paper reference:

Applying LSTM for Stock Price Prediction with Sentiment Analysis

Alexandre Heiden
Graduate Program in Applied Computing
Santa Catarina State University (UDESC)
Joinville, SC - Brazil

Rafael Stubs Parpinelli
Graduate Program in Applied Computing
Santa Catarina State University (UDESC)
Joinville, SC - Brazil

***Abstract*—Financial news has been proven to be valuable source of information for the evaluation of stock market volatility. Most of the attention has been given to social media platforms, while news from vehicles such as newspapers are not as widely explored. Newspapers provide, although in a smaller volume, more reliable information than social media platforms. In this context, this research aims to examine the influence of financial news within the stock price prediction problem, by using the VADER sentiment analysis model to process the news and feed the sentiments as a feature into a LSTM-based stock price prediction model, along with the historical data of the assets. Experiments indicate that the model has better results when the news' sentiments are considered, and the model demonstrates potential to accurately predict stock prices up to around 60 days into the future.**

Keywords*—*stock price prediction, sentiment analysis, financial news.

I. Introduction

The financial market is an environment for trading assets, which are divided into monetary (currency and exchange), public and private bonds, commodities and stocks. The stock market is the most popular among retail investors due to its high earning potential.

Shares are titles corresponding to a share of ownership in a company, so the shareholder of a given company has a right to assets and profit sharing [1]. The moment an investor acquires ownership of a set of shares in one or more companies, he becomes the owner of an investment portfolio.

While the stock market is profitable, it also proves to be highly volatile. This behavior is due to the fact that the stock price depends directly on the decisions taken by the companies [2], decisions that are unpredictable to some extent, supported by the high competitiveness of the market. Therefore, the investor's objective is to build a portfolio that can balance the risk and return factors. For this, investors have a framework of mathematical models, which help them to build their portfolio.

In order to abstract the responsibility of the analysis and decision-making of the investor, the Portfolio Optimization Problem (POP) emerged. The classic approach to the problem comprises the optimal allocation of investor capital based on historical asset series. This construction is relatively famous, and several models have already been implemented to solve this problem [3].

Recently, models based on Machine Learning strategies have been widely studied to take an even more ambitious step in the area of investment portfolio optimization: stock price prediction. Building an artificial neural network and inserting the historical data of assets as a feature has the potential to predict prices with high accuracy [4].

However, analyzing just the historical series of prices is somewhat superficial, as there are multiple other factors that influence the quotation of an asset [5]. Recent studies indicate that one of the main factors responsible for the variation in stock prices on the stock exchange is news related to the respective proprietary companies [6], [7], [8]. Adding more elements to the analysis is extremely important to build a balanced portfolio, and the sentimental analysis of the news shows positive results even during the COVID-19 [9] pandemic, a fact that destabilized the economy on a global scale and made the equity investment environment absolutely unstable.

The COVID-19 global pandemic has changed the world's views on a wide array of subjects, including the stock market investments. Strengthened by reliable financial information sources and investment tools easily available to the general public, there has been a significant growth in the interest of retail investors in the stock market. Although there are financial technology companies, called Fintechs, that offer portfolio management methods for these small investors, the services provided by the Fintechs are usually quite expensive. Therefore, the small investors are increasingly seeking their own investment strategies [10], reducing their dependence on a Fintech or a bank in their applications on the market.

However, the retail investor has significantly fewer tools than a large financial institution. Leaving all decision making in the investor's hands can become overwhelming for them, due to the massive amount of indicators in the market and factors that contribute to the volatility of asset prices.

The solution to the problem is to abstract difficult decision making using computational intelligence. Machines have a much greater processing power than an investor alone, being able to absorb a large amount of information such as news from companies in the financial sector and prices of their respective assets. Minimizing human interaction, in addition to improving decision-making, also leads to less conflict of interest and, on occasion, better market efficiency [11].

This research aims to contribute to the portfolio optimization literature by developing a stock price prediction strategy using a Machine Learning model based on the Long Short-Term Memory (LSTM) artificial recurrent neural network. The model will be paired with a sentiment analysis strategy, which will classify news from the newspaper The New York Times to help prediction prices in the period studied.

The remainder of this paper is organized as follows. Section 2 briefly discusses the relevancy of the study and further details on the stock price prediction approaches; Section 3 displays related work on the area; Section 4 elaborates the architecture and characteristics of the proposed model; Section 5 discusses about the validation and results of the model; and Section 6 points out the conclusions, limitations of this study and future work.

II. Background

A. *Stock Price Prediction*

There is an important theory that must be understood when working with quotation prediction, the *Efficient Market Hypothesis* (EMH) [12]. EMH indicates that all available information immediately reflects the current state of the market, suggesting that no predictions for future changes can be made. This hypothesis was proven false by its own creator [2], but even overturned, the hypothesis served as a catalyst for numerous researches in the area. Currently, there are two main types of approaches to predicting market behavior: technical analysis and fundamental analysis.

