CENTRO UNIVERSITÁRIO FEI
DIEGO EDUARDO SILVA

REDES NEURAIS ARTIFICIAIS E PROCESSAMENTO DE LINGUAGEM NATURAL APLICADOS À PREVISÃO DO MINICONTRATO FUTURO DO ÍNDICE IBOVESPA

São Bernardo do Campo
2021

DIEGO EDUARDO SILVA

REDES NEURAIS ARTIFICIAIS E PROCESSAMENTO DE LINGUAGEM NATURAL APLICADOS À PREVISÃO DO MINICONTRATO FUTURO DO ÍNDICE IBOVESPA

Dissertação de mestrado apresentada ao Centro Universitário FEI, para obtenção do título de Mestre em Engenharia Elétrica, orientada pelo Prof. Dr. Reinaldo Augusto da Costa Bianchi.

São Bernardo do Campo

2021

Eduardo Silva, Diego.
Redes Neurais Artificiais e Processamento de Linguagem Natural Aplicados à Previsão do Minicontrato Futuro do Índice Ibovespa / Diego Eduardo Silva. São Bernardo do Campo, 2021.
115 f. : il.

Dissertação - Centro Universitário FEI.
Orientador: Prof. Dr. Reinaldo Augusto da Costa Bianchi.

1. Redes Neurais Recorrentes. 2. Processamento de Linguagem Natural. 3. Bovespa. 4. Análise Técnica. I. Augusto da Costa Bianchi, Reinaldo, orient. II. Título.

Elaborada pelo sistema de geração automática de ficha catalográfica da FEI com os dados fornecidos pelo(a) autor(a).

centro universitário FEI

APRESENTAÇÃO DE DISSERTAÇÃO
ATA DA BANCA EXAMINADORA

Programa de Pós-Graduação Stricto Sensu em Engenharia Elétrica

Mestrado

PGE-10

Aluno: Diego Eduardo Silva **Matrícula:** 119106-3

Título do Trabalho: REDES NEURAIS ARTIFICIAIS E PROCESSAMENTO DE LINGUAGEM NATURAL APLICADOS À PREVISÃO DO MINICONTRATO FUTURO DO ÍNDICE IBOVESPA..

Área de Concentração: Inteligência Artificial Aplicada à Automação e Robótica

Orientador: Prof. Dr. Reinaldo A. C. Bianchi

Data da realização da defesa: 28/06/2021

ORIGINAL ASSINADA

Avaliação da Banca Examinadora:

A banca foi realizada no dia 28 de junho de 2021 às 09:00 horas, e se iniciou com a apresentação do aluno, que foi satisfatória, e seguiu para a arguição, onde o aluno respondeu às questões de forma adequada, com algumas ressalvas. Foram sugeridas melhorias em relação ao texto e aos experimentos que devem ser realizadas pelo aluno para a versão final. A aprovação foi por unanimidade.

São Bernardo do Campo, 28 / 06 /2021.

MEMBROS DA BANCA EXAMINADORA	
Prof. Dr. Reinaldo A. C. Bianchi	Ass.: _____
Prof. Dr. Guilherme Wachs	Ass.: _____
Prof. Dr. Valdinei Freire da Silva	Ass.: _____

A Banca Julgadora acima-assinada atribuiu ao aluno o seguinte resultado:

APROVADO ☒ REPROVADO ☐

VERSÃO FINAL DA DISSERTAÇÃO

APROVO A VERSÃO FINAL DA DISSERTAÇÃO EM QUE FORAM INCLUÍDAS AS RECOMENDAÇÕES DA BANCA EXAMINADORA

_____

Aprovação do Coordenador do Programa de Pós-graduação

_____

Prof. Dr. Carlos Eduardo Thomaz

Dedico este trabalho a Nossa Senhora da Aparecida e a Deus; Eles foram essenciais para essa jornada e superação para os momentos difíceis que passamos, e sem eles não conseguiria capacidade para desenvolver este trabalho.

AGRADECIMENTOS

A Deus por me guiar e iluminar os caminhos, pelo dom de viver.

A Nossa Senhora da Aparecida que me guia e dá forças, principalmente nos momentos mais difíceis dessa trajetória.

Ao meu orientador, Prof. Dr. Reinaldo Augusto da Costa Bianchi, pelo apoio, paciência e valiosos conhecimentos colaborados ao longo de todo o mestrado.

Ao Centro Universitário FEI pelo apoio institucional.

Aos colegas do grupo de estudos FINFEI, ao professor Dr. Guilherme Wachs, e aos colegas durante o curso, pelas discussões, ideias, conselhos e sugestões no trabalho.

À FAPESP, CAPES e CNPq, pelo apoio financeiro. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Aos meus pais, Antônio e Penha, pelo esforço e dedicação prestados nos momentos mais difíceis de minha vida, e principalmente da minha formação como ser humano.

À minha esposa, Daniela, pelo apoio e pela compreensão nos momentos de ausência durante as horas de trabalho de pesquisa.

“No que diz respeito ao empenho, ao compromisso, ao esforço, à dedicação, não existe meio termo. Ou você faz uma coisa bem feita ou não faz.”

Ayrton Senna

“Agradeço todas as dificuldades que enfrentei; não fosse por elas, eu não teria saído do lugar. As facilidades nos impedem de caminhar. Mesmo as críticas nos auxiliam muito.”

Chico Xavier

RESUMO

Estudos científicos recentemente realizados comprovam que existe relação das informações divulgadas nas Redes Sociais com as variações dos preços dos ativos negociados na Bolsa de Valores Brasileira (BOVESPA). Nesses estudos são utilizadas técnicas de Processamento de Linguagem Natural (PLN) para o tratamento de dados textuais que possibilitam a compreensão da linguagem humana pelas máquinas que, enriquecida com as informações históricas dos ativos, geram indicações para tomada de decisão nas negociações da Bolsa. Os relacionamentos entre a Rede Social Twitter e a Bovespa são abordados através do uso de PLN na base de dados da Rede Social com *Word Embedding*, realizando uma classificação dicotômica para tomadas de decisões, não se atendo para as práticas de maiores retornos com os ganhos das variações dos ativos nos pequenos intervalos entre o dia. A proposta deste trabalho é a criação de um modelo para tomada de decisão no mercado financeiro apoiada nas mensagens relativas à BOVESPA na Rede Social Twitter, tratadas por técnicas de Processamento de Linguagem Natural (PLN). Neste ponto é usado frases completas para vetorização do *Word Embedding* e classificadas com uma Rede Neural Recorrente (LSTM) para indicar negociações do ativo mini-índice da BOVESPA com atuação regida pela tendência do mercado acrescida da classificação do *Word Embedding*, agregadas em 5, 15 e 30 minutos, para atuações nos minutos sequências de operações *day trade*. Os experimentos realizados neste trabalho demonstraram a validade da hipótese de que mensagens de uma rede social podem apoiar decisões no mercado financeiro, permitindo obter lucros neste domínio.

Palavras-chave: Processamento de Linguagem Natural (PLN), Twitter, Bovespa, Redes Neurais Recorrentes (LSTM), *Word Embedding*, *Day Trade*.

ABSTRACT

Recently conducted scientific studies prove that there is a relationship between the information published on Social Networks and the variations in the prices of assets traded on the Brazilian Stock Exchange (BOVESPA). In these studies, Natural Language Processing (PLN) techniques are used for the treatment of textual data that enable the understanding of human language by machines, which, enriched with the historical information of the assets, generate indications for decision-making in the stock market negotiations. The relationships between the Social Network Twitter and Bovespa are approached through the use of PLN in the Social Network database with Word Embedding, performing a dichotomous classification for decision making, not taking into account the practices of greater returns with the earnings of the variations of asset in the small intervals between the day. The purpose of this work is to create a model for decision making in the financial market supported by messages related to BOVESPA on the Twitter, handled by Natural Language Processing (PLN) techniques. At this point, complete sentences are used for Word Embedding vectorization and classified with a Recurrent Neural Network (LSTM) to indicate trades in the BOVESPA mini-index asset with performance governed by the market trend plus the Word Embedding classification, aggregated into 5, 15 and 30 minutes, for actions in the sequence of minutes of day trade operations. The experiments carried out in this work demonstrated the validity of the hypothesis that messages from a social network can support decisions in the financial market, allowing to obtain profits in this domain.

Keywords: Natural Language Processing (PLN), Twitter, Bovespa, Recurring Neural Networks (LSTM), Word Embedding, Day Trade.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS

API	Interface de programação de aplicativo - *Application Programming Interface*
BOVESPA	Bolsa de Valores do Estado de São Paulo
CBOW	Saco contínuo de palavras - *Continuos Bag of Word*
CNN	Rede Neural Convolucional- *Convolutional Neural Network*
CSV	Valores Separados Por Vírgula - *Comma-separated-values*
FEI	Centro Universitário FEI
IA	Inteligência Artificial
LSTM	Memória Longa de Curto Prazo - *Long Short-Term Memory*
MLP	Perceptron multicamadas - *Multi-layer Perceptron*
NLP	*Natural Language Processing*
PLN	Processamento de Linguagem Natural
PMC	Perceptron Multicamadas
RNN	Rede Neural Recorrente - *Recurrent Neural Network*
RNNTW-TC	RNN com Twitter para classificação de operações trader - *Recurrent Neural Network with Twitter Trader Classifier*
TF-IDF	Frequência do termo - frequência inversa do documento - *Term frequency – Inverse document frequency*

SUMÁRIO

1 INTRODUÇÃO

Dentre as diversas técnicas para se operar na bolsa de valores destacam-se as fundamentadas pela Teoria de Dow, baseadas no princípio de que o preço do ativo já contém toda a informação das variações do mercado, e que o mercado se move em tendências, segundo Pring (2002). Algumas dessas técnicas são encontradas nos trabalhos Peng e Jiang (2016) e Li e Shah (2017), e relacionam a variação do preço do ativo com os dados contidos nas Redes Sociais para realizar uma classificação binária, praticando assim outro fundamento de Dow que é a confirmação no qual é usado outras variáveis para confirmação do preço do ativo, utilizada na tomada de decisão de compra ou venda na Bolsa de Valores.

As propostas de classificação do texto de forma binária são utilizadas para vinculação aos dados das variações de preço diário dos ativos da bolsa de valores, com o objetivo de auxiliar na tomada de decisão nas negociações de *swing trade* – operação de compra ou venda de um ativo em dias distintos – e se limita no tempo por não considerar que as mensagens são espalhadas durante o dia, e também na limitação de mensagens sobre um único tema ou ativo.

Durante um pregão, que é o período de negociação de ações e outros ativos na bolsa, são realizadas ordens a todo instante, representando diversas operações de compra e venda feitas por pessoas físicas e jurídicas, que através de suas ordens demonstram escolhas sobre operar em um ou outro ativo. Essas negociações, pelo modelo de análise técnica, são agrupadas para representar gráficos que possibilitam observar as tendências que um determinado ativo tem no mercado financeiro.

Normalmente, os gráficos da análise técnica aglutinam as ordens de negociações em: 1 minuto, 5 minutos, 15 minutos, 60 minutos ou 24 horas para análises para operações com mais de um dia de duração. Os agrupamentos mais comuns para as operações de *Day Trade*, operações de compra ou venda de um ativo no mesmo dia, são de 1 a 30 minutos, sendo o primeiro normalmente utilizado para operações chamadas de *scalping*, que possibilita obter maior lucro com as pequenas variações dos valores dos ativos, mas também que expõem o investidor a riscos maiores e por isto, necessita de mais conhecimento e técnica para operar.

Dentro desse contexto, este trabalho tem como proposta a criação de um modelo que auxilie na tomada de decisão nas operações de *scalping* no mercado financeiro, utilizando as mensagens postadas em uma rede social relacionadas à Bovespa, bolsa de valores brasileira, tratadas por PLN. Porém, distinto de outros trabalhos que realizam uma classificação binária das semânticas da rede social, realizei a proposta da vetorização das frases completas através

de *Word Embedding* gerando possibilidades de atuação no mercado financeiro com o uso de uma Rede Neural Recorrente (LSTM).

Será utilizado como base de dados o Twitter, que é uma Rede Social utilizada para postagens de pequenos textos, o que inclui discussões e opiniões sobre economia, fatos políticos, informações sobre ativos presentes na bolsa, dentre outros assuntos. Essa base de dados, que é aberta para consultas públicas através de uma API, é predominantemente textual, o que facilita na adoção de técnicas de Processamento de Linguagem Natural (PLN) e Inteligência Artificial.

1.1 DECLARAÇÃO DO PROBLEMA

Os trabalhos de Peng e Jiang (2016) e Li e Shah (2017) demonstram que o uso de técnicas de Processamento de Linguagem Natural em informações publicadas no Twitter auxilia na criação de modelos para a tomada de decisão na Bolsa de Valores. Essas técnicas realizam uma pré-classificação dos textos do Twitter de forma binária, demonstrando se o texto publicado pode ser bom ou ruim, em relação ao seu contexto, para indicação de compra ou venda do ativo, de forma a relacionar aos dados da variação do preço diário dos ativos e criar um modelo que auxilie nas tomada de decisão em operações de *Swing Trade,* o que reduz o espaço de informações a serem tratadas, o que reduz o processamento computacional, porém também reduz as possibilidades de ganho e atuação nos ativos pela própria simplificação dos estados.

1.2 PERGUNTA DE PESQUISA

As Redes Neurais Recorrentes (LSTM) são capazes de processar e classificar grandes variedades de informações sequenciais de modo automatizado. Com isso, elas são ideais para o tratamento e classificação de dados sequenciais temporais, que é o caso da variação de valores dos ativos da bolsa, com sua correlação com as informações do Twitter.

A hipótese desenvolvida é o uso de Redes Neurais Recorrentes do tipo LSTM para classificação de dados da série temporal do ativo mini-índice, acrescidos de dados coletados na rede social Twitter, com o intuito de criar um modelo que possa ajudar na tomada de decisões de operações de compra e venda do ativo mini índice.

Para tal, foi feita as perguntas abaixo, que buscarei responder com o desenvolvimento do presente estudo:

a) A criação de um modelo para tomada de decisão no mercado financeiro, apoiada nas mensagens relativas à Bovespa na rede social Twitter e tratadas pela técnica de Processamento de Linguagem Natural em um determinado período e considerando um determinado ativo, pode ter seus retornos comparados com propostas conservadoras de investimentos?
b) Comparando o modelo que propomos desenvolver com outros algoritmos já existentes, fazendo uma análise das métricas estáticas (acurácia, precisão, revocação e F1 *Score*), e analisando as mensagens por agrupamento de 5, 15 e 30 minutos qual modelo apresenta melhor desempenho?
c) Há interferência no resultado de acordo com o período em que houve a postagem no Twitter, ou seja, com relação ao agrupamento das mensagens em 5, 15 e 30 minutos é possível verificar se há maior influência quando as mensagens são postadas próximos ao fechamento do mini-índice?

1.3 OBJETIVO

O objetivo desse trabalho é a aplicação de técnicas de Inteligência Artificial correlacionando dados entre o Twitter e as informações da bolsa de valores para predição do valor de ativos do mercado financeiro brasileiro, visando atuações oportunas de compra ou venda nos investimentos no mercado de ações brasileiro.

1.4 JUSTIFICATIVA

Os rendimentos de renda fixa no mercado brasileiro não estão alcançando as expectativas dos investidores, por isso as pessoas estão buscando aplicações em renda variável, tal como a Bolsa de Valores, com o objetivo de aumentar seus rendimentos. Porém muitas pessoas já perderam, e continuam perdendo, suas economias com aplicações incorretas no mercado de ações, conforme Machado (2020) e Bittencourt *et al*. (2018). Dentre essas aplicações de alto risco podemos citar as aplicações em renda variável de *Swing trade* e *Day Trade*.

As técnicas de Inteligência Artificial como Redes Neurais e Aprendizado de Máquina, quando correlacionadas às notícias e publicações em Redes Sociais, apresentam-se como mecanismos com possibilidade de reduzir esses riscos, apontando as oportunidades de

investimentos mais assertivas de acordo com o histórico de eventos sociais e do comportamento dos ativos, conforme Peng e Jiang (2016) e Li e Shah (2017).

A atuação com pouco conhecimento no mercado de ações possibilita que especuladores experientes, ou operadores mal intencionados, levem todas as reservas financeiras de famílias, que por inexperiência acabam entrando em operações como se fosse uma simples loteria. Por isso, é necessária a criação de um mecanismo, neste trabalho concretizado como um algoritmo, que permita investimentos com menores riscos e maiores retornos.

1.5 METODOLOGIA

A metodologia utilizada no trabalho é a de experimentação, o que inclui os motivos pela escolha de PLN e LSTM, a descrição da base de dados do Twitter e da Bovespa utilizada nas simulações, a arquitetura da topologia e as aplicações utilizadas nas simulações dos algoritmos, e os critérios de medição e comparação de desempenho entre o algoritmo proposto e os Peng e Jiang (2016) e Li e Shah (2017), encontrados na literatura.

1.5.1 Motivação para uso de Processamento de Linguagem Natural (PLN)

Conteúdos coletados na Rede Social Twitter são usados na criação de modelos de auxílio na tomada de decisão no mercado financeiro, através da coleta dos textos postados. A escolha dessa se dá devido a maior parte das postagens feitas serem através de textos e disponibilizados pela rede. Esses textos demonstram a opinião do usuário sobre algum assunto, ou mesmo uma informação noticiada por algum meio de comunicação.

Com o intuito de usar esses textos no modelo desenvolvido, é necessário deixá-los interpretáveis pela máquina e para isso será necessário a realização de seu tratamento. O Processamento de Linguagem Natural (PLN) aparece como a ferramenta mais apropriada para esse procedimento, pois estudos como Bengio *et al.* (2003) e Mikolov *et al.* (2013) demonstram os avanços e importância da PLN dentro da área de processamento textual.

Conforme resultados de Bengio *et al.* (2003) e Mikolov *et al.* (2013) verifica-se que técnicas de PLN geram vetores a partir de palavras de uma base de dados, e esses vetores, quando colocados em um plano cartesiano, podem representar relação semântica.

Temos então que a PLN é de suma importância devido às suas características na realização de processamento textual, além de proporcionar uma base de conhecimento muito eficiente através de estudos variados realizados na área.

1.5.2 Motivação para uso de Redes Neurais Recorrentes (LSTM)

A análise dos preços de ativos no Bovespa em geral é feita através do agrupamento de ordens de negociação, gerando gráficos de períodos, permitindo fazer uso na demonstração gráfica dos períodos conforme escala desejada, o que pode ser usado para identificar tendências de alta ou baixa conforme movimentação dos preços dos ativos, de acordo com Pring (2002). O resultado desse agrupamento por um longo período é a geração de uma série temporal, principal variável a ser analisada pelo modelo de auxílio na tomada de decisão proposta por esse trabalho.

De acordo com estudos feitos por Hochreiter e Schmidhuber (1997), é possível constatar que a Rede Neural Recorrente (LSTM) tem como principal papel resolver problemas que ocorriam com a Rede Neural convencional, como por exemplo a dissipação do gradiente (*Gradient Vanishing*) e perda de memória durante o processamento de séries temporais, diante do avanço nos períodos temporais analisados, característica presente nos ativos da Bovespa.

A utilização da Rede Neural Recorrente (LSTM) na série temporal das informações da Bovespa é recomendável, visto suas características de processamento desse tipo de dados. Dessa maneira, os mecanismos propostos pela LSTM são de grande importância para se evitar problemas como a perda da memória de dados durante o processamento da série temporal, e todos esses mecanismos contidos em uma rede neural possibilitarão maior diversidade em sua aplicação nos dados da série temporal da Bovespa.

1.5.3 Metodologia de Simulação

A verificação da eficiência do modelo para auxílio na tomada de decisão será feito através dos princípios da Teoria de Dow, tendo a confirmação das tendências do ativo Bovespa com os resultados deste modelo. Com o intuito de melhorar essas indicações usando apenas a tendência, o modelo proposto neste trabalho vem como medida de confirmação, ou seja, de acordo com a tendência realizaremos a confirmação de compra e venda conforme o modelo proposto.

Este modelo traz recursos de Inteligência Artificial (IA), como Rede Neural Recorrente (LSTM), somada a aplicação de análise sobre as postagens feitas na Rede Social Twitter com recursos de Processamento de Linguagem Natural (PLN). Dessa maneira, como a Bovespa funciona em um determinado horário, deverão ser analisadas as postagens do Twitter feitas durante esse período, utilizados na análise de modelos *intraday*, ou seja, feitas em um mesmo

dia, de modo a confirmar se postagem feitas durante o dia inferem diretamente nos valores dos ativos.

A estratégia consiste na realização do processamento dos dados coletados no Twitter em períodos de 5, 15 e 30 minutos, somados aos dados da série temporal. Por exemplo, os 15 minutos de histórico irão ser utilizados para aprendizado para atuação no minuto subsequente (cego), com método de sucesso da operação de acordo com o histórico de transações realizadas no momento da operação cega, com a definição dos valores de referência para o investimento e somatória dos resultados.

1.5.4 Critério de Avaliação dos Resultados

Pode se dizer que o número de operações será elevado, pois as indicações de compra e venda ocorrerão em curtos períodos, ocorre então a necessidade de contemplar a somatória dos custos referentes às operações realizadas por este modelo, usando como base o valor médio cobrado pelas corretoras no Brasil para esse tipo de operação.

Ao final será possível identificar, através dos resultados das comparações entre essa somatória de operações com resultados apresentados em outros trabalhos, como Peng e Jiang (2016) e Li e Shah (2017). Através disso, será possível calcular a eficiência do modelo desenvolvido com base nas operações *intraday*, em relação a outros modelos propostos por outros autores.

