits for LSTM, DT and KNN (in US $) [14]

	Combined data	Numerical data	Sentimental data
LSTM	139.06	222.82	140.01
DT	-92.40	-74.90	-13.61
KNN	-66.23	-131.00	-25.61
LSTM with strategy	144.21	167.97	128.98
DT with strategy	-115.94	-113.79	-24.45
KNN with strategy	-45.85	-128.15	-33.56

Table 3 AAPL Accuracy of LSTM, DT and KNN [14]

	Combined data	Numerical data	Sentimental data
LSTM	0.48	0.58	0.59
DT	0.36	0.40	0.49
KNN	0.44	0.38	0.46

testing dataset the days of the following 2 years, namely, 01.01.2018 until 31.12.2019.

Sequential Data

It is required to build sequential data from our current datasets to use the LSTM and CNN models properly. This is a critical stage because, to achieve better results, we must include data from the previous week and not only from the previous day. As a result, each data point in our scenario is formed by concatenating the data of five days. If our dataset is made up of n days, our sequential data are as follows:

$$\{[x_1, x_2, x_3, x_4, x_5], [x_2, x_3, x_4, x_5, x_6], \ldots, [x_{n-4}, x_{n-3}, x_{n-2}, x_{n-1}, x_n]\},$$

where x includes all the features (except the labels) of each day of the original dataset. Each new data point takes the label of the element corresponding to the last day i.e., the label of $[x_1, x_2, x_3, x_4, x_5]$ is the label of x_5, the label of $[x_2, x_3, x_4, x_5, x_6]$ is the label of x_6, the label of $[x_3, x_4, x_5, x_6, x_7]$ is the label of x_7 and so on.

Table 4 AAPL profits for CNN, SVC and MLP (in US $)

	Combined data	Numerical data	Sentimental data
CNN	136.46	151.32	124.53
SVC	54.91	130.31	− 13.61
MLP	− 57.63	81.34	138.49
CNN with strategy	131.66	138.79	128.84
SVC with strategy	77.99	126.95	133.69
MLP with strategy	− 31.50	78.02	108.42

Table 5 AAPL accuracy of CNN, SVC and MLP

	Combined data	Numerical data	Sentimental data
CNN	0.59	0.59	0.58
SVC	0.53	0.58	0.58
MLP	0.48	0.54	0.55

Machine and Deep Learning Models

Last step is to feed our data into the machine and deep learning models. We used the following machine learning models that we used in our experiments to predict stock market trends. The models that we used include k-nearest neighbours (KNN), decision trees, support vector machines (SVM), and multi-layer perceptron (MLP).

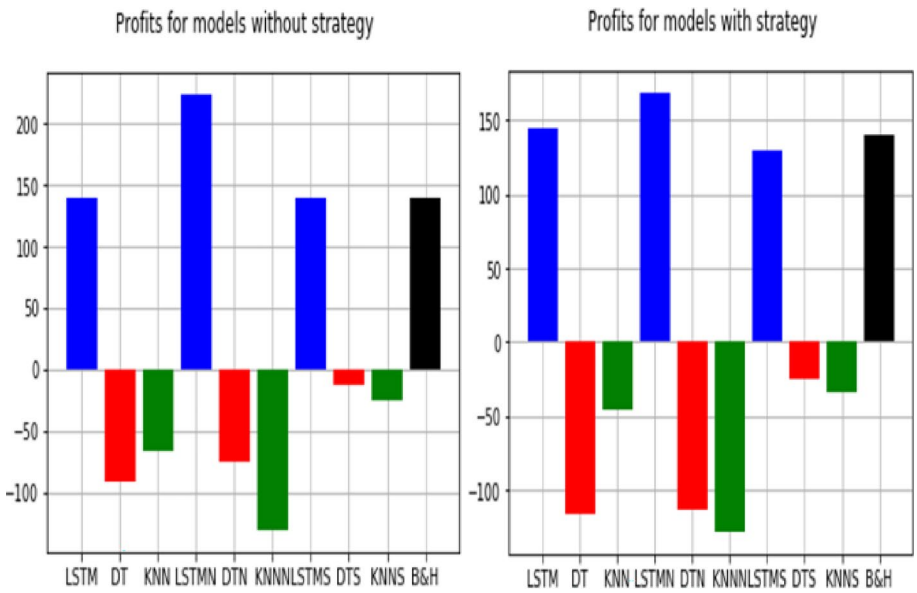

Fig. 1 AAPL Profits for LSTM, DT and KNN without strategy (left) and with strategy (right) [14]

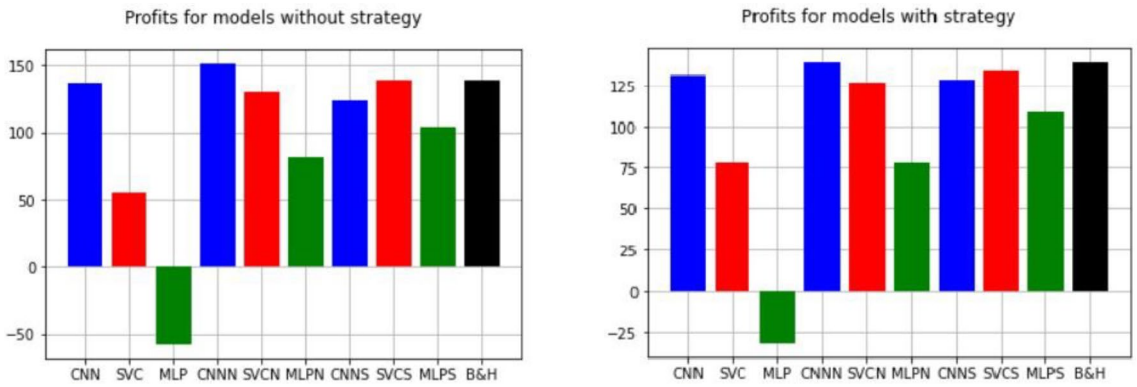

Fig. 2 AAPL profits for CNN, SVC and MLP without strategy (left) and with strategy (right)

