Predição de valores das 100 maiores empresas da Nasdaq baseado em títulos de notícias extraídos do twitter e preços das ações

A Hayek, Bruno Elie. Zhu, Xuan.

Link do repositório:

https://github.com/BrunoEAH/Tech-Stock-Market-Analysis

Faculdade de Computação e Informática
Universidade Presbiteriana Mackenzie (UPM) – São Paulo, SP – Brazil

brunoelie.hayek@mackenzista.com.br, xuan.zhu@mackenzista.com.br,

Abstract. *The article concerns a description of a model that intends to predict the price prediction of the stocks from the 100 biggest companies listed on the Nasdaq, based on the stock prices and the sentiment analysis from financial news headlines scrapped from Twitter. For the sentiment analysis a pre-trained model named FinBERT will be used, which will analyse a dataset containing the tweets headlines. After that , there will be a Deep Learning model, using the Keras library, that will make the stock price prediction, based on a dataset containing 10 years of data from those companies. Finally, both part will be unified in order to have a model that can estimate stock market prices from these companies.*

Resumo. *O artigo em questão trata sobre a descrição de um modelo que faz a predição dos valores das 100 maiores empresas da Nasdaq, baseado no preço das ações e na análise de sentimento dos títulos de notícia extraídos do twitter. Para a análise de sentimento será empregado o modelo pré-treinado FinBERT, o qual analisará o dataset contendo os tweets com as headlines das notícias. Posteriormente será feito um modelo com base em Deep Learning, utilizando a biblioteca Keras, para a predição dos preços das ações dessas 100 maiores empresas, com base no dataset contendo mais de 10 anos sobre informações de preços de ações dessas empresas. Ao final ambas as partes serão integradas para fornecer uma estimativa sobre os preços das ações dessas empresas.*

1. Introdução

1.1. Contextualização

Em um mundo globalizado e com grande fluxo de informações, muitos indivíduos começaram a se interessar na área de investimento, estudando assim empresas de diversos setores e tamanhos para investir seu dinheiro, na expectativa de um retorno financeiro.

Um dos mercados mais importantes atualmente é o de tecnologia, caracterizado por sua grande fluidez e constante evolução, mostrando que novas tecnologias podem prosperar rapidamente, estimulando assim um grande processo de pesquisa e

desenvolvimento para acompanhar mudanças no mercado por parte dos investidores.

Uma peça chave para o mercado financeiro na área da tecnologia é a Nasdaq, acrônimo para Associação Nacional de Corretores de Títulos de Cotações Automáticas, um mercado de ações norte-americano com mais de 3000 ações de diferentes empresas, todas incluídas no índice Nasdaq *Composite* (ROBINHOOD,2025). As maiores empresas de tecnologia do mundo se encontram na Nasdaq, como a Apple, Amazon, Microsoft, T-Mobile, Alphabet, Tesla e Meta. (NASDAQ, 2021)

No mundo de mercado financeiro muitas notícias podem influenciar o sentimento dos investidores, afetando diretamente no preço das ações e como consequência, no lucro e rendimento das grandes empresas. O sentimento de mercado, como assim é denominado, tende muitas vezes a orientar o mercado, por isso é especulada e explorada por analistas de mercado (IMF, 2019).

1.2. Justificativa

Com esse grande volume de indivíduos com interesse em estudar e investir no mercado financeiro, surgiram *startups* e propostas que visam auxiliar iniciantes a começarem no mercado de ações, sendo um exemplo a corretora americana *Robinhood*, que ensina novos investidores à como aplicar o dinheiro no mercado de ações, dando oportunidades de grandes empresas a pequenos investidores (ROBINHOOD,2025).

A principal justificativa é o fato do mercado de ações de tecnologia ser diversificado com várias empresas diferentes que atuam em ramos distintos desse campo, causando muitas vezes um excesso de informações que pode ser intimidador aos investidores.

1.3. Objetivo

O propósito central do projeto é fornecer uma visão baseada em dados sobre o futuro de empresas do mercado de tecnologia, para investidores com prática analisarem e investirem e para iniciantes do mercado financeiro aprenderem como o mercado tecnológico funciona.

1.4. Opção do problema

A opção para a resolução do problema é a de um *framework* , em que são empregadas técnicas do *Deep Learning* integradas a um modelo pré-treinado denominado de FinBERT. Os *frameworks* para o *Deep Learning* são a biblioteca Keras e Tensorflow, ambas empregadas. O FinBERT é um modelo pré-treinado em textos financeiros para a análise de sentimento em meios de comunicação sobre mercado financeiro, enquanto que os outros *frameworks* serão usados para a medição.