Technical analysis denotes the study of past prices, using charts as the main tool. This analysis assumes that market reactions to the news are instantaneous and therefore does not take them into account in its prediction attempts. The objective of technical analysis is to identify patterns in the historical series, in order to anticipate market changes [13].

Fundamental analysis looks at the indicators that affect market supply and demand [14]. The idea is to collect and process information before it reflects its consequences on the market. This in-between time represents an opportunity to dispose of stocks that are about to go down or buy stocks that are about to go up in value. This type of analysis uses data about companies to predict market movements, with news as the main source.

The big difference between technical and fundamental analysis is the inclusion of news in prediction models, generally represented by artificial neural networks. News carries precious information about the market and represents a great impact on the prediction of stock prices [15]. The problem is that automating the interpretation of news proves to be a very complex task, and this task characterizes a large area of study, known as sentiment analysis.

B. *Sentiment Analysis*

Sentiment Analysis is a classification problem that aims to discover the polarity of an opinion expressed in textual form, that is, whether a text represents something positive, negative or neutral [16]. There are two main ways to automate the sentiment analysis process: dictionary-based methods and machine-learning application-based methods. Using sentiment analysis, it is possible to perceive the intentions of companies and investors in real time, something extremely important for stock exchange decision-making.

Sentiment analysis models that apply machine learning rely on techniques to learn from automatically received textual bodies. No manual rules need to be developed. The neural network training process is entirely based on feature extraction.

Measuring investor sentiment aims to verify if financial news have impact on market movements and to identify patterns of sentiments paired to these movements. On a volatile environment like the stock market, news sentiments is one of the many pointers that can be useful in stock price prediction.

C. *Long Short-Term Memory (LSTM)*

a type of RNN that has memory so that past data contributes to the output

The biggest difficulty of applying an ordinary neural network architecture on a stock price prediction scenario is their inability to handle long term dependencies, a factor of extreme importance for any problem that consists of a time series. Introduced by Hochreiter and Schmidhuber [17], the LSTM is a special type of recurrent neural network explicitly designed to solve the long term dependency problem, by introducing a new concept of memory cell to replace the traditional artificial neurons on the hidden layers.

The cell state in an LSTM is designed to be able to remember information for a long period of time [18]. This information is carefully regulated by three structures called gates, which will control what information will be thrown away from the cell, what information will be inserted on the cell and what will be the cell output. With this complex structure of memory cell, the LSTMs are able to assimilate the structure of the data dynamically over a time span with high prediction ability.

Due to the ability of processing sequential data in remarkable manner, LSTMs prove to be an efficient mechanism in many different fields, including but not limited to computer science, statistics, linguistics and medicine. All of these areas have complex tasks revolving around prediction, classification and analysis of sequential data, tasks for which LSTM is known to perform well [19].

III. Related Work

Price prediction has always been an extremely complex task due to the volatility of the investment market [20]. The area shows a lot of interest from researchers and presents several different approaches to approach the problem.

Du and Tanaka-Ishii [21] developed their own sentiment analysis model to investigate news drawn from the Wall Street Journal (WSJ) and Reuters & Bloomberg (R&B) database, categorizing news according to a metric for calculation of the weight of each news in relation to its respective action. They used Multi Layer Perceptron (MLP) to perform price prediction and fed an optimization model based on the classic Markowitz model [22]. Observing 18 stocks selected from the American market, the authors compared their strategy with others presented in the literature. Their model obtained better results for the R&B dataset, but it was outperformed for the WSJ dataset by a Word2Vec model that only considers news to prediction the prices.

Creamer [23] used Automated Text Analysis (ATA) to classify news extracted from the TR News Archive. With the result, the author assembled the representation of the investor's vision in the Black-Litterman portfolio optimization model [24]. The proposed model, featuring 27 stocks from the European market, indicated better performance compared to the market index, which is generally a good indicator. However, the model assumes that, for the investor's view generated, the error co-variance is low and the confidence is high. This assumption is somewhat arbitrary, as the view generated cannot be generalized to all investors: some are more cautious, others more impulsive.

Xu and Cohen [6] developed a sentiment analysis model for posts collected from Twitter heavily based on the Gated Recurrent Unit (GRU) principle, which featured a layer in their original model for predicting stock prices, called StockNet. The application scenario was purely theoretical, featuring 88 benchmark stocks. The authors also needed to implement a strategy to filter tweets, implying that many comments are discarded. This work does not predict the gross closing value for the assets, but rather predicts the movement of the asset (binary prediction: the asset price will raise or fall) on that day. The authors did not use the result to build portfolios, they just compared the accuracy of the model with other models in the literature.

Nguyen *et al.* [25] evaluated news from the Yahoo Finances forum on 18 stocks in the US market. The authors' sentiment analysis categorizes each news item into one of 5 classes (Strong Buy, Buy, Hold, Sell, and Strong Sell) to feed the prediction model using an SVM. This work also presents binary prediction and also only compares the accuracy between models. The biggest problem with using an SVM to make this prediction is the fact that you can't tell for sure the length of the prediction window, i.e. how many days to take into account in training to consolidate a prediction. The authors used a window of size of 2, therefore, to predict the movement of the day d, the news and closing values of the days $d-2$ and $d-1$ are used.

