CURSO DE CIÊNCIA DA COMPUTAÇÃO – TCC	
( ) PRÉ-PROJETO (x) PROJETO	ANO/SEMESTRE:2021/1

USO DE REDES NEURAIS ARTIFICIAIS PARA PREVISÃO DO PREÇO DAS AÇÕES NA BOLSA DE VALORES POR MEIO DE NOTÍCIAS

Fabrício Oliveira Bezerra

Prof. Dra. Andreza Sartori – Orientadora

1 INTRODUÇÃO

Segundo Trisotto (2020), o ano de 2020 fechou com 3,2 milhões de pessoas físicas cadastradas como investidoras na bolsa de valores brasileira, somando R$ 424 bilhões em ações. Um dos motivos citados foi a democratização no acesso à informação. Assim, parte dos novos investidores tomam suas decisões de compra ou venda de ações, por intermédio da Internet, por meio de canais de influenciadores digitais, ou por conta própria, a partir da análise de dados e informações de fontes distintas. A volatilidade, que se associava com o afastamento dos pequenos investidores, teve menor impacto a partir de 2020. Trisotto (2020) revela que o presidente da Bolsa de Valores (B3 - Gilson Finkelsztain) ficou surpreso com o volume e a velocidade dos investimentos durante a crise.

No entanto, a entrada desses novos investidores constitui também um desafio. Como assinalam Li et al. (2020), as situações que desencadeiam a volatilidade da alta ou baixa no preço das ações estão relacionadas ao movimento de oferta e demanda, à tendência das empresas como lançamento de algum produto inovador, eventos aleatórios como desastres naturais ou morte de algum líder político. É mencionado também o sentimento do mercado, em que informações relevantes à empresa tornam-se publicamente disponíveis e segundo Pagolu (2016, p. 15), “existe uma forte correlação entre subida e queda em preços de ações de uma empresa em relação às opiniões ou emoções públicas”.

No entanto, para avaliar essas variáveis, que envolvem diferentes meios de comunicação, aspectos psicológicos e subjetivos, unindo dados fundamentais das empresas e notícias do mercado financeiro para previsão do movimento das ações, é necessária a utilização de Inteligência Artificial (IA) para o processamento de dados. Neste sentido, trabalhos recentes como os apresentados por Li et al. (2020), Li e Pan (2020) e Vargas et al. (2018) obtiveram bons resultados na resposta a estes desafios. A estratégia dos autores foi de utilizar técnicas de predição não lineares de Redes Neurais Artificiais (RNA) e de Processamento de Linguagem Natural, também chamado de PLN e NLP (do inglês, Natural Language Processing). As RNA são sistemas de computação com nós interconectados que funcionam como os neurônios do cérebro humano. Usando algoritmos, elas podem reconhecer padrões escondidos e correlações em dados brutos, agrupá-los e classificá-los, e, com o tempo, aprender e melhorar continuamente esses processos. Conforme afirma Ferro (2013), devido a tais características, a utilização de RNA para a criação de modelos preditivos é adequada, pois simulam eventos cotidianos com a ajuda de algoritmos e de sistemas computacionais. O PLN permite que uma máquina entenda o significado de frases ditas e escritas por humanos, seja por texto ou por voz. Nesse cenário, o PLN se apresenta para resolver um problema, unindo áreas linguísticas e técnicas.

Comentado [MH1]: Iniciar um segundo parágrafo com segunda negação deixa o leitor confuso.

Assim, este trabalho propõe a construção de um protótipo com técnicas de predição das RNA e dos PLN, com uma web scraper para obtenção das cotações das ações e para buscar as notícias do mercado financeiro para análise, utilizando modelos e métricas para prever o movimento das ações e assertividade das previsões e disponibilizar os resultados para investidores interessados. Dessa forma, esse trabalho propõe a criação de uma ferramenta que a partir de uma série temporal, encontre padrões e recomende ao investidor qual é o momento mais adequado para comprar ou vender determinada ação, visando maximizar seus lucros.

1.1 OBJETIVOS

Este trabalho tem como objetivo disponibilizar um protótipo usando Redes Neurais Artificiais e Processamento de Linguagem Natural para sugerir a compra ou venda de ativos negociados na Bolsa de Valores brasileira por meio de notícias do mercado financeiro, a fim de auxiliar investidores nas suas transações no mercado de renda variável.

Os objetivos específicos são:

a) disponibilizar um *web scraper* para obtenção das notícias do mercado financeiro e das notícias relacionadas às empresas;

b) consultar API da Bolsa de Valores brasileira para obter as cotações das ações das empresas;

c) avaliar modelos preditivos de Aprendizado de Máquina baseados em Redes Neurais Artificiais e modelos de Aprendizado de Máquina baseados em Processamento de Linguagem Natural para realizar a recomendação do momento de compra ou venda de ações;

d) avaliar as métricas dos modelos de aprendizado de máquina.

2 TRABALHOS CORRELATOS

Esse capítulo apresenta os trabalhos com características semelhantes aos principais objetivos do estudo proposto. O primeiro trabalho correlato é o estudo realizado por Vargas et al. (2018), que utiliza técnicas de aprendizado profundo unindo Rede Neural Recorrente e Rede Neural Convolucional. O segundo, realizado por Li et al. (2020), utilizou Rede Neural Multimodal orientada a eventos para previsão do valor das ações da bolsa de valores. O terceiro, uma pesquisa realizada por Li e Pan (2020), propõe um método para usar duas Redes Neurais Recorrentes seguidas de uma Rede Neural Totalmente Conectada (Fully Connected Neural Network - FCNN) para prever o movimento das ações. Para melhor compreender os trabalhos correlatos, detalha-se cada um dos estudos.