2. Descrição do problema

O mercado de tecnologia, representado por empresas listadas na Nasdaq,

enfrenta desafios significativos devido ao grande volume de dados gerados diariamente. Notícias, relatórios financeiros, análises de especialistas, redes sociais e eventos globais produzem um fluxo contínuo de informações que impactam diretamente o sentimento dos investidores e, consequentemente, os preços das ações. Entretanto, processar e interpretar esse volume massivo de dados de forma manual ou com ferramentas tradicionais torna-se inviável devido à complexidade e velocidade das informações.

3. Uso ético da Inteligência Artificial e responsabilidades

Durante o desenvolvimento desse modelo foram levados em conta diversos fatores a respeito do uso ético da inteligência artificial e suas responsabilidades, sendo a integridade dos dados um ponto de extrema importância. Na elaboração do projeto foram utilizados dados com grande confiabilidade e integridade, de fontes seguras e confiáveis, fazendo com que as predições possam respeitar e se assemelhar às regras do mercado financeiro.

No aspecto de responsabilidade pode-se afirmar que o intuito deste projeto não é de fornecer com certeza absoluta em qual ação um determinado investidor deve arriscar seu dinheiro, mas sim dar uma perspectiva baseada em tecnologia e dados que pode auxiliar o investidor a ter mais percepções, auxiliando também um iniciante a entender como o mercado de ação funciona e como as notícias podem afetá-lo.

4. Dataset

Durante a elaboração do projeto, foram usados dois *datasets*, um que contém os dados a respeito das ações das 100 maiores empresas da Nasdaq [1] e outro contendo *tweets* com os títulos das notícias do mercado financeiro e um sentimento atribuído a elas[2].

O *dataset* sobre as empresas da Nasdaq contém 8 colunas: *Date, Low, Close, Open, High, Volume, Adj_Close* e *Name*. As colunas *Date* e *Name* representam a data e nome, respectivamente, de cada ação. A coluna *High* se refere ao preço máximo durante o dia e a coluna *Low* se refere ao preço mínimo atingido durante o mesmo dia. Já as colunas *Open* e *Close* se referem aos preços em que a ação começou e terminou (SMIGEL,2023). A coluna *Volume* se refere à quantidade de ações movimentadas durante o dia (HAYES, 2023) e por fim, a coluna *Adj_Close* trata sobre um cálculo feito sobre o valor de fechamento, levando em conta diversos fatores da empresa (GANTI, 2020).

No segundo *dataset* constam duas colunas: *text* e *label*. A coluna *text* contém os títulos das notícias e a coluna *label* possui três números que simbolizam os sentimentos do mercado financeiro em relação à notícias, as quais são: 0 que significa *Bearish*, 1 que simboliza *Bullish* e 2 que significa *Neutral.* O sentido de *Bearish* é negativo, o qual costuma prever uma queda dos preços das ações, enquanto que *Bullish* tem uma conotação positiva, indicando um aumento nos preços das ações. Por fim, *Neutra* tem a

[1] https://www.kaggle.com/datasets/kalilurrahman/nasdaq100-stock-price-data
[2] https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment

definição de um sentimento neutro (IMF, 2019).

Primeiramente foi aplicado o pré-processamento dos *datasets*, em que linhas duplicadas e sem valores foram removidas. Após isso, colunas que não fornecem informações significativas foram removidas.

No caso do *dataset* contendo os títulos de notícias do mercado financeiro retiradas do twitter, foi usado um tipo de pré-processamento específico, em que foram removidos URL's, menções, *hashtags*, pontuação, caracteres especiais e *stopwords*. Posteriormente o texto foi deixado em letras minúsculas e foi feito o processo de tokenização, usando o *tokenizer* do FinBERT, que ao final converteu o texto em *tokens ID* (RAJKUMAR,2019).

O *dataset* contendo os dados sobre as 100 maiores empresas foi modificado de outra maneira, em que foi inserida uma nova coluna com o SMA (*Simple Moving Average*), uma média aritmética que estipula quanto foi movimento a cada 15 dias, em que os valores são do *Adjusted Close* ou fechamento ajustado do dia (HAYES,2023). Além disso também foi criada a coluna de retorno, em que se enquadra o retorno diário, sendo feita uma comparação entre o valor do dia anterior com o do atual, e dando um percentual da diferença desse valor. Foi adicionada a coluna com RSI também, *Relative Strength Index*, que mede a magnitude da ação (FERNANDO, 2024).