Table 6 AAPL profits with the extended dataset for LSTM and CNN (in US $)

	Combined extended data	Numerical extended data	Sentimental extended data
LSTM	186.84	161.17	197.81
CNN	182.74	138.31	141.14
LSTM with strategy	184.58	156.69	190.63
CNN with strategy	172.61	129.49	131.90

Table 7 AAPL accuracy with the extended dataset of LSTM and CNN

	Combined extended data	Numerical extended data	Sentimental extended data
LSTM	0.61	0.58	0.57
CNN	0.58	0.59	0.59

As for the deep learning models we used an LSTM and a CNN architecture. Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that is commonly used for sequential data such as time series data. While, CNNs are a type of deep neural network that are commonly used for image recognition tasks, but they can also be used for time series data.

In the KNN method, we use three nearest neighbors as the k parameter and for the Decision Trees method, we set the max depth equal to 5. For the SVC method, we set the parameter C equal to 1 and for the MLP we use as activation function the ReLu and hidden_layer_sizes is 100.

The LSTM model consists of 3 stacked layers with the activation function the ReLu. In addition, the Dropout function is used to avoid the phenomenon of over fitting. The CNN model 3 layers of Conv-1D were used with kernel size 7, 5 and 3 respectively and with activation function the ReLu. Total parameters of the model were about 300,000.

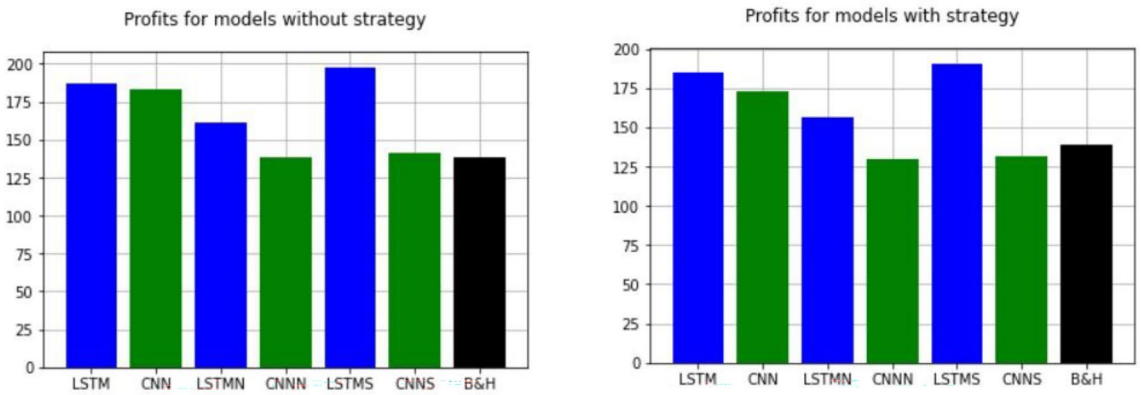

Fig. 3 AAPL profits with the extended dataset for LSTM and CNN without strategy (left) and with strategy (right)

Table 8 GOOG Profits for LSTM, DT and KNN (in US $) [14]

	Combined data	Numerical data	Sentimental data
LSTM	1287.43	2498.33	1813.72
DT	2446.37	2334.23	990.72
KNN	2497.72	1914.16	1090.37
LSTM with strategy	1847.12	1989.89	1310.00
DT with strategy	1478.73	1410.16	1681.82
KNN with strategy	1892.21	1501.27	1426.08

Table 9 GOOG accuracy of LSTM, DT and KNN [14]

	Combined data	Numerical data	Sentimental data
LSTM	0.50	0.53	0.45
DT	0.50	0.50	0.54
KNN	0.48	0.46	0.49

For the LSTM and CNN model, each of the three datasets are trained for 30 epochs with batch size 64, learning rate 10^{-4} and the Adam optimizer was used.

Strategy

When the 5-day holding period expires, the positions are closed (i.e., buy and sell decisions are made). We apply a simple approach in conjunction with the LSTM, Decision Trees, and KNN algorithms to deal with the stock market's high volatility in the best possible way. The following is the strategy: When the 5-day holding period has expired, or the percentage of stop loss or take profit has been exceeded, the positions are closed. These percentages are usually – 5% and 7%, respectively, however, they might vary depending on the asset. It might potentially be used as a trailing take-profit tool, although backtesting is not possible owing to the nature of the data. However, it is a very good tool for real-time use and is highly recommended.

Table 10 GOOG profits for CNN, SVC and MLP (in US $)

	Combined data	Numerical data	Sentimental data
CNN	1920.92	3685.22	2574.27
SVC	1876.20	1847.98	1319.16
MLP	1182.08	2038.14	1319.26
CNN with strategy	1959.91	3294.99	2654.54
SVC with strategy	1735.30	1707.08	1642.55
MLP with strategy	1306.37	1598.03	1642.73

Table 11 GOOG accuracy of CNN, SVC and MLP

	Combined data	Numerical data	Sentimental data
CNN	0.55	0.57	0.54
SVC	0.58	0.58	0.57
MLP	0.55	0.57	0.57

Buy and Hold Strategy

Our findings are compared to the Buy and Hold (B&H) strategy, which is a popular stock market approach. Investors purchase assets (stocks, ETFs, Indices, and so on) and

Fig. 4 GOOG profits for LSTM, DT and KNN without strategy (left) and with strategy (right) [14]