Song *et al.* [7] implements a relatively simple sentiment analysis but covers the need imposed by the prediction model - they only calculate two values: shock and news trend taken from Thomson Reuters News Analytics. The Learning-to-rank model is composed of a neural network with learning guided by a descending gradient that has a cross-entropy loss function. This article does not prediction prices, but a ranking of 512 assets in the American market and makes simulations using 4 different investment strategies, without using an optimizer. The strategy that generated the greatest return was to buy the best-ranked assets and hold them for the entire time period studied.

The work of Xing *et al.* [8] is quite unique in relation to the others studied. The prediction model, an MLP network, is applied to predict whether the price of the US Dollar and Euro will appreciate or not. The analysis made by the authors is extremely complete and confirms the predictive power of a model using high frequency news without any kind of technical analysis.

Mittal and Goel [26] used Twitter as a database. The sentimental analysis model implemented only categorized the news into 4 groups, representing the user's mood when posting their tweet: calm, happy, alert or generous. The authors implemented 4 prediction models: linear regression, logistic regression, SVM and Self-Organizing Fuzzy Neural Network (SOFNN). According to the Mean Absolute Percentage Error (MAPE) metric, the best predictor was SOFNN. This network has the need to use a time window value, the authors arbitrarily chose 3 days. With the predictor result, the authors implemented a greedy strategy to select the best portfolio, based solely on predicted closing values: there was no risk calculation.

Beraldi [27] used the 7 *Exchange Traded Funds* (ETFs) represented by the B3 segment indexes of the IBOVESPA as assets and Twitter as the news base. The exercise is purely theoretical, as the Brazilian ETF market is not robust. Sentiment analysis is done using the Valence Aware Dictionary for Sentiment Reasoning (VADER) technique, which is quite interesting for quantifying news sentiments in a way that facilitates the creation of a time series. The author implemented the Black-Litterman model for portfolio optimization and fed it with feelings representing the investor's view, similar to Creamer's approach [23], so this work does not use feelings to prediction prices . The author's model performs better than the IBOVESPA benchmark, but the portfolio allocation is unbalanced, due to the characteristic of the Black-Litterman model. The author only measures the portfolio's return and does not pay attention to risk, which could potentially be high on the poorly diversified portfolios, especially in scenarios with high volatility such as during the COVID-19 pandemic.

Faizan [28] used stock exchanges, Bitcoin and even the price of gold in his experiments. The news was taken from the NY Times and the sentimental analysis was done using IBM's Watson API, the author does not inform in detail the method used by the API. The author used the Reference Point Method (RPM) to solve a bi-objective model (return maximization and risk minimization) of portfolio optimization, considering the information from sentimental analysis in the form of constraints in the model, a singular approach in in relation to the other works studied.

Roondiwala *et al.* [4] implemented a Long Short-Term Memory (LSTM) model to prediction prices in the Indian stock market. By feeding only historical asset data into the model, the authors were able to predict closing prices with formidable accuracy. The authors leave open questions for not using sentimental analysis in their model nor exploring the predictions generated with a portfolio optimization algorithm.

A. *Methodologies and Applied Techniques*

Both the state of the art in sentiment analysis and in stock price prediction is dominated by ANNs. Table I summarizes the literature review presented. It is noticed that, although always equipped with ANNs, the literature does not tend to a preference for a particular model, neither for sentiment analysis nor for the prediction of stock prices per se. It is also noticed that not all researches use the predicted prices to actually select the investment portfolios: in many cases, the researchers only assess the prediction accuracy without worrying about the application of its results in a real scenario of the investment market.

No sentiment analysis model proves to be more effective than the others. The fact is that sentiment analysis cannot be treated as a detached component, considering that the result of the model is applied on the prediction model. This makes it difficult to evaluate this component by itself. Furthermore, the models are usually generic, with no evidence of a model

TABLE I: Main Approaches Identified in Literature

Reference	Scenario	Sentiment Analysis	Prediction Model	Portfolio Selection Model
[21]	US	Original	MLP	Markowitz
[23]	Europe	Automated Text Analysis	-	BL
[6]	Theorical	Original*	Original	-
[25]	Us	Joint Sentiment/Topic	SVM	-
[7]	US	Indicators	Learning-to-rank	-
[8]	Exchange	BERT	MLP	-
[26]	US	Original	SOFNN	Guloso
[27]	Theorical	VADER	-	BL
[28]	US/BTC	Watson	-	RPM
[4]	India	-	LSTM	-

specifically aimed at helping to predict quotes. However, the news database is a factor that influences the choice, as there are faster models (for databases with large volume, such as Twitter) and more robust models (for databases with small volume, such as newspapers).

Unlike sentiment analysis, the predictive model has an approach that stands out. Recent studies indicate that the LSTM network, an architecture specially designed to process data in the form of historical series, has better performance than the other identified models [4], but there is a lack of evidence to prove that the results are good enough to generate efficient portfolios scenarios with real data.