1.2 APRENDIZADO PROFUNDO PARA PREVISÃO DO MERCADO DE AÇÕES USANDO INDICADORES TÉCNICOS E ARTIGOS DE NOTÍCIAS FINANCEIRAS[1]

O estudo de Vargas *et al*. (2018) utilizou quatro modelos para prever os movimentos direcionais diários do preço das ações da *Chevron Corporation* (CVX) utilizando títulos de notícias financeiras e indicadores técnicos como entrada. Os dois conjuntos de indicadores técnicos, são: Conjunto 1: Estocástico %K, Estocástico %D, Momentum, Taxa de Variação, Williams %R, Acumulação Distribuição (A/D) Oscillator and Disparity. O Conjunto 2: Média Móvel Exponencial, Média Móvel Convergente e Divergente, *Relative Strength Index* (RSI), *On Balance Volume* (OBV) e Bandas de *Bollinger*. Uma comparação é feita entre os quatro modelos utilizados que recebem um dos conjuntos de indicadores técnicos em execuções diferentes. O primeiro é o modelo de Rede Neural *Long Short-Term Memory* (LSTM) apenas com o conjunto 1 de indicadores, I-RNN; o segundo é o modelo Rede Neural LSTM, apenas com o conjunto 2 de indicadores, I-RNN-2; o terceiro é o modelo híbrido composto por Rede Neural Convolucional (*Convolutional Neural Network* - CNN) e LSTM com o conjunto 1 de indicadores, SI-RCNN; e, finalmente, o quarto é o modelo híbrido composto por CNN e LSTM com o conjunto 2 de indicadores, SI-RCNN-2.

O banco de dados usado neste trabalho consiste em 106.494 artigos de notícias do site da *Reuters*, correspondentes ao período de 20 de outubro de 2006 a 21 de novembro de 2013. O tema principal de todos esses artigos são notícias do mercado financeiro. Cada notícia consiste em seu título, conteúdo e data de publicação. A data de publicação é empregada para o alinhamento das notícias com uma série temporal financeira correspondente. A acurácia do modelo capaz de prever os preços das ações conseguiu atingir 56.84% nos dados de teste, conforme mostra a Tabela 1.

Tabela 1 – Comparação da acurácia dos resultados

Modelos	Treino Acurácia (%)	Validação Acurácia (%)	Teste Acurácia (%)
I-RNN	55.22	55.97	52.52
SI-RCNN	84.08	60.45	56.84
I-RNN-2	59.08	50.74	48.92
SI-RCNN-2	88.31	61.19	51.08

Fonte: Vargas *et al*. (2018, p. 6).

"É importante observar que os modelos de aprendizado profundo requerem uma grande quantidade de dados, mas ao lidar com previsões de preços do mercado de ações, ocorrências passadas não necessariamente têm correlação com o comportamento futuro" (VARGAS *et al*., 2018, p. 8, tradução nossa)[2], o que justifica a acurácia do teste. Para aperfeiçoamento do modelo, os autores sugerem a inclusão de algum tipo de estratégia de negociação capaz de eliminar pequenas variações e focar somente nas variações significativas de preços.

O modelo híbrido proposto no estudo de Vargas *et al*. (2018), apresentou os melhores resultados utilizando como entrada um conjunto de indicadores técnicos extraídos das empresas relacionadas e títulos de notícias financeiras publicados na véspera do dia da previsão. É aplicado um processo de duas etapas para representar cada notícia no conjunto de dados: primeiro, um modelo que usa técnica de Processamento de Linguagem Natural, Word2vec, usado para gerar uma representação de palavra e, segundo, a média de todos os vetores de palavras do mesmo título é executado, abordando a dispersão em entradas baseadas em palavras. O modelo RCNN visa obter vantagens de ambos os modelos: CNN e Rede Neural Recorrente (*Recurrent Neural Network* - RNN). O CNN tem uma capacidade superior de extrair informações semânticas de textos em comparação com RNN e RNN é melhor para capturar as informações de contexto e na modelagem de características temporais complexas. Finalmente,

[1] Título do trabalho de Vargas *et al*. (2018). No original: *Deep learning for stock market prediction using technical indicators and financial news articles*.

[2] No original: *It is important to note that deep learning models require a large amount of data, but when dealing with stock market prices predictions past occurrences does not necessarily have correlation with future behavior.*

quando os modelos criam as previsões dos movimentos direcionais do preço das ações, um agente de negociação decide quando comprar ou vender uma ação.

2.2 UM MODELO LSTM MOVIDO POR EVENTO MULTIMODAL PARA PREVISÃO DE MERCADO DE AÇÕES USANDO NOTÍCIAS ON-LINE[3]

O trabalho de Li *et al*. (2020) realizou previsões dos preços de ações levando em consideração que as informações sobre os fundamentos (volume de negócios, preços de abertura e volumes de negociação) e notícias das empresas afetam o movimento das ações caracterizando assim um problema multimodal. Para lidar com esses desafios, foi proposto um modelo de Rede *Neural Long Short-Term Memory* (LSTM) orientado a eventos para atender os diferentes tempos de amostragem. Isso é, fundindo os dados dos fundamentos da empresa que são em intervalos iguais e as notícias, que são em intervalos não iguais.

No experimento foram utilizados dados de ações da *China Securities Index* (CSI 100), fornecidos por Li *et al*. (2014). Adicionalmente, o rastreador (*web crawler*) de Li *et al.* (2020), buscou 45.021 notícias das 100 companhias listadas na CSI 100, entre 01/01/2015 e 31/12/2015, do site www.eastmoney.com, que é um dos portais de informações financeiras da China.

Foram consideradas no número k de dias à frente que a notícia influencia a ação. No Quadro 1 observa-se que Target 1 compara o preço de abertura das ações no dia i + k com o preço de abertura no dia i. Target 2 segue a mesma lógica, mas com base no fechamento das ações ao invés do preço de abertura. Para a Target 3, compara-se o preço de fechamento no dia i + k com o preço de abertura no dia i (ou seja, do dia anterior).

