A comparative analysis between machine learning models applied to stock market forecasting

Guilherme Coelho Minervino
Departamento de Ciência da Computação (CIC)
Universidade de Brasília
Brasília, DF
guiminervino@gmail.com

***Abstract*—The stock market is extremely important for the economy of a country, with billions of dollars flowing daily. In this area, the task of forecast a stock price turned to be an important issue and very discussed on literature. It is necessary to have contact with a massive quantity of noisy data. Besides, a stock price pattern is very dynamic and non linear. To deal with that, two types of analysis is presented on literature: fundamental and technical. The proposed study aims to make a fundamental analysis using news headlines to forecast the behavior of the stock market price Dow Jones Industrial Average (DJIA). In order to accomplish that task, two types of models will be used: traditional supervised machine learning and deep learning.**

***Index Terms*—Dow Jones, stock market, NLP, traditional machine learning, fundamental analysis, deep learning**

I. Introduction

Financial markets are fundamental for the thriving of a country. It is by this type of market that companies receive investment in exchange for a part of themselves, contributing to the financial growth of both the company and the investor. But, because of the influence of factors like the global economy, politics, policies, disasters, wars and global diseases, the task of forecast a stock price behavior is not trivial. Many works have been made in the area of machine learning (ML) and Natural Language Processing (NLP) to deal with that problem. As said, there are two approaches in the literature to analyze that task: fundamental and technical analysis [2].

The fundamental analysis estimates the stock price based on factors like news, social media, politics, internal and external management of the company and its profile. Technical analysis uses only the historical price of a stock to predict its future behavior and it tries to find patterns and trends that repeat. In this research, only the fundamental analysis will be used news headlines as a resource.

Some studies used the traditional machine learning method applying Support Vector Machine (SVM) for the prediction of stock behaviors [10,13]. Moreover, some studies used ANNs [11,12]. The DL model LSTM is widely used too [8,10]. In addition, some used BERT, the state-of-art language model tool [8], to recognize sentiments [8,9,10]. Based on [10] as inspiration, the present study proposes a comparative analysis between different models. The novel is the addition of the models KNN, Logistic Regression and Random Forest to the analysis and the study of a different domain: the stock market.

The reminder of this article is organized as follows: Section II shows related works, Section III shows the methodology to accomplish the task, Section IV describes the experimental results and finally, Section V is the conclusion.

II. Related Works

There are several sentimental analysis works in literature to predict stock market price behavior. Despite the progress made, so far was not found a comparative analysis of the models mentioned with the task to predict the Dow Jones Industrial Average Index (DJIA).

Starting with the lexical approach, [5] uses pre-build open source libraries like TextBlob and VADER to recognize sentiments on tweets. Their results show that the use lexical approach (unsupervised learning) alone is not accurate. They suggest that a supervised model complements the analysis for better accuracy.

Furthermore, [7] proposed a traditional machine learning method using SVM and Naive Bayes to deal with the stochastic nature of stock market data and its forecasting. Using texts from news and articles, this paper proposes a comparative analysis between Naive Bayes, KNN and SVM for prediction. The results show an accuracy range between 75% to 91.2%, 65.30% to 83.80%, 74.70% to 85% for KNN, SVM and Naive Bayes, respectively. Thus, the KNN model was the most accurate for this analysis.

Moreover, [11] proposes a method that compares different sources of information, using news and tweet written in Portuguese, and evaluates it based on the movement of the Brazilian stock market Ibovespa (BVSP) between January and March using ANNs. The model used was MLP and the results show that sentiment from news has more impact than tweets when forecasting open price stock on the next day. On the other hand, sentiments from tweets have more impact than news when forecasting the closing price of a stock.

In the Deep Learning (DL) area, [8] proposed a hybrid method composed of LSTM and BERT to forecast stock prices. The state-of-the-art natural language model tool was used to recognize text sentiments and the long short-term memory neural network was used to analyze time series data from historical transactions. Their proposed method had a 12.05 accuracy improvement on the root mean square error metric (RMSE). The given task was performed on a collection of texts from the PTT forum and news articles. Was discovered that these texts can reduce the RMSE, proving that sentiments from news and forums play an important role in the stock market forecast.

Information about the models and dataset that will be used is described below.