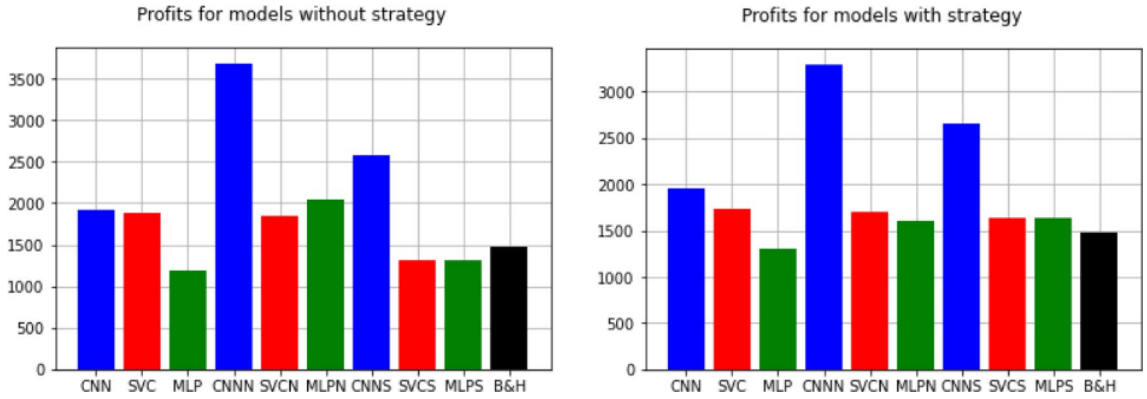

Fig. 5 GOOG profits for CNN, SVC and MLP without strategy (left) and with strategy (right)

Table 12 GOOG profits with the extended dataset for LSTM and CNN (in US $)

	Combined extended data	Numerical extended data	Sentimental extended data
LSTM	3289.26	2156.42	2167.28
CNN	2493.02	4055.27	3070.58
LSTM with strategy	3146.54	1947.51	1662.87
CNN with strategy	2144.49	3959.72	2609.32

Table 13 GOOG accuracy with the extended dataset of LSTM and CNN

	Combined extended data	Numerical extended data	Sentimental extended data
LSTM	0.54	0.56	0.55
CNN	0.56	0.57	0.57

hold them for the long term. Keep in mind that this strategy does not rely on technical analysis tools and is hence fairly straightforward. Hedge fund managers and many investors aim to "beat the market" as much as possible. When investors say they have "beat the market," they are referring to the fact that they have outperformed the B&H approach in terms of cumulative returns.

Results

In "Statistics", we discuss the profits and the accuracy of the methods when applied on the stock's tickers of AAPL, GOOG, NVDA and SPIS, while in "Comparing passive investor's and LSTM method's returns with the original dataset" we compare the returns of the B&H strategy to the ones of the most profitable method, the LSTM method applied on the numerical data with the original dataset. In "Improvements in profit with the extended datasets", we

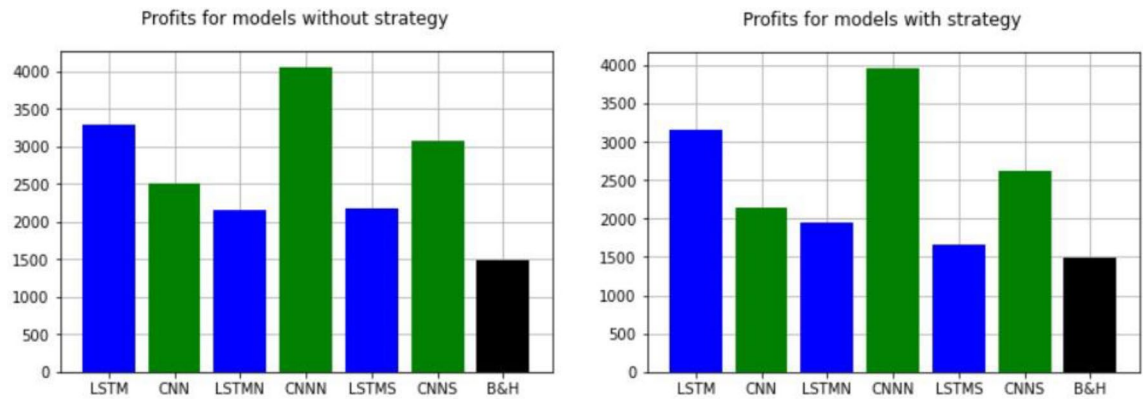

Fig. 6 GOOG profits with the extended dataset for LSTM and CNN without strategy (left) and with strategy (right)

Table 14 NVDA profits for LSTM, DT and KNN (in US $) [14]

	Combined data	Numerical data	Sentimental data
LSTM	323.49	1149.00	236.44
DT	376.05	-253.68	-150.64
KNN	-194.61	148.53	-177.15
LSTM with strategy	99.63	719.11	269.74
DT with strategy	580.89	150.54	200.54
KNN with strategy	-138.65	-8.61	237.33

Table 15 NVDA Accuracy of LSTM, DT and KNN [14]

	Combined data	Numerical data	Sentimental data
LSTM	0.40	0.56	0.54
DT	0.52	0.51	0.52
KNN	0.49	0.52	0.51

show the improvements in profit with the Extended Datasets when applied to our algorithms instead of the original datasets. Finally, in "LSTM vs CNN with the extended datasets", we compare the profits of LSTM and CNN with the Extended Datasets.

Statistics

AAPL The profit of the Buy and Hold strategy for AAPL was 139.89$.

Table 2 [14] presents the AAPL profits of LSTM, DT and KNN with and without our strategy for each one of the three datasets in US dollars. Best case scenario for the AAPL ticker was the LSTM on numerical data.

Table 3 [14] presents the accuracy of each method for the three datasets. Although LSTM on numerical data provided us with more profit, the accuracy of LSTM on sentimental data was slightly better (59%).

Figure 1 [14] provides us with the information about AAPL profits for LSTM, Decision Trees and KNN for the 3 datasets with or without our strategy. We use the suffix "N" ("S") for each method to denote that the method is applied to numerical data (resp., sentimental data). When no suffix is used, the method is applied to the combined dataset (numerical & sentimental data). B&H is the abbreviation of the Buy-and-Hold strategy.