Com isso, a avaliação dos resultados deste modelo será feita considerando um cálculo dos valores positivos e negativos nos períodos avaliados, comparado com os ganhos e perdas de outros modelos e algoritmos. Outra comparação a ser realizada é a análise de tendência apenas da série temporal, de forma a verificar se o modelo contribui de forma positiva para as pesquisas encontradas na literatura.

1.6 ESTRUTURA DO TRABALHO

Este trabalho está organizado em seis capítulos. No primeiro capítulo há um resumo do tema abordado e os objetivos propostos. No segundo é apresentada a fundamentação teórica das metodologias em que esse trabalho está baseado, como os conceitos básicos de Mercado Financeiro, Redes Neurais, técnicas de Processamento de Linguagem Natural (PLN) e sua contextualização com a vetorização de textos, *Word Embedding*. No terceiro há uma apresentação dos trabalhos correlatos. No quarto é apresentada a modelo proposto. No quinto

apresentamos os resultados obtidos, e finalmente, no sexto e último capítulo apresentamos a conclusão.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta as bases, conceitos e metodologia usados nesse trabalho e foi segmentado por Mercado Financeiro, Processamento de Linguagem Natural e Redes Neurais Recorrentes.

2.1 MERCADO FINANCEIRO

O sistema financeiro desempenha papel essencial dentro da sociedade econômica moderna, pois é a partir dele que é feito o estudo de alocação dos recursos da economia, baseado nos desejos de consumo da sociedade, sendo que nele temos as decisões econômicas de como destinar os recursos para o consumo, poupança e investimento (CVM, 2019).

As primeiras sociedades tinham como objetivo apenas bens de consumo e não existia o fluxo financeiro com moedas como há atualmente. Com isso, os negócios aconteciam por trocas entre os bens, método conhecido como escambo. O foco dessas negociações era o de suprir a alimentação e necessidades básicas, visando a sobrevivência, além de lidar com precariedade das formas de armazenamento, o que tornava muito frequente essas negociações. Outro ponto é o fato de que a população da época não conseguia estocar ou conservar os alimentos, isso fazia com que as negociações fossem muitas vezes para consumo em pouco tempo ou até mesmo imediato, fato que deixou ainda mais forte a prática de escambo (RUDGE; CAVALCANTE, 1996).

Com o desenvolvimento da sociedade e consequentemente do comércio, começaram a se desenvolver as moedas, os comerciantes, os banqueiros e as intuições para negociações e transações financeiras. Isso ajudou a formar as primeiras bolsas de valores na Europa e, a partir do século XIX e a ampliação dos empreendimentos comerciais possibilitou as negociações através do comércio. Essa prática ganhou cada vez mais força tendo importância internacional, e sendo essa prática uma mediadora de grandes negociações (RUDGE; CAVALCANTE, 1996).

Atualmente destacam-se como principais bolsas de valores no cenário mundial as bolsas de Nova York, Londres, Paris, Tóquio e Shanghai. No Brasil temos a Bovespa, ou B3, resultante da fusão das bolsas de valores brasileiras.

2.1.1 Bolsa de Valores - Bovespa

No Brasil, a Bovespa é instituição no segmento de intermediação para operações de mercado de capitais, controle e gerenciamento para negociações de ativos, derivativos, títulos de renda fixa, títulos públicos federais, derivativos financeiros, moedas à vista e commodities agropecuárias. Além disso, a Bovespa exerce o papel de fomentar o mercado de capitais brasileiro.

Já o Índice Bovespa avalia o desempenho dos ativos de companhias na B3 através da criação de uma média de desempenho dos ativos mais negociados pela Bolsa de Valores (CVM, 2019). Atualmente, desde grandes empresas de investimentos até pessoas físicas podem realizar negociações na bolsa de valores, para isso basta ter um cadastro devidamente aprovado em um banco ou corretora que tenha operações de compra e venda na B3, e seja o intermediador desses clientes permitindo que tenham acesso aos pregões e operações (CVM, 2019).

2.1.2 Análise sobre o mercado de ativos

Caso uma pessoa desejar aplicar ou investir no mercado de ativos, é necessário preparação e estudos voltadas a essa área, de forma que a obtenção desse conhecimento fará com que o investidor reduza os riscos possíveis quando realizar uma operação com determinado ativo. Isso porque os riscos estão sempre presentes nesse tipo de transação, pois os mercados são muito voláteis e estão expostos a vários fatores econômicos e políticos que podem influenciar na alta ou baixa de um ativo (CVM, 2017). Para tanto, existem métodos de realizar as transações que fazem com que o investidor tenha seu rumo traçado na busca de lucros, esses métodos vão desde uma análise do ativo antes da compra, sua liquidez, a procedência da empresa, sua saúde financeira, entre outros.

A falta desse tipo de compreensão por parte dos investidores acaba sendo um dos principais motivos de fracasso, pois muitas vezes algumas pessoas pensam que apenas olhar os gráficos referentes ao ativo ou a comparação desse ativo com outros da Bolsa de Valores irá embasar para ser um bom "*trader*" (termo usado para definir o investidor do mercado financeiro que busca ganhar dinheiro com operações de curto prazo, aproveitando-se da volatilidade do mercado e buscando retornos financeiros através da compra e a venda de ativos) e com bons resultados de retorno (CVM, 2017).

Dessa maneira, uma companhia que tem seus ativos bem valorizados ao longo do tempo e demonstra uma boa gestão, possui maior confiança por parte dos investidores, que acreditam

que a companhia é estável e continuará sempre crescendo. Com isso, é relevante a importância de analisar de forma ampla uma empresa, juntamente com seus resultados e seus objetivos ao longo do tempo, caso o foco dos resultados seja para longo prazo.

2.1.3 Tipos de Análise de Investimento

Os investidores em geral utilizam duas formas que objetivam avaliar um ativo e descobrir pontos de compra e venda, a análise técnica e a análise fundamentalista.

2.1.3.1 Análise Fundamentalista

A Análise Fundamentalista, ou sobre os fundamentos da empresa, utiliza como variáveis os dividendos, o lucro e o valor contábil de uma empresa. Essas variáveis são fundamentais neste estudo e normalmente são excluídas de uma Análise Técnica. Constata-se na Análise Fundamentalista que não somente os padrões de preços do passado podem influenciar o destino de um ativo, mas sim que é possível analisar o que acontecerá com os ativos de uma empresa através dos resultados entregues pela empresa detentora dos ativos (SIEGEL, 2015).

De acordo com Damodaran (2015), um dos objetivos da análise fundamentalista é se evitar a compra de ativos a um preço superior ao seu valor potencial ou valor justo, o que gera oportunidades baseado na saúde financeira da empresa e seu real valor, podendo com isso reduzir riscos ao se encontrar ativos “baratos”, considerando-se o valor do ativo e o preço de mercado da empresa.

Reilly e Brown (1997) afirmam que com base nos fundamentos de uma empresa há possibilidades de entender seu passado financeiro e com isso realizar projeções para seu futuro. Com isso é possível escolher empresas que possuam um maior potencial de crescimento e um menor risco de perda, ou seja, entender o real valor da empresa e sua projeção de rendimento, e não apenas o preço de seus ativos no momento.

A Análise Fundamentalista é em geral usada para investidores que pensam em estratégias de médio e longo prazo, comprando empresas com valor abaixo do analisado e esperando uma valorização futura, esse tipo de investidor é também conhecido como “*buy and hold*”, do inglês “*comprar e segurar*” (CVM, 2017). De acordo com CVM (2019), esse tipo de análise possui alto grau de complexidade, pois requer desde a avaliação da saúde financeira como também verificar outros detalhes como sua gestão, visão, participação e reputação da marca no mercado.

Segundo Gitman (2001), pode-se afirmar que o valor de um ativo é igual ao valor atual de todos os benefícios futuros que se espera que ele ofereça, em outras palavras, todos os seus retornos baseados em uma entrega futura. Esses benefícios podem incluir a distribuição de dividendos e a valorização, dentro de um horizonte temporal infinito. Complementando o conceito da análise fundamentalista Damodaran (2015) ressalta que um ativo deve ser avaliado tomando por base seu fluxo de benefícios futuros, considerando a influência do ambiente interno e externo à empresa.

2.1.3.2 Análise Técnica

A Análise Técnica considera que sempre haverá repetições de comportamentos no preço de um ativo ao longo do tempo. Assim, os investidores acreditam que os ativos irão se mover em tendências e os preços dos ativos devem seguir essa tendência (CVM, 2017). Essa repetição é frequentemente atrelada a conceitos psicológicos do mercado, permitindo medições nas previsões com base em emoções e entusiasmo, constatando então que a realização de operações com o uso dessa análise pode maximizar ganhos e diminuir riscos (SIEGEL, 2015).

É um tipo de análise geralmente recomendada para o retorno no curto prazo buscando ganhos de forma rápida, tendo suas análises e observações feitas através de gráficos que associam as imagens geradas com as movimentações dos preços dos ativos, identificando determinados padrões em termos como bandeiras, flâmulas, ombro-cabeça-ombro, entre outros. De acordo com Siegel (2015), muitos economistas e acadêmicos têm repúdio quanto a análise técnica, pois acreditam que esse tipo de análise não implica em um respaldo econômico do cenário em que se encontra a empresa, associando esse tipo de análise a astrologia.

Temos então que a premissa básica da análise técnica é de que todas as informações estão representadas nos gráficos referentes as movimentações de preço de um ativo e que, na medida em que eles traduzem o comportamento do mercado, os investidores conseguem realizar suas operações conforme as oportunidades que aparecem, podendo essa operação ter a duração que investidor desejar (CVM, 2019). O uso de gráficos de preços, ligado com indicadores técnicos, é de fato, a base das maiorias das transações dos investidores. Através desta união, analistas e investidores investigam a trajetória passada dos preços visando encontrar algum padrão, para que, futuramente, este padrão seja capaz de ser replicado em novas operações.

2.1.4 Operações *Day Trade*

Os conceitos de Análise Técnica também são empregados em operações financeiras num curto período. Habitualmente, chamado de *day trade*, esta modalidade de negociação permite ao investidor a possibilidade de lucros rápidos a partir da volatilização dos preços através de operações de compra e venda obrigatoriamente encerradas no mesmo dia.

Considerando-se que as operações *intraday / day trade*, operações dentro de um mesmo dia, o objetivo é trabalhar com variações mínimas do preço de um ativo, o que ocasiona um maior risco e devem ser feitas por investidores com um certo nível de conhecimento e experiência no mercado. O período em que é usado para esse tipo de operação pode variar de acordo com as configurações usadas por cada investidor, no qual costumam usar os tempos gráficos, períodos de 1 a 60 minutos para análise do ativo, propondo uma tomada de decisão para o período seguinte ao analisado. Dessa forma as operações de *day trade* são operações de movimentos curtos, rápidos e sequenciais (CVM, 2017).

O lucro de uma operação *day trade* é medido entre a diferença do valor médio de venda em relação ao valor médio de compra subtraído dos gastos para realizar a operação. Sendo assim, em minutos, o investidor procura captar o maior número de oportunidades a partir da especulação do ativo financeiro. O investidor pessoal que realiza esta modalidade de operações de trading é comumente denominado de *day trader* (CVM, 2017). Dentre os ativos que mais são usados em operações *day trade*, há destaque maior para os minicontratos futuros de dólar e do próprio índice Bovespa, isso devido ao baixo custo necessário para se operar nesses ativos. Neste trabalho, o ativo usado para as operações *day trade* é o Minicontrato Futuro Ibovespa, também conhecido como mini-índice (CAMPOLINA, 2019).

2.1.5 Minicontrato Futuro Ibovespa

O Minicontrato Futuro Ibovespa é um ativo derivado do próprio Índice Bovespa, mede o desempenho das empresas mais negociadas no mercado, e pode ser negociado em bolsa de valores diferentemente do próprio Índice diretamente, sendo apenas uma referência para os investidores. O Índice Bovespa funciona como um termômetro da bolsa de valores e o Minicontrato é uma fração desse ativo, daí o termo mini-índice. Ambos têm as mesmas características e podemos citar duas delas: a primeira refere-se ao vencimento, pois a cada 2 meses vence um código de negociação, e a segunda refere-se a análise feita pelos investidores

que fazem uso de ambos para previsões futuras de acordo com as expectativas do mercado (CAMPOLINA, 2019).

A partir desse indicador, o mercado cria expectativas sobre movimentos futuros do mercado de ativos, e quando negociamos o mini-índice estamos negociando essa expectativa. Muitos investidores usam esse ativo para criar proteções para a seus investimentos, mas também é usado para operações *day trade* devido à alta volatilidade e baixo custo exigido pelas corretoras para que possa ser negociado (CAMPOLINA, 2019).

O Minicontrato Futuro Ibovespa é a porta de entrada para os investidores ingressarem no mercado de Bolsa de Valores, principalmente pessoas físicas e pequenas empresas, devido ao valor baixo exigido pelas corretoras para operar nesse ativo. Para se operar esse ativo o investidor precisa depositar um valor na corretora, que se refere a uma garantia dos mecanismos estabelecidos nas regras da Bolsa de Valores, baseado na média das variações dos ativos e dos conceitos do *circuit break,* estrutura que em momentos muito turbulentos age de forma rígida parando todas as operações por alguns minutos até que as ordens de compra e venda se normalizem (CAMPOLINA, 2019).

Como há esse mecanismo de alavancagem, a bolsa de valores realiza o ajuste diário evitando grandes distorções no valor de um dia para o outro. No entanto, para que o investidor use o ajuste diário é necessário ter depositado como garantia uma quantia equivalente ao valor do contrato do mini-índice. Com isso a maioria das operações com o Minicontrato acontece da forma *day trade*, evitando assim o ajuste diário.

Oportuno ressaltar que tanto o Índice Bovespa quanto o Minicontrato recebem tratamento diferenciado dos outros ativos da Ibovespa. Quando pensamos na compra e venda de ações sabemos que a transação é negociada pelo valor em moeda, o que não ocorre com o mini-índice que tem o valor fixado por meio de pontos. Existe algumas diferenças entre eles, por exemplo que cada ponto no índice tem valor de R$ 1,00, e o valor do mini-índice é de 20% do índice, ou seja R$ 0,20 por ponto. Outra característica é que as negociações para o contrato do índice precisam ter um lote de cinco contratos, já para as operações do mini-índice, é necessário apenas um contrato. Isso facilita a entrada dos investidores através do minicontrato do índice Bovespa (BM&FBOVESPA, 2017).

2.1.6 Teria de Dow

Há diversas maneiras de realizar operações no mercado financeiro usando análise técnica, formas que vão desde métodos mais antigos até métodos mais recentes. Ao se analisar no contexto histórico um ativo da bolsa de valores, é possível realizar agrupamentos por períodos de acordo com os objetivos do investidor, através da demonstração gráfica dos movimentos realizados com as negociações de mercado, tendo como objetivo central a identificação da movimentação dos agrupamentos de negociações.

Charles Henry Dow, no início do século XX, usou o comportamento do mercado de ações como um barômetro para checar às condições de mercado, sendo que seu objetivo não era usar isso como base para a previsão dos preços de um ativo, mas sim constatar que a maioria dos ativos segue uma tendência, na maior parte das vezes. Todos esses métodos foram estudados e aperfeiçoados dando origem ao atual índice Dow Jones, usado para medir o mercado de ações de empresas industriais e de transportes (antes apenas ferroviária, mas com os avanços da área de transportes se adequou com outras empresas) (RHEA, 2013).

De acordo com Pring (2002), a Teoria de Dow é o método mais antigo que busca identificar para onde os agrupamentos de negociações estão apontando, demonstrando tendências no mercado de ativos. Conforme o autor, após uma tendência ter sido estabelecida só será parada com uma reversão. Com isso, sua preocupação não é apenas com a direção da tendência, que pode ser de alta ou baixa, sua duração ou tamanho final, mas sim determinar mudanças no movimento primário (PRING, 2002).

Para isso, é necessário ter os preços de um determinado período juntamente com o volume das transações a serem analisadas (PRING, 2002). Com base nisso, seu entendimento exige interpretação e entender a aplicação dos seis princípios que regem a Teoria de Dow, sendo:

1. As médias dos preços descontam tudo: o primeiro princípio da Teoria de Dow é que o preço é predominante e desconta tudo, isso inclui que mesmo em momentos de mudanças do movimento causado pela emoção dos investidores ou qualquer outro ponto, tudo será descontado pelo preço, estando nele agregado tudo que é conhecido e previsível no mercado, sendo que as únicas exceções desse princípio são os atos da natureza, como as tragédias naturais, e ao longo do tempo esses fatos também serão avaliados e considerados no preço dos ativos, conforme Pring (2002).
2. O mercado tem três movimentos e se move em tendências: esse princípio considera que o mercado tem três tendências principais, de acordo com o período avaliado, o

movimento primário, secundário e terciário. O movimento primário é o mais importante, sendo os longos períodos de alta ou baixa que predominam no mercado e que podem durar anos, conhecidos como "*bear market"* quando se referem ao movimento de baixa, e "*bull Market"* quando se referem ao movimento de alta. Já no movimento secundário, é possível constatar uma movimentação intermediária à primária que refaz todo o movimento primário anterior, o que gera dificuldade aos analistas de identificarem se uma tendência continua em movimento primário ou está iniciando o secundário. Por fim, o movimento terciário é o de menor duração, não tendo influência sobre o longo prazo, sofrendo influência direta de fontes externas, como notícias, por exemplo (PRING, 2002).

3. Linhas indicam movimento: Segundo Rhea (2013), ao analisar os estudos de Dow, foi constatado que existe uma linha com o movimento de preços de duas a três semanas ou mais, e que os preços dessas médias se movem dentro de uma faixa de cerca de 5%. Ocorre que essa indicação mostra acúmulo, principalmente para os conhecedores do mercado dentro dessa média, o que pode favorecer as negociações.
4. Relação preço/volume fornece antecedentes: Esse princípio demonstra que há sinalização de maiores volumes de negociações indicando corridas, momentos de extrema compra ativos, e menores volumes indicando contração e declínios, fato que deve ser usado apenas como uma espécie de informação base, pois a Teoria de Dow define que ocorre a relação do preço controlando tudo, inclusive possíveis reversões de tendência (RHEA, 2013).
5. Ação de preço determina a tendência que dura até ser substituída por outra: esse princípio reforça que o preço determina a tendência, indicações de alta acontecem quando corridas sucessivas adentram os topos, diante do fundo de queda de uma tendência secundária acima do fundo anterior. Ou seja, várias altas seguidas podem ser interrompidas por uma reversão de tendência. Diante disso, é possível afirmar que uma tendência é predominantemente feita pelos agrupamentos de preços de comercialização de um ativo, só sendo revertida com uma nova tendência (PRING, 2002).
6. As médias devem confirmar: de acordo com esse princípio é necessário um segundo elemento para confirmação da tendência do primeiro, assim existe relação entre o índice industrial e o índice de transportes do Dow Jones, onde caso haja uma tendência de alta em um desses índices, pode-se usar o outro para confirmação da tendência (PRING, 2002).

7. Constatando por exemplo que se há uma tendência de alta em um desses índices pode-se usar o outro índice para confirmar essa tendência. Ocorre que todos esses mecanismos sinalizam para que caso desejável verificar se há uma tendência de alta ou baixa em um ativo, o uso de uma segunda variável pode ser aplicada com o intuito de confirmação dessa tendência (PRING, 2002).

2.2 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

O Processamento de Linguagem Natural (PLN), do inglês *Natural Language Processing* - (NLP), é um campo de pesquisa da Inteligência Artificial (IA) considerada como uma sub área da IA, cujo objetivo é explorar a forma como os computadores e máquinas entendem e manipulam textos, ou a fala em linguagem natural do ser humano. Ao realizar essa manipulação e posterior interpretação, o computador poderá realizar operações (CHOWDHURY, 2003).

Segundo Covington, Barker e Szpakowicz (1994) a PLN consiste em modelos computacionais para tarefas que dependem de informações que são expressas em alguma língua natural. Como exemplo podemos citar tarefas associadas à tradução de textos, interpretação de textos e busca de informações em documentos. É possível afirmar que PLN é um conjunto de atividades que visa a extração de informações importantes de bases de dados não estruturadas, objetivando o entendimento de conteúdo textual incomum, de acordo com Feldman e Sanger (2007).

Constata-se uma das importâncias da PLN é a compreensão automática de textos em linguagem natural humana, de forma que possa ser traduzido para máquina e repassado para o humano de forma compreensível. Com isso, temos que PLN é um mecanismo para capturar qualquer tipo de interação que seja feita através de uma língua humana, como textos ou via voz, realizando o processamento dessa em algo entendível pelo computador e retornar a resposta para o humano.

Segundo Benevenuto, Ribeiro e Araújo (2015), atualmente PLN é usado em diversas aplicações, como exemplo:

a) Conversação iterativa em *call centers*, respondendo perguntas padrão de requisições de usuários;
b) Tradução de idiomas, como "*Google Translator*";
c) Processamento de palavras, que empregam PLN para a correção gramatical;
d) Assistentes pessoais, como "OK Google", "Siri", "Cortana" e "Alexa".