A. Traditional Machine Learning

Machine learning is a branch of Artificial Intelligence and is very useful in the task of forecast a stock market price. Train, validation and test data can be used. The problem can be classification or regression. In this research, classification will be used to categorize sentence polarity as positive, neutral or negative. Because of that, it is very important that the training data is labeled. After an analysis in literature, its known that the best models for that type of task is SVM, KNN and Naïve Bayes [3,4,7]. Thus, they will be used to represent this category's performance.

B. Deep Learning

Is a subset of machine learning and incorporates computational models and algorithms that imitate the architecture of the biological neural networks in the brain. 'Deep' is a technical term, and refers to the number of layers in an ANN. A deep ANN differs from a superficial ANN by having a large number of hidden layers. Therefore, it is able to perform much more complex tasks [1]. This study will analyze two types of DL models: BERT and LSTM. The reason for that decision is that they are the most popular models for SA in respect of DL.

C. Dataset

This study will use a public dataset from Kaggle that contains News Headlines related to Dow Jones Industrial Average Index [14]. Besides the news column, Each row has the information if the stock price goes up (1) or down (0), serving as a label for the analysis. The decision to choose this dataset is that the text data is already in tabular form, making the use of crawlers not necessary for this research.

III. Methodology

For this study, the models Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Long-Short Term Memory (LSTM) was used in order to find the best traditional model to the task of forecast the behavior of DJIA stock price.

The proposed method 1 has the following steps:

- Data splitting
- Pre-processing
- Bag of words
- Training and Evaluation

A. Data splitting

Firstly, the dataset is split into training and testing. The best split found was 90% for train and 10% for test. At first, samples were chosen randomically using a function from sklearn library. But, because of bad metric results, train data was taken based on news with data below 2015/01/01 (3972 samples) and test data was taken with data above 2014/12/31 (378 samples)

B. Pre-processing

Due to its unstructured nature, the text data is processed before being put into the training model. The pre-process steps are:

- Concatenation of all news headlines columns into only one for each row
- Regular expression to allow only lower and camel case letters
- Convert to lower case
- Stop word removal
- Tokenization
- Stemming

TABLE I
CountVectorizer Bag of words example

	art	boat	soccer	england
0	2	0	1	1
1	0	0	1	0
2	1	4	0	1

C. Bag of words

Due to the fact that only numbers is allowed to models input, is necessary to convert all the text data to a numerical form. For that purpose, the libraries CountVectorizer and TFIDF from nltk is utilized.

1) CountVectorizer: the result is a table in which each column is a word and each row is a group of headlines of a specific date. In this study, was decided to use a contiguous sequence of n items from a given sample of text (ngrams) equal to one because showed more interesting results. Table 1 illustrates an example of the method.

2) TF-IDF: stands for term frequency-inverse document frequency. It is a method used to represent a corpus numerically. Unlike the CountVectorizer, in which the occurrence of a word increments the value of the column of the specific row, the TF-IDF value of each cell is the result of the following equation:

$$TF - IDF = TF(t, d) * IDF(t, d) \tag{1}$$

Where TF(t,d) is the number of times that a term appears in a document and IDF(t,D) is the inverse of the frequency that a term appears in all documents of the corpus. Finally, TF-IDF varies from zero to one. Zero indicates that the term is absent and one indicates that the term is extremally frequent on a given document and absent on the rest of the corpus. Similarly to CountVectorizer, was decided to ngram = 1 due to more satisfactory results.