Table 4 presents the AAPL profits of CNN, SVC and MLP with and without our strategy for each one of the three datasets in US dollars. Best method for the AAPL ticker was the CNN on numerical data with a profit 151.32$.

Table 5 presents the accuracy of each method for the three datasets. LSTM on numerical data provided us with the most profit and its accuracy was 59%.

Figure 2 is a graphical representation of the APPL's profits for CNN, SVC and MLP for the 3 datasets with or without our strategy.

Table 6 presents the AAPL profits of LSTM and CNN for the extended dataset with and without our strategy for each one of the three extended datasets in US dollars. Best

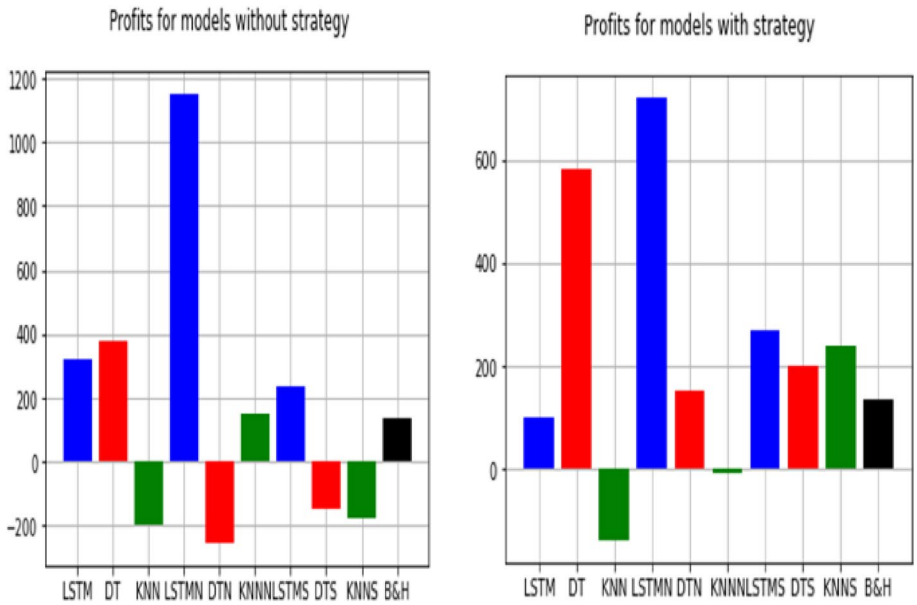

Fig. 7 NVDA Profits for LSTM, DT and KNN without strategy (left) and with strategy (right) [14]

Table 16 NVDA profits for CNN, SVC and MLP (in US $)

	Combined data	Numerical data	Sentimental data
CNN	373.74	716.95	64.03
SVC	357.78	275.24	140.45
MLP	353.13	228.83	9.02
CNN with strategy	537.92	551.98	360.02
SVC with strategy	585.77	559.40	373.84
MLP with strategy	420.61	507.10	316.94

Table 17 NVDA profits for CNN, SVC and MLP (in US $)

	Combined data	Numerical data	Sentimental data
CNN	0.54	0.58	0.54
SVC	0.57	0.57	0.55
MLP	0.55	0.55	0.50

method for the AAPL ticker was the LSTM on Sentimental data.

Table 7 presents the accuracy of each method for the three datasets. CNN on combined data had an accuracy of 61%, which was the largest one.

Figure 3 is a graphical representation of the APPL's profits for LSTM and CNN for the 3 extended datasets with or without our strategy.

GOOG The profit of the Buy and Hold strategy for the GOOG sticker was 1480.85$. Table 8 [14] shows that the best-case scenario for the GOOG ticker was the LSTM on numerical data, with a very close difference to the KNN on combined data. It is impressive that every scenario is profitable.

Table 18 NVDA profits with the extended dataset for LSTM and CNN (in US $)

	Combined extended data	Numerical extended data	Sentimental extended data
LSTM	444.03	449.19	475.02
CNN	377.08	665.39	400.63
LSTM with strategy	470.42	390.89	164.80
CNN with strategy	469.19	615.75	400.44

Table 19 NVDA profits with the extended dataset for LSTM and CNN (in US $)

	Combined extended data	Numerical extended data	Sentimental extended data
LSTM	0.52	0.47	0.45
CNN	0.55	0.55	0.44

Table 9 [14] shows that the accuracy of LSTM on numerical data was 53% while the accuracy of Decision Trees on sentimental data was slightly better (54%).

Figure 4 [14] is a graphical representation of the GOOG's profits for LSTM, Decision Trees and KNN for the 3 datasets with or without our strategy.

Table 10 presents the GOOG profits of CNN, SVC and MLP with and without our strategy for each one of the three datasets in US dollars. Best method for the GOOG ticker was the CNN on Numerical data.

Table 11 presents the accuracy of each method for the three datasets. SVC on combined and numerical data had an accuracy of 58%.

Figure 5 is a graphical representation of the GOOG's profits for CNN, SVC and MLP for the 3 datasets with or without our strategy.