Quadro 1 - Os três períodos: abertura, fechamento do preço da ação e influência da notícia

Tracks	Targets formula
Target 1	$price_{i+k}^{open} - price_{i+k-1}^{open}$
Target 2	$price_{i+k}^{close} - price_{i+k-1}^{close}$
Target 3	$price_{i+k}^{close} - price_{i+k-1}^{open}$

Fonte: Li *et al*. (2020, p.10).

Para avaliação de desempenho geral da abordagem proposta foram usados vários métodos clássicos para comparação, incluindo Máquina de Vetor de Suporte (*Support Vector Machine* - SVM), Árvore de Decisão (*Decision Tree* - DT), Rede Neural *Backpropagation* (BP), Rede Neural LSTM e o modelo TeSIA que é um dos métodos mais modernos para prever movimentos de ações. Como métricas, foram selecionadas a *Directional Accuracy* (DA), que é uma métrica para tarefas de classificação de ações e a *Matthews Correlation Coefficient* (MCC), que leva em consideração verdadeiros e falsos positivos e negativos e é geralmente considerado uma medida equilibrada que pode ser usada mesmo se as classes forem de tamanhos muito diferentes. DA tende a apresentar viés quando as classes não são balanceadas, ou seja, se apresentarem tamanhos muito diferentes. Um exemplo seria uma base com 100 amostras em que 98% seriam oriundos da classe positiva e 2% da classe negativa. Se o classificador julgar que todos os valores são da classe positiva, então a DA alcançada é de 98%. No entanto, esse classificador poderia falhar em reconhecer as amostras da classe negativa. Assim, MCC também foi escolhida para evitar esse viés causado por dados desbalanceados. Para ambas as métricas, um valor maior indica melhor desempenho. A Tabela 2 exibe os resultados utilizando os períodos alvos usados no Quadro 1.

Tabela 2 – Comparação de desempenho entre modelos

	Target 1		**Target2**		**Target 3**	
Model	DA	MCC	DA	MCC	DA	MCC
SVM	0.547	0.1956	0.519	0.0679	0.528	0.1374
DT	0.562	0.2244	0.537	0.1277	0.524	0.1195
BP	0.542	0.1438	0.551	0.1997	0.539	0.1422
LSTM	0.571	0.2354	0.583	0.2594	0.601	0.3058
TeSIA	0.584	0.2775	0.576	0.2371	0.597	0.2789
Our model	0.617	0.3516	0.614	0.3304	0.694	0.4472

[3] Título do trabalho de Li *et al.* (2020). No original: *A multimodal event-driven LSTM model for stock prediction using online news.*

Fonte: Li *et al*. (2020, p. 11).

Em termos das métricas DA, MCC, SVM e os modelos DT alcançaram o melhor desempenho para o Target 1, entre os três targets considerados. O modelo BP alcançou seu melhor desempenho para o Target 2, enquanto que LSTM, TeSIA e a abordagem proposta por Li *et al*. alcançou seu melhor desempenho para o Target 3.

2.3 MODELO DE APRENDIZAGEM PROFUNDA PARA PREVISÃO DE MERCADO DE AÇÕES BASEADO EM PREÇOS DE AÇÕES E NOTÍCIAS[4]

No que se refere ao trabalho de Li e Pan (2020), verifica-se o uso de análise de sentimento para extrair informações úteis de texto de múltiplas fontes de dados e um modelo de Aprendizado Profundo Combinado para prever o movimento futuro de ações. Esse Modelo Combinado possui dois níveis. O primeiro nível contém duas Redes Neurais Recorrentes (*Recurrent Neural Network* - RNN), uma Rede Neural de Memória de Longo Prazo (*Long Short-Term Memory* - LSTM) e uma unidade de Rede Neural Recorrente Bloqueada (*Gated Recurrent Unit* - GRU). O segundo nível conta com uma Rede Neural Totalmente Conectada (*Fully Connected Neural Network* - FCNN). Os modelos RNNs, LSTM e GRU podem capturar efetivamente os eventos de série temporal nos dados de entrada e a Rede Neural Totalmente Conectada é usada para reunir vários resultados de predições individuais para melhorar ainda mais a precisão da previsão.

Os dados usados por Li e Pan (2020) foram retirados do estudo de Li *et al*. (2019) e foram divididos em: dados de notícias, obtidos de CNBC.com, Reuters.com, WSJ.com, Fortune.com, com datas no período de dezembro de 2017 até o fim de junho de 2018, e, dados de ações que são do índice S&P Index 500, no mesmo intervalo de datas dos dados de notícias. O S&P 500 é um índice do mercado de ações que mede o desempenho das ações das 500 maiores empresas de capital aberto dos Estados Unidos.

Conforme apresentado na Tabela 3, o Modelo Combinado (*Blending Ensemble*) supera todos os outros modelos em cada uma das métricas utilizadas. Ao analisar isoladamente o Erro Quadrático Médio (MSE) que é uma métrica usada em regressões para calcular o erro nas previsões (SAMMUT; WEBB, 2010), o Modelo Combinado tem uma melhoria significativa ao reduzir o erro, sobretudo, ao comparar com o modelo proposto no trabalho anterior de Li *et al*. (2019), o DP-LSTM.

Tabela 3 – Comparação de desempenho entre modelos

Evaluation Metrics	LSTM	DP-LSTM	GRU	Averaging Ensemble	Weighted Average Ensemble	Blending Ensemble
MSE	438.94	330.97	249.34	231.16	229.52	186.32
MPA	99.29%	99.48%	99.57%	99.57%	99.57%	99.65%
Precision	25%	20%	40%	25%	40%	60%
Recall	25%	25%	50%	25%	50%	75%
F1-Score	25%	22.22%	44.44%	25%	44.44%	66.67%
MDA	33.33%	22.22%	44.44%	33.33%	33.33%	66.67%

Fonte: Li *et al*. (2020, p. 10).