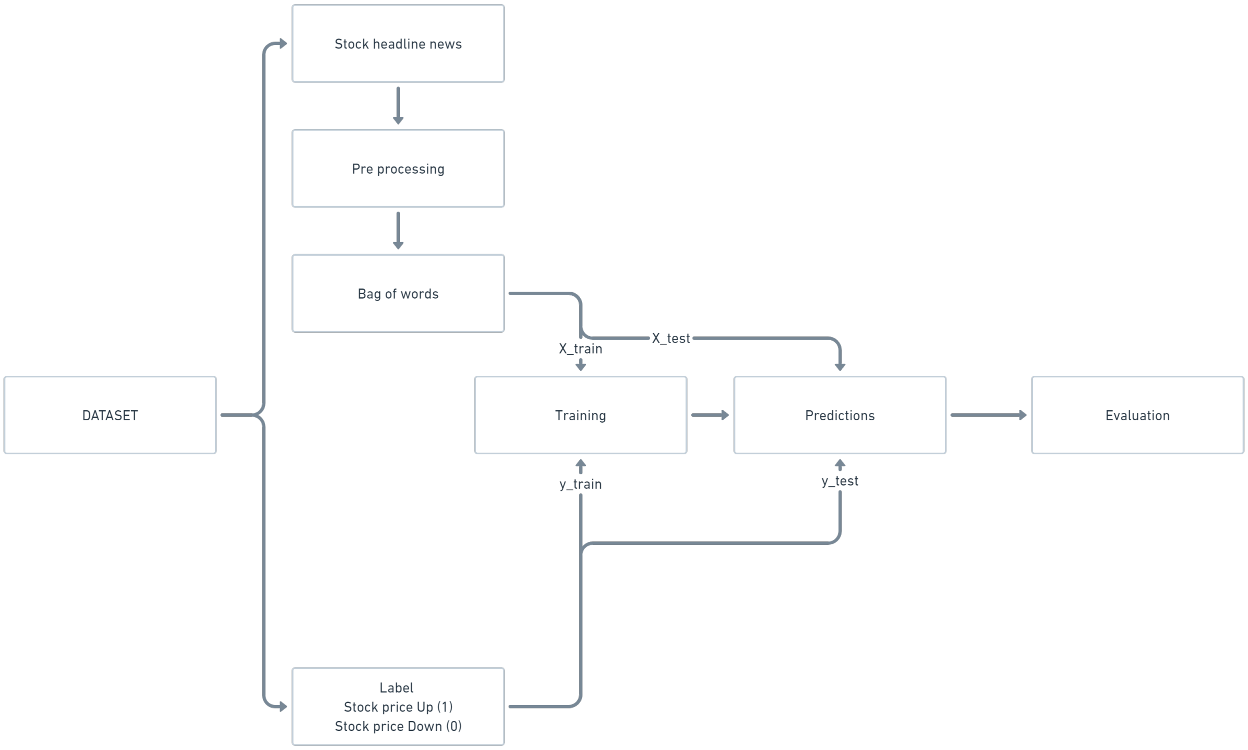

Fig. 1. Fluxo de execução da metodologia proposta

TABLE II
RESULTS WITH COUNTVECTORIZER BAG OF WORDS

	NB	LR	RF	DT	SVM	KNN
Accuracy	84.92%	83.60%	83.86%	84.13%	82.80%	52.91%
Precision	81.10%	82.50%	80.75	84.38%	82.90%	51.90%
Recall	91.67%	85.94%	89.58%	84.38%	83.33%	99.48%
F1-score	86%	84.18%	84.94%	84.38%	83.12%	68.21%

TABLE III
RESULTS WITH TF-IDF BAG OF WORDS

	NB	LR	RF	DT	SVM	KNN
Accuracy	52.12%	81.75%	83.86%	84.13%	82.8%	52.91%
Precision	51.47%	76.39%	81.78%	83.07%	77.42%	62.50%
Recall	100%	92.71%	91.15%	81.77%	87.50%	70.31%
F1-score	67.96%	83.76%	86.21%	82.41%	82.15%	66.18%

D. Training and Evaluation

After pre-processing, the data is ready to be input in order to train and evaluate the models. The metrics Accuracy, Precision, Recall and F1-score are used for the evaluation.

IV. EXPERIMENTAL RESULTS

Throughout the experiment, many parameters have been changed.

In case of the traditional machine learning, was the percentage of split for training and test, type of bag of words (bow), number of words of bow vocabulary, ngram size and types of pre-processing

On deep learning, changes in preprocessing, types and numbers of layers and neurons, epochs, padding maximum length and loss type were made.

It is important to emphasize that the importance of predicting correctly is the same for both label cases: stock price goes up (1) or prices go down (0). Thus, accuracy will be utilized to decide which model is the best.

Based on Table 2, it can be seen that Naive Bayes was the best model when the bag of words of type CountVectorizer was utilized and KNN was the worse one.

On Table 3, it can be seen that Decision Tree was the best model when the bag of words type TF-IDF was utilized and Naive Bayes was the worst one.