Table 12 presents the AAPL profits of LSTM and CNN for the extended dataset with and without our strategy for

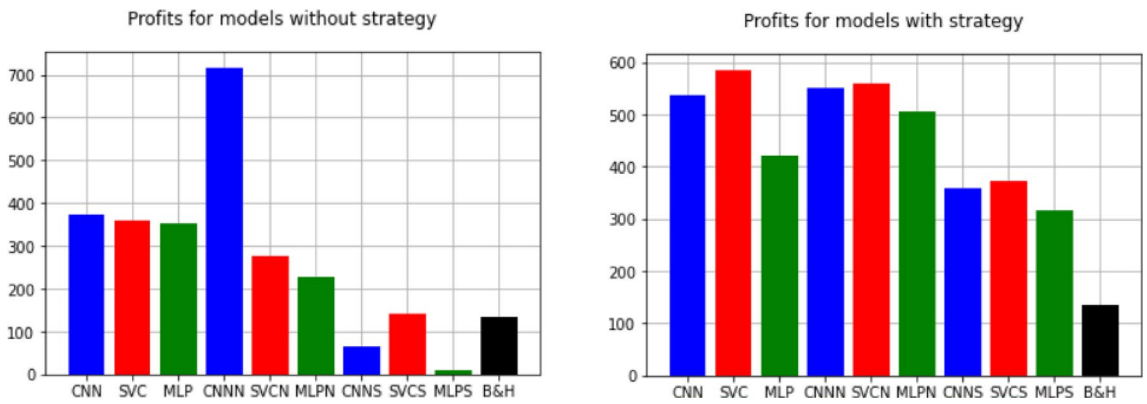

Fig. 8 NVDA Profits for CNN, SVC and MLP without strategy (left) and with strategy (right)

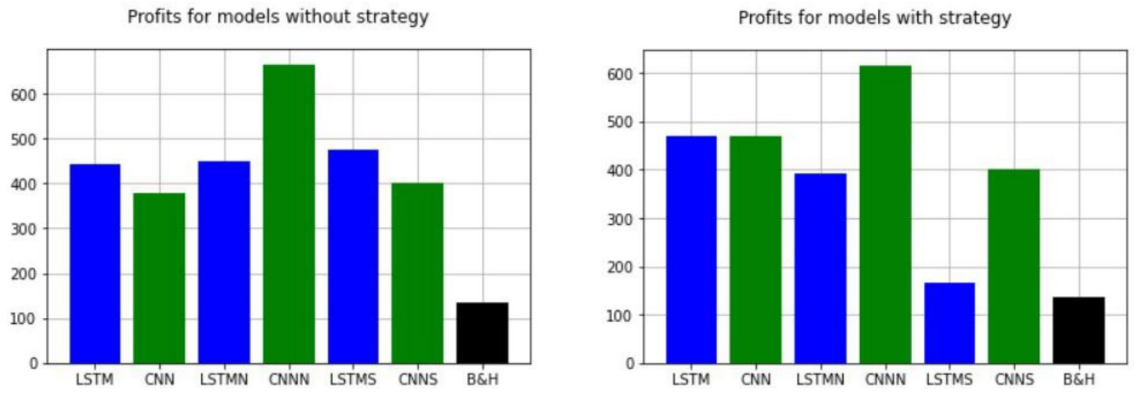

Fig. 9 NVDA Profits with the extended dataset for LSTM and CNN without a strategy (left) and with strategy (right)

Table 20 SPIS Profits for LSTM, DT and KNN (in US $) [14]

	Combined data	Numerical data	Sentimental data
LSTM	1903.76	3101.80	1772.00
DT	-190.10	1741.06	1193.60
KNN	1289.34	1794.93	183.84
LSTM with strategy	1676.49	3226.39	1544.71
DT with strategy	-65.51	1513.80	1080.50
KNN with strategy	1179.80	1685.38	27.38

Table 21 SPIS Accuracy of LSTM, DT and KNN [14]

	Combined data	Numerical data	Sentimental data
LSTM	0.60	0.60	0.60
DT	0.51	0.60	0.57
KNN	0.57	0.59	0.55

each one of the three extended datasets in US dollars. Best method for the GOOG ticker was the CNN on Numerical data, which was far superior to any other method.

Table 13 presents the accuracy of each method for the three datasets. CNN on numerical data provided us with the most profit and its accuracy was 57%.

Figure 6 is a graphical representation of the GOOG's profits for LSTM, Decision Trees and KNN for the 3 Extended datasets with or without our strategy.

NVDA The profit of the Buy and Hold strategy for the NVDA sticker was 135.50$. Table 14 [14] shows that the best-case scenario for the NVDA ticker was the LSTM on numerical data.

Table 15 [14] shows that for the NVDA the most accurate method is LSTM on numerical data while the accuracy of LSTM on sentimental data is slightly worse.

Figure 7 [14] is a graphical representation of the NVDA's profits for LSTM, Decision Trees and KNN for the 3 datasets with or without our strategy.

Table 16 presents the NVDA profits of CNN, SVC and MLP with and without our strategy for each one of the three datasets in US dollars. Best method for the NVDA ticker was the CNN on numerical data.

Table 17 presents the accuracy of each method for the three datasets. CNN on numerical data provided us with the most profit and its accuracy was 58%.

Figure 8 is a graphical representation of the NVDA's profits for CNN, SVC and MLP for the 3 datasets with or without our strategy.

Table 18 presents the NVDA profits of LSTM and CNN for the extended dataset with and without our strategy for each one of the three extended datasets in US dollars. Best method for the NVDA ticker was the CNN on Numerical data.

Table 19 presents the accuracy of each method for the three extended datasets. CNN on Numerical data provided us with the most profit and its accuracy was 55%. Interestingly enough, with percentages less than 50% we can obtain profitable strategies.

Figure 9 is a graphical representation of the NVDA's profits for LSTM and CNN for the 3 Extended datasets with or without our strategy.