Ao *dataset* da Nasdaq também foi adicionada a coluna que contém a flutuação diária, uma operação de diferença entre o preço máximo e o preço mínimo do dia. Posteriormente, foi feita a normalização dos dados, em que foi empregado o método de Mínimo e Máximo, o qual normaliza os dados em um intervalo de 0 a 1, e utiliza a fórmula abaixo (RANGI,2025):

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Durante a análise exploratória de dados, foram elaborados oito gráficos que contribuíram para o entendimento do *dataset* e as suas respectivas relações às variáveis. O primeiro gráfico foi criado como um método de verificação da normalização feita durante o pré-processamento, para comparar se tanto os valores não escalados como os escalados estavam com a mesma proporção.

Os próximos gráficos são matrizes de multicolinearidade das variáveis do *dataset*, em que foram empregados três métodos de correlação: Pearson um método linear, Kendall e Spearman, ambos métodos não-lineares. Pearson é um método para verificar a relação linear entre as variáveis, Kendall mensura a dependência entre as variáveis e Spearman avalia a associação entre as variáveis. A partir das matrizes criadas, conseguimos interpretar que há uma grande correlação entre as variáveis, porém entre outras não há relação (STATISTICS SOLUTIONS, 2025).

Posteriormente foi feito um gráfico que selecionou as 10 maiores empresas pelo retorno cumulativo de 11 anos, um cálculo que leva em conta quanto um investidor teria ganhado e perdido ao longo desses 11 anos, pelo valor do fechamento ajustado normalizado. Percebe-se que as empresas AMZN, GOOG e NFLX possuíram o maior crescimento em comparação com as outras empresas.

Foi feito também um gráfico do tipo *boxplot*, em que foram selecionadas as 10 empresas com maior média de flutuação diária, assim no gráfico foram plotados os valores da flutuação diária. Apesar das *boxes* abrangerem um intervalo grande valor, foram observados vários valores que são caracterizados como *outliers*, os quais em vários casos podem ocorrer por uma queda brusca das ações devido a notícias e medidas financeiras de diversos países.

Outro gráfico elaborado foi o de *scatterplot* de todas as empresas, levando em conta a média do volume diário e a média da flutuação dos preços. Neste gráfico foi possível concluir que no geral, tanto empresas que possuem uma grande média do volume diário quanto que possuem pequena média, elas não possuem uma média muito grande de flutuação de preços, com exceção de poucas empresas.

O gráfico de histograma foi feito para visualizar as empresas com base em quatro intervalos de valores do volume diário médio: *Low, Medium-Low, Medium-High* e *High*. Posteriormente foi feito uma contagem e plotagem como histograma, em que é visível que há uma distribuição quase que uniforme das empresas por categoria de volume.

Por fim, foi feito um gráfico do tipo *violinplot* em que se dividiu novamente as empresas em quatro categorias tomando como base o volume diário médio, plotando junto a isso os preços dos fechamentos ajustados. É perceptível que quase todos esse violinos possuem a mesma densidade para valores semelhantes de fechamentos ajustados, porém há categorias como *Medium-Low* e *Medium-high* que possuem valores de densidade pequena em preços de fechamentos ajustados altos, dando assim características de empresa *outliers*.

5. Metodologia e resultados esperados

Durante este projeto, será empregado o modelo pré-treinado FinBERT[3], que junto ao *dataset* contendo os tweets será direcionado a classificar tais títulos de notícias. Posteriormente, será usada uma API do *Twitter*[4] para receber os títulos das notícias e assim o FinBERT classifica o título com base nas três categorias de *Bearish, Bullish e Neutral*.

Com relação aos dados da Nasdaq, será usado técnicas de *Deep Learning* utilizando a biblioteca Keras e Sklearn do python para a predição dos valores, assim ao final dessas técnicas a API do *Yahoo Finance* [5]será usada para receber os dados de uma ação específica e com base nisso a inteligência artificial será capaz de criar uma

[3] https://huggingface.co/ProsusAI/finbert
[4] https://developer.x.com/en/docs/x-api

[5] https://python-yahoofinance.readthedocs.io/en/latest/api.html

estipulação a respeito do preço.

Como resultado, é esperado conseguir integrar ambas as partes de predições de preços junto à parte de análise de sentimento, para assim ao modelo receber as informações a respeito das notícias e dos dados das ações, ele conseguir realizar uma estimativa da predição dos preços da empresa selecionada.