2.2.1 Pré-Processamento de Texto

Na era da informação que vivemos, na qual a quantidade de dados é crescente, a aplicação de aprendizado de máquina nesses dados é muito importante para processamento e interpretação dos dados. Quando pensamos em redes sociais, o aumento na quantidade de dados é ainda maior principalmente devido ao uso exponencial das redes sociais, destaque para os

dados gerados na rede social Twitter que são usados neste trabalho. Esses dados são, em sua maioria, expressos em linguagem natural, isso torna indispensável o trabalho com esses dados para torná-los entendíveis pela máquina criando padrões que possam ser utilizados para tomada de decisões.

Ao capturar textos, ou qualquer linguagem natural para processamento da máquina, muitas vezes essa informação se encontra de forma não estruturada, o que pode dificultar a criação de modelos de predição com o PLN. Quando isso ocorre, é necessário um tratamento prévio dos dados, conhecido como pré-processamento sintático, que funcionará como uma filtragem, tornando esses dados estruturados para possibilitar a ação da máquina e organizando as sentenças para que se obtenha sentido gramatical (BENEVENUTO; RIBEIRO; ARAÚJO, 2015).

O pré-processamento de sintaxe reduzirá grande quantidade de dados inúteis, possibilitando análise e interpretação da linguagem natural repassada para o processamento, onde permanecerão apenas informações úteis para o processamento computacional. Portanto, a tarefa do PLN é importante para estruturação de uma fonte de informação não-estruturada, devido a pontos como a complexidade e diversidade da linguagem humana (DALE; MOISL; SOMERS, 2000).

Algumas das funções que são feitas pelo PLN e são aplicados neste trabalho são:

a) *Tokenização*: Procedimento que transforma cada palavra em um token decompondo o documento em cada termo que o compõe. Para demarcar cada termo, comumente são usadas quebras de linhas, espaço em branco, símbolos, entre outros. Neste trabalho a demarcação foi através do espaço em branco entre as palavras das frases coletadas no Twitter.
b) Normalização: Processo que realiza ajustes como a transformação de letras maiúsculas em minúsculas, remoção de caracteres especiais, remoção de pontuação. Esses ajustes na base de dados ajudam a otimizar o modelo.
c) Remoção de *Stopwords*: *Stopwords* (traduzindo do inglês, “palavras irrelevantes”) são palavras que possuem sua presença elevada na frequência dos textos. Semanticamente, elas não acrescentam valor e por isso são removidas da análise. Como exemplo temos pronomes, artigos indefinidos e definidos, preposições, verbos auxiliares e conjunções.
d) *Stemização* (do inglês “derivação de algo”): Sua função é a de transformar a palavra que está de seu estado flexionado para seu radical. Como exemplo as palavras “estudar”, “estudos”, “estudarei”, que estão na forma flexionada, se transformariam em “estud”.

Esse mecanismo usado em PLN está atrelado diretamente ao estudo de morfologia, área que refere à composição das palavras e sua natureza.

e) TF-IDF (*Term Frequency–Inverse Document Frequency*, traduzindo do inglês “termo frequência - inverso da frequência nos documentos”*)*: Conforme Joachims (1997), essa técnica avalia a frequência local e a frequência geral de determinado termo e também um fator de ponderação de forma para que aqueles que mais surgem nos documentos tenham uma representatividade menor. O “*Term Frequency*” (TF) realiza o cálculo da frequência referente a presença de um termo é demonstrado no documento. Dessa forma, de acordo com a frequência é possível notar a importância do termo.

$$TF\ (i,j\ ) = \frac{Ocorrências\ de\ palavras\ no\ documento}{Total\ de\ palavras\ no\ documento} \quad (1)$$

No qual i é o índice para a palavra e j é o índice para todo o corpus do texto.

Já o *Inverse Document Frequency* (IDF), calcula a frequência com que o termo está presente em todos os documentos da base de dados. Para esse caso quanto maior o IDF teremos menor importância para o termo.

$$IDF\ (i) = \frac{Número\ total\ de\ documentos\ presente\ no\ corpus}{Número\ de\ documentos\ contendo\ a\ palavra\ analisada} \quad (2)$$

No qual i é o índice para a palavra.

Assim, para o cálculo final do TFIDF do documento, é necessário realizar o produto entre os resultados de TF e IDF, conforme abaixo:

$$TFIDF = \ TF(i,j) * IDF(i) \quad (3)$$

No qual i é o índice para a palavra e j é o índice para todo o corpus do texto.

2.2.2 Word Embedding

Word Embedding, também conhecido como vetorização, é uma das técnicas mais importante do processo vetorização de textos, pois é a partir dela que é possível a representação de palavras, termos ou frases em forma numérica, o que permite a aplicação de algoritmos para processamento ou operações matemáticas, objetivando a extração de informações importantes. Nesse processo vetorização de textos, é feito a transformação da palavra em um vetor contendo números reais que representam palavras e ou documentos em um espaço vetorial predefinido (DALE; MOISL; SOMERS, 2000).

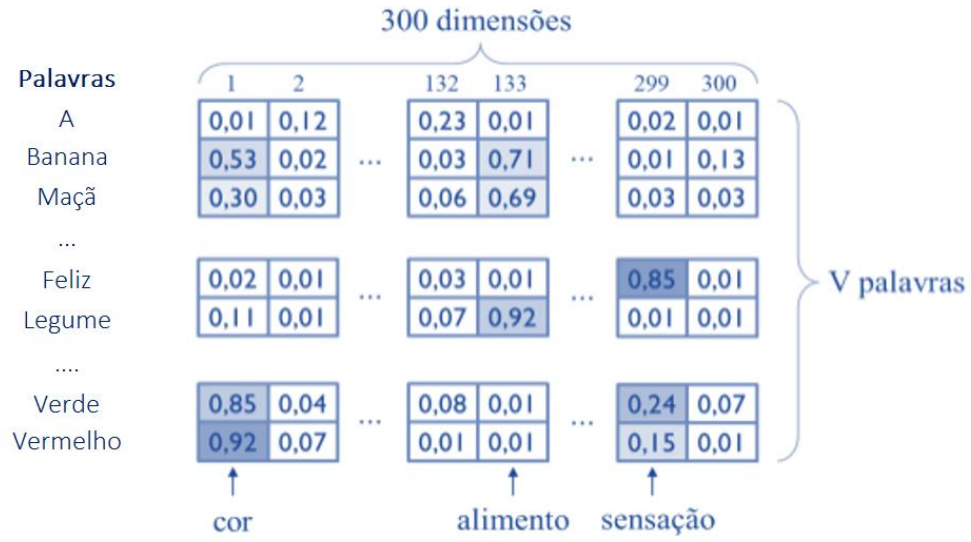
Esses vetores, nos quais as palavras são transformadas nesse processo vetorização de textos, tem seus valores ajustados por uma rede neural. Neste processo vetorização de textos de aprendizado as palavras são treinadas e definidas de acordo com o contexto no qual são inseridas. O resultado disso é o conhecimento morfológico, sintático e semântico, que são capturados pelos *Embedding* (HARTMANN et al., 2017).

Os estudos para a estruturação do conceito de *Word Embedding* relacionando os vetores com conhecimento morfológico, sintático e semântico foram conduzidos inicialmente por (BENGIO et al., 2003), que realizaram a vetorização de uma forma esparsa (conhecida como "*One Hot Encoding*"), pois cada palavra representava um índice no vetor e com isso quanto maior o dicionário, maior seria a dimensão do vetor, e posteriormente evoluídos com os estudos de (MIKOLOV; LE; SUTSKEVER, 2013), que por sua vez usaram a vetorização de forma esparsa, conforme (BENGIO et al., 2003), para a aplicação de modelos matemáticos para transformar esses vetores em uma estrutura densa.

Essa capacidade dos *Word Embeddings* para modelar o contexto das palavras, capturando características linguísticas tão importantes como a semântica, permite que encontremos palavras próximas no espaço vetorial usando cálculos simples, tal como a medida de similaridade de cosseno. Dessa forma, palavras semelhantes são representadas por vetores quase semelhantes colocados muito próximos em um espaço vetorial (GOLDBERG; HIRST, 2017).

Na Figura 1 é demonstrada a representação distribuída de palavras, onde cada dimensão representa uma característica (*feature*), permitindo através de cálculos de distanciamento, relatar uma similaridade entre palavras que pertencem ao mesmo grupo. Por exemplo: fruta, banana e maçã possuem valores elevados na mesma dimensão, o que poderia estar representando o grupo dos alimentos, assim como cada palavra possuí seu valor conforme sua classe gramatical.

Figura 1 - Demonstração de palavras distribuídas em Word Embedding.



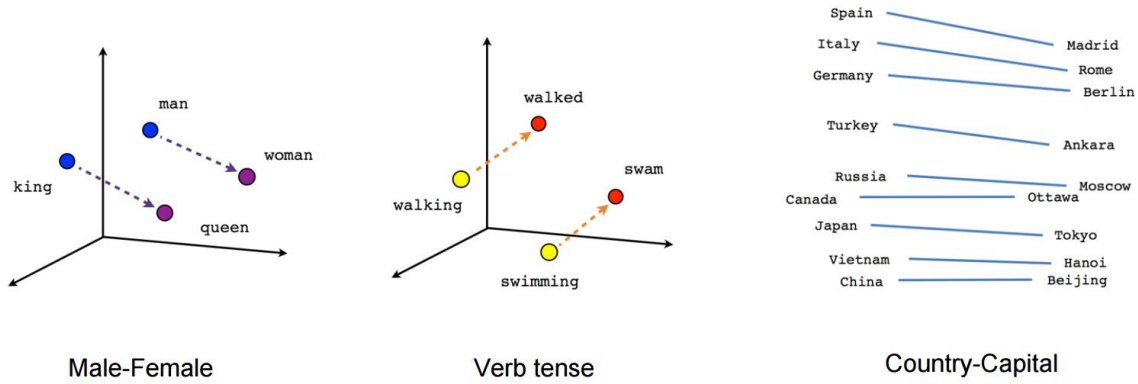
Fonte: Elaborado pelo autor.

Segundo Wittgenstein (1953), para buscar a extração semântica dos textos é necessário entender que o significado da palavra está no uso dela na linguagem, e não no que ela está de forma individual rotulando. Baseado nisso, constata-se que para descobrir o sentido da palavra é necessário descobrir o uso da palavra dentro do contexto e interpretar o uso que é feito dessa palavra, pois as palavras podem ter diferentes significados em diferentes regiões, dentro de culturas e até dentro de áreas profissionais diferentes.

Para alcançar e extrair valores semânticos dos textos presentes nas postagens do Twitter, há a necessidade de se realizar o processo de extração e vetorização dos textos, o que possibilita a aplicação de modelos matemáticos sob esses textos convertidos em vetores (DALE; MOISL; SOMERS, 2000). Existem diferentes processos de criação de *Word Embeddings* como o *GloVE*, *Word2Vec*, *FastText* e entre outros.

A Figura 2 demostra o resultado da representação vetorial num espaço gráfico, onde fica exposto em uma primeira demonstração as relações entre gêneros masculino e feminino, e ao lado, na segunda demonstração, a relação de tempo verbal, passado e presente, de palavras treinadas e posteriormente representadas pela técnica de *Word2vec* de (MIKOLOV; LE; SUTSKEVER, 2013).

Figura 2 - Demonstração de relação de palavras treinadas por Word2Vec.



Fonte: TensorFlow, 2016.

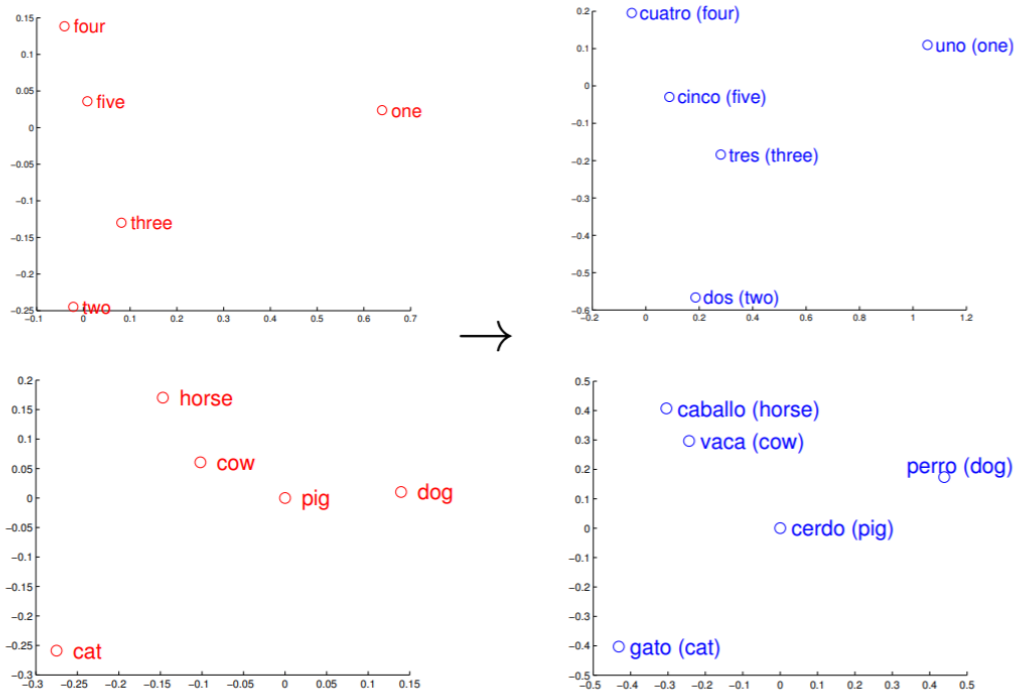
2.2.3 Word2Vec

Conforme Mikolov, Le e Sutskever (2013), o *Word2Vec* é um método que tem sua concentração em estatística para o *Word Embedding* de forma eficaz a partir de um corpo de texto. O objetivo do *Word2vec* é construir essa representação vetorial para textos, de forma não supervisionada, isto é, o conjunto de dados para treinamento não precisa ser rotulado. Com isso o algoritmo terá o objetivo de associar cada texto que estará representado num vetor esparso a um vetor denso correspondente. O objetivo do *Embedding* é utilizar palavras próximas para garantir que as representações vetoriais possuam informações semânticas e sintáticas.

Essa representação permite que apareça alguns resultados interessantes quando fazemos aritméticas com vetores que representam as palavras, estamos mencionando por exemplo que se tivermos duas palavras do mesmo gênero, nesse caso gênero masculino, como “REI” e “HOMEM”, e se caminharmos numa direção no espaço vetorial, podemos encontrar a relação entre a palavra “HOMEM” para “MULHER” e consequentemente de “REI” para “RAINHA”, ambas alterando o sentido para o gênero feminino (MIKOLOV; LE; SUTSKEVER, 2013). Isso faz sentido pois é retirado o sentido masculino da palavra e adicionando o sentido feminino, com isso a relação semântica se torna muito forte em *Word2Vec*.

Isso permite uma ampliação no planejamento e objetivos do desenvolvimento, pois a relação entre duas palavras pode fornecer informações sobre a relação entre outras duas palavras. Na Figura 3, há outros exemplos e demonstração da inserção desses vetores em um outro idioma, caso em que a similaridade permanece entre as posições dos vetores.

Figura 3 - Demonstração da representação vetorial de palavras.



Fonte: Mikolov, Le e Sutskever, 2013.

Além do funcionamento descrito acima, o *Word2Vec* possui duas formas de treinamento: *Skip-Gram* e *Continuous Bag of Words* (CBOW), ambos apresentados abaixo.

2.2.3.1 Continuos Bag of Word (CBOW) e Skip-Gram

O modelo *Continuos Bag of Word* (CBOW) realiza de forma contínua o treinamento do sequenciamento das palavras e tem o objeto de realizar a predição da palavra atual que melhor se encaixe, com base em um contexto. Teoricamente funciona como uma espécie de preenchimento de lacunas em uma frase. A dimensão da camada oculta e da camada de saída permanecerá a mesma. Somente a dimensão da camada de entrada e o cálculo das ativações da camada oculta serão alterados. Se tivermos 5 palavras de contexto para uma única palavra-alvo, teremos 5 vetores de entrada 1x V. Cada um será multiplicado com a camada oculta V x E, retornando vetores 1 x E . Todos os 5 vetores 1x E serão calculados como elementos para obter a ativação final que será, então, alimentada na camada softmax (MIKOLOV; LE; SUTSKEVER, 2013).

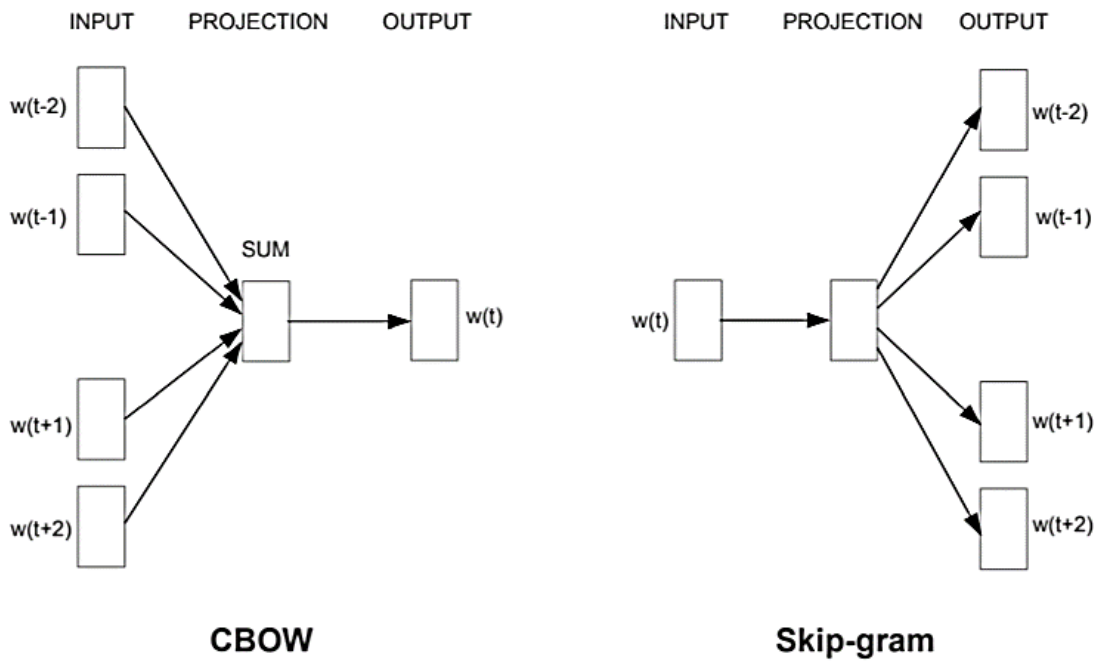
Já o segundo modelo é o *Skip-Gram*, que estima a probabilidade de um contexto passando para a rede apenas uma palavra, funcionando de forma oposta ao funcionamento do

CBOW, ou seja, uma palavra pedirá seu contexto com base no modelo de dados que esteja sendo estudado. Entendendo a matemática desse modelo, temos a diferença logo no início do treinamento, onde se usa o vetor "*One Hot Encoding*", com dimensão 1 x V da palavra como entrada onde V é o número de palavras do vocabulário.

A única camada oculta terá a dimensão V x E, onde E é o tamanho da representação numérica da palavra e é um hiper parâmetro. A saída da camada oculta será da dimensão 1 x E, com uma camada softmax. As dimensões da camada de saída serão de 1 x V, onde cada valor no vetor será a pontuação de probabilidade da palavra-alvo nessa posição. Em seguida, calcularemos os vetores de erros correspondentes a cada palavra-alvo (no exemplo, 4 vetores) e executaremos uma soma por elemento, reduzindo para 1 vetor 1 $x\ V$. Em seguida os pesos da camada oculta serão atualizados com base nesse vetor de erro acumulado (MIKOLOV; LE; SUTSKEVER, 2013).

Mikolov, Le e Sutskever (2013) constataram a superioridade do *Word2Vec* relacionado aos demais algoritmos existentes na época, e que o modelo *Skip-Gram* é mais eficiente em previsão de contextos quando há base de dados maiores, apresentando resultados melhores em comparação com o CBOW quanto a semântica e análise sintática, o que foi confirmado por Hartmann et al. (2017) em sua pesquisa de *Word Embeddings* em português.

Figura 4 - Arquitetura CBOW e Skip-Gram.



Fonte: Mikolov, Le e Sutskever, 2013.

Na Figura 4 é demonstrado a arquitetura de CBOW, explicado no item 2.2.3.1, no qual recebe várias palavras de forma vetorizadas, que seriam o contexto, com isso é feito uma média dos vetores referentes a esse contexto, e com isso o processamento e tentativa de predição da próxima palavra. Já na arquitetura de *Skip-gram* há entrada de apenas uma palavra, realizando assim a predição das palavras que estão ao seu redor (MIKOLOV; LE; SUTSKEVER, 2013).

A Figura 4 também permite o entendimento de CBOW e *Skip-Gram* como $W(t-1)$ e $W(t-2)$, como sendo a primeira e a segunda palavra que antecede a palavra a ser prevista ou palavra passada como entrada na rede, nessa ordem. Entendendo-se $W(t+1)$ *e* $W(t+2)$ como as próximas duas palavras, que no caso do CBOW, que entrarão na rede e no caso do *Skip-Gram*, as duas próximas palavras posteriores a palavra que entrou na rede que serão preditas no contexto (MIKOLOV; LE; SUTSKEVER, 2013).

Além dessas diferenças entre os dois modelos dentro da proposta do *Word2Vec*, o modelo como um todo requer menos memória que outras representações, além do fato de levar em consideração as palavras da vizinhança, faz com que a representação das palavras possua mais informação. Apesar das vantagens de usar o *Word2Vec*, encontrar os hiper parâmetros é difícil e, além disso, a função de softmax é custosa do ponto de vista computacional e, consequentemente, o treino dos modelos leva muito tempo.