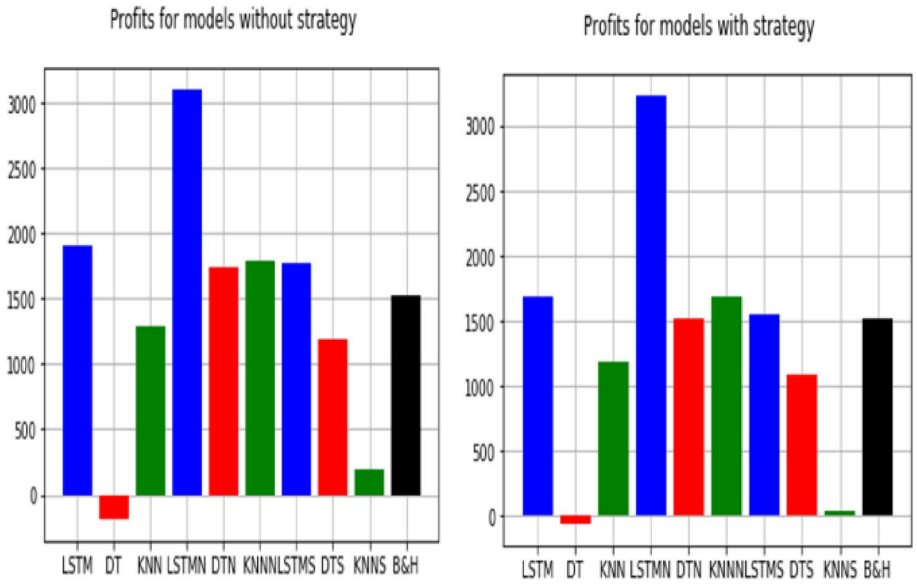

Fig. 10 SPIS profits for LSTM, DT and KNN without strategy (left) and with strategy (right) [14]

Table 22 SPIS Profits for CNN, SVC and MLP (in US $)

	Combined data	Numerical data	Sentimental data
CNN	1530.06	1792.19	1790.77
SVC	2003.18	2003.18	1705.67
MLP	2146.14	2486.08	1397.89
CNN with strategy	1060.93	1295.40	1293.97
SVC with strategy	1476.76	1476.76	1208.87
MLP with strategy	1825.66	2749.23	1006.22

Table 23 SPIS Accuracy of CNN, SVC and MLP

	Combined data	Numerical data	Sentimental data
CNN	0.52	0.54	0.54
SVC	0.54	0.54	0.53
MLP	0.55	0.56	0.53

S&P Information Technology The profit of the Buy and Hold strategy for the SPIS index was 1521.35$. According to Table 20 [14], the best-case scenario for the SPIS ticker was the LSTM with a strategy on numerical data.

Table 21 [14] shows that for the SPIS ticker, the accuracy of LSTM on all types of data is 60%. The Decision Trees method on numerical data is 60% accurate.

Figure 10 [14] is a graphical representation of the SPIS's profits for LSTM, Decision Trees and KNN for the 3 datasets with or without our strategy.

Table 22 presents the SPIS profits of CNN, SVC and MLP with and without our strategy for each one of the three datasets in US dollars. Best method for the SPIS ticker was the MLP on numerical data with strategy.

Table 23 presents the accuracy of each method for the three datasets. MLP on numerical data had an accuracy of 56%.

Figure 11 is a graphical representation of the SPIS's profits for CNN, SVC and MLP for the 3 datasets with or without our strategy.

Table 24 presents the SPIS profits of LSTM and CNN for the extended dataset with and without our strategy for each one of the three extended datasets in US dollars. Best method for the SPIS ticker was the LSTM on Sentimental data.

Table 25 presents the accuracy of each method for the three datasets. LSTM on Sentimental data provided us with the most profit and its accuracy was 62%.

Figure 12 is a graphical representation of the APPL's profits for LSTM and CNN for the 3 extended datasets with or without our strategy.

Comparing Passive Investor's and LSTM Method's Returns with the Original Dataset

The results of our previous work [14] show that the LSTM method applied to numerical data behaves better than the LSTM on combined or sentimental data as well as the KNN and the DT methods on any type of data. The following table shows the profits and the returns of the LSTM method

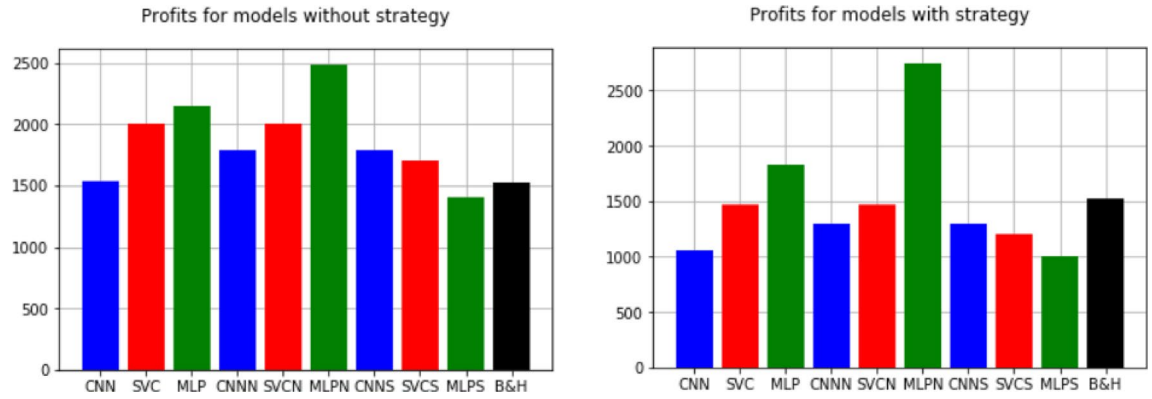

Fig. 11 SPIS Profits for CNN, SVC and MLP without strategy (left) and with strategy (right)

Table 24 SPIS profits with the extended dataset for LSTM and CNN (in US $)

	Combined extended data	Numerical extended data	Sentimental extended data
LSTM	1984.13	2090.33	2730.72
CNN	2408.55	2117.13	1879.86
LSTM with strategy	1470.81	1595.66	2271.49
CNN with strategy	1883.15	1446.18	1440.18

Table 25 SPIS accuracy with the extended dataset of LSTM and CNN

	Combined extended data	Numerical extended data	Sentimental extended data
LSTM	0.60	0.60	0.62
CNN	0.55	0.53	0.59

on numerical data of the original dataset and the Buy and Hold Strategy.