Os resultados apontados pelo modelo estudado por Li *e Pan* (2020) reduzem o erro quadrático médio (MSE) em 57,55%, aumentando a taxa de precisão em 40%, *Recall* em 50%, pontuação F1-*score* em 44,78%, direção do movimento em 33% e precisão (MDA) em 34%. Validando assim, o modelo aplicado, principalmente em termos de compensação entre o retorno e o risco.

3 PROPOSTA DO PROTÓTIPO

Este capítulo apresenta a justificativa para elaboração do projeto, os requisitos principais para o protótipo proposto por este estudo, assim como a metodologia adotada.

3.1 JUSTIFICATIVA

Investir na bolsa de valores pode representar excelentes oportunidades financeiras, o que impulsiona pesquisadores e investidores a predizer o mercado financeiro. Possuir a capacidade de antecipar-se ao movimento de um mercado com tantas variáveis pode representar vantagem em relação aos demais investidores e grande lucratividade. Estudos relacionados à inteligência artificial têm ganhado destaque nesse contexto. No entanto, cada mercado reflete um contexto social específico, não somente devido às propriedades econômicas de sua organização, mas também em relação ao padrão predominante de circulação de informação. Machado (2017, p.

[4] Título do trabalho de Li e Pan (2020). No original: *A novel ensemble deep learning model for stock prediction based on stock prices and news*.

23) afirma que “acionistas e economistas financeiros se interessam profundamente na análise da relação entre o risco de um ativo financeiro e a segurança do seu retorno”. Tais agentes, sempre buscam estudos e opiniões fidedignas que possam embasar suas decisões sobre a venda, compra ou a renegociação de títulos de dívidas do setor público ou privado. Apesar da centralidade destes fatores, ainda não existe um dispositivo que esteja adaptado ao cenário nacional brasileiro. Elagamy, Stanier e Sharp (2018) ressaltam que o mercado financeiro representa um papel crucial no crescimento do comércio e da indústria. Por isso, encontrar formas eficientes de analisar e visualizar os dados deste setor é tarefa significante para a economia moderna. Desta forma, o desenvolvimento deste estudo constitui uma oportunidade para preencher a lacuna existente neste setor e com isso, cria a possibilidade de acompanhamento das relações existentes entre investidores e o contexto financeiro e midiático brasileiro, o que justifica sua aplicação social e teórica.

Este projeto aponta um conjunto de fatores relacionados ao mercado de capitais, mostrando que as estratégias de monitoramento e investimento vêm sendo amparadas por mecanismos automatizados, considerando o efeito combinado de expansão e volatilidade desse mercado. Nesse cenário, destacam-se as Redes Neurais Artificiais. Dessa forma, o projeto é justificado prática e cientificamente, porque propõe a configuração de um protótipo usando Redes Neurais Artificiais para sugerir a compra ou venda de ativos negociados na Bolsa de Valores brasileira, por meio de notícias do mercado financeiro, a fim de auxiliar investidores nas suas transações no mercado de renda variável.

Quanto aos trabalhos correlatos apresentados por este estudo, é preciso estabelecer uma comparação entre eles e considerar que, apesar de os estudos possuírem características em comum, a aplicação e conclusões guardam diferenças, que são evidenciadas pelo Quadro 2.

Quadro 2 – Comparativo dos trabalhos correlatos

Trabalhos Correlatos / Características	Vargas *et al*. (2018)	Li *et al*. (2020)	Li e Pan (2020)
Rede Neural Convolucional	Sim	Não	Não
Rede Neural Recorrente / LSTM	Sim	Sim	Sim
Ativos analisados	Ações da Chevron Corporation (CVX)	Ações da CSI 100	Ações do S&P500
Fonte das notícias	www.reuters.com	www.eastmoney.com	www.cnbc.com, www.reuters.com, www.fortune.com, www.wsj.com.
Período das notícias	Outubro de 2006 até Novembro de 2013	Janeiro de 2015 até Dezembro de 2015	Dezembro de 2017 até Junho de 2018
Frequência das notícias	Diária	Não informado	Não informado
Métricas	Acurácia	Directional Accuracy, Matthews Correlation Coefficient	Erro Quadrático Médio, Mean Prediction Acurácia, Precisão, Recall, F1-Score, Movement Direction Accuracy

Fonte: elaborado pelo autor.

Assim, enquanto o trabalho de Li *et al*. (2020) utilizou Rede Neural *Long Short-Term Memory* (LSTM) orientado a eventos para lidar com os diferentes tempos de amostragem, o trabalho desenvolvido por Li e Pan (2020) utilizou técnicas *Ensemble* no intuito de conseguir um melhor desempenho nas previsões da movimentação dos preços. Neste sentido, verifica-se que apesar de todos os trabalhos utilizarem Redes Neurais Recorrentes e LSTM, somente o estudo de Vargas *et al*. (2018) fez uso de Redes Neurais Convolucionais. Isto é particularmente importante porque a aplicação de Redes Neurais Convolucionais tem se mostrado muito mais eficaz na captura de semântica de dados textuais.

Referentes aos ativos utilizados, o trabalho de Li *et al*. (2020) utilizou ações da Bolsa de Valores da China, enquanto os trabalhos propostos por Vargas *et al*. (2018) e Li e Pan (2020) utilizaram ações do mercado ocidental. Esse fato, adicionando os diferentes períodos de obtenção das notícias relacionadas às ações, e também às diferentes métricas escolhidas, torna difícil a comparação entre os modelos apresentados. No entanto, cada um contribui para aprimorar e avaliar as métricas dos modelos de aprendizado de máquina, usando redes neurais artificiais em relação às ações da bolsa de valores.