Com relação ao desempenho de processamento, o custo do treinamento de CBOW pode ser dado pela fórmula:

$$Q = N * D + D * \log 2(V) \quad (4)$$

No qual V é a quantidade de palavras distintas no vocabulário, *N* o tamanho da camada de projeção e D a dimensão dos vetores.

Já o custo do treinamento de *Skip-Gram* pode ser dado pela fórmula:

$$Q = C * (D + D * \log 2(V) \quad (5)$$

O tempo gasto para treinamento de todo o corpus coletado para este trabalho através do *Word2Vec* foi de aproximadamente 3 a 4 horas, sendo que o tamanho do vocabulário é de 49.220 e a dimensão do vetor é de 200.

2.3 ALGORITMOS DE APRENDIZADO DE MÁQUINA

A regressão logística é uma metodologia estatística que visa estimar probabilidades de ocorrências em variáveis dependentes do tipo binário. Segundo Arango (2001), a regressão logística é um instrumento da estatística útil para casos nas quais se deseja predizer a presença ou ausência de uma determinada característica ou resultado, baseado em valores de um conjunto de variáveis independentes. A regressão logística é amplamente usada em ciências médicas e sociais, e tem outras denominações, como modelo logístico, modelo logit, e classificador de máxima entropia.

Já o *Random Forest* (árvores de decisão) representa uma das formas mais simplificadas de um sistema de suporte à decisão. É um método estatístico, de aprendizagem supervisionada, podendo ser utilizado em problemas de classificação e na realização de previsões. As Árvores de Decisão são um dos modelos mais práticos e mais usados em inferência indutiva. Este método representa funções como árvores de decisão. Estas árvores são treinadas de acordo com um conjunto de treino (exemplos previamente classificados) e posteriormente, outros exemplos são classificados de acordo com essa mesma árvore.

O *Support Vector Machine* (SVM), também conhecido como Máquina de Suporte Vetorial, foi elaborado com o estudo proposto por Boser, Guyon e Vapnik em 1992. Ele é um algoritmo de aprendizado supervisionado, cujo objetivo é classificar determinado conjunto de pontos de dados que são mapeados para um espaço de características multidimensional usando uma função kernel, abordagem utilizada para classificar problemas. Nela, o limite de decisão no espaço de entrada é representado por um hiperplano em dimensão superior no espaço (VAPNIK et al., 1997 e SARADHI et al., 2005).

O KNN (K-*Nearest Neighbor*) é um algoritmo que pode ser usado tanto para classificação como regressão. Seu objetivo é determinar a qual grupo uma determinada amostra vai pertencer com base nas amostras vizinhas. Ele é um algoritmo simples e de fácil implementação, os exemplos de treinamento são armazenados e a previsão é feita somente quando um novo registro precisa ser classificado. Ao contrário dos outros algoritmos ele não constrói um, ele faz somente o cálculo de distância. Por conta dessa característica, ele é considerado um método do tipo preguiçoso. Dado um novo registro ele vai calcular a distância desse registro com todas as amostras da base de dados de treinamento, informando qual vai ser o número de vizinhos que serão comparados.

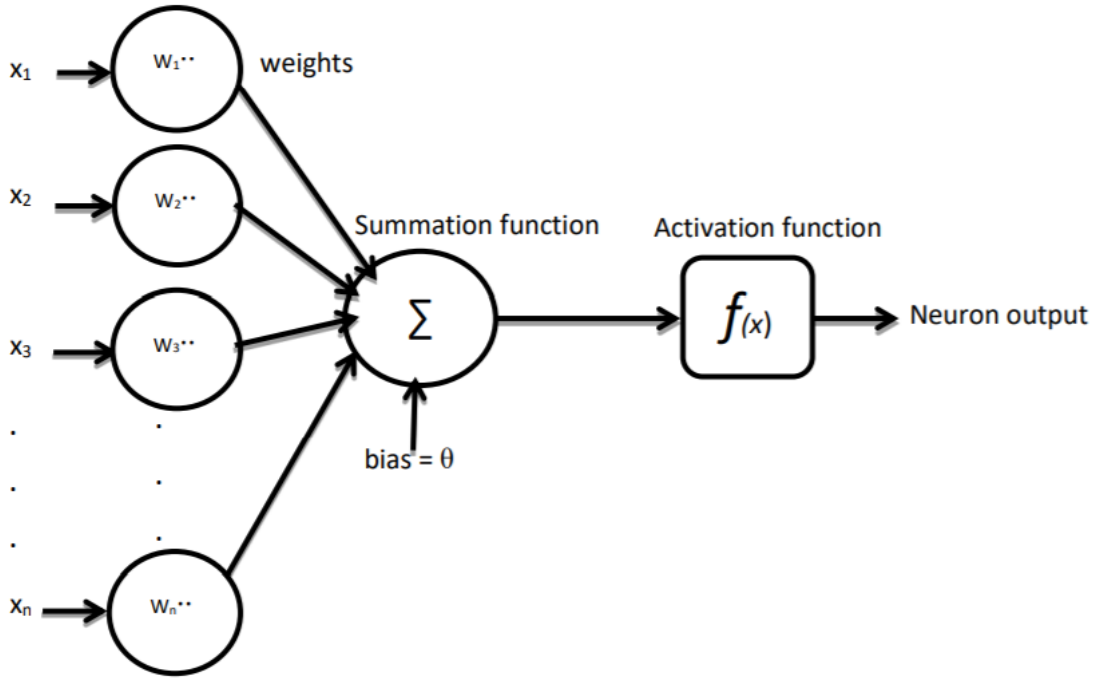
2.4 REDES NEURAIS ARTIFICIAIS (RNA)

Quando há a simulação de neurônios de forma artificial sendo usados juntos para processar dados há uma Rede Neural, sendo que seu funcionamento é basicamente cada um dos neurônios presentes na rede recebendo sinais das variáveis de entrada e passando para os próximos neurônios o resultado do processamento desse dado. Cada entrada em um neurônio é a saída do processamento do neurônio anterior, e essa saída possuí uma a do dado processado conforme o modelo que está sendo analisado (FERNEDA, 2006).

Russell e Norvig (2010), demonstram a ideia de que as redes neurais artificiais são modelos de um cérebro biológico e esse fato dissemina a expressão “rede neural artificial” entre os acadêmicos em Inteligência Artificial. Atualmente as Redes Neurais Artificiais (RNA) estão presentes em vários trabalhos que tratam do campo da Inteligência Artificial (SILVER ET AL., 2018) e (RADFORD ET AL., 2019).

Embora a expressão “Redes Neurais Artificiais” possa incluir um amplo espaço de abordagens, todas estas ressaltam, de alguma forma, um modelo da atividade cerebral envolvida na tarefa a que o sistema exerce, ou seja, simula comportamento e funções do neurônio biológico através de um modelo matemático (YACIM; BOSHOFF, 2018), perspectiva refletida tanto na pesquisa acadêmica sobre redes neurais quanto em suas aplicações a problemas específicos.

Figura 5 - Demonstração de um neurônio artificial.



Fonte: adaptado de (YACIM; BOSHOFF, 2018).

onde:

- x_m são as entradas da rede.
- w_{km} são os pesos, ou pesos sinápticos, associados a cada entrada.
- $\sum$ é o somatório dos sinais calculados pelo produto das entradas pelo peso.
- *f(x)* é a função de ativação.

Conforme Russell e Norvig (2010), quando é feito a replicação do neurônio biológico através do modelo matemático, há as funções do cérebro sendo adaptadas pelos conceitos da RNA. Por exemplo há os dendritos, célula nervosa que recebe sinais no neurônio biológico, são substituídos pelas entradas da rede, e tem as ligações feitas pelos pesos adaptando as sinapses neurais. Já a função de soma ponderada das entradas e os "bias" recebem e processam os estímulos que são captados pelas entradas. Por fim, o disparo do neurônio biológico se traduz no modelo matemático para uma função de ativação e isso é repassado para outras partes da RNA simulando um axônio, local onde se transmite o influxo nervoso.

Conforme Haykin, 2008, a função da RNA se dá pela seguinte equação:

$$Y = Activation\ (\ \sum(\text{weight} * \text{input}\ ) + bias\ ) \tag{6}$$

onde:

- $\sum$ é o somatório dos sinais calculados pelo produto das entradas pelo peso.

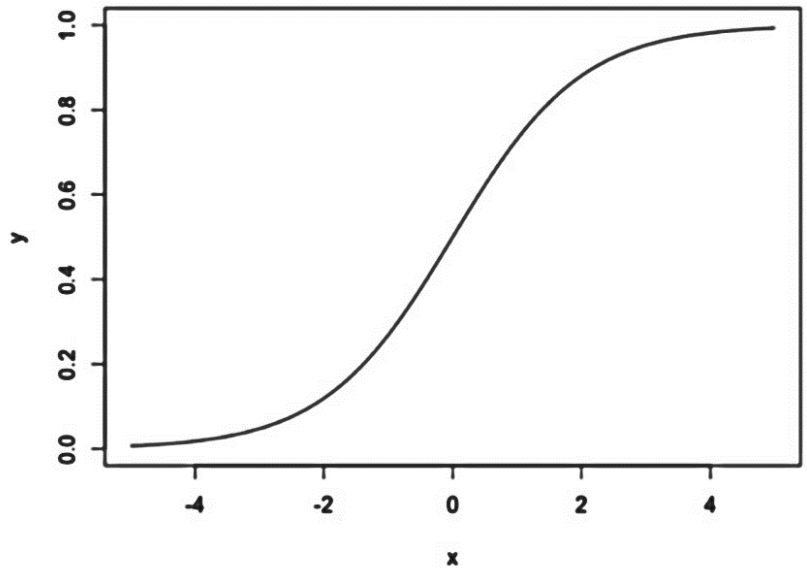
Pode-se dizer que nos pesos há concentração de todo o conhecimento obtido pela rede. Os pesos são os parâmetros ajustáveis que mudam e se adaptam à medida que o conjunto de treinamento é apresentado à rede. Assim, o processo de aprendizado supervisionado em uma RNA com pesos, resulta em sucessivos ajustes dos pesos sinápticos, de tal forma que a saída da rede seja a mais próxima possível da resposta desejada (HAYKIN, 2008).

A função de ativação é um componente matemático que é incluído na rede, abaixo as principais funções que são usadas neste trabalho:

a) ***Sigmóide***: é uma função de ativação muito usada, ela varia de 0 a 1 tendo um formato S, com isso não sendo linear:

$$f(x) = \frac{1}{1+ e^{-x}} \tag{7}$$

Figura 6 - Gráfico da função Sigmóide.

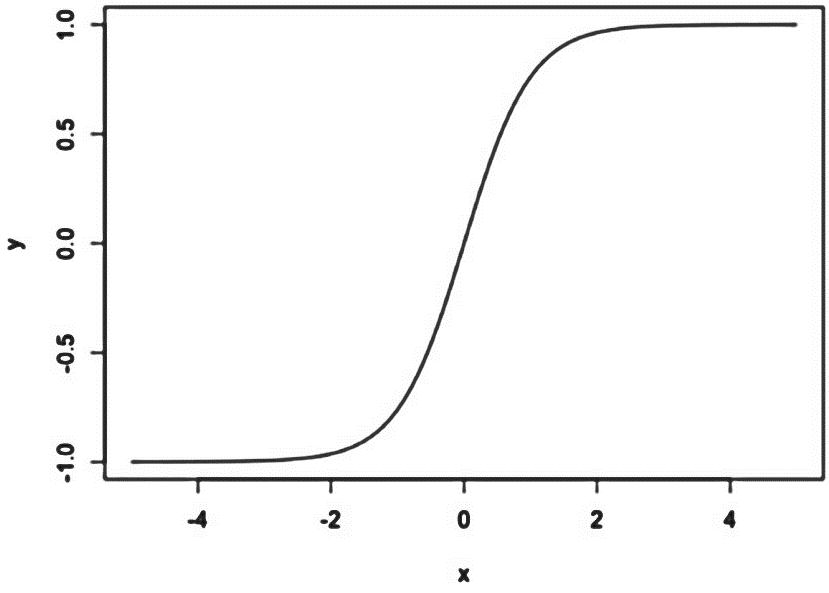


Fonte: Elaborado pelo autor.

b) ***Tangente Hiperbólica (TanH)***: A função TanH é muito semelhante à função sigmóide, no entanto ela é apenas uma versão escalonada da função sigmóide, e varia de -1 a 1.

$$f(x) = \frac{1-\exp -x}{1+\exp -x} \quad (8)$$

Figura 7 - Gráfico da Função Tangente Hiperbólica - TanH



Fonte: Elaborado pelo autor.

c) Softmax: A função softmax também é um tipo de função sigmóide, é bastante útil para lidarmos com problemas de classificação onde precisamos trabalhar com mais de duas classes na classificação. A função softmax transforma as saídas para cada classe para valores entre 0 e 1 e também e divide pela soma das saídas, fazendo dessa forma com que sua saída apresente a probabilidade dos dados.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_j}} \qquad for\ j = 1, \ldots, K. \tag{9}$$

Dentre as Redes Neurais mais comuns pode-se destacar a *Perceptron Multicamadas* (PMC), que possuí uma arquitetura composta por uma camada de entrada, que recebe os dados, um ou mais camadas escondidas, onde os neurônios irão realizar o processamento dos dados, e a camada de saída. Essa arquitetura consegue aprender e resolver bem problemas atrelados a *machine learning*. No entanto, essa arquitetura há dificuldades no processamento de sequenciamento de dados de entrada, como por exemplo séries temporais, processamento de linguagem e vídeos (FERNEDA, 2006).

2.4.1 Rede de Memória de Curto Longo Prazo (Long Short Term Memory - LSTM)

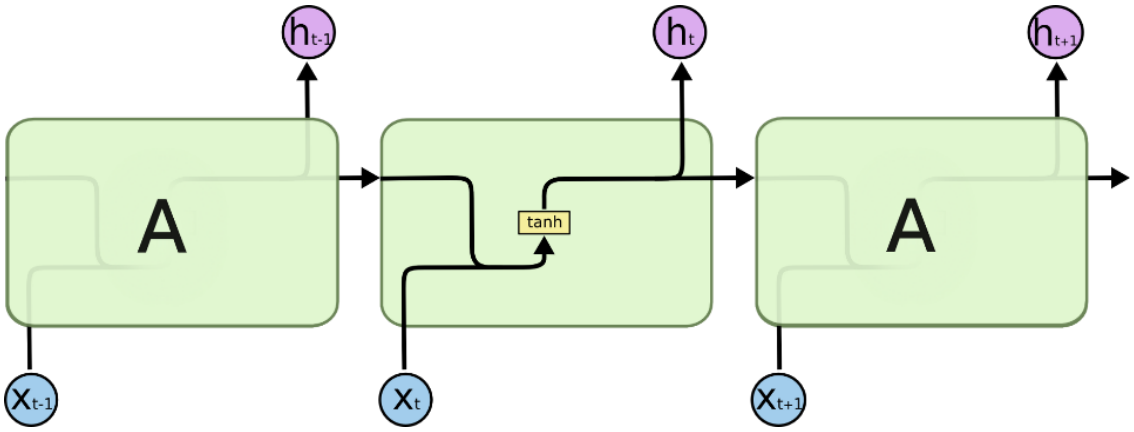
As Redes Neurais Recorrentes (*Recurrent Neural Networks* - RNN) foram desenvolvidas principalmente para resolver problemas associados à predição de séries temporais e sequências de dados, além de ser muito útil também para predição de palavras em um texto de entrada e reconhecimento de fala, conforme Sak, Senior e Beaufays (2014).

De acordo com as afirmações feitas para a Rede Neural Recorrente, há uma saída para neurônio da rede que pode ser usado como entrada para o próximo nó da rede. Devido a este funcionamento há uma realimentação de informações com os próprios dados passados da rede. Com isso a decisão de uma RNN na etapa *t – 1* terá impactos diretamente na sequência da rede no tempo *t* (GOODFELLOW; BENGIO; COURVILLE, 2016).

Conforme Hochreiter e Schmidhuber (1997), a rede *Long Short Term Memory* (LSTM) é um tipo de específico de RNN que consegue manter o aprendizado por período indeterminado. Seu principal objetivo é de resolver o problema dissipação do gradiente (do inglês *Gradient Vanishing*), também conhecido como o problema de explosão ou desvanecimento do gradiente nas RNNs.

Segundo Amidi e Amidi (2018), o problema de dissipação do gradiente ocorre quando os gradientes tendem a zero por serem continuamente multiplicados por números inferiores a um. Isso ocorre devido à dificuldade de conseguir manter suas dependências de longo prazo, onde os autores afirmam que não é possível aprendê-las devido ao gradiente multiplicativo que pode aumentar ou diminuir exponencialmente conforme o número de camadas da rede.

Figura 8 - Rede Neural Padrão.



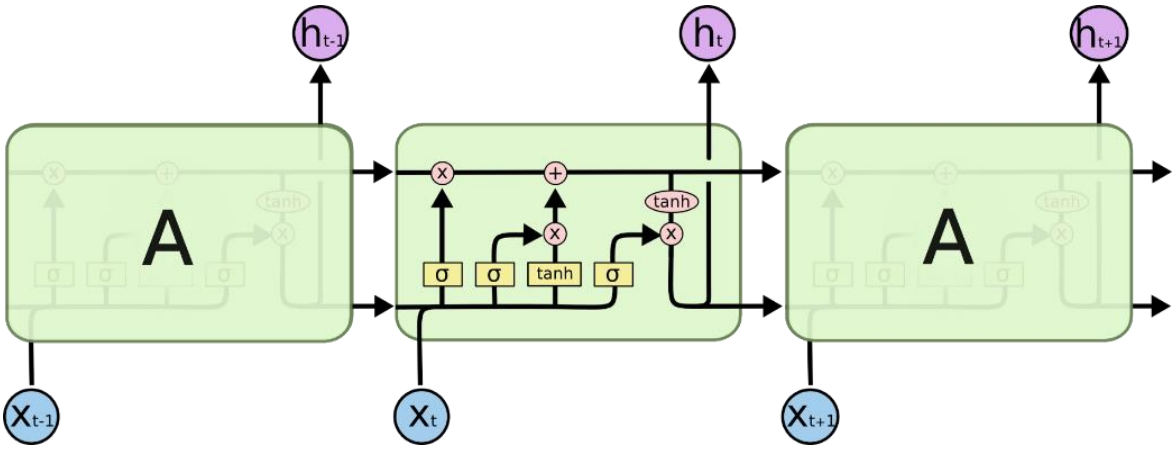
Fonte: Olah, 2015.

Estruturalmente, todas as redes neurais recorrentes possuem como forma uma espécie de sequência em cadeia com repetição dos módulos na rede neural. A estrutura em RNNs padrão, diferentes de LSTM, é muito simples, contendo apenas uma camada que realiza os cálculos e com uma tangente hiperbólica de ativação, conforme Figura 8 (HOCHREITER; SCHMIDHUBER, 1997).

Já as redes LSTMs, assim como as RNNs padrão, contém uma estrutura em cadeia sequencial, no entanto, suas células contém uma estrutura especial no qual há uma interação mais robusta que realiza mais cálculos, através de suas quatro camadas de rede neural, e possui portões (*gates)*, que faz a decisão de retirada e permanência de informações dentro das células, no qual terá sua explicação no item 2.3.2.1, e sua estrutura pode ser visualizada na Figura 9 (HOCHREITER; SCHMIDHUBER, 1997).

Na estrutura demonstrada na Figura 9, há informações passam através da linha superior horizontal, no qual é percorrido toda a cadeia dentro de uma célula e em sua cadeia de sequenciamento. Essas informações, chamadas de estado da célula *(cell state)*, possui interações lineares menores ao longo do percurso na célula, isso faz com que as informações tenham menos alterações ao longo do percurso (HOCHREITER; SCHMIDHUBER, 1997).

Figura 9 - Redes Neurais Recorrentes LSTM.



Fonte: Olah, 2015.

2.4.2 Processamento da LSTM

Internamente, em cada célula existem equações matemáticas que definem o comportamento da rede. São esses mecanismos que realizam os cálculos dos pesos e as interações dentro da própria célula, esses mecanismos são chamados de portões. Para cada uma das células de uma Rede Neural Recorrente (LSTM) há três portões e cada um deles é um vetor de valores entre *(0,0) e (1,0)*, usados para determinar o quanto da informação será esquecido ou lembrado em cada processamento de uma célula, ao longo do ciclo de entrada e saída de toda a rede (HOCHREITER; SCHMIDHUBER, 1997).

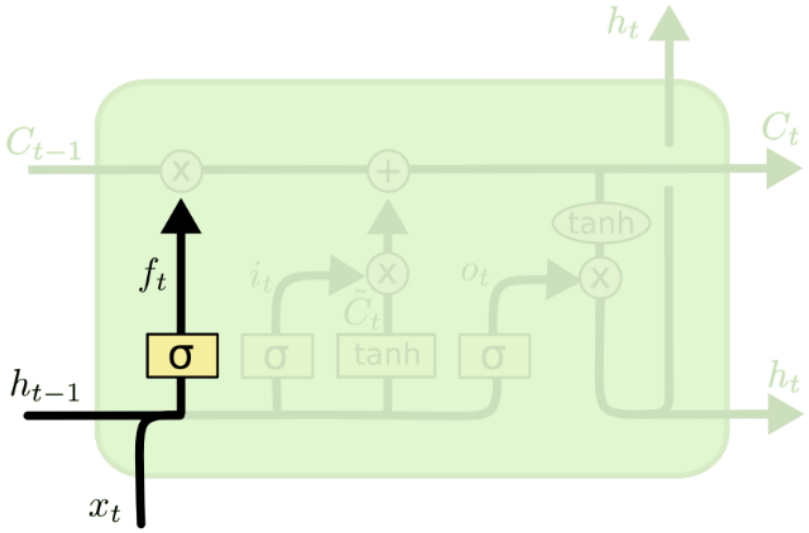
As células de uma rede LSTM usam $h(t-1)$ e $c(t-1)$ como entradas, no qual $t-1$ significa a etapa de tempo anterior, c representa o estado da célula e $h(t)$ é a saída de cada uma das células. Portanto, $h(t-1)$ e $c(t-1)$ são os valores de saída e os valores de estado anteriores.