IV. Proposed Model

The approach presented in this work separates the model into two stages: sentiment analysis and price prediction, as shown in Figure 1. Both stages have different requirements and will end up working together.

The sentiment analysis stage aims to process the news of the companies studied and define the position of investors during the period studied, based on the sentiment obtained. Sentiment analysis tasks generally requires some kind of data pre-processing in order to accomplish good results.

The prediction of prices stage, which is supported by the result of the previous stage and by the historical series of the prices of assets in the period studied, has the objective of developing a model that is intelligent enough to generalize the prediction to all studied assets, which will represent a portion of the universe of assets on a stock exchange. The price prediction task requires a robust model to be developed, which has the necessity of validation with data and parameter tuning.

1) Sentiment Analysis Model: VADER

The VADER sentiment analysis model does not require a large database to achieve high accuracy [29], which is extremely important when working with a lean database such as a newspaper. Although newspaper news are extremely more reliable than comments on a social network, the amount of news is much more limited, meaning the database is small.

The potential of the VADER model is demonstrated in [29] by comparing it with several other strategies. The comparison indicates that the model obtains better results than all of them, both for an analysis of comments on social networks and for analysis of newspaper news.

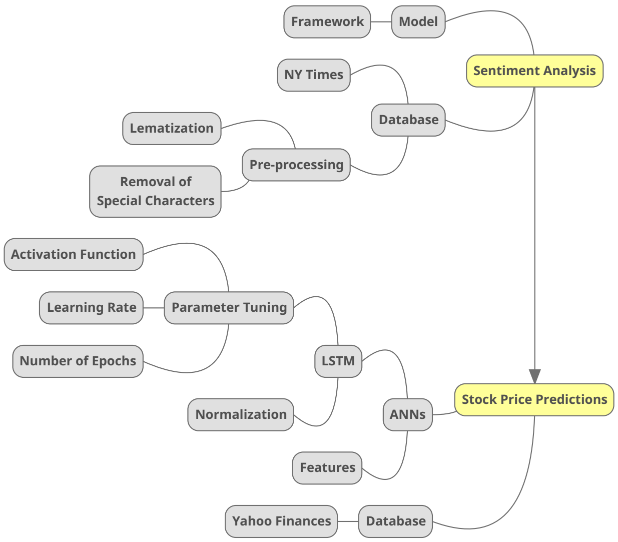

Fig. 1: Model Requirements

Although the VADER model was not specially developed to work in the investment market scope, the generalization of the model is proven to be great. The application of the model on this paper will verify VADER's efficiency on a purely financial context. The fact that it is a relatively fast model compared to other sentiment analysis models is also convenient, allowing news processing to be done quickly and efficiently.

2) Stock Price Prediction Model: LSTM

The LSTM artificial neural network is capable of identifying long-term dependencies, as long as the training data is properly segmented into sub-sequences with a well-defined beginning and end [18]. This is proven to be true when analyzing historical series of asset prices, as any sequential subset of the series can be effectively listed as training data. It is up to the researcher to define the size of these subsets, considering that a window that is too small can inhibit the model's memorization capacity and, on the other hand, a window that is too large will cause an execution bottleneck during model training.

The ability to keep information from the past is what guarantees the potential application of LSTM to the problem of predicting asset prices, as the previous price is a crucial element to predict the price in the future.

3) Database Descriptions: To analyze the efficiency of the model in a real investment scenario, the experiments will be carried out with the historical series of assets that make up the Top 10 of the S&P 500 index[1], which is designed to include companies with the largest market capitalization, with indicators for the first quarter of 2021. Market capitalization indicates the market value of a company, reflected by the share price and the number of shares available. The assets considered in the experiments are shown in Table II.

The historical series of assets were extracted from the Yahoo Finances platform and the news were extracted from the API of The New York Times newspaper. The period considered was from January 1, 2016 to December 31, 2020. For each asset in each year, the 500 most relevant news from the section with the theme "finance" were considered.

TABLE II: Stocks Analyzed

Ticker	Corresponding Company
AAPL	Apple
AMZN	Amazon
BRK-B	Berkshire Hathaway
FB	Facebook
GOOG	Alphabet
JNJ	Johnson & Johnson
MSFT	Microsoft
PG	Procter & Gamble
V	Visa
WMT	Walmart

4) Problem Characterization: In the sentimental analysis stage, the news of each asset was individually submitted to the VADER model. The aim is to determine the news sentiment for each closing day in the 2016-2020 period for each of the 10 assets. Sentiment s will be a value in the range $[-1, 1]$, $s = -1$ representing the most negative sentiment possible (only bad news) and $s = 1$ representing the most positive sentiment possible (only good news).

As 500 news per year are considered, some days have more than one news for the asset. For these cases, the sentiment of the day in question is given by the average of the sentiments of each news. Furthermore, some days do not have any news for the asset. For these cases, the sentiment is given as neutral, $s = 0$.

At the end of the step, a sentiment value will have been assigned for each closing day for each analyzed asset. This value will be used as one of the features of the prediction model in the next step. The model will have a total of two features, the sentiment values and the historical stock prices for every asset.