3.2 REQUISITOS PRINCIPAIS DO PROTÓTIPO

Para alcançar os objetivos, o protótipo proposto deve:

Comentado [MH2]: Deveria considerar algum requisito para atender ao objetivo específico b) consultar API da Bolsa de Valores brasileira para obter as cotações das ações das empresas;

a) permitir ao usuário selecionar o ativo a ser analisado (Requisito Funcionais - RF);

b) permitir ao usuário selecionar o *timeframe* desejado (RF);

c) permitir ao usuário visualizar o gráfico de *candlesticks* (RF);

d) permitir ao usuário visualizar as notícias e a análise de sentimento delas (RF);

e) permitir ao usuário visualizar os resultados obtidos com as sugestões de compra ou venda (RF);

f) ser implementado no *framework* Django no ambiente de desenvolvimento *PyCharm* (Requisito Não-Funcional - RNF);

g) utilizar um modelo preditivo baseado em técnicas de Aprendizado de Máquina (RNF);

h) utilizar uma Rede Neural Artificial para Processamento de Linguagem Natural (RNF);

i) utilizar um *web scraper* para buscar as notícias da internet (RNF);

j) utilizar a linguagem de programação *Python* (RNF);

k) utilizar banco de dados PostgreSQL (RNF).

3.3 METODOLOGIA

O trabalho será desenvolvido observando as seguintes etapas:

a) Levantamento bibliográfico: pesquisar trabalhos relacionados e estudos sobre bolsa de valores, modelos preditivos e suas ferramentas: *web scraper*, *Machine Learning*, Deep Learning, Redes Neurais Profundas como *Recurrent Neural Network* (Rede Neural Recorrente – RNN), a Rede Neural Recorrente *Long Short-Term Memory* (LSTM) e a *Convolutional Neural Network* (Rede Neural Convolucional – CNN), entre outros temas e ferramentas observadas no levantamento teórico;
b) elicitação de requisitos: reavaliar os requisitos da etapa anterior e especificar outros mediante necessidade identificada durante a revisão bibliográfica;
c) coleta de dados: buscar dados com valores de abertura e fechamento das ações (www.b3.com.br) e notícias do mercado financeiro (br.investing.com);
d) especificação: formalizar por meio da *Unified Modeling Language* (UML) a diagramação das classes e dos casos de uso com a ferramenta *Microsoft Visio*;
e) implementação: desenvolver o protótipo utilizando o *framework* Django para desenvolvimento *web* no ambiente de desenvolvimento *PyCharm*;
f) testes: em conjunto com a etapa anterior, realizar testes do protótipo para validação dos resultados obtidos, confiabilidade dos dados e performance.

As etapas serão realizadas nos períodos relacionados pelo Quadro 3.

Quadro 3 – Cronograma

	2021									
	Ago.		Set.		Out.		Nov.		Dez.	
Etapas / quinzenas	1	2	1	2	1	2	1	2	1	2
Levantamento bibliográfico										
Elicitação de requisitos										
Coleta de dados										
Especificação										
Implementação										
Testes										

Fonte: elaborado pelo autor.

4 REVISÃO BIBLIOGRÁFICA

Neste capítulo são apresentados os principais assuntos que fundamentam e constroem a proposta de trabalho deste projeto.

4.1 MERCADO FINANCEIRO E INTELIGÊNCIA ARTIFICIAL

Segundo Mueller e Massaron (2016), o acesso à informação, a complexidade e volatilidade do mercado financeiro têem levado investidores a buscar auxílio na Inteligência Artificial, explorando a capacidade dos computadores de aprender, se adaptar a novas circunstâncias, detectar padrões, criar novos comportamentos e tomar decisões a partir de um conjunto de dados. Para Santos *et al*. (2016) a busca de ferramentas que auxiliem a tomada de decisões têem utilizado modelos baseados em Redes Neurais Artificiais (RNA), que conseguem reter aprendizado e realizar procedimentos de controle, reconhecimento de padrões e classificação, contribuído

diretamente para avanços em relação ao desenvolvimento de técnicas que propiciam um melhor entendimento dos padrões presentes em dados financeiros e na construção de modelos preditivos.

4.2 MODELOS PREDITIVOS E BOLSA DE VALORES

Embora aplicáveis em diversos contextos, os modelos preditivos têm alcançado espaço em um universo dinâmico e variável: a bolsa de valores. Giacomel (2016) explica que a bolsa de valores é de existência necessária a países que possuem diversas empresas de capital aberto, como um lugar em comum para a negociação das ações dessas empresas e das instituições participantes, devidamente credenciadas. Silva (2017) afirma que no Brasil a única bolsa de valores que realiza a negociação de valores mobiliários é a B3 (Brasil, Bolsa, Balcão – antiga BM&FBOVESPA), e que esse órgão é autorregulador e responsável por todos os registros de negociações de ativos do mercado brasileiro.

4.3 FERRAMENTAS PARA MODELOS PREDITIVOS

Para criar um modelo preditivo capaz de analisar os dados do mercado financeiro e auxiliar investidores nas suas transações no mercado de renda variável é preciso criar ferramentas capazes de coletar grandes quantidades de informações da *web*. Um dos *scripts* automatizados mais utilizados para isso é o

> *Web Scraper* (Raspador *Web*), um programa que realiza a extração automática de dados específicos de uma página *web*, na linguagem *Python*, utilizando como base uma biblioteca chamada *Beautiful Soup*. Essa biblioteca realiza a leitura e a extração de dados de textos HTML ou XML, permitindo a busca por *strings*, *tags*, *ids*, classes e qualquer outro atributo que possa servir de identificação para um elemento (MAZINI; SATO, 2020, p. 1-2).

Comentado [MH3]: Ficou estranho usar essa definição, pois nem todo web scraper precisa ser em Python e usar essa biblioteca.

Depois de extrair os dados é possível transformá-los em informação estruturada, criando um padrão e um formato necessários para a análise que se pretenda fazer. Em seguida, os dados são carregados no sistema onde serão cruzados, relacionados, tratados, analisados e visualizados. Essa capacidade que a inteligência artificial tem de aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana é chamado de Aprendizado de Máquina (em inglês, *Machine Learning*).