Seus portões de processamento funcionam da seguinte maneira:

a) Portão de esquecimento (*forget gate*): Responsável por decidir quais partes são importantes o suficiente para continuar no progresso da rede fazendo o esquecimento do que não é relevante (HOCHREITER; SCHMIDHUBER, 1997). Através de uma camada sigmóide o portão de esquecimento recebe os dados de entrada para h_{t-1} e x_t e gera resultado entre 0 e 1 para cada número no estado da célula C_{t-1}, no qual dão entrada no estado da célula, fórmula representada abaixo.

$$f_t = \sigma(W_f * [h_t, x_t] + b_f) \quad (10)$$

Figura 10 - Demonstração do Portão de Esquecimento.



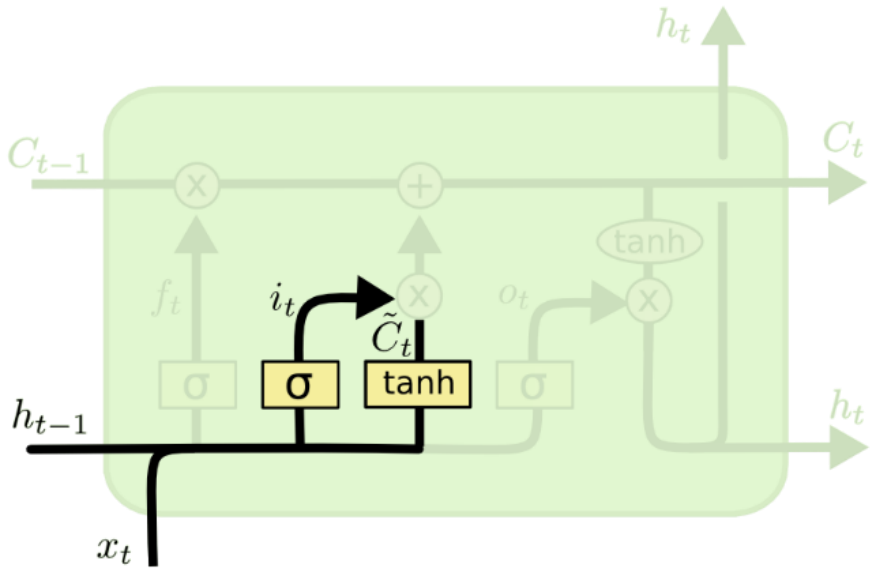
Fonte: Olah, 2015.

b) Portão de Entrada (*input gate*): Responsável por decidir quais informações são importantes para a memória de curto prazo, ou seja, o que cada célula irá armazenar, fazendo a adição dessas ao estado da célula (HOCHREITER; SCHMIDHUBER, 1997). A primeira parte dos cálculos realiza através de uma camada sigmóide quais informações serão atualizadas. Em seguida uma camada com TanH gera um vetor de novos valores candidatos, $\check{C}_t$.

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \quad (11)$$

$$\check{C}_t = tanh(W_c * [h_{t-1}, x_t] + b_c) \quad (12)$$

Figura 11 - Demonstração do Portão de Entrada.



Fonte: Olah, 2015.

Os resultados dos dois cálculos são combinados gerando uma nova entrada no estado da célula.

$$C_t = f_t * C_{t-1} + i_t * \check{C}_t \quad (13)$$

c) Portão de saída (*output gate*): Responsável por decidir quais partes do estado da célula são importantes no instante atual para gerar a saída (*output*) da célula. Para isso, o primeiro cálculo usa uma camada sigmóide que fará decisão de quais informações do estado da célula serão produzidos. Já o segundo cálculo usa o próprio estado da célula passando por uma camada TanH. O resultado do primeiro cálculo é multiplicado pelo resultado do segundo cálculo gerando assim a saída de uma célula.

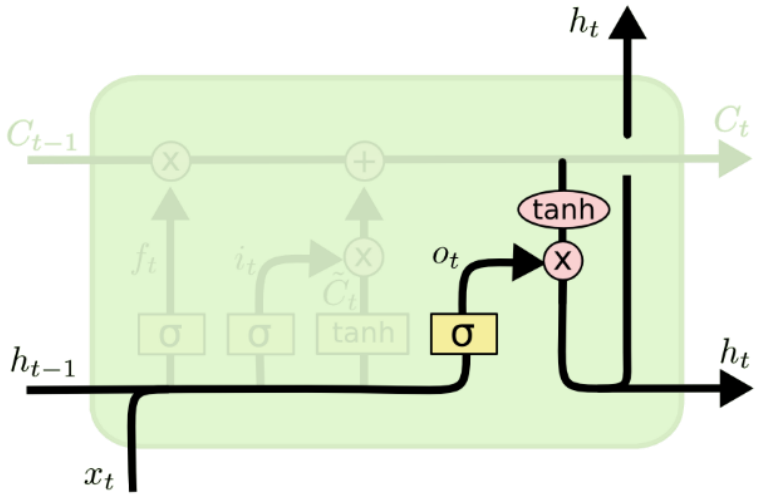
$$O_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \quad (14)$$

$$h_t = o_t * \tanh(C_t) \quad (15)$$

Conforme (GOODFELLOW; BENGIO; COURVILLE, 2016), é possível notar através da Figura 13 que temos a célula de uma RNN fazendo a reutilização de informações que já foram processadas na rede a cada novo registro, o que se verifica com as demonstrações à esquerda de cada célula. Já à direita de cada célula temos o funcionamento da rede que se resume em uma série de neurônios se comunicando entre si ao longo do tempo. A importância

desse mecanismo se resume que ao longo do processamento o problema de dissipação de gradiente é descartado.

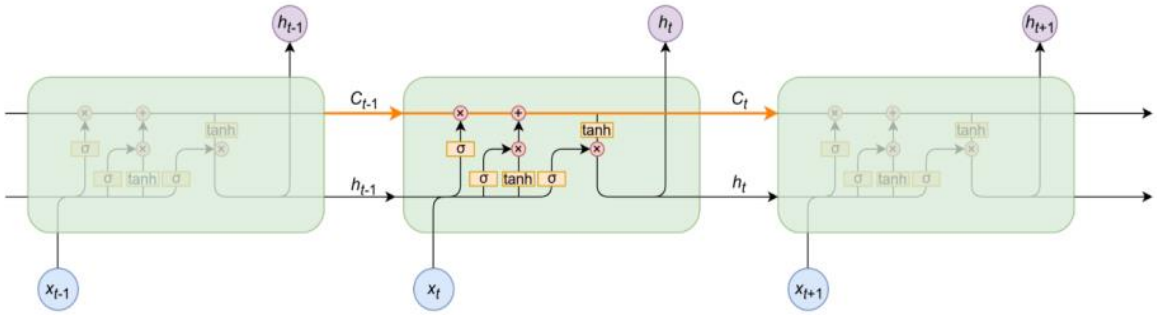
Figura 12 - Demonstração do Portão de Saída.



Fonte: Olah, 2015.

De acordo com os recursos existentes em uma Rede Neural Recorrente (LSTM), é possível afirmar que aplicá-la à uma série temporal pode trazer bons resultados, pois ela pode processar dados temporais com eficácia. Diante disso, o uso da LSTM para processamento de dados coletados da série temporal dos ativos da bolsa de valores pode ser positivo. Mesmo com um longo tempo após seu desenvolvimento a LSTM ainda é muito usada para problemas que envolvam series temporais.

Figura 13 - Arquitetura LSTM de Sequenciamento.



Fonte: Olah, 2015.

3 TRABALHOS CORRELATOS - O ESTADO DA ARTE

Esta seção demonstra alguns trabalhos que tratam de problemas similares ao proposto, no qual há por objetivo o entendimento e levantamento dos diversos problemas relacionados a este trabalho, podendo analisar quais propostas, mecanismos e aplicações foram utilizados na sua abordagem.

Nos últimos anos, o uso de tecnologias voltadas para IA e aplicadas à estudos no mercado financeiro, somado ao estudo de análise de sentimento e processamento textual de notícias ou redes sociais, vem ganhando espaço no meio científico, isso devido a ser um problema atual e por possibilitar a geração de apoio às decisões que visam lucros e melhores escolhas no mercado financeiro.

Os avanços tecnológicos dos últimos anos, também permitiram o processamento desses dados, através do aumento do poder de processamento dos equipamentos e diminuição do seu custo, o que também influenciou na disseminação desses estudos.

O trabalho de Vargas et al. (2018) usa modelos de aprendizado profundo como Rede Neural Convolucional (CNN) e Redes Neurais Recorrentes (RNN) para o dia a dia na previsão de movimentos direcionais do preço dos ativos usando notícias e indicadores técnicos como entrada. É realizada ainda uma comparação entre dois conjuntos diferentes de indicadores técnicos.

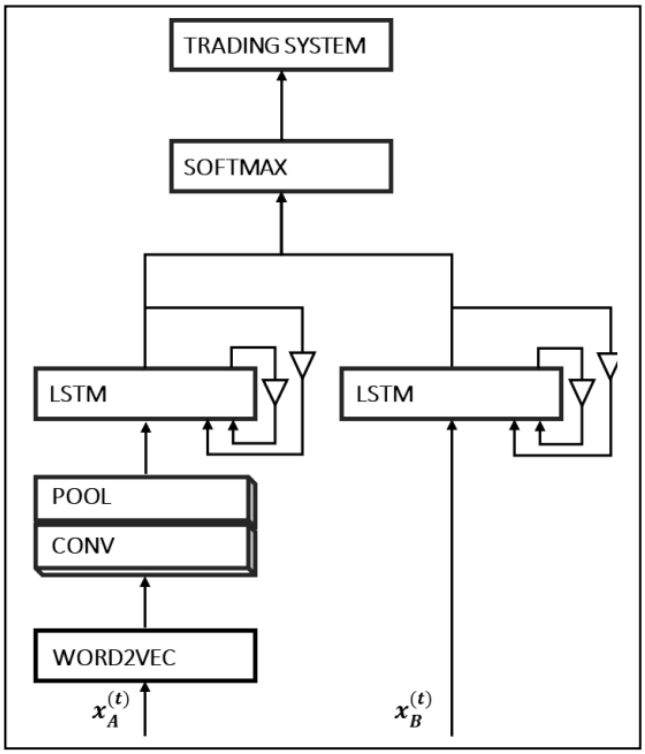
Já na arquitetura proposta por Vargas et al. (2018), presente na Figura 14, demonstra-se a aplicação do *Word2Vec* criando as *features* de entrada para a rede. Os textos são sequenciados e, em um primeiro modelo, os autores realizam a aplicação de um modelo pré treinado em CBOW, nomeado por *Google News*. Já num segundo modelo, Vargas et al. (2018) realiza a média de todos os vetores de palavras presentes em uma notícia, chamado de “vetor frase” e a partir disso, cria-se uma tabela para entrada na rede.

Os autores detalham também que a saída de cada modelo é usada como entrada para um agente de negociação, que compra ativos no dia atual e vende no dia seguinte, quando o modelo prevê que o preço vai para cima, caso contrário, o agente vende ativos no dia atual e compra no próximo dia. O método proposto mostra um papel importante de notícias financeiras na estabilização dos resultados e não apresenta quase nenhuma melhoria ao comparar diferentes conjuntos de técnicas com indicadores.

Em sua pesquisa, os autores Vargas et al. (2018) visaram evidenciar que o estudo dessas variáveis pode detectar e analisar padrões complexos e interações nos dados permitindo um processo de negociação mais preciso. Contudo os resultados dessa aplicação ficaram em torno

de 57% de acurácia e segundo os autores, uma das causas desse resultado está atrelado aos dados usados, indicando a melhor seleção do conteúdo a ser analisado, e os dados usados nem sempre eram exclusivamente voltados para o mercado financeiro.

Figura 14 - SI-RCNN - Modelo de arquitetura.



Fonte: Vargas et al., 2018.

Já no trabalho de Ferreira et al. (2019) é realizada uma caracterização e análise de dados da *Stocktwits*, uma rede social voltada para o mercado financeiro, a fim de obter gatilhos e visualizações que podem ser aplicadas aos mercados financeiros e negociação algorítmica. Os autores afirmam que há uma forte consideração de informações de sentimentos em mensagens para criar um indicador social, usado como um modelo de previsão e apoio nas decisões como estratégia de atuação em bolsas de valores. Essa caracterização revela o comportamento dos usuários e padrões de conteúdo na rede, onde ainda são utilizados três conceitos de avaliação, sendo eles análise estática, temporal e de correlação. Como resultados, os autores informam que existe uma alta correlação positiva entre o número de mensagens diárias e o volume de negócios diários nos ativos.

Isso mostra que quando há uma alta volatilidade nos preços dos ativos, o número de mensagens trafegadas na rede tende a aumentar também. Neste estudo, foram calculadas a correlação das variáveis *Open* (valor de abertura), *High* (valor máximo), *Low* (valor mínimo), *Close* (valor de fechamento) *e Adj Close* (valor de ajuste de fechamento) com a quantidade de

mensagens postadas durante o dia sobre cada ativo. A maioria dos ativos observados têm uma correlação alta (maior que 50%) ou uma correlação muito alta (maior que 70%) entre o número de mensagens postadas e o volume negociado, o que pode nos dizer que o número de postagens acompanha o volume de ações negociadas.

Os autores Carosia, Coelho e Silva (2019) realizaram um trabalho em que consideraram como fontes de dados tanto notícias como *tweets*, ou seja, duas fontes de dados. Construíram um analisador de sentimentos para língua portuguesa do Brasil com base em uma técnica de Aprendizado de Máquina, o *Multilayer Perceptron (MLP)* e em sua arquitetura objetivaram o relacionamento entre Análise de Sentimentos (AS) e o mercado financeiro, com comparações da linha temporal prevista pelo treinamento dos sentimentos e linha temporal dos ativos. Os resultados desse trabalho foram de F1-Score com uma média de 70 a 76% e acurácia de cerca de 80 a 82%.

No trabalho desenvolvido por Mern, Anderson e Poothokaran (2017), foram utilizados indicadores econômicos como *Down Jones Industrial Average*, S&P 500, assim como dados analíticos do *Google Trends* para o termo "*Bitcoin*". Esses dados foram utilizados para se prever o preço da moeda do dia seguinte e como melhor performance tiveram um modelo que utilizou uma Rede Neural Convolucional. Foi apresentado nesse trabalho resultado de acurácia de 66,7%.

Dentre as pesquisas observadas até o momento, um dos parâmetros que frequentemente é utilizado pelos autores na previsão de ativos é a opinião pública em relação a determinado ativo, ou seja, analisando diretamente o texto em relação ao seu sentimento, sendo esse positivo ou negativo. Por exemplo em Mittal (2011), foi realizada previsão do valor do *Down Jones Industrial Average* (DJIA), sendo que os parâmetros utilizados foram o valor do DJIA nos últimos 3 dias, assim como o sentimento de *tweets* no mesmo período. Também são encontrados projetos em que o sentimento do título da notícia é utilizado, como visto em 2018 (2018), enquanto outros utilizam o conteúdo da notícia, como visto em Daultani (2017).

Com isso, verifica-se que muitas características deste projeto a outros que estão acontecendo nos últimos anos, há por exemplo o uso do Twitter como fonte de dados, sendo possível destacar Xu e Cohen (2018) e Souza, Lucena e Queiroz (2019). Além desses há outros trabalhos ainda baseados na fonte de dados comum deste trabalho, como Li e Shah (2017), que realiza uma análise léxica sobre os *tweets* coletados através de *Word Vector*, o que ressalta que a transformação de palavras em vetor tem grande valor para semântica e também Kraaijeveld e De Smedt (2020) e Alzazah e Cheng (2020), que realizam através dessa mesma fonte de dados

predição de preço de criptomoedas. Essa abordagem léxica também é estudada e analisada sobre o mercado financeiro por Turner, Labille e Gauch (2020).

Ainda sobre criptomoedas, há uma série de trabalhos recentes, esses trabalhos envolvem desde análise textual do Twitter e de alguns portais de notícias diversos, há destaque para os correlatos a este trabalho como Mern, Anderson e Poothokaran (2017) e Anup et al. (2018), por exemplo.

Além desses também é possível correlacionar outros trabalhos com a área de criptomoedas usando Redes Neurais, Hachicha e Hachicha (2020) e Uras et al. (2020), Jain et al. (2018), Atashian e Hrachya (2018), Tianyu Ray Li et al. (2019), Kim et al. (2016) e Sattarov et al. (2020) e o autor Raju e Tarif (2020) que usa em tempo real a predição das criptomoedas com a análise de sentimentos como relação.

Voltado diretamente para predição ou relação de mercados financeiros, podemos destacar trabalhos como Peng e Jiang (2016), que utiliza *Word Embedding* somando redes neurais, que também estão muito presentes nos trabalhos para a área financeira de Nelson, Pereira e Oliveira (2017), Chen, Dautel et al. (2020) e Yu e Yan (2020), Mehtab e Sen (2019) e Jiang (2020) e Zhang e Lou (2020).

Outra característica deste trabalho são as Redes Neurais Recorrentes – LSTM e essa também possuí uma série de trabalhos em andamento voltados para o mercado financeiro, destaque para Qiu, Wang e Zhou (2020), além de outros trabalhos como Zoen et al. (2019) e Xinyi Li et al. (2019) que usam notícias para aplicação da LSTM e posterior predição. Ainda sobre LSTM, Caux, Bernardini e Viterbo (2020) usa-a para um trabalho de predição de criptomoedas.

O termo análise de sentimento é bastante usado e vem sendo bastante empregado e aprofundado pelos autores, sendo possível ainda ressaltar alguns trabalhos nessa área, com pretensões no mercado financeiro como Nguyen, Shirai e Velcin (2016), Joshi, N e Rao (2016) e Xing, Cambria e Welsch (2018). Esses trabalhos e estudos vem sendo feitos em diversos países, relacionando seus mercados financeiros, sendo possível destacar o trabalho de Nti, Adekoya e Weyori (2020) que estuda predição do mercado usando análise de sentimento em Gana. Já no mercado brasileiro há como destaque alguns trabalhos como Carosia, Coelho e Silva (2019) e Medeiros e Borges (2019).

Na Tabela 1 há um resumo da quantidade de artigos, conforme temas apresentados acima, por ano de publicação, e na Tabela 2 a relação de alguns trabalhos que vem sendo publicados nos últimos anos.

Tabela 1 - Quantidade de publicações por ano.

Ano	Quantidade de Publicações
2011	1
2016	4
2017	4
2018	9
2019	8
2020	6

Fonte: Elaborado pelo autor.

Tabela 2 - Estado da Arte.

Autor	Ano	Metodologia
Mittal	2011	Stock Prediction Using Twitter Sentiment Analysis
Peng and Jiang	2016	Leverage Financial News to Predict Stock Price Movements Using Word Embeddings and Deep Neural Networks
Kim et al	2016	Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies
Nguyen et al	2016	Sentiment analysis on social media for stock movement prediction
Joshi et al	2016	Stock Trend Prediction Using News Sentiment Analysis
Li e Shah	2017	Learning Stock Market Sentiment Lexicon and Sentiment-Oriented Word Vector from StockTwits
Nelson et al	2017	Stock market's price movement prediction with LSTM neural networks
Daultani	2017	Stock predictions through news sentiment analysis
Mern et al	2017	Using Bitcoin Ledger Network Data to Predict the Price of Bitcoin
Vargas et al	2018	Deep Leaming for Stock Market Prediction Using Technical Indicators and Financial News Articles
Jain et al	2018	Forecasting Price of Cryptocurrencies Using Tweets Sentiment Analysis
Xing et al	2018	Predictive analytics, Text mining, Natural language processing, Knowledge engineering, Financial forecasting, Computational finance
Coinanalysis	2018	Predict cryptocurrency prices based on news and historical price data
Pant et al	2018	Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis
Atashian e Hrachya	2018	Sentiment Analysis To Predict Global Cryptocurrency Trends
Xu e Cohen	2018	Stock Movement Prediction from Tweets and Historical Prices
Olivier e Johannes	2018	The predictive power of public Twitter sentiment for forecasting cryptocurrency prices
Velay e Daniel	2018	Using NLP on news headlines to predict index trends
Mehtab e Sem	2019	A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing
Zoen et al	2019	BERT-based Financial Sentiment Index and LSTM-based Stock Return Predictability
Ferreira et al	2019	Data Science in Financial Markets: Characterization and Analysis of Stocktwits
Li et al	2019	DP-LSTM: Differential Privacy-inspired LSTM for Stock Prediction Using Financial News
Souza et al	2019	O Efeito do Sentimento do Investidor Expresso via Twitter sobre o Comportamento do
Anup et al	2019	Sentiment-Based Prediction of Alternative Cryptocurrency Price Fluctuations Using Gradient Boosting Tree Model
Carosia et al	2019	The Influence of Tweets and News on the Brazilian Stock Market through Sentiment Analysis
Medeiros e Borges	2019	Tweet Sentiment Analysis Regarding the Brazilian Stock Market
Chen e al	2020	Applications of deep learning in stock market prediction: recent progress
Dautel et al	2020	Forex exchange rate forecasting using deep recurrent neural networks
Alzazah e Cheng	2020	Recent Advances in Stock Market Prediction Using Text Mining: A Survey
Turner et al	2020	Lexicon-Based Sentiment Analysis for Stock Movement Prediction
Caux et al	2020	Short-Term Forecasting in Bitcoin Time Series Using LSTM and GRU RNNs
Yu e Yan	2020	Stock price prediction based on deep neural networks

Fonte: Elaborado pelo autor.