In the price prediction stage, the historical series of assets and their respective sentiment values are used as input in the prediction model. An ordinary LSTM model contains three layers: an input layer, an LSTM layer of size n, and a dense layer with a node, responsible for consolidating the input received from the LSTM layer into a final prediction value. The model proposed in this work, in turn, has an additional LSTM layer, as outlined in Figure 2. The stacking of layers, in this scenario two LSTMs, provides greater abstraction capacity to the model [30], [31], allowing the representation of more complex patterns. This strategy, on the other hand, increases the computational cost of the model. The architecture was defined empirically.

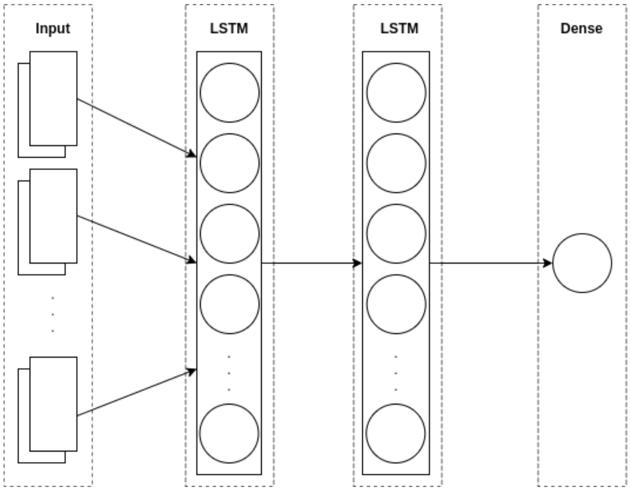

Fig. 2: Prediction Model Architecture

Input data is formatted according to one parameter, the window size. This parameter indicates how many days will be considered "dependent" for the stock price prediction. For example, with a window size $t = 30$, the 30 days prior to the day d will be considered in the prediction. The larger the window, the greater the model's view of past information, but the higher the model's computational cost. With the window size t, the historical prices and sentiment news series are split in $n - t$ (n being the series' length) sequential groups and paired together, creating the input layer.

The separation between training and validation data is not done randomly, a strategy commonly applied in artificial neural networks. As the data is a historical series, the values are dependent on its predecessors, for example: the price of the day d depends directly on the price of the days $d-1$, $d-2$ and so on, therefore, the separation of sets is carried out according to the date of the prices. The training/validation split is given at $80\%/20\%$, culminating in the years 2015-2019 being used for training and the year 2020 being used for validation.

5) Evaluation Metrics: The metric used to assess the accuracy of the model is the Root Mean Squared Error (RMSE), which represents the square root of the mean squared error. The use of RMSE is quite common and is considered an excellent general purpose error metric in the [32] numerical prediction scenario. Equation 1 represents the RMSE metric. Min-max normalization will be applied to inhibit the different scales of asset prices.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_n - y_n)^2} \quad (1)$$

With n being the size of the series, y_n the real stock price on

[1] Live chart with all the 500 tickers can be found at https://www.slickcharts.com/sp500

the day n and $\hat{y}_n$ the stock price predicted by the model on the day n.

V. Results Obtained and Analysis

The calculation of sentiments was performed with the VADER model, implemented exactly as exposed by the creators in [29]. With the sentiments calculated and paired with the historical series of asset prices, the LSTM model was executed. The model execution parameters were empirically set at 100 epochs and a window of 60 days for building the input samples. The model was implemented using the Python language and all experiments were carried out in a controlled environment, in equipment with the following specifications: Intel i7 Core™ i7-4770 processor operating at 3.9 GHz, with 16 Gigabytes of RAM and running a GNU/Linux operational system with kernel 4.8.10. The model's architecture is composed of four layers, as shown earlier in Figure 2. The first LSTM layer has 64 units and the second LSTM layer has 32. Both of the LSTM layers use the Hyperbolic Tangent as activation function. The dense layer has a singular unit and the activation function is Rectified Linear Unit (ReLU).

The first tests aim to verify if the sentiment of the news have an impact on the performance of the model. The model was ran 50 times without the sentiment feature and another 50 times with the sentiment feature, and the average and standard deviation of the RMSE was calculated at the end of every run for every set of stocks, as well as the average of the normalized RMSE, represented by RMSE-N. The goal of normalizing the RMSE is to tone down the discrepancies of the stock prices: a company with higher valued stocks will show a higher RMSE than a company with lower valued stocks, even in some scenarios where the proportional error is smaller. The results are shown in Table III.