Comentado [MH4]: Será dessa forma que fará a análise de sentimento das notícias extraídas (requisito d)?

Embora existam muitos métodos de extrair *insights*, padrões e relações que podem ser usados nas tomadas de decisão, eles possuem abordagens e capacidades diferentes. O objetivo do *Machine Learning* é entender a estrutura dos dados e encaixar essas distribuições teóricas em dados bem entendidos. Há uma comprovação matemática por trás de modelos estatísticos, que deve atender a certos pressupostos. O teste para um modelo de *Machine Learning* é um erro de validação em dados novos e não um teste teórico. Hiransha *et al*. (2018) destaca que *Machine Learning* geralmente usa uma abordagem iterativa para aprender com os dados, executando etapas até encontrar um padrão, e esse aprendizado pode ser automatizado.

Para analisar grandes quantidades de dados complexos, podem ser utilizadas as técnicas de *Deep Learning* (DL), que usam Redes Artificiais Neurais Profundas. Para Hiransha *et al*. (2018), as técnicas de DL são ferramentas importantes para a análise de dados não categorizados, fazendo uso das redes neurais em processamento de imagens, reconhecimento de imagens, de áudio, reconhecimento facial, de caracteres, mineração de dados, classificação de doenças *etc*.

As Redes Neurais Artificiais Profundas são definidas por Luger (2013), como um modelo em camadas em que as novas informações são geradas ou informações existentes são adaptadas por meio de conexões entre as camadas. Assim, as camadas anteriores têm relação com as camadas posteriores, criando a mesma ideia de informações ancestrais apresentada na definição da predição estatística. São projetadas para reconhecer padrões em sequências de dados, como texto, genomas, caligrafia, palavra falada ou dados de séries numéricas que emanam de sensores, bolsas de valores e agências governamentais. Esses algoritmos consideram tempo e sequência, eles têm uma dimensão temporal. Entre as Redes Neurais Profundas destacam-se a *Recurrent Neural Network* (Rede Neural Recorrente – RNN), a Rede Neural Recorrente *Long Short-Term Memory* (LSTM) e a *Convolutional Neural Network* (Rede Neural Convolucional – CNN).

As Redes Neurais Recorrentes (RNN) sofrem de memória de curto prazo. Elas têm dificuldade em transportar informações muito longas das etapas anteriores para as posteriores, pois sofrem com o problema da dissipação do gradiente. Haykin (2008, p. 24) explica que “gradientes são valores usados para atualizar os pesos das redes neurais”. Se um valor de gradiente se torna extremamente pequeno, não contribui com o aprendizado e para de aprender. Como essas camadas não aprendem, as RNN podem esquecer o que foi visto em sequências mais longas, caracterizando assim uma memória de curto prazo. Uma solução para a memória de curto prazo é a Rede Neural Recorrente LSTM, que possui mecanismos internos chamados portões que podem regular o fluxo de informações.

Já as Redes Convolucionais, segundo Goodfellow, Bengio e Courville (2016) realizam o reconhecimento óptico de caracteres para digitalizar texto e tornar possível o processamento de linguagem natural em documentos analógicos e manuscritos, arquivos de áudio quando estes são representados visualmente, análise de texto, bem como dados gráficos. Assim, as imagens são símbolos a serem transcritos. Para os autores, a etapa de convolução caracteriza-se pela passagem do núcleo (*kernel*) pela imagem (*input*) e o resultado desse processamento é denominado mapa de características (*output*), que permitirá o reconhecimento de padrões, inclusive de outros padrões da rede.

REFERÊNCIAS

ELAGAMY, Mazen Nabil; STANIER, Clare; SHARP, Bernadette. Sistema de mineração de texto florestal aleatório do mercado de ações minerando indicadores críticos de movimentos do mercado de ações. **2018. 2ª Conferência Internacional sobre Linguagem Natural e Processamento de Fala (ICNLSP)**, 2018, pp. 1-8,. Doi: 10.1109 / ICNLSP.2018.8374370. Disponível em: https://ieeexplore.ieee.org/document/8374370. Acesso em: 08/06/2021.

FERRO, Luciano. Aplicação da rede neural MLP (Multilayer Perceptron) em indústria de pisos e revestimentos do Polo Cerâmico de Santa Gertrudes - SP. 2013. 143 f. **Tese** (Doutorado) - Universidade Estadual Paulista Júlio de Mesquita Filho, Instituto de Geociências e Ciências Exatas, 2013. Disponível em: Acesso em: 09/06/2021.

GIACOMEL, Felipe dos Santos. Um método algorítmico para operações na bolsa de valores baseado em *ensembles* de redes neurais para modelar e prever os movimentos dos mercados de ações. 2016. 92 f. **Dissertação** (Mestrado) - Curso de Programa de Pós-graduação em Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016. Disponível em: https://lume.ufrgs.br/handle/10183/134586. Acesso em: 09/06/2021.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep learning**. The MIT Press, 2016, 800 pp, ISBN: 0262035618. Disponível em: https://www.researchgate.net/publication/320703571_Ian_Goodfellow Yoshua_Bengio_and_Aaron_Courville_Deep_learning_The_MIT_Press_2016_800_pp_ISBN_0262035618/link/5b880b494585151fd13c8b95/download. Acesso em: 08/06/2021.

HAYKIN, Simon. **Neural Networks and Learning Machines**. Third edition. New York: Pearson Education, 2008. Disponível em: http://dai.fmph.uniba.sk/courses/NN/haykin.neural-networks.3ed.2009.pdf**.** Acesso em: 07/06/2021.

HIRANSHA, M. et al. Nse stock market prediction using deep-learning models. *Procedia Computer Science*, v. 132, p. 1351 – 1362, 2018. ISSN 1877-0509. **International Conference on Computational Intelligence and Data Science**. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050918307828>. Acesso em: 09/06/2021.