4. MODELO PROPOSTO

Nesta seção será apresentada a metodologia para a criação do modelo de tomada de decisão desenvolvido, com destaque para a coleta dos dados do Twitter e Bovespa, juntamente com o tratamento desses dados e posterior treinamento e classificação pelo modelo.

Para o desenvolvimento da pesquisa, foi utilizado o procedimento metodológico de coleta de dados pelo método quantitativo. Para tanto, é coletado dados da plataforma Twitter e feito coleta de informações fornecidas pela própria Bovespa através do Software Profit. Feita a captura das informações, realiza-se o tratamento dos dados coletados através de técnicas de Processamento Linguagem Natural (PLN) de modo que, ao final do nosso estudo, tenha-se um modelo que auxilie na tomada de decisões de compra e venda de ações.

4.1 ARQUITETURA DO SISTEMA

Neste trabalho, é proposta uma abordagem aplicando a arquitetura de Rede Neural Recorrente do tipo Memória de Curto Longo Prazo (*Long Short Term Memory* - LSTM) com o objetivo de identificar os melhores pontos de operação, tanto de compra e venda, para o ativo mini-índice do Ibovespa.

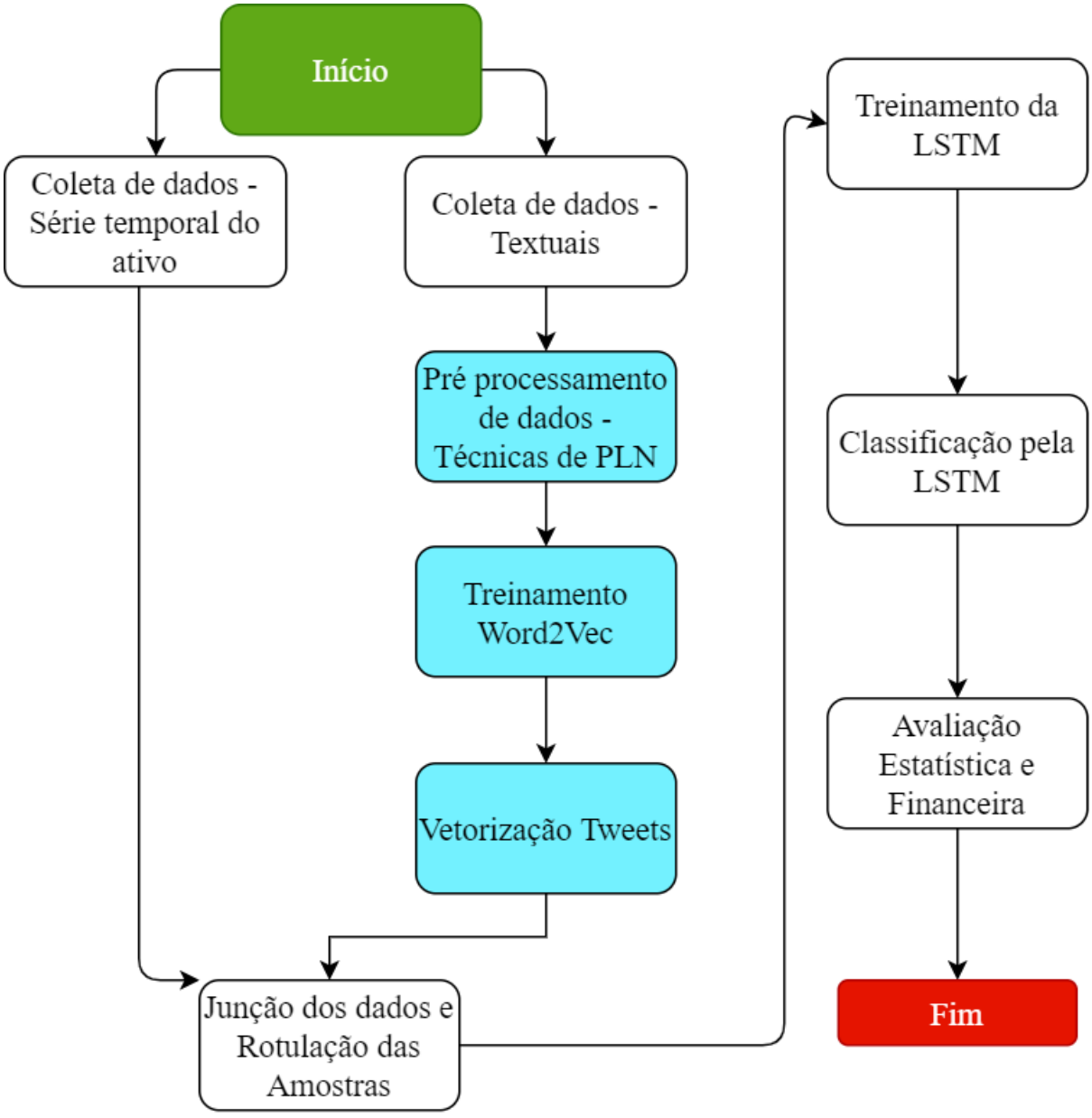
O modelo desenvolvido neste trabalho é nomeado como “RNNTW-TC”, sigla para *Recurrent Neural Network with Twitter to Trading Classifer*, pois é feito, como já dito, uso de dados tanto da plataforma Twitter para postagens que possam direcionar operações no ativo mini-índice, além de uso de Rede Neural Recorrente (RNN). O intuito é o de identificar padrões para indicação de operações de compras e vendas em operações *day trade,* operações iniciadas e finalizadas num mesmo dia.

O modelo proposto é apresentado na Figura 15. Os dados do Twitter e dados históricos do ativo mini-índice são coletados e passam por um pré-processamento. Neste processo, destaque para a aplicação, nos textos capturados do Twitter, de técnicas de Processamento de Linguagem Natural (PLN) para normalização dos dados, e posterior treinamento e geração dos *Embedding,* processo explicado na seção 2.2, das frases contidas em cada uma das postagens.

Com a conclusão do processo de *Embedding* é feita a junção dessa base de dados aos dados de fechamento do ativo no período em que aconteceu a postagem. Essas amostras são rotuladas utilizando o método da seção 4.3 e, após isso, são treinados pela rede LSTM. O resultado é a realização da classificação e sinalização de compra, venda ou aguardar (*buy, sell,*

hold) através de uma camada softmax, dependendo da tendência de mercado pelo modelo, tendência essa que pode ser explicada pela Análise Técnica, explicado na seção 2.1.3.2.

Figura 15 - Modelo de Proposta de Pesquisa.



Fonte: Elaborado pelo autor

4.2 DADOS DE ENTRADA

Justifica-se o uso do ativo mini-índice, pois ele reflete toda a variação do cenário do mercado financeiro no Brasil, representando a variação dos principais ativos presentes no índice Bovespa, de modo que, ao analisar as informações acerca do mini-índice, temos condições de visualizar um panorama geral das ações presentes na Bovespa, como por exemplo, Petrobrás, Vale, Itaú, Bradesco, etc.

Usando a plataforma Twitter, foi possível coletar uma quantidade maior de dados referentes às postagens que se referenciam ao Bovespa, fazendo o uso de filtros (como será melhor informado em item próprio que analisaremos na sequência), de modo a permitir que os dados coletados sejam tratados de forma segura e correta, garantindo um resultado confiável ao final do estudo.

Cabe salientar que as informações coletadas na plataforma Twitter são agregadas as informações fornecidas pela própria Bovespa referentes ao ativo mini-índice, que é um dos mais usados em operações *day trade,* pois em regra é liquidado no mesmo dia pelos investidores, e com isso conseguimos ter melhores informações para opinarmos sobre as negociações *day trade.*

Conforme já discorrido no item 2.3.2, o processamento e classificação da informações coletadas é feito pela Rede Neural Recorrente (LSTM) e justifica-se o uso dessa arquitetura de rede, pois através dos estudos de Hochreiter e Schmidhuber (1997) há demonstração que a LSTM possuí melhores resultados, e evita problemas como o já mencionado *Gradient Vanishing* (desvanecimento do gradiente) quando aplicada à series temporais, problema que ocorre devido a muitas multiplicações com valores muito próximos a zero.

4.2.1 Obtenção da Base de dados do Bovespa

Os dados do ativo mini-índice nos sequenciamentos, de 5, 15 e 30 minutos, foram coletados através do Software Profit, usado nesse ponto para poder acessar os dados em períodos *intraday*, dentro do mesmo dia, e fazer o download para um arquivo “csv” que posteriormente é facilmente integrado ao modelo através da biblioteca pandas em Python.

Os dados coletados foram: data (DD/MM/YYYY HH:MM), valor de abertura, valor máximo, valor mínimo e valor do fechamento no período. Ao total, foram obtidos 2.298 pontos de fechamento para o sequenciamento de 5, 770 pontos de fechamento para o sequenciamento de 15 minutos e 385 pontos de fechamento para o sequenciamento de 30 minutos.

É considerado todo o período de execução de ordens no Ibovespa, visto que são cerca de 8 horas diárias de pregão e durante esse período os dados ficam à disposição para serem coletados. Ademais também coletamos dados dos períodos pré e pós-pregão, que costumam durar 15 minutos e cujas informações são fornecidas pela Ibovespa de modo imediato conforme vão acontecendo os fechamentos.

O ativo mini-índice tem um ciclo de validade de 2 meses, conforme explicado no item 2.1.5, sendo assim, no nosso estudo analisamos os dados coletados do ciclo ativo partindo de 04-03-2021 até 09-04-2021. De acordo com o período usado neste estudo, o código de ativo com data de vencimento mais próximo foi o "WINJ21", ativo do índice com vencimento em 15/04/2021, e por esse motivo adotamos o código "WINJ21" como referência para nossa pesquisa. O período e ativo foi influenciado pelas limitações de busca dos dados textuais da rede social, conforme seção 4.2.3.

4.2.2 Obtenção Base de dados do Twitter

Não foi possível encontrar nas buscas realizadas um conjunto de dados que possuísse informações de *tweets* postados sobre o mercado financeiro, e por essa razão para o desenvolvimento do estudo tivemos que elaborar o desenvolvimento da busca dos dados do Twitter. Isso levou ao desenvolvimento próprio de mecanismos para criação das bases de dados que utilizamos neste estudo, considerando o ciclo desde a coleta até o tratamento dos dados do Twitter, e assim conseguimos desenvolver a pesquisa e atingir os objetivos desejados.

Para a base de dados do Twitter, a coleta foi feita em períodos de no máximo 7 dias, através da API "Tweepy" em Python (disponibilizada pela própria rede social para pesquisas de forma gratuita), divisão em períodos também sinalizada na seção 5.1.

A API "Tweepy", é associada a uma conta ativa na rede social, necessitando assim de que o proprietário da conta realize uma requisição pedindo permissão de uso, informando o que pretende realizar com os tweets capturados. Essa captura possui limitações de requisições por hora em sua versão gratuita, tendo também uma versão paga para uso comercial. Após a liberação de permissão de uso, o Twitter libera uma chave que estará associada ao usuário que realizar a requisição e deve ser usada sempre que a API for utilizada. O uso de buscas nessa API é limitado no máximo a 7 dias, é delimitado pela própria API, por isso fora feito a classificação em períodos, com o intuito de explorar desempenho entre esses períodos.

Para a coleta dos dados foi feita uma busca utilizando os filtros dados de dados abaixo indicados, que se atrelam ao índice Bovespa, ou seja, palavras chaves que representam uma

forte importância dentro da variação do índice, refletindo assim no mini-índice Bovespa. Os termos utilizados foram divididos em 2 categorias:

a) Símbolos referentes a ativos com boa liquidez no índice Bovespa:
   ABEV3, B3SA3, BBAS3, BBDC4, CSNA3, GGBR3, ITSA4, ITUB4, KLBN11, LAME4, LREN3, PETR3, PETR4, RAIL3, SUZB3, USIM5, VALE3, VVAR3, WEGE3.
b) Símbolo do ativo mini-índice usado nesta pesquisa e termos que se associem a bolsa de valores:
   WINJ21, BOVESPA, IBOVESPA, B3, CDB, CDI, AFTER MARKET, ALAVANCAGEM, ANÁLISE GRÁFICA, ANÁLISE TÉCNICA, BALANÇA COMERCIAL, BENCHMARK, BLUE CHIPS, BOLSA DE VALORES, CANDLE, COMMODITIES, COPOM, CORRETORAS DE VALORES, CVM, DAY TRADE, DAY TRADER, ESPECULAÇÃO, HEDGE, HOME BROKER, MERCADO DE CAPITAIS, MERCADO FUTURO, MINI-ÍNDICE, SCALPING, SMALL CAPS, SWING TRADE, TRADER.

Considerando esses filtros, foram capturados um total de 272.220 tweets contendo os seguintes dados: Data e hora (DD/MM/YYYY HH:MM) e texto postado no Twitter (tweet). A diante será apresentada a divisão dessa quantidade de tweets capturados de acordo com seus períodos.

Após a coleta, é feito o agrupamento das postagens vetorizadas por um período pré-determinado (5, 15 e 30 minutos), de acordo com o sequenciamento de negociação do ativo, conforme os testes que foram realizados neste trabalho, divisão explicada na seção 5.1.

4.2.3 Pré-Processamento da base de dados

Para os dados coletados no Twitter, foram aplicadas as seguintes técnicas de pré-processamento textual, já explicadas no item 2.2.1, que em síntese consiste na: normalização, remoção de pontuação, remoção de palavras muito pequenas, remoção de caracteres especiais, a remoção de *stopwords* e por fim, a tokenização palavras contidas nas mensagens.

Em continuidade aplicamos o método *Word2Vec*, mecanismo que já possui o TD-IDF embutido em seu processamento, e trabalhamos com duas bibliotecas distintas, as bibliotecas *nltk e gensim*, ambas em Python, transformando as palavras em vetores conforme já mencionado no item 2.2.3, e para tanto utilizamos a base de dados treinada, usando o próprio

corpus de textos para aprender a vetorizar através de ambas as bibliotecas. Em outras palavras foi feito uso das palavras, coletas para uso no treinamento a fim de aprender a vetorização com a própria base de dados. Na tabela 3 são apresentados os parâmetros usados para treinamento *Word2Vec.*

Tabela 3 - Parâmetros do treinamento com Word2Vec.

Parâmetro	**Descrição**	**Valor**
size (tamanho)	O número de dimensões do *embedding*, por exemplo, o comprimento do vetor denso para representar cada *token* (palavra).	200
window (janela)	A distância máxima entre uma palavra-alvo e palavras ao redor da palavra-alvo.	5
min_count	A contagem mínima de palavras a considerar ao treinar o modelo; palavras com uma ocorrência menor que essa contagem será ignorada.	2
workers	O número de *threads* a serem usados durante o treinamento.	32
sg	Aplicação do "*skip-gram model*".	1
epochs	Quantidade de épocas usadas para o treino.	60

Fonte: Elaborado pelo autor.

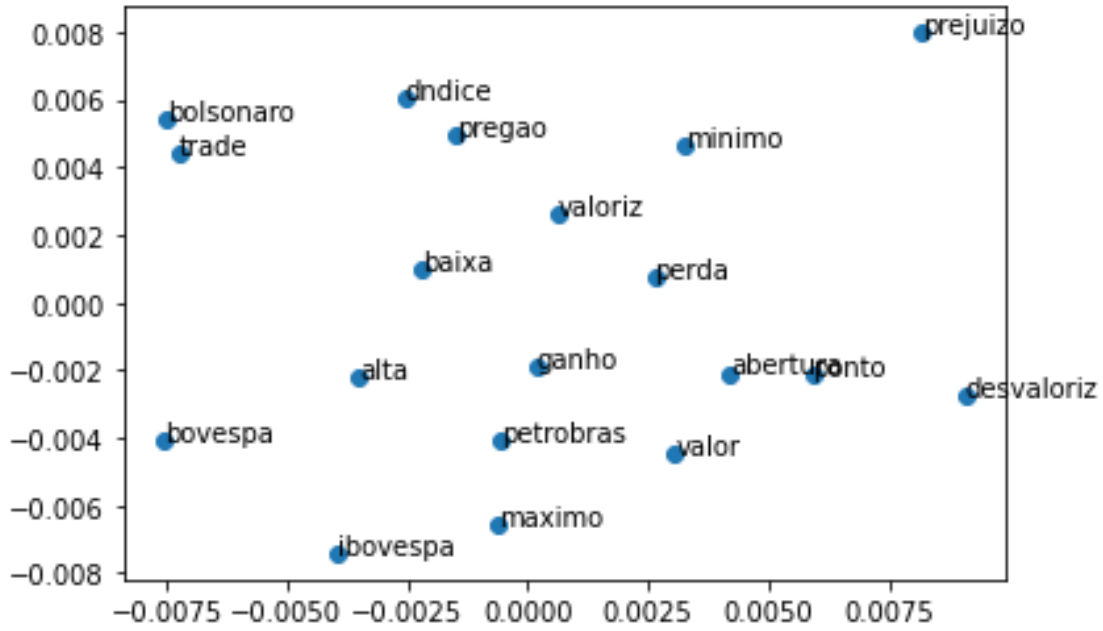
O *Word2Vec* faz uso de todos esses tokens para criação interna de um vocabulário ou conjunto de palavras, e após sua construção é feito o treinamento do modelo com *Word2Vec*, que nada mais é senão treinamento de uma rede neural simples com uma única camada oculta, cujo objetivo é o de aprender os pesos dessa camada, pesos que representam os vetores das palavras que o modelo está tentando aprender.

O treinamento desse modelo levou cerca de 3 a 4 horas com base em 272.220 tweets, gerando um total de 49.220 vetores, que são os códigos de cada palavra no espaço vetorial, após o treinamento. Como exemplo da relação semântica, podemos verificar, as palavras que são mais próximas no espaço vetorial da palavra 'BOVESPA' e seu distanciamento gráfico na Figura 16, para essa visualização foi usado a biblioteca gensim em python.

É fato que próximo à palavra 'BOVESPA' há palavras bem relacionadas com operações em bolsa de valores. Dentro da demonstração apresentada, no qual foram separadas 20 palavras mais próximas à "BOVESPA", é possível notar que palavras como "ganho", "máximo", "alta" estão totalmente relacionadas e próximas no espaço vetorial gerado posterior ao treinamento

com *Word2Vec*. Ainda é possível notar que as palavras “mínimo”, “prejuízo”, “perda”, localizadas no canto superior direito, estão também localizadas bem próximas no espaço vetorial de duas dimensões para representação gráfica.

Figura 16- Demonstração gráfica com treinamento Word2Vec.

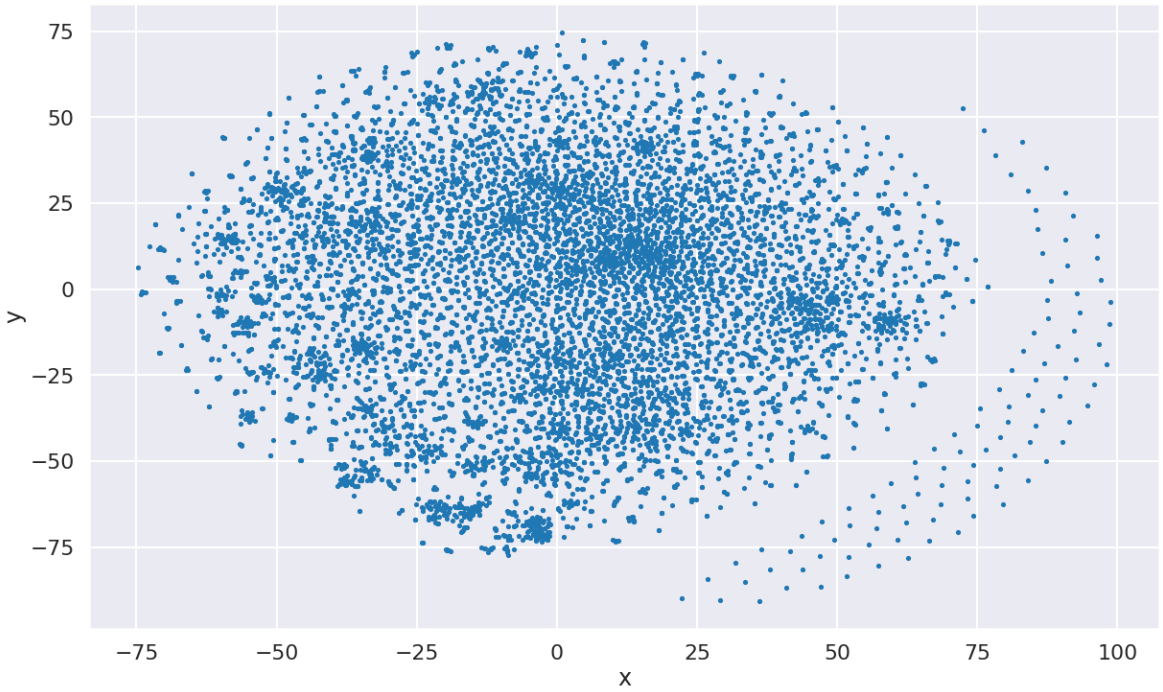


Fonte: Elaborado pelo autor.