From Table 26 [14] we calculate the average returns of each method:

The average returns of B&H: 33, 52% and average returns of LSTM on numerical data: 80, 42%

So, the LSTM method on numerical data offers about 2.5 times more profit on average than the passive investor's strategy.

Improvements in Profit with the Extended datasets

Table 27 shows that the extended dataset is far better than the original one that we applied at [14], with the only exception of the Numerical data with the LSTM method where we would have 33.72% profit loss. In any other case,/ the profits on average were augmented amazingly. The best

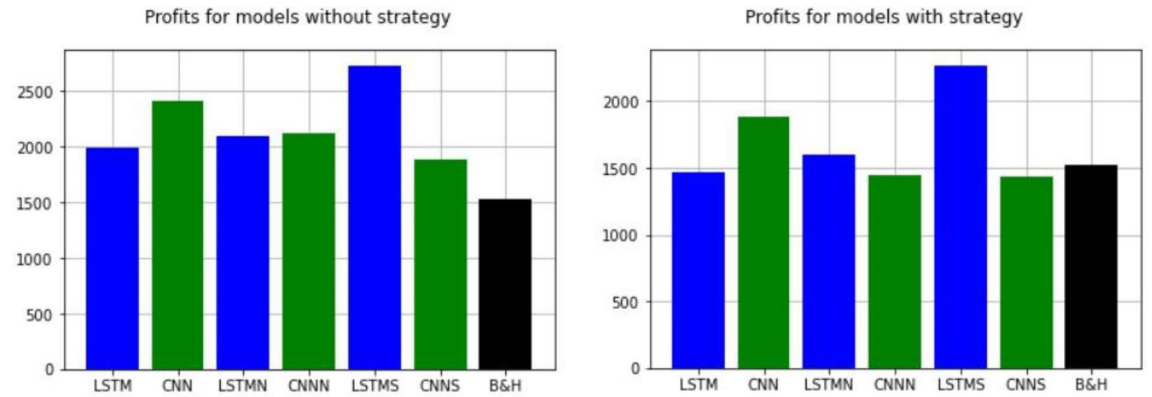

Fig. 12 SPIS profits with the extended dataset for LSTM and CNN without strategy (left) and with strategy (right)

Table 26 Returns of B&H strategy and LSTM on numerical data [14]

	Buy & hold	LSTM (numerical data)	Returns on B&H (%)	Returns on LSTM (numerical data) (%)
AAPL	139.89	222.82	65	103.4
GOOG	1480.85	2498.33	28.4	47.8
NVDA	135.50	1149.00	13.6	115.2
SPIS	1521.35	3101.80	27.1	55.3

improvement was with the CNN sentimental Dataset reaching a 140% increase on average.

LSTM vs CNN with the Extended Datasets

From Table 28, we conclude that, in a 2-year testing period returns from all Deep Learning Methods with the extended datasets outperformed the B&H strategy and the best method (returning the most profits) was CNN Numerical Data with 61.55% Return on Investment (ROI). Furthermore, the improvement from our previous results [14] on Sentimental data is significant.

Conclusions

Having developed a system in our previous research [14] that had the potential to capture stock market trends, we extended it further in this paper, making it more accurate, trustworthy and approachable from different perspectives. In our first attempts, the numerical data in combination with the LSTM model proved to be the most profitable answer. Knowing this we tried to make the different types of data work better together. Our extended research boosted the accuracy and the profits of the textual and the combined data while maintaining the high precision of the numerical, thus giving more options on how to accurately predict stock market movements. Providing more alternative methods to make such predictions, whether it is only textual data deriving from a Machine Learning method under an LSTM model or only numerical data under a CNN model, and so on, showed that it can only increase the forecasting capabilities available.

Regarding the contribution of this paper, we utilized a combination of numerical data and textual/sentimental data to predict stock market trends, whereas previous work was mainly focused on either one or the other. We used a variety of machine learning and deep learning models to process both types of data, including LSTM, CNN, KNN, decision trees, SVM, and MLP. Furthermore, we evaluated the performance of our approach on multiple datasets and also used

Table 27 Average % improvement in profits when applying the extended dataset in comparison to the original one

	LSTM combined data (%)	LSTM numerical data (%)	LSTM sentimental data (%)	CNN combined data (%)	CNN numerical data (%)	CNN sentimental data (%)
AAPL	34.36	− 27.67	41.29	33.91	− 8.6	13.34
GOOG	155.49	− 13.69	19.49	29.78	111.11	19.28
NVDA	37.27	− 60.91	100.91	0.89	− 7.19	525.65
SPIS	4.22	− 32.61	54.1	57.42	18.13	4.97
Average per method	**57.83**	**− 33.72**	**53.95**	**30.5**	**28.36**	**140.81**

Bold is the average percentage per method when taking into account all tickers

Table 28 Return on investment with the extended datasets for LSTM and CNN and comparison to B&H strategy

	Returns on LSTM combined data (%)	Returns on LSTM numerical data (%)	Returns on LSTM sentimental data (%)	Returns on CNN combined data (%)	Returns on CNN numerical data (%)	Returns on CNN sentimental data (%)	Returns on buy & hold (%)
AAPL	86.7	74.79	91.8	84.8	64.18	65.5	65
GOOG	62.93	41.26	41.47	47.7	77.59	58.75	28.4
NVDA	44.52	45.04	47.63	37.81	66.71	40.17	13.6
SPIS	35.37	37.27	48.68	42.94	37.74	33.51	27.1
Average per method	57.38	49.59	57.39	53.31	61.55	49.48	**33.52**

as metrics not only accuracy but also profitability. There are cases where high accuracy does not mean high profitability. Finally, we provided insights into how sentiment analysis can be used to strengthen technical analysis in predicting stock market trends, but also the significant improvement with the extended dataset that we utilized.