TABLE III: Comparison between Model With and Without Sentiments

	Without Sentiments		With Sentiments	
Ticker	RMSE	RMSE-N	RMSE	RMSE-N
AAPL	31.21 ± 0.33	0.383	4.42 ± 0.10	0.051
AMZN	735.39 ± 5.85	0.396	149.96 ± 4.43	0.083
BRK-B	26.99 ± 0.17	0.377	7.27 ± 0.35	0.098
FB	52.86 ± 0.19	0.330	12.32 ± 1.84	0.091
GOOG	243.71 ± 1.48	0.319	60.14 ± 2.41	0.076
JNJ	9.55 ± 0.08	0.204	4.48 ± 0.23	0.094
MSFT	32.49 ± 0.24	0.336	8.06 ± 0.22	0.083
PG	15.56 ± 0.09	0.325	3.50 ± 0.09	0.073
V	21.27 ± 0.16	0.256	8.51 ± 0.14	0.103
WMT	17.38 ± 0.13	0.352	3.78 ± 0.17	0.079

The results show clear superiority of the model using sentiments as a feature, having a better performance for all stocks analyzed. The relatively low standard deviation for the RMSE values indicate that the model is well calibrated, showing small difference for the results of every independent run. The RMSE values are also very similar for every stock, therefore the model has a good generalization and works well for every stock.

To validate the hyper-parameters tuning, the training and validation loss curves of the model were investigated. These curves, although simply-looking, provide very valuable information about the model performance. Figure 3 shows the average training and validation loss curves of the 50 runs of the final model.

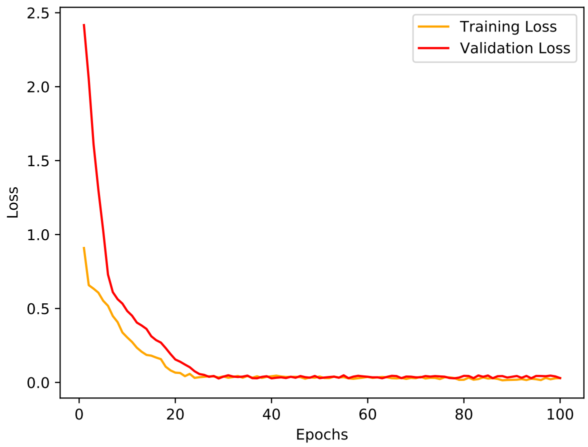

Fig. 3: Training and Validation Loss

The results show that both curves decrease to a point of stability and have a small gap between them. The first fact indicates that the model is trained enough to generalize well with the data provided, and further training will eventually lead to overfitting. The second fact is expected, because the loss of the model will almost always be lower on the training dataset compared to the validation dataset, this is only a problem if the gap is large. The results indicate that the model is well adjusted, however adding more data to it would require further analysis. It is very important not to overcomplicate the model's architecture for it to be able to generalize predictions on unseen data, which is the final test performed on the model.

A. *Predicting Stock Prices into the Future*

The model's biggest challenge is to predict the stock prices into the future, a scenario which the model has to operate with limited points of data. As explained before, the LSTM model works with a window size to represent the memorization capacity. This is crucial to understand why is it so difficult to predict prices into the future.

The question to be answered is how long into the future is possible to extract accurate predictions for a stock price. The following analysis will be based on the ticker which presented the worst normalized RMSE on the Table III, AMZN. The illustration of the results for the AMZN ticker demonstrates that even the ticker with the worst results have a good performance for the model. The Figure 4 shows the model validation for the AMZN stock.

To answer the question, the model will attempt to predict the prices after the testing period, which ended in 2020. Using the same approach as described before to generate the input samples, the first 90 closing days prices of 2021 are generated and compared to the real closing price. The results are shown in Figure 5.

As expected, the predictions have a good accuracy at the beginning of the series and suffer a noticeable decrease on

i dont quite get how they can do this without any news articles

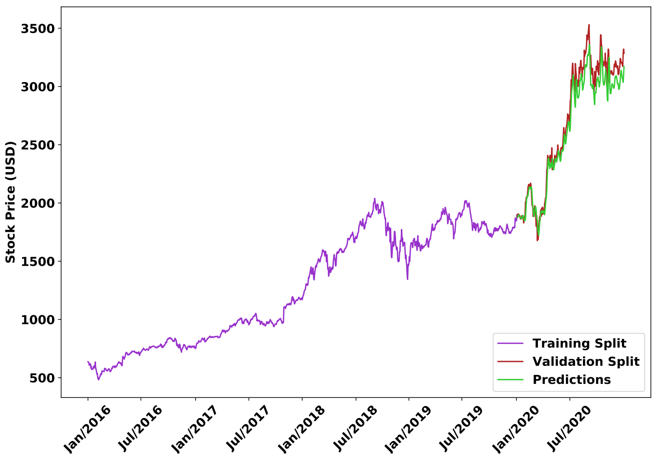

Fig. 4: Model Validation for AMZN

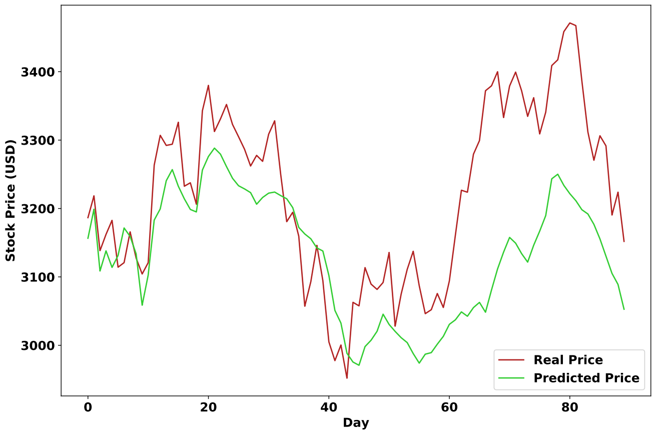

Fig. 5: Future Predictions for AMZN

accuracy as the time passes. As the time passes, the model has to predict the price using more data coming straight from predictions and less real data. As soon as the accuracy starts to decrease, the model will start to use poorly predicted values for the future predictions, resulting in even less accurate prices. To identify the threshold of when the predictions start to lose accuracy, the series of 90 days are divided into splits of 10 days and the RMSE of each period is calculated. The results are shown in Figure 6.