6. Referências e Bibliografia

SMIGEL, Leo. What Is Open High Low Close in Stocks?. **Analyzing Alpha**, 13 de out. 2023. Disponível em: https://analyzingalpha.com/open-high-low-close-stocks Último acesso em: 7 de abr 2025.

BAKER, Brian. *What is the Nasdaq Composite?*. **Bankrate**, 28 jan. 2025. Disponível em: https://www.bankrate.com/investing/what-is-nasdaq-composite/. Acesso em: 7 abr. 2025.

CHEN, James. *What Does the Nasdaq Composite Index Measure?*. **Investopedia**, 2 ago. 2023. Disponível em: https://www.investopedia.com/terms/n/nasdaqcompositeindex.asp. Acesso em: 7 abr. 2025.

INTERNATIONAL MONETARY FUND. *The Power of Text: How News Sentiment Moves Markets*. **IMF Blog**, 16 dez. 2019. Disponível em: https://www.imf.org/en/Blogs/Articles/2019/12/16/blog-the-power-of-text. Acesso em: 7 abr. 2025.

NASDAQ. *What is the Nasdaq Composite and What Companies Are In It?*. 12 maio 2021. Disponível em: https://www.nasdaq.com/articles/what-is-the-nasdaq-composite-and-what-companies-are-in-it-2021-05-12. Acesso em: 7 abr. 2025.

ROBINHOOD. *What is the Nasdaq?*. **Robinhood Learn**, 4 jan. 2025. Disponível em: https://robinhood.com/us/en/learn/articles/1GXPrecRfkSNp0aBblBSsL/what-is-the-nasdaq/. Acesso em: 7 abr. 2025.

THE STREET. *What Is the Nasdaq Composite Index? Definition, Sectors....* 15 jan. 2023. Disponível em: https://www.thestreet.com/dictionary/nasdaq-composite. Acesso

em: 7 abr. 2025.

HAYES, Adam. What Is Volume of a Stock, and Why Does It Matter to Investors? . **Investopedia Volume** , 07 de mai.2023. Disponível em: **https://www.investopedia.com/terms/v/volume.asp.** Último acesso em: 7 de abr 2025.

GANTI, Akhilesh. Adjusted Closing Price Definition. **Investopedia**, 28 de dez. 2020. Disponível em: https://www.investopedia.com/terms/a/adjusted_closing_price.asp Último acesso em: 7 de abr 2025.

Market Sentiment. **CFI**. Disponível em: https://corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/market-sentiment/. Último acesso em: 7 de abr 2025.

CHEN, James. What Does Cumulative Return Say About a Stock's Performance? **Investopedia**, 10 de out.2020. Disponível em: **https://www.investopedia.com/terms/c/cumulativereturn.asp**

RANGI, Kamlesh Kumar. How Min-Max Scaler Works. 25 de ago. 2025. Disponível em: https://scribe.rip/@iamkamleshrangi/how-min-max-scaler-works-9fbebb9347da . Acesso em: 7 abr. 2025.

HAYES, Adam. *Simple Moving Average (SMA): What It Is and the Formula*. **Investopedia**, 13 de jun 2023. Disponível em: https://www.investopedia.com/terms/s/sma.asp. Acesso em: 7 abr. 2025.

FERNANDO, Jason. *Relative Strength Index (RSI) Indicator Explained With Formula*. **Investopedia**,19 de nov 2024. Disponível em: https://www.investopedia.com/terms/r/rsi.asp. Acesso em: 7 abr. 2025.

RAJKUMAR, Sudalai. *Getting started with Text Preprocessing,* aug 2019. Disponível em: https://www.kaggle.com/code/sudalairajkumar/getting-started-with-text-preprocessing. Acesso em: 7 abr. 2025.

STATISTICS SOLUTIONS. *Correlation: Pearson, Kendall, Spearman*. Disponível em: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/. Acesso em: 07 abr. 2025.

Datasets:

https://www.kaggle.com/datasets/kalilurrahman/nasdaq100-stock-price-data

https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment

Documentação usada:

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

https://stackoverflow.com/questions/40060842/moving-average-pandas

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rolling.html

https://ta-lib.github.io/ta-lib-python/doc_index.html

https://technical-analysis-library-in-python.readthedocs.io/en/latest/ta.html

https://huggingface.co/ProsusAI/finbert

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html

https://seaborn.pydata.org/

https://matplotlib.org/stable/index.html

https://www.kaggle.com/code/sudalairajkumar/getting-started-with-text-preprocessing#Removal-of-Punctuations

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.qcut.html

https://python-yahoofinance.readthedocs.io/en/latest/api.html

https://developer.x.com/en/docs/x-api