TABLE IV
RESULTS USING DEEP LEARNING MODEL LSTM

	LSTM	BiLSTM
Loss	0.6919	0.6926
Accuracy	0.5366	0.5366
Val Loss	0.6924	0.6925
Val Accuracy	0.5226	0.5226

Finally, On Table 4, despite the distinct changes to hyperparameters, the Loss stayed very high and Accuracy very low. Thus, LSTM was not satisfactory to that task. Despite that, more attempts is needed to claim that this kind of model is inefficient.

Surprisingly, Naive Bayes was the best and the worse on the bag of words CountVectorizer and TF-IDF, respectively.

V. CONCLUSION AND FUTURE WORKS

In the development of this study, it was clear that news headlines have an impact on the price of a stock. This work proposed a comparative analysis of machine learning models to find the one that best suits the given task. Naive Bayes and KNN was the best models to forecast the behavior of DJIA stock price using news headlines. The deep learning model LSTM did not show interesting results, making it necessary a more elaborate study of the hyperparameters and the data.

As future works, is intended to improve the proposed deep learning LSTM model and attempt to use other successful models in the literature, like BERT and CNN. Finally, is intended to use sentiment analysis on news, social media or forums in order to achieve better results.

REFERENCES

[1] Jakhar, D., Kaur, I. (2019). Artificial intelligence, machine learning deep learning: Definitions and differences. Clinical and Experimental Dermatology. doi:10.1111/ced.14029

[2] Nti, I.K., Adekoya, A.F. Weyori, B.A. A systematic review of fundamental and technical analysis of stock market predictions. Artif Intell Rev 53, 3007–3057 (2020). https://doi.org/10.1007/s10462-019-09754-z

[3] Zulfadzli Drus, Haliyana Khalid, Sentiment Analysis in Social Media and Its Application: Systematic Literature Review, Procedia Computer Science, Volume 161, 2019,

[4] Shiliang Sun, Chen Luo, and Junyu Chen. 2017. A review of natural language processing techniques for opinion mining systems. Information Fusion 36 (2017), 10–25.

[5] Zahoor, S., Rohilla, R. (2020). Twitter Sentiment Analysis Using Lexical or Rule Based Approach: A Case Study. 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). doi:10.1109/icrito48877.2020.9197910

[6] Wongkar, M., Angdresey, A. (2019). Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter. 2019 Fourth International Conference on Informatics and Computing (ICIC). doi:10.1109/icic47613.2019.8985884

[7] Kalra, S., Prasad, J. S. (2019). Efficacy of News Sentiment for Stock Market Prediction. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). doi:10.1109/comitcon.2019.8862265

[8] Ko CR, Chang HT. LSTM-based sentiment analysis for stock price forecast. PeerJ Comput Sci. 2021 Mar 11;7:e408. doi: 10.7717/peerj-cs.408. PMID: 33817050; PMCID: PMC7959635.

[9] Sousa, M. G., Sakiyama, K., Rodrigues, L. de S., Moraes, P. H., Fernandes, E. R., Matsubara, E. T. (2019). BERT for Stock Market Sentiment Analysis. 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)

[10] Alaparthi, S., Mishra, M. BERT: a sentiment analysis odyssey. J Market Anal 9, 118–126 (2021). https://doi.org/10.1057/s41270-021-00109-8

[11] Arthur E. de O. Carosia, Guilherme P. Coelho, and Ana E. A. da Silva. 2019. The influence of tweets and news on the brazilian stock market through sentiment analysis. In Proceedings of the 25th Brazillian Symposium on Multimedia and the Web (WebMedia '19). Association for Computing Machinery, New York, NY, USA, 385–392. https://doi.org/10.1145/3323503.3349564

[12] Nti, Isaac Adekoya, Adebayo Weyori, Benjamin. (2020). Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence From Ghana. Applied Computer Systems. 25. 33-42. 10.2478/acss-2020-0004.

[13] Mahtab, S. Arafin, N. Islam, and M. Mahfuzur Rahaman. (2018, 21-22 Sept. 2018). "Sentiment Analysis on Bangladesh Cricket with Support Vector Machine", in the 2018 International Conference on Bangla Speech and Language Processing (ICBSLP).

[14] https://www.kaggle.com/datasets/lykin22/stock-headlines?select=Stock+Headlines.csv

[15] https://www.kaggle.com/datasets/mnassrib/dow-jones-industrial-average