The accuracy decreases significantly after the $6th$ 10-day interval split, interval which the stock has a steep increase on it's price, but also the interval that represents the end of real data usage, since the window size used was 60. Both factors culminated on the quick decay of the model's accuracy, resulting in bad predictions for the intervals 7, 8 and 9.

In the same manner, the RMSE for the intervals is calculated for the remaining assets. The results are exposed on Table IV. The highlighted intervals demonstrate the point of steep accuracy decrease, which revolves around the $7th$ interval, excluding the tickers FB and WMT. This means that the predictions start to become unreliable after the day 60 due to the predictions' quality loss. On a real investment scenario, inaccurate predictions are useless, therefore the model's predictions are useful only up to day 60.

the ability for the model to make predictions farther and farther into the future

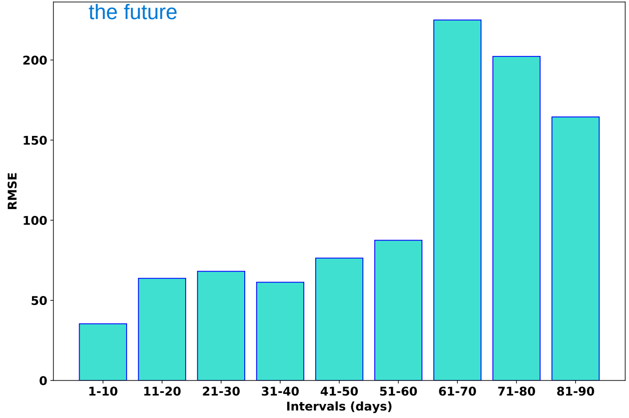

Fig. 6: RMSE for the 10-day Intervals

VI. Conclusions and Future Work

Stock price prediction is a very ambitious and complicated task, mostly because of the many different factors that affect a stock price. This paper presents a strategy to aggregate one of these major factors: the news associated to the company. We proposed the sentiment analysis of news collected from the New York Times using the VADER framework and used the results as one of the features, along with the historical data of the stocks, of a stock price prediction model based on the LSTM architecture.

The model was validated using the assets that make up the Top 10 of the S&P 500 index, on a period of 5 years. The results shows that the model has potential to predict with good accuracy the stock prices within a time window of 50 days into the future, which has a lot of relevancy for any kind of trading strategy.

Even after a display of good results, there is a lot of room for improvement. The model is working with a somewhat limited dataset. Hence, all the datasets can be improved to feed a longer time span into the model. With more data points, the model could explore different window sizes and attempt to predict into a more distant future with good accuracy. The model could also be focused to day trading strategies, in which the investors usually give more value to short term predictions.

Lastly, the scenario which the model worked with, just 10 assets, is theoretical. A more realistic scenario would have a broader selection of assets, fact that implicates in more data to be processed and, in theory, the necessity of a more robust model architecture, with a good parameter tuning strategy that could be supported by the use of an Auto-Machine Learning framework. Subjecting the model to a bigger universe of possibilities would definitely evidence increase on the model's complexity, but also would have potential to create even more forecasting "distance"

TABLE IV: Calculated RMSE for all Tickers

	Intervals								
Ticker	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90
AAPL	1.20	2.23	2.51	2.54	3.13	3.78	**10.95**	11.55	14.59
AMZN	35.388	63.738	68.123	61.294	76.354	87.489	**224.944**	202.243	164.474
BRK-B	2.43	4.38	4.68	4.21	5.25	6.02	**15.48**	13.92	14.32
FB	2.87	5.17	5.32	4.95	6.15	**15.76**	18.90	19.40	21.34
GOOG	17.47	31.11	35.93	32.91	43.74	58.58	**122.78**	115.26	109.60
JNJ	1.73	2.83	3.14	3.82	3.31	6.04	**9.36**	9.32	10.58
MSFT	2.46	5.21	6.50	5.05	6.76	8.78	**19.87**	18.37	26.87
PG	1.95	2.67	2.88	2.60	3.23	3.61	**9.29**	10.35	9.79
V	2.22	4.20	4.66	3.91	4.55	6.63	**16.47**	15.01	16.58
WMT	1.54	2.72	2.93	3.62	3.26	4.74	5.62	**8.51**	7.53

accurate predictions. Using a more realistic scenario would also open up the possibility of unbiased comparison of the model with state-of-the-art technology.