Taking all of the above into consideration, the question that arises is: "If there are such alternatives to forecast stock market trends, is there always a way to predict a company's stock trends?". If this is true then there is an even bigger question arising against the efficient market hypothesis. While many researchers and we ourselves have shown that there are certain patterns that are recognizable and exploitable, to what extent is it still debatable. More research should be done with data on different companies of different media and sizes before finalizing a statement. That means that more samples of different scales are needed. Furthermore, while we have implemented certain methods and ways, there are still many different approaches available, which need to be tested and developed (e.g. the use of ontologies [28]).

In the future, we hope to evolve our technology into an autonomous system that can forecast the stock ticker's trend every day. To make this work in the long run, it is necessary to train the system online over time to keep it up to date. We'll also experiment with different strategies for utilizing different types of data in order to increase forecast accuracy.

Funding This study has received no funding.

Data availability The economic data utilized in this study was sourced from Yahoo Finance. Economic data used in this research is publicly available and can be accessed through Yahoo Finance's platform. The financial_phrasebank dataset referenced in this study was also utilized. The dataset is publicly available.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Liberti JM, Petersen M. Information: hard and soft. Rev Corp Finance Stud. 2019;8(1):1–41.
2. Chong E, Han C, Park FC. Deep learning networks for stock market analysis and prediction: methodology, data representations, and case studies. Expert Syst Appl. 2017;83:187–205.
3. Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions. Eur J Oper Res. 2018;270(2):654–69.
4. Long W, Lu Z, Cui L. Deep learning-based feature engineering for stock price movement prediction. Knowl Based Syst. 2019;164:163–73.
5. Zhong X, Enke D. Predicting the daily return direction of the stock market using hybrid machine learning algorithms. Financ Innov. 2019;5(1):1–20.
6. Vignesh CK. Applying machine learning models in stock market prediction. EPRA Int J Res Dev. 2020;5(4):395–8.
7. Nabipour M, Nayyeri P, Jabani H, Mosavi A, Salwana E. Deep learning for stock market prediction. Entropy. 2020;22(8):840.
8. Ferreira F, Gandomi A, Cardoso R. Artificial intelligence applied to stock market trading: a review. IEEE Access. 2021;9:30898–917.
9. Sun A, Lachanski M, Fabozzi F. Trade the tweet: social media text mining and sparse matrix factorization for stock market prediction. Int Rev Financ Anal. 2016;48:272–81.
10. Shapiro AH, Sudhof M, Wilson D. Measuring news sentiment, Federal Reserve Bank of San Francisco Working Paper 2017-01. 2017. https://doi.org/10.24148/wp2017-01.
11. Pagolu VS, Reddy KN, 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, Paralakhemundi, India, 3–5 October 2016, pp 1345–1350. https://doi.org/10.1109/SCOPES.2016.7955659
12. Batra R, Daudpota SM. Integrating StockTwits with sentiment analysis for better prediction of stock price movement. In: Proceedings of International Conference on Computing, Mathematics and Engineering Technologies; 2018. pp. 1–5.
13. Tabari N, Seyeditabari A, Peddi T, Hadzikadic M, Zadrozny W. A comparison of neural network methods for accurate sentiment analysis of Stock Market Tweets. In: ECML PKDD 2018 Workshops. MIDAS 2018, PAP 2018. LNCS, vol 11054. 2019; Springer.
14. Chatziloizos G, Gunopulos D, Konstantinou K. Forecasting stock market trends using deep learning on financial and textual data. Proceedings of the 10th International Conference on Data Science, Technology and Applications (DATA 2021). SciTePress. pp. 105–114.
15. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.
16. Ussama Y, Soon C, Vijayalakshmi A, Jaideep V. Sentiment-based analysis of tweets during the US Presidential Elections. 2017. pp. 1–10. https://doi.org/10.1145/3085228.3085285.
17. Rao T, Srivastava S. Analyzing stock market movements using twitter sentiment analysis. Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM); 2012. pp. 119–123.
18. Lounnapha S, Zhongdong W, Sookasame C. Research on stock price prediction method based on convolutional neural network. Proceedings of the international conference on virtual reality and intelligent systems (ICVRIS). IEEE; 2019. pp. 173–6.
19. Yan X, Zhao J. Application of improved convolution neural network in financial forecasting. Proceedings of the 4th IEEE international conference on cloud computing and big data analytics. 2019. pp. 321–6.
20. Cao J, Wang J. Stock price forecasting model based on modified convolution neural network and financial time series analysis. Int J Commun Syst. 2019;32:e3987.
21. Malo P, Sinha A, Korhonen P, Wallenius J, Takala P. Good debt or bad debt. J Assoc Inf Sci Technol. 2014;65:782–96. https://doi.org/10.1002/asi.23062.

22. Hutto CJ, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the 8th international conference on weblogs and social media. ICWSM; 2015.
23. Loughran T, McDonald B. When is a liability not a liability? textual analysis, dictionaries and 10-Ks. J Finance. 2011;66:35–65.
24. Fiol-Roig G, Miró-Julià M, Isern-Deyà AP. Applying data mining techniques to stock market analysis. In: Trends in practical applications of agents and multiagent systems. Advances in intelligent and soft computing, vol 71. Springer; 2010.
25. chart-formations.com. http://www.chart-formations.com/indicators/atr.aspx?cat=volatility
26. Larry Williams CTI Publishing. https://williamspercentr.com/the-original-percent-r
27. Pedregosa F, et al. Scikit-learn: machine learning in Python. JMLR. 2011;12:2825–30.
28. Kontopoulos E, Berberidis C, Dergiades T, Bassiliades N. Ontology-based sentiment analysis of twitter posts. Expert Syst Appl. 2013;40(10):4065–74.