LI, Yang; PAN, Yi. A Novel Ensemble Deep Learning Model for Stock Prediction Based on Stock Prices and News. **Semantic Scholar**. 23/07/2020. Disponível em: https://www.semanticscholar.org/paper/A-Novel-Ensemble-Deep-Learning-Model-for-Stock-on-Li-Pan/d5aaa87a737c4ff98e0955b951b9892d03d221af. Acesso em: 08/06/2021.

LI, Qin; *et al*. **A Multimodal Event-driven LSTM Model for Stock Prediction Using Online News**. 2020. Disponível em: https://www.researchgate.net/publication/338783254_A_Multimodal_Eventdriven_LSTM Model_for_Stock_Prediction_Using_Online_News. Acesso em: 08/06/2021.

LI, Shihua; *et al*. Object-oriented method combined with deep convolutional neural networks for land-use-type classification of remote sensing images. **Journal of the Indian Society of Remote Sensing**, 47(6), pp. 951-965, Jan., 2019. Disponível em: https://www.researchgate.net/publication/330429024_ObjectOriented_Method_ Combined_with_Deep_ConvolutionalNeural_Networks_for_LandUseType_Classification_of_Remote_Sensing Images/link/5cb049a64585156cd 79176ad/download. Acesso em: 10/06/2021.

LI, Tao; *et al*. Mapping Near-surface Air Temperature, Pressure, Relative Humidity andWind Speed over Mainland China with High Spatiotemporal Resolution. **Advances In Atmospheric Sciences**, vol. 31, September 2014, 1–9. Disponível em: https://www.researchgate.net/publication/267633552_Li_et_al_2014_AAS/link/ 545598b40cf2bccc490cce7c/download. Acesso em: 09/06/2021.

LUGER, George F. **Inteligência Artificial**. 6. ed. São Paulo: Pearson Education, 2013.

MACHADO, Daniel José. Comparando modelos alternativos de precificação de ativos: uma análise para o mercado brasileiro. 2017. 178 f. **Tese** (Doutorado em Administração de Empresas) - Universidade Presbiteriana Mackenzie, São Paulo, 2017. Disponível em: http://tede.mackenzie.br/jspui/handle/tede/3444**.** Acesso em: 09/06/2021.

MAZINI, Dhaniel Nunes; SATO, Renato César. **Extração de dados financeiros com um *web scr*aper: um estudo sobre a rentabilidade dos dividendos**. Universidade Federal de São Paulo – São José dos Campos, SP., 2020. Disponível em: http://www.comp.ita.br/labsca/waiaf/papers/DhanielMazini_paper_20.pdf**.** Acesso em: 09/06/2021.

MUELLER, John Paul; MASSARON, Luca. **Machine Learning for Dummies**. Nova Jersey: John Wiley & Sons, Inc, 2016. 399 p.

PAGOLU, Venkata Sasank; *et al*. (2016), Sentiment analysis of Twitter data for predicting stock market movements**. International Conference On Signal Processing, Communication, Power And Embedded System** (SCOPES), IEEE., 2016. http://dx.doi.org/10.1109/scopes.2016.7955659**.** Disponível em: https://www.semanticscholar.org/paper/Sentiment-

Código de campo alterado

analysis-of-Twitter-data-for-predicting-Pagolu-Challa/fcbba03b6156295a5738f9f03d157f67f665365c**.** Acesso em: 07/06/2021.

SAMMUT, Claude; WEBB, Geoffrey. (Editors). Erro Médio Absoluto. In: Sammut C., Webb GI (Eds.) **Encyclopedia of Machine Learning and Data Mining**. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7687-1_953. Disponível em: https://link.springer.com/referenceworkentry/10.1007%2F978-1-4899-7687-1_953#howtocite. Acesso em: 06/06/2021.

SANTOS, Murilo Alves; *et al*. Aplicação de redes neurais no Brasil: um estudo bibliométrico. **Biblionline**, v. 12, n. 2, p. 101-116, 2016. Disponível em: https://www.researchgate.net/publication/303957049_APLICACAO _DE_REDES NEURAIS_NO_BRASIL_UM_ESTUDO_BIBLIOMETRICO. Acesso em: 07/06/2021.

SILVA, Anderson Rodrigues da. Aspectos regulatórios da bolsa de valores no Brasil. 2017. 135 f. **Dissertação** (Mestrado) - Curso de Direito, Programa de Estudos Pós-graduados em Direito, Pontifícia Universidade Católica de São Paulo, São Paulo, 2017. Disponível em: https://tede2.pucsp.br/handle/handle/20876. Acesso em: 09/06/2021.

TRISOTTO, Fernanda. Apesar da pandemia, Bolsa de Valores teve ano de recordes. E o que esperar de 2021? **Gazeta do Povo**. 30/12/2020. Disponível em: https://www.gazetadopovo.com.br/economia/bolsa-de-valores-ano-2020-recordes-pandemia/. Acesso em: 10/06/2021.

VARGAS, M. R., ANJOS, C. E. M. dos; BICHARA, G. L. G., EVSUKOFF, A. G. Deep Learning for Stock Market Prediction Using Technical Indicators and Financial News Articles, 2018. **International Joint Conference on Neural Networks** (IJCNN), IEEE, 2018. Disponível em: https://www.researchgate.net/profile/AlexandreEvsukoff/publication/328400101_Deep_Leaming_for_Stock_Market_Prediction_Using_Technical_Indicators_and_Financial_News_Articles/links/5c6ab8294585156b57036c91/Deep-Leaming-for-Stock-Market-Prediction-Using-Technical-Indicators-and-Financial-News-Articles.pdf. Acesso em: 09/06/2021.