References

[1] G. Hanaoka, Seleção de carteiras de investimentos através da otimização de modelos restritos multiobjetivos utilizando algoritmos evolutivos, Programa de Mestrado em Modelagem Matemática e Computacional (2014).

[2] E. F. Fama, Efficient capital markets: Ii, The journal of finance 46 (5) (1991) 1575–1617.

[3] K. Liagkouras, K. Metaxiotis, Efficient portfolio construction with the use of multiobjective evolutionary algorithms: Best practices and performance metrics, International Journal of Information Technology & Decision Making 14 (03) (2015) 535–564.

[4] M. Roondiwala, H. Patel, S. Varma, Predicting stock prices using lstm, International Journal of Science and Research (IJSR) 6 (4) (2017) 1754–1756.

[5] B. S. Blichfeldt, P. Eskerod, Project portfolio management–there's more to it than what management enacts, International Journal of Project Management 26 (4) (2008) 357–365.

[6] Y. Xu, S. B. Cohen, Stock movement prediction from tweets and historical prices, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1970–1979.

[7] Q. Song, A. Liu, S. Y. Yang, Stock portfolio selection using learning-to-rank algorithms with news sentiment, Neurocomputing 264 (2017) 20–28.

[8] F. Xing, D. H. Hoang, D.-V. Vo, High-frequency news sentiment and its application to forex market prediction, in: Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS), 2020.

[9] D. Valle-Cruz, V. Fernandez-Cortez, A. López-Chau, R. Sandoval-Almazán, Does twitter affect stock market decisions? financial sentiment analysis during pandemics: A comparative study of the h1n1 and the covid-19 periods, Cognitive Computation (2021) 1–16.

[10] F. Cheng, C. Chiao, C. Wang, Z. Fang, S. Yao, Does retail investor attention improve stock liquidity? a dynamic perspective, Economic Modelling 94 (2021) 170–183.

[11] M. Beketov, K. Lehmann, M. Wittke, Robo advisors: quantitative methods inside the robots, Journal of Asset Management 19 (6) (2018) 363–370.

[12] E. Fama, The distribution of the daily differences of the logarithms of stock prices, PhD diss., University of Chicago. Reprinted in the Journal of Business as "The Behavior of Stock Market Prices 38 (1) (1964) 34–105.

[13] R. Schumaker, H. Chen, Textual analysis of stock market prediction using financial news articles, AMCIS 2006 Proceedings (2006) 185.

[14] N. Naik, B. R. Mohan, Optimal feature selection of technical indicator and stock prediction using machine learning technique, in: International Conference on Emerging Technologies in Computer Engineering, Springer, 2019, pp. 261–268.

[15] Y.-c. Chan, A. C. Chui, C. C. Kwok, The impact of salient political and economic news on the trading activity, Pacific-Basin Finance Journal 9 (3) (2001) 195–217.

[16] A. B. Pawar, M. Jawale, D. Kyatanavar, Fundamentals of sentiment analysis: concepts and methodology, in: Sentiment analysis and ontology engineering, Springer, 2016, pp. 25–48.

[17] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[18] F. A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: Continual prediction with lstm, Neural computation 12 (10) (2000) 2451–2471.

[19] K. Smagulova, A. P. James, A survey on lstm memristive neural network architectures and applications, The European Physical Journal Special Topics 228 (10) (2019) 2313–2324.

[20] K. Adam, A. Marcet, J. P. Nicolini, Stock market volatility and learning, The Journal of Finance 71 (1) (2016) 33–82.

[21] X. Du, K. Tanaka-Ishii, Stock embeddings acquired from news articles and price history, and an application to portfolio optimization, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 3353–3363.

[22] H. Markowitz, Portfolio selection, The journal of finance 7 (1) (1952) 77–91.

[23] G. G. Creamer, Can a corporate network and news sentiment improve portfolio optimization using the black–litterman model?, Quantitative Finance 15 (8) (2015) 1405–1416.

[24] F. Black, R. Litterman, Asset allocation: combining investor views with market equilibrium, Goldman Sachs Fixed Income Research 115 (1990).

[25] T. H. Nguyen, K. Shirai, J. Velcin, Sentiment analysis on social media for stock movement prediction, Expert Systems with Applications 42 (24) (2015) 9603–9611.

[26] A. Mittal, A. Goel, Stock prediction using twitter sentiment analysis, Standford University, CS229 (2011) 15 (2012).

[27] M. V. Beraldi, Robôs de investimento a partir de dados de redes sociais, Ph.D. thesis (2020).

[28] M. A. Faizan, Multiobjective portfolio optimization including sentiment analysis (2019).

[29] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 8, 2014.

[30] J. Schmidhuber, Learning complex, extended sequences using the principle of history compression, Neural Computation 4 (2) (1992) 234–242.

[31] R. Pascanu, C. Gulcehre, K. Cho, Y. Bengio, How to construct deep recurrent neural networks, arXiv preprint arXiv:1312.6026 (2013).

[32] W. Rodi, N. Fueyo, Engineering turbulence modelling and experiments 5, Elsevier, 2002.