Na Figura 17 é apresentado a demonstração gráfica de todas as palavras do corpus e sua representação vetorial, criando um gráfico de dispersão com os pontos de cada uma das palavras, sendo que esses resultados podem variar devido à natureza estocástica do algoritmo, procedimento de avaliação ou diferenças na precisão numérica, isso foi notado quando tentamos executar a demonstração gráfica em um outro momento.

Após a aplicação do *Word2Vec*, foi utilizado o *Doc2vec*, no qual conceitua-se como é um modelo que representa cada documento como um vetor, é uma extensão do *Word2Vec* também disponível na biblioteca *nltk* em Python, para realizar a vetorização de cada uma das frases das postagens em um único vetor, usando os próprios vetores gerados pelo *Word2Vec*, sendo o resultado um vetor para cada uma das postagens coletadas (frase completa). Esse passo é importante em se tratando de processamento de linguagem natural, pois o vetor capta o sentido individual de cada palavra, e em seguida seu contexto, de forma a captar o sentido semântico.

Figura 17 - Demonstração gráfica do corpus coletado.



Fonte: Elaborado pelo autor.

4.3 ROTULAÇÃO DOS DADOS

Com o intuito de supervisionar as amostras, é feita a rotulação como compra, venda e neutro conforme Algoritmo 1, usando o histórico do preço do ativo mini-índice.

No exemplo demonstrado no algoritmo, foi dividido o conjunto de dados em amostras de 5 minutos, de forma que ele possa agregar um maior número de informações. A rotulagem das amostras foi feita com base no método da janela deslizante, através de uma função interna do pacote *scipy.signal* e a função interna *arglextrema* que busca os pontos máximos e mínimos locais, como pontos de compra ou venda para servir como rótulos dos modelos.

Algoritmo 1 - Rotulação de Dados

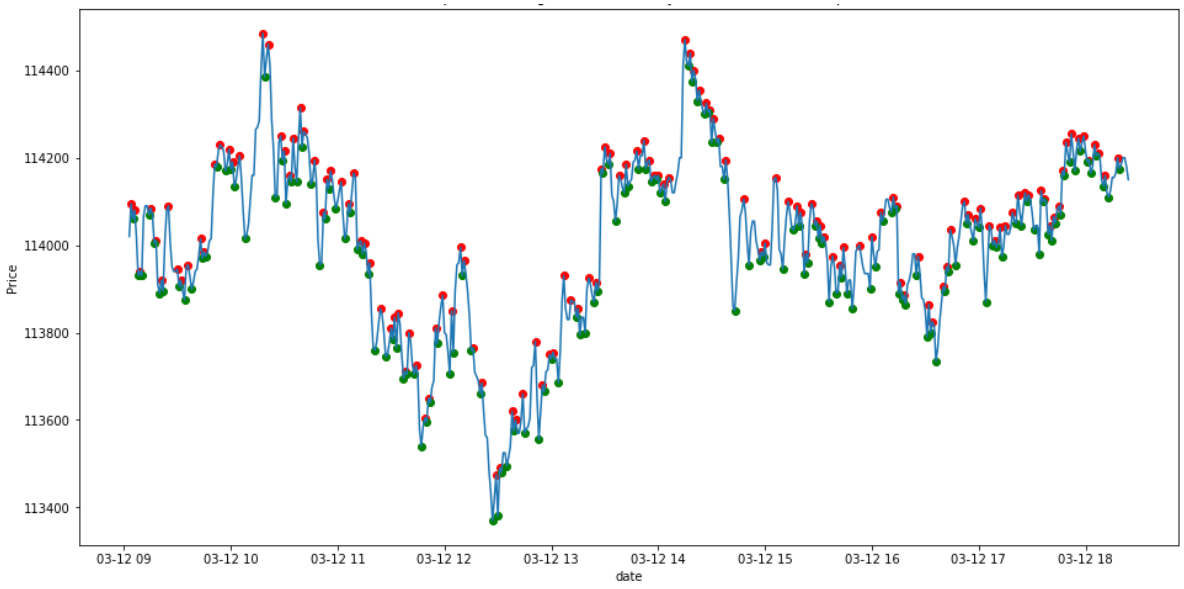
```
Var
        Entrada: pontosAtivo
        Saída: sinal
        JanInicio <- 0
        JanTamanho <- 5 (conforme desejar)
        janFinal <- JanInicio + JanTamanho
        ComprimentoTotal <- Comprimento(pontosAtivo)
Inicio
        Se (JanTamanho!= JanTamanho) ou JanTamanho < 1) então
          erro("JanTamanho precisa ser numero inteiro.")
        Senão
                enquanto x <= ComprimentoTotal faça
                        janAtual[] <- JanInicio:janFinal
                        vlrMin  <= Min(pontosAtivo[janAtual[]])
                        vlrMax  <= Max(pontosAtivo[janAtual[]])

                        para i in pontosAtivo[janAtual[]] faça
                            se pontosAtivo[i] = vlrMin então
                                sinal[i + 1] = "Compra";
                            fim
                            senão se pontosAtivo[i] = vlrMax então
                                sinal[i + 1] = "Venda";
                            fim
                            JanInicio <= janFinal + 1
                            janFinal <= JanInicio + JanTamanho
                            x <= janFinal
                 fim
        fim
        return Saída
FimAlgoritmo
```

Em resumo é feita a comparação entre os preços, e conforme foram encontrados os pontos de máximo e mínima do ativo foi feito a sua inserção em uma lista de mínimos e uma lista de máximas. O objetivo dessa rotulação é encontrar topos e fundos na série temporal, conforme demonstrado na Figura 18 com rotulagem de dados mais aberta feita com base em dias, utilizando uma janela de 11 períodos.

Já na Figura 18 são apresentados os sinais de compra e venda nos dados do ativo mini-índice conforme rotulação realizada através do algoritmo. O método de rotulagem divide a série temporal em segmentos menores, e é apresentado de sua concentração por minutos, no exemplo abaixo estamos utilizando uma janela de 5 períodos. O resultado da Figura 18 demonstra claramente os pontos de virada de tendência do preço do ativo e após essa virada um ponto de entrada, conforme sinalizado pela Teoria Dow, explicada na seção 2.1.6, e da Análise Técnica, explicada na seção 2.1.3.2, ambos usados neste trabalho. Os pontos verdes da Figura 18 são pontos de compra do ativo e os pontos vermelhos são pontos de venda.

Figura 18 - Rotulagem com 5 períodos.



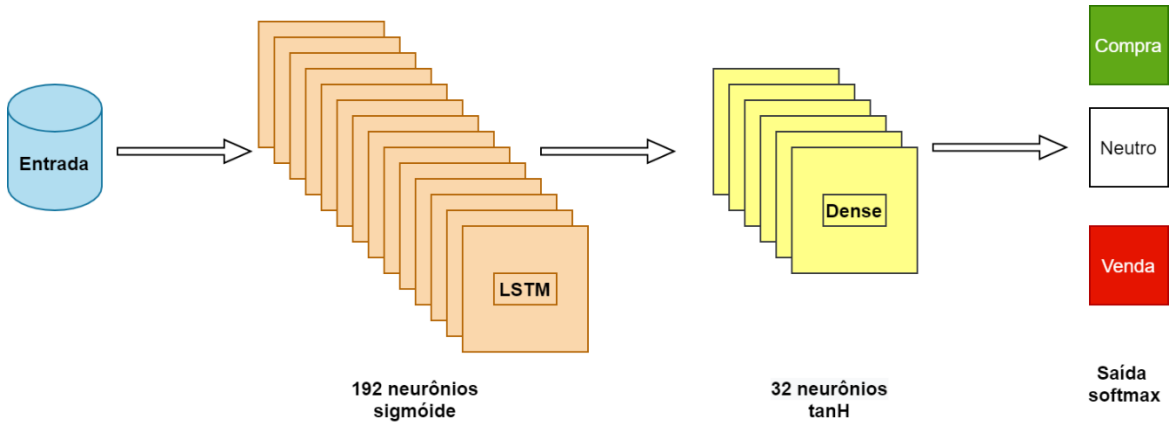
Fonte: Elaborado pelo autor.

4.4 ARQUITETURA DA REDE LSTM USADA NO MODELO

O modelo de predição foi feito com base no modelo sequencial usando o *framework* Keras (CHOLLET et al., 2015) para *deep learning* e utilizando a camada *Long Short Term Memory* (LSTM).

A LSTM está estruturada da seguinte maneira: A rede possui uma dimensão de 3 camadas, sendo a primeira LSTM com formato de 192 neurônios de recorrência servindo de entrada para a camada densa e com ativação sigmóide. A segunda camada densa, totalmente conectada, possui 32 neurônios normais e ativação com tangente hiperbólica. E a terceira camada *dense* é a saída do modelo e possui 3 saídas usando a ativação so*ftmax*, isso porque são consideradas 3 classes para previsão: compra, venda e aguarda (*hold*).

Figura 19 - Demonstração da arquitetura de rede do modelo proposto.



Fonte: Elaborado pelo autor.

Na entrada da rede é feita a junção das duas bases de dados, a do mini-índice e dos dados vetorizados do Twitter. No período em que não havia *tweets* foram considerados os *tweets* do período sequencial anterior. Ainda como entrada para delimitação de peso, considera-se o número de *retwetts* e o número de favoritos, pois com eles é possível delimitar um peso para as mensagens que são mais importantes.

Para desenvolvimento do trabalho utiliza-se a linguagem de programação Python, além de bibliotecas e API's fornecidas para a linguagem, auxiliando assim o desenvolvimento e posterior análise dos resultados. O uso dessa linguagem será feito para coleta de dados do Twitter e também nas partes relacionadas ao treinamento e testes do modelo. A linguagem foi escolhida devido a fácil familiaridade junto aos recursos de auto processamento e sua maturidade na área de desenvolvimento.

A arquitetura usada consiste tanto no uso do Anaconda, utilizada para instalação e gerenciamento de pacotes, suportando o uso do "*Jupyter notebook*" como também o uso do "*Google Cloud*" com o uso do "*Google Colab*".

Dentre as bibliotecas ou API disponibilizadas para Python, as principais utilizadas foram: "Numpy", "Pandas", "Matplotlib", "*Yfinance*", "*Scikit-learn*", "NLTK", "Keras" entre outras.

5. EXPERIMENTOS REALIZADOS

Nesta seção será apresentada os detalhes dos experimentos realizados conforme a metodologia utilizada. A seção 5.1 apresenta a divisão dos dados conforme sequenciamento periódico de negociação do ativo e divisão por períodos diários das análises feitas. A seção 5.2 apresenta o modelo proposto neste trabalho sendo comparado com a execução de outros algoritmos de classificação. A seção 5.3 demonstra o método de treinamento usado neste trabalho. Na seção 5.4 apresenta as métricas estatísticas usadas neste trabalho e por fim na 5.5 a avaliação financeira realizada neste trabalho.

5.1 CONJUNTO DE DADOS POR SEQUENCIAMENTO E PERIODOS

As coletas dos dados de postagens da rede social Twitter foram feitas conforme limite de 7 dias, demonstrado na seção 4.2.2. Com isso, foram feitas avaliações em 4 períodos diferentes, sendo esses períodos analisados da seguinte forma:

- Período 1: de 04-03-2021 até 12-03-2021.
- Período 2: de 15-03-2021 até 19-03-2021.
- Período 3: de 29-03-2021 até 01-04-2021.
- Período 4: de 05-04-2021 até 09-03-2021.

Esses períodos são analisados também de forma acumulativa em avaliações feitas ao longo do trabalho. Há um período menor devido ao feriado que aconteceu, mas o intuito foi manter sempre uma semana por completa na avaliação.

Com o intuito de explorar qual sequenciamento da série temporal pode obter melhores resultados relacionado as postagens da rede social, foi feita a coleta dos dados da série temporal do ativo mini-índice, apresentado na seção 2.1.5 e 4.2.1, em vários sequenciamentos, desde sequenciamentos mais curtos como de um 1 minuto até mais longos como o de 60 minutos. No entanto os dois extremos foram descartados da pesquisa, pois optamos por analisar os sequenciamentos de 5, 15 e 30 minutos, uma vez que entendemos que a coleta de dados do sequenciamento de 5, 15 e 30 minutos nos traria informações mais precisas sobre os dados coletados.

O sequenciamento nada mais é que agrupamentos por determinado período, em operações diárias, por exemplo, o sequenciamento é feito através do preço do ativo fechado no dia. Isso depende muito da forma como o investidor deseja operar. Como o intuito do modelo é realizar o estudo no período "*intraday*", os sequenciamentos foram feitos com base em negociações feitas dentro de um mesmo dia.

Juntamente aos dados do ativo, foi feita a junção dos dados da base de dados textual, Twitter, e criado três amostras de experimentos diferentes, a primeira amostra para o sequenciamento de 5 minutos, o segundo para o sequenciamento de 15 minutos e o terceiro para o sequenciamento de 30 minutos.

5.2 COMPARAÇÕES DO MODELO

O modelo proposto foi comparado com outros modelos de sistemas de classificação como Regressão Logística, Floresta Aleatória (do inglês *Random Forest*), Máquina de vetores de suporte (SVM do inglês *Support Vector Machine*) e Vizinhos Mais Próximos (KNN do inglês *KNearest Neighbors*), a fim de comparar o desempenho do modelo com outros algoritmos.

Esta comparação tem intuito testar a eficácia desta abordagem de classificação de operação com outros modelos através de métricas estatísticas das estratégias. Dentre as métricas estatísticas usadas para comparação temos: F1, Acurácia, Precisão e Revocação (*recall).*

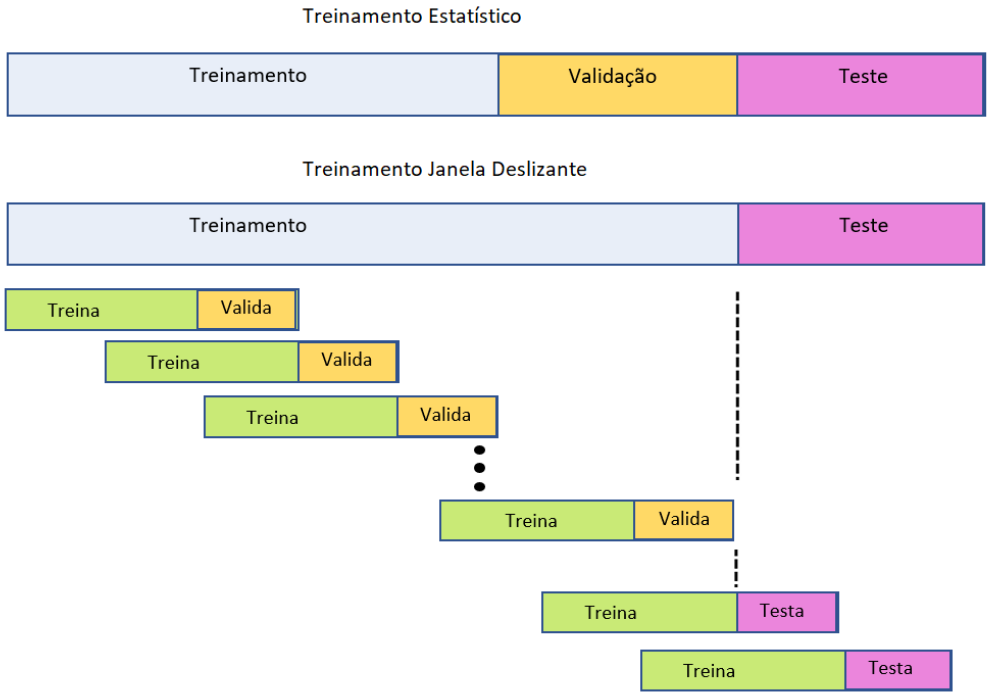
Em se tratando de análises financeiras, foi feito a comparação com a estratégia "*buy and hold*", comprar e segurar, com o intuito de avaliar o desempenho do modelo em relação a estratégias conservadoras de investidores.

A implementação do RNNTW-TC teve bons resultados e consegue se destacar de outros modelos na avaliação estatística e também com a estratégia *buy and hold* no que trata de retornos financeiros. Esses resultados não foram gerados logo na primeira implementação, e houve vários ajustes feitos nos hiper parâmetros (earlyStopping, modelCheckpoint, reduceLROnPlateau, metrics, sparse, batch_size, epoch) para que pudesse atingir resultados melhores para este modelo.

5.3 PROCESSO DE TREINAMENTO

Para o processo de treinamento usamos o modelo da janela deslizante (*sliding window*), onde de acordo com Dietterich (2002) é o modelo mais adequado a ser usado para série de dados temporais, pois permite que o modelo se ajuste aos eventuais ciclos e tendências da série de dados. A divisão dos dados foi feita separando 50% para treinamento, 25% para validação e 25% para testes, método K-Fold no qual é feito validação cruzada dividindo o conjunto total de dados em k subconjuntos e, a partir daí, um subconjunto é utilizado para teste e os k-1 restante são usados para estimação dos parâmetros fazendo o cálculo da acurácia do modelo.

Figura 20 - Treinamento do Modelo.



Fonte: Elaborado pelo autor

A técnica é realizada definindo uma janela de treinamento realizando o treinamento do modelo com os dados, e avançando uma quantidade de registros, fazendo o treinamento com os dados e avançando até terminar essa base de dados. A validação por janela deslizante resulta no treinamento ilustrado na Figura 20.

5.4 AVALIAÇÃO ESTATÍSTICA

O modelo foi avaliado através da utilização das seguintes métricas: Acurácia, Precisão *(precision)* e Revocação (*Recall)* e F1. Abaixo é demonstrado as métricas estatísticas usadas neste trabalho. A seguir é apresentado o método de cálculo base para cada métrica para melhor compreensão (POWERS, 2020), no qual:

VP: verdadeiro positivo, por exemplo é classificado compra em um dia que foi rotulado como compra na base de dados;

VN: verdadeiro negativo, por exemplo não é classificado venda/neutro em um dia que foi rotulado como compra;

FP: falso positivo, por exemplo é classificado compra em um dia que foi rotulado como venda na base de dados;

FN: falso negativo, por exemplo a classificação deveria ser compra, mas foi classificado venda.

Este estudo também utiliza essas mesmas métricas para comparar o modelo proposto com outros algoritmos conforme mencionado na seção 5.2.

a) **Acurácia**: é a métrica que pode ser usada para demonstrar o desempenho geral do modelo. Usa em seu cálculo o número de acertos (positivos) dividido pelo número total de exemplos, ou seja, é o percentual de acerto com base na quantidade total de registros, a fórmula desta métrica é apresentada pela Equação (16).

$$acurácia = \frac{VP+VN}{VP+FP+FP+FN} \tag{16}$$

b) **Precisão:** é a métrica que avalia dentro os registros classificados como certos, quais realmente estavam corretos. É feito a divisão dos positivos verdadeiros pela soma desse valor e o número de exemplos classificados como falsos positivos. A fórmula desta métrica é apresentada pela Equação (17).

$$precision = \frac{VP}{VP+FP} \tag{17}$$

c) **Revocação *(Recall)*:** é a métrica que avalia sensibilidade, dentre os positivos reais qual a proporção de acerto. Sua fórmula, conforme Equação (18), faz a divisão do

número de positivos verdadeiros pela quantidade total de exemplos que pertencem a esta classe somado aos falsos negativos.

$$revocação = \frac{VP}{VP+FN} \tag{18}$$

d) **F1 Score**: Assume que a precisão e a revocação tem a mesma importância, logo seria o melhor modelo de avaliação para base de dados com classes desproporcionais. Sua fórmula é apresentada na conforme Equação (19).

$$F_1 = 2 * \frac{precision*recall}{precision+recall} \tag{19}$$

5.5 AVALIAÇÃO FINANCEIRA

Com o objetivo de avaliar financeiramente o modelo proposto neste trabalho, foram simuladas operações de negociação no ativo mini-índice, código WINJ21, nos dados de teste a cada sessão de treinamento, conforme os períodos avaliados.

Esta avaliação realiza simulações de compra e venda no momento da classificação, conforme demonstrado através da rotulação na seção 4.3, utilizando o valor de fechamento do ativo no sequenciamento avaliado, baseado na Análise Técnica, demonstrado na seção 2.1.3.2.

Para cada um dos quatro períodos avaliados é feita a simulação capturando o resultado e armazenando, com isso é possível avaliar o comportamento do modelo, sobre retornos financeiros, dentro de cada um dos períodos avaliados. Além da avaliação, por cada um dos períodos, é feita uma avaliação geral, realizando o acumulado da movimentação dos pontos ganhos e perdidos para as operações simuladas durante os 4 períodos avaliados. O acumulado de pontos é contabilizado conforme valor de cada ponto do ativo mini-índice.

O valor inicial usado para a simulação proposta para este trabalho é de R$ 22.267,00, conforme mencionado na seção 2.1.5, que se refere a um valor equivalente a 20% do fechamento do índice na data anterior ao início das simulações deste trabalho. Diante disso, as variações de lucro e prejuízo serão cálculos sobre esse valor.