ASSINATURAS

(Atenção: todas as folhas devem estar rubricadas)

Assinatura do Aluno: _____

Assinatura da Orientadora: _____

Assinatura do(a) Coorientador(a) (se houver): _____

Observações do orientador em relação a itens não atendidos do pré-projeto (se houver):

FORMULÁRIO DE AVALIAÇÃO – PROFESSOR TCC I

Acadêmico: Fabrício Oliveira Bezerra

Avaliador(a): _____

	ASPECTOS AVALIADOS[1]	Atende	Atende parcialmente	Não atende
ASPECTOS TÉCNICOS	1. INTRODUÇÃO O tema de pesquisa está devidamente contextualizado/delimitado?			
	O problema está claramente formulado?			
	2. OBJETIVOS O objetivo principal está claramente definido e é passível de ser alcançado?			
	Os objetivos específicos são coerentes com o objetivo principal?			
	3. JUSTIFICATIVA São apresentados argumentos científicos, técnicos ou metodológicos que justificam a proposta?			
	São apresentadas as contribuições teóricas, práticas ou sociais que justificam a proposta?			
	4. METODOLOGIA Foram relacionadas todas as etapas necessárias para o desenvolvimento do TCC?			
	Os métodos, recursos e o cronograma estão devidamente apresentados?			
	5. REVISÃO BIBLIOGRÁFICA (atenção para a diferença de conteúdo entre projeto e pré-projeto) Os assuntos apresentados são suficientes e têm relação com o tema do TCC?			
ASPECTOS METODOLÓGICOS	6. LINGUAGEM USADA (redação) O texto completo é coerente e redigido corretamente em língua portuguesa, usando linguagem formal/científica?			
	A exposição do assunto é ordenada (as ideias estão bem encadeadas e a linguagem utilizada é clara)?			
	7. ORGANIZAÇÃO E APRESENTAÇÃO GRÁFICA DO TEXTO A organização e apresentação dos capítulos, seções, subseções e parágrafos estão de acordo com o modelo estabelecido?			
	8. ILUSTRAÇÕES (figuras, quadros, tabelas) As ilustrações são legíveis e obedecem às normas da ABNT?			
	9. REFERÊNCIAS E CITAÇÕES As referências obedecem às normas da ABNT?			
	As citações obedecem às normas da ABNT?			
	Todos os documentos citados foram referenciados e vice-versa, isto é, as citações e referências são consistentes?			

PARECER – PROFESSOR DE TCC I OU COORDENADOR DE TCC

(PREENCHER APENAS NO PROJETO):

O projeto de TCC será reprovado se:

- Qualquer um dos itens tiver resposta NÃO ATENDE;
- Pelo menos **4 (quatro)** itens dos **ASPECTOS TÉCNICOS** tiverem resposta ATENDE PARCIALMENTE; ou
- Pelo menos **4 (quatro)** itens dos **ASPECTOS METODOLÓGICOS** tiverem resposta ATENDE PARCIALMENTE.

PARECER: ( ) APROVADO ( ) REPROVADO

Assinatura: _____ Data: _____

[1] Quando o avaliador marcar algum item como atende parcialmente ou não atende, deve obrigatoriamente indicar os motivos no texto, para que o aluno saiba o porquê da avaliação.

FORMULÁRIO DE AVALIAÇÃO – PROFESSOR AVALIADOR

Acadêmico(a): Fabrício Oliveira Bezerra

Avaliador(a): Marcel Hugo

	ASPECTOS AVALIADOS[1]	atende	atende parcialmente	não atende
ASPECTOS TÉCNICOS	10. INTRODUÇÃO O tema de pesquisa está devidamente contextualizado/delimitado?	x		
	O problema está claramente formulado?	x		
	11. OBJETIVOS O objetivo principal está claramente definido e é passível de ser alcançado?	x		
	Os objetivos específicos são coerentes com o objetivo principal?	x		
	12. TRABALHOS CORRELATOS São apresentados trabalhos correlatos, bem como descritas as principais funcionalidades e os pontos fortes e fracos?	x		
	13. JUSTIFICATIVA Foi apresentado e discutido um quadro relacionando os trabalhos correlatos e suas principais funcionalidades com a proposta apresentada?	x		
	São apresentados argumentos científicos, técnicos ou metodológicos que justificam a proposta?	x		
	São apresentadas as contribuições teóricas, práticas ou sociais que justificam a proposta?	x		
	14. REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO Os requisitos funcionais e não funcionais foram claramente descritos?		X	
	15. METODOLOGIA Foram relacionadas todas as etapas necessárias para o desenvolvimento do TCC?	x		
	Os métodos, recursos e o cronograma estão devidamente apresentados e são compatíveis com a metodologia proposta?	x		
	16. REVISÃO BIBLIOGRÁFICA (atenção para a diferença de conteúdo entre projeto e pré-projeto) Os assuntos apresentados são suficientes e têm relação com o tema do TCC?		X	
	As referências contemplam adequadamente os assuntos abordados (são indicadas obras atualizadas e as mais importantes da área)?	X		
ASPECTOS METODOLÓGICOS	17. LINGUAGEM USADA (redação) O texto completo é coerente e redigido corretamente em língua portuguesa, usando linguagem formal/científica?	X		
	A exposição do assunto é ordenada (as ideias estão bem encadeadas e a linguagem utilizada é clara)?	X		

PARECER – PROFESSOR AVALIADOR:
(PREENCHER APENAS NO PROJETO)

O projeto de TCC ser deverá ser revisado, isto é, necessita de complementação, se:

- qualquer um dos itens tiver resposta NÃO ATENDE;
- pelo menos **5 (cinco)** tiverem resposta ATENDE PARCIALMENTE.

PARECER: (X) APROVADO ( ) REPROVADO

Assinatura: _____ Data: 25/06/2021

[1] Quando o avaliador marcar algum item como atende parcialmente ou não atende, deve obrigatoriamente indicar os motivos no texto, para que o aluno saiba o porquê da avaliação.