Ainda, todas essas avaliações são feitas para cada um dos três sequenciamentos usados neste trabalho (5, 15 e 30 minutos). Com o intuito de realizar avaliação de qual dos três sequenciamentos pode ser mais rentável é feita uma comparação entre a rentabilidade de cada

um destes sequenciamentos. Por fim, é acrescido a comparação entre os três sequenciamentos de operações a comparação com a estratégia *buy and hold* a fim de notar o desempenho dos modelos quanto a essa estratégia ao longo dos quatro períodos avaliados e no acumulado dos períodos.

6. ANÁLISE DOS RESULTADOS

Dos 272.220 *tweets* obtidos entre os dias 04-03-2021 e 09-04-2021 através do pacote Tweepy e considerando os filtros já apresentados, chegamos as seguintes informações de acordo com os períodos estudados e o agrupamento por tempo de sequenciamento de 5, 15 e 30 minutos.

Conforme já mencionado o valor usado como base para os cálculos partirá de R$ 22.267,00, evitando assim saída inesperada da operação em caso de movimentação brusca no ativo. Para o montante referente ao valor inicial, iremos apresentar seus resultados acumulados nos Resultados Gerais, seção 6.5.

6.1 RESULTADOS NO PERÍODO 1

Esse período vai do dia 04/03/2021 ao 12/03/2021, totalizando 7 dias úteis, no qual coletamos 20542 *tweets* que foram agrupados com os dados do mini-índice. Quanto ao período da análise dos *tweets*, agrupamos por períodos de 5, 15 e 30 minutos, conforme sequenciamento de movimentação do ativo mini-índice. Nesse período tivemos uma geração de 790 *candles*, também chamados de *candlestick* e consiste na representação gráfica do preço de um ativo ao longo de determinado período, por exemplo, 30 minutos, um dia ou um mês, de pontos fechamento para o sequenciamento de 5 minutos, 266 *candles* de pontos fechamento para o sequenciamento de 15 minutos e 133 *candles* pontos de fechamento para o sequenciamento de 30 minutos.

Foi dividido a análise em resultados estatísticos (acurácia, precisão, revocação e F1) em comparação com outros algoritmos e também com o próprio modelo LSTM sem os dados do Twitter. Por fim foi feito a análise dos resultados financeiros, vejamos:

6.1.1 Resultados Estatísticos no Período 1

Da análise dos gráficos estatísticos e financeiros que elaboramos no desenvolvimento da nossa pesquisa, e que abaixo passamos a apresentar de modo detalhado, entendemos ser oportuno apresentar de antemão os principais resultados que constatamos para o período, deixando desde já evidente que o método RNNTW-TC tem performance com valores em destaque comparada aos outros algoritmos de classificação.

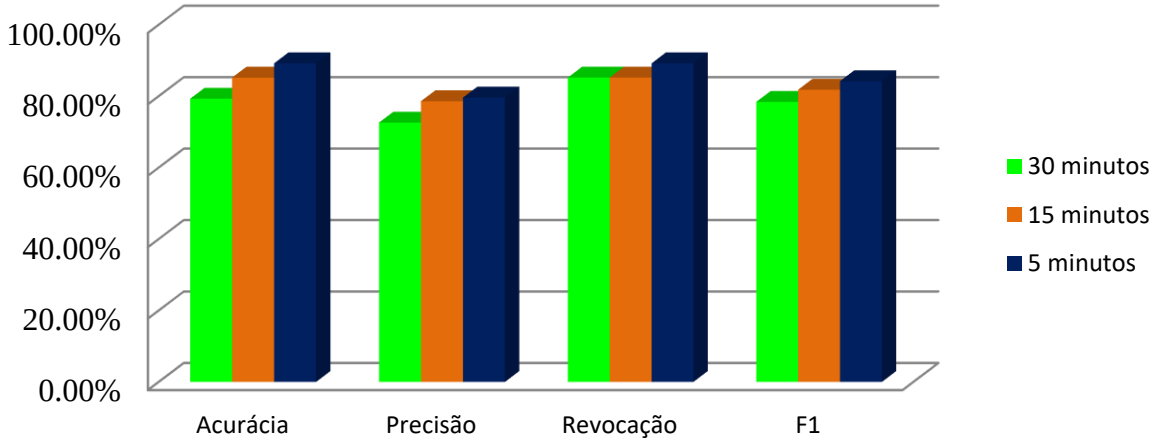
Abaixo apresentamos tabela e gráfico com os dados mais relevantes dos minutos agrupados apenas do modelo RNNTW-TC para esse Período 1.

Tabela 4 - Métricas de avaliação estatística o Período 1.

Sequenciamento	Acurácia	Precisão	Revocação	F1
30 minutos	79.29%	72.60%	85.20%	78.40%
15 minutos	85.19%	78.60%	85.20%	81.80%
5 minutos	89.20%	79.60%	89.20%	84.20%

Fonte: Elaborado pelo autor.

Figura 21 - Resultado estatístico em barras no Período 1.



Fonte: Elaborado pelo autor.

Conforme desenvolveremos com melhor atenção quando da análise final, é válido neste momento apontar que quando o agrupamento dos *tweets* é feito por período de 5 minutos os resultados estatísticos são melhores e conforme o período do agrupamento aumenta os resultados pioram. O notável desempenho é apresentado através da visão por um gráfico linear com a comparação de sequenciamentos entre os resultados estatísticos do modelo RNNTW-TC.

6.1.1.1 Sequenciamento 5 minutos no Período 1

Quanto ao sequenciamento de 5 minutos, no qual tivemos 790 *candles* de pontos de fechamentos. Passamos a apresentar os resultados estatísticos em comparação com outros algoritmos de classificação:

Tabela 5 - Avaliação estatística no sequenciamento de 5 minutos para o Período 1.

Modelo	Acurácia	Precisão	Revocação	F1
RNNTW-TC	**89.20%**	**79.60%**	**89.20%**	**84.20%**
LSTM	85.10%	75.50%	84.40%	80.10%
Regressão Logística	87.40%	69.72%	63.10%	61.31%
Random Forest	77.22%	69.63%	62.15%	70.83%
SVM	87.85%	69.75%	63.30%	71.43%
KNeighbors	87.97%	69.63%	62.15%	70.84%

Fonte: Elaborado pelo autor.

Verifica-se pela Tabela 5, que o modelo RNNTW-TC apresentou melhor desempenho nos quatro quesitos analisados (Acurácia, Precisão, Revocação e F1), e fazendo uma análise geral podemos constatar que o desempenho do sistema desenvolvido em alguns casos tem resultado com Precisão, Revocação e F1 bem superior ao dos outros algoritmos. O modelo RNNTW-TC, que usa o modelo desenvolvido com uma rede LSTM acrescida aos dados da rede social Twitter, foi comparado também com a execução dos dados com a rede LSTM sem os dados do Twitter (LSTM na tabela), e nessa comparação o modelo apresentou uma superioridade também.

6.1.1.2 Sequenciamento 15 minutos no Período 1

Quanto ao sequenciamento de 15 minutos, no qual tivemos 266 *candles* de pontos de fechamentos. Passamos a apresentar os resultados estatísticos em comparação com outros algoritmos de classificação:

Tabela 6 - Avaliação estatística no sequenciamento de 15 minutos para o Período 1.

Modelo	Acurácia	Precisão	Revocação	F1
RNNTW-TC	**85.19%**	**78.60%**	**85.20%**	**81.80%**
LSTM	83.99%	76.40%	83.00%	79.60%
Regressão Logística	86.47%	69.63%	73.33%	71.37%
Random Forest	89.10%	69.41%	71.25%	70.30%
SVM	86.84%	69.62%	71.30%	73.30%
KNeighbors	86.47%	69.33%	70.55%	69.90%

Fonte: Elaborado pelo autor.

Verifica-se pela Tabela 6, que o modelo RNNTW-TC apresentou melhor desempenho em três quesitos analisados (Precisão, Revocação e F1). Por exemplo, o quesito Acurácia teve valor menor em algumas comparações, mas ainda assim teve um valor bem próximo aos outros algoritmos. Ao comparar com a rede LSTM sem dados do Twitter, podemos ver que os dados do Twitter conseguem uma leve melhora sobre os resultados obtidos em todos os critérios.

6.1.1.3 Sequenciamento 30 minutos no Período 1

Quanto ao sequenciamento de 30 minutos, no qual tivemos 133 *candles* de pontos de fechamentos. Passamos a apresentar os resultados estatísticos em comparação com outros algoritmos de classificação.

Verifica-se pela Tabela 7, que o modelo RNNTW-TC apresentou melhor desempenho em três quesitos analisados (Precisão, Revocação e F1). O quesito Acurácia teve valor menor para todas as comparações e ainda assim teve um valor bem próximo aos outros algoritmos, mas teve a distância dentro desse quesito maior que o sequenciamento de 15 minutos. Ao comparar com a rede LSTM sem dados do Twitter, podemos ver que os dados do Twitter conseguem uma leve melhora sobre os resultados obtidos em todos os critérios.

Tabela 7 - Avaliação estatística no sequenciamento de 30 minutos para o Período 1.

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1**
RNNTW-TC	**79.29%**	**72.60%**	**85.20%**	**78.40%**
LSTM	78.53%	71.84%	84.44%	77.64%
Regressão Logística	84.21%	68.30%	67.73%	71.90%
Random Forest	84.96%	68.39%	67.30%	72.66%
SVM	84.96%	61.21%	71.10%	71.20%
KNeighbors	85.71%	68.40%	78.90%	74.67%

Fonte: Elaborado pelo autor.

6.1.2 Avaliação financeira

Os resultados da avaliação financeira para o Período 1 demonstra a comparação entre os três sequenciamentos usados neste trabalho. Além disso é feito a comparação com a estratégia *buy and hold*, comprar e segurar. É feito a demonstração primeiramente dos resultados referentes a pontos ganhos por contrato do mini-índice e na sequência a

demonstração dos valores referentes a cada ponto movimentado. Cada ponto movimentado do contrato do mini-índice equivale a R$ 0,20 (vinte centavos de reais).

Passamos a apresentar os resultados da avaliação financeira para o Período 1:

Tabela 8 - Demonstração de pontos movimentados no Período 1.

Data	Buy/Hold	5min	15min	30min
4-Mar-21	1730	420	1590	200
5-Mar-21	655	3750	2765	500
6-Mar-21	-3445	-425	-805	800
7-Mar-21	1045	1530	2740	500
8-Mar-21	2295	4030	-555	800
11-Mar-21	-10	1430	1605	700
12-Mar-21	130	1410	300	500
	2400	**12145**	**7640**	**4000**

Fonte: Elaborado pelo autor.

Na Tabela 8 é apresentada a quantidade de pontos movimentados, tanto para cima quanto para baixo, para cada um dos sequenciamentos e para a estratégia de *buy and hold* para todos os dias presentes neste período.

Na análise financeira para o Período 1, o modelo RNNTW-TC obteve valores maiores relacionados a comparação com a estratégia *buy and hold*, comprar e segurar, em todos os sequenciamentos gerados pelo modelo.

Tabela 9 - Demonstração do valor financeiro diário no Período 1.

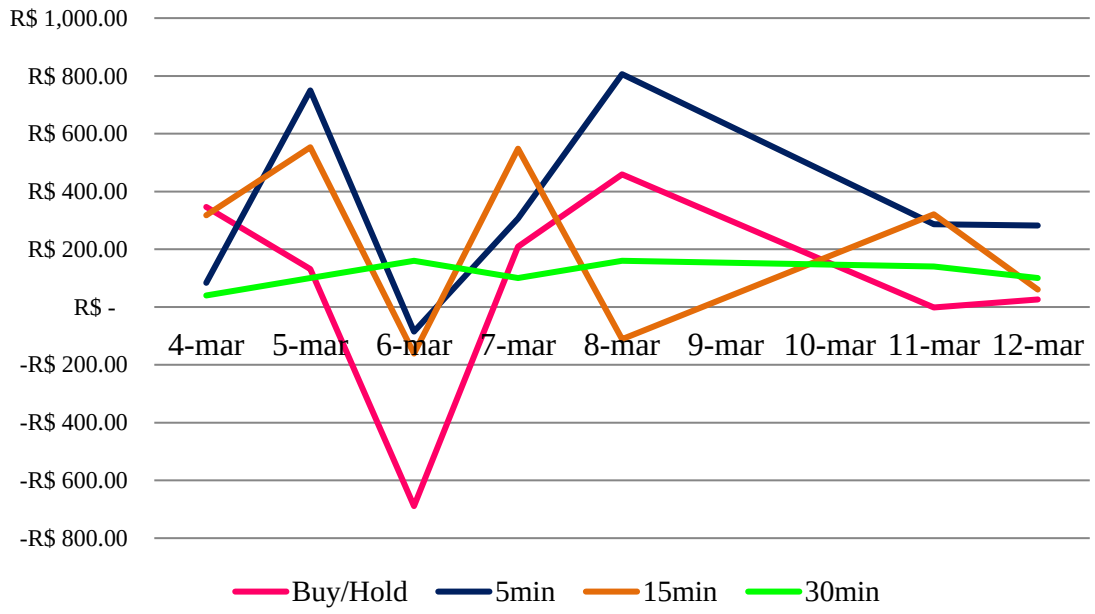
Data	Buy/Hold	5min	15min	30min
4-Mar-21	R$ 346.00	R$ 84.00	R$ 318.00	R$ 40.00
5-Mar-21	R$ 131.00	R$ 750.00	R$ 553.00	R$ 100.00
6-Mar-21	-R$ 689.00	-R$ 85.00	-R$ 161.00	R$ 160.00
7-Mar-21	R$ 209.00	R$ 306.00	R$ 548.00	R$ 100.00
8-Mar-21	R$ 459.00	R$ 806.00	-R$ 111.00	R$ 160.00
11-Mar-21	-R$ 2.00	R$ 286.00	R$ 321.00	R$ 140.00
12-Mar-21	R$ 26.00	R$ 282.00	R$ 60.00	R$ 100.00
	R$ 480.00	**R$ 2,429.00**	**R$ 1,528.00**	**R$ 800.00**

Fonte: Elaborado pelo autor.

Na comparação entre os sequenciamentos, assim como na avaliação estatística deste período é possível notar que os valores para o sequenciamento de 5 minutos são melhores em

relação aos outros sequenciamentos com tempo maior, em seguida temos o sequenciamento de 15 minutos e de 30 minutos nessa ordem. Abaixo é demonstrado através do gráfico em linhas que apresenta uma visão na comparação entre os dias presentes neste período.

Figura 22 - Demonstração da rentabilidade gerada para os dias do Período 1.



Fonte: Elaborado pelo autor.

6.2 RESULTADOS NO PERÍODO 2

Esse período vai do dia 15/03/2021 ao 19/03/2021, totalizando 5 dias úteis, no qual coletamos 6.941 *tweets* que foram agrupados com os dados do mini-índice. Quanto ao período da análise dos *tweets*, agrupamos por períodos de 5, 15 e 30 minutos, conforme sequenciamento de movimentação do ativo mini-índice. Nesse período tivemos uma geração de 535 *candles* de fechamento para o sequenciamento de 5 minutos, 180 *candles* de fechamento para o sequenciamento de 15 minutos e 90 *candles* de fechamento para o sequenciamento de 30 minutos.

Foi dividido a análise em resultados estatísticos (acurácia, precisão, recall e F1) em comparação com outros algoritmos e em resultados financeiros, vejamos:

6.2.1 Resultados Estatísticos no Período 2

Assim como nos resultados apresentados para o período 1, da análise dos gráficos estatísticos e financeiros que elaboramos no desenvolvimento da nossa pesquisa, e que abaixo passamos a apresentar de modo detalhado, entendemos ser oportuno apresentar de antemão os principais resultados que constatamos, deixando desde já evidente que o método RNNTW-TC tem performance com valores em destaque comparada aos outros algoritmos de classificação.

Abaixo apresentamos tabela e gráfico com os dados mais relevantes dos minutos agrupados apenas do modelo RNNTW-TC para esse Período 2.

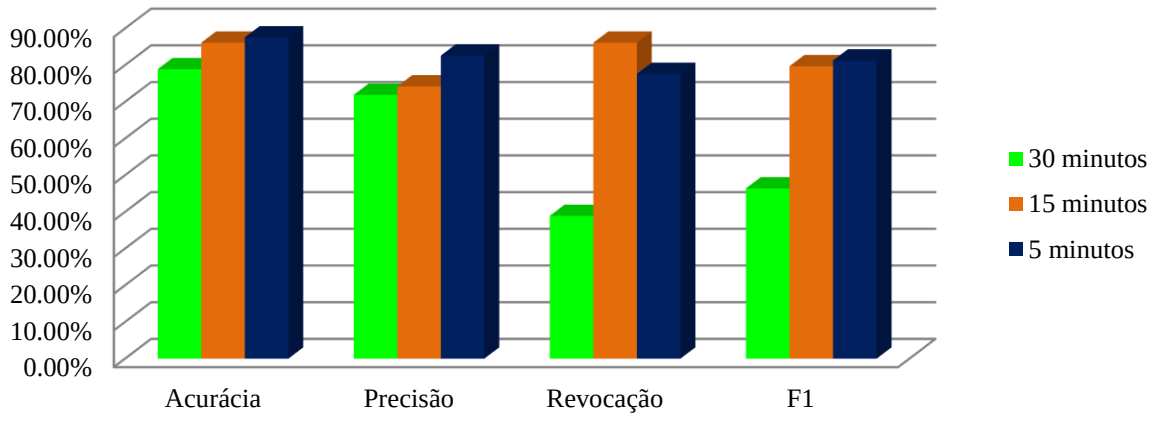
Tabela 10 - Métricas de avaliação estatística o Período 2.

Sequenciamento	**Acurácia**	**Precisão**	**Revocação**	**F1**
30 minutos	78.89%	71.90%	38.90%	46.40%
15 minutos	86.11%	74.20%	86.10%	79.70%
5 minutos	87.57%	82.60%	77.60%	81.26%

Fonte: Elaborado pelo autor.

Assim como no Período 1, neste Período 2 estaremos desenvolvendo com melhor atenção quando da análise final, valido neste momento apontar que quando o agrupamento dos *tweets* é feito por período de 5 minutos os resultados estatísticos são melhores e conforme o período do agrupamento aumenta os resultados pioram. O notável desempenho é apresentado através da visão por um gráfico linear com a comparação de sequenciamentos entre os resultados estatísticos do modelo RNNTW-TC.

Figura 23 - Resultado estatístico em barras no Período 2.



Fonte: Elaborado pelo autor.

6.2.1.1 Sequenciamento 5 minutos no Período 2

Quanto ao sequenciamento de 5 minutos, no qual tivemos 535 *candles* de pontos de fechamentos. Passamos a apresentar os resultados estatísticos: Passamos a apresentar os resultados estatísticos em comparação com outros algoritmos de classificação:

Tabela 11 - Avaliação estatística no sequenciamento de 5 minutos para o Período 2.

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1**
RNNTW-TC	**87.57%**	**82.60%**	**77.60%**	**81.26%**
LSTM	83.70%	78.73%	73.73%	77.39%
Regressão Logística	87.66%	71.90%	73.30%	71.86%
Random Forest	86.64%	70.27%	70.27%	70.27%
SVM	87.85%	70.53%	71.87%	73.33%
KNeighbors	88.60%	77.80%	72.99%	71.69%

Fonte: Elaborado pelo autor.

Verifica-se pela Tabela 11, que o modelo RNNTW-TC apresentou melhor desempenho nos quatro quesitos analisados (Acurácia, Precisão, Revocação e F1), e fazendo uma análise geral podemos constatar que o desempenho do sistema por nos desenvolvido em alguns casos tem resultado com Precisão, Revocação e F1 bem superior ao dos outros algoritmos. O modelo RNNTW-TC, que usa o modelo desenvolvido com uma rede LSTM acrescida aos dados da rede social Twitter, foi comparado também com a execução dos dados com a rede LSTM sem os dados do Twitter (LSTM na tabela), e nessa comparação o modelo apresentou uma superioridade também.

6.2.1.2 Sequenciamento 15 minutos no Período 2

Quanto ao sequenciamento de 15 minutos, no qual tivemos 180 *candles* (pontos de fechamentos). Passamos a apresentar os resultados estatísticos em comparação com outros algoritmos de classificação:

Verifica-se pela Tabela 12, que o modelo RNNTW-TC apresentou melhor desempenho em quatro quesitos analisados (Acurácia, Precisão, Revocação e F1). Ao comparar com a rede LSTM sem dados do Twitter, podemos ver que os dados do Twitter conseguem uma leve melhora sobre os resultados obtidos em todos os critérios.

Tabela 12 - Avaliação estatística no sequenciamento de 15 minutos para o Período 2.

Modelo	Acurácia	Precisão	Revocação	F1
RNNTW-TC	**86.11%**	**74.20%**	**86.10%**	**79.70%**
LSTM	85.00%	73.09%	84.57%	78.59%
Regressão Logística	86.78%	68.70%	69.62%	70.84%
Random Forest	83.22%	71.25%	75.90%	70.53%
SVM	84.80%	64.30%	79.84%	73.30%
KNeighbors	82.89%	70.27%	72.73%	70.84%

Fonte: Elaborado pelo autor.

6.2.1.3 Sequenciamento 30 minutos no Período 2

Quanto ao sequenciamento de 30 minutos, no qual tivemos 90 *candles* (pontos de fechamentos). Passamos a apresentar os resultados estatísticos em comparação com outros algoritmos de classificação:

Tabela 13 - Avaliação estatística no sequenciamento de 30 minutos para o Período 2.

Modelo	Acurácia	Precisão	Revocação	F1
RNNTW-TC	**78.89%**	**71.90%**	**38.90%**	**46.40%**
LSTM	78.18%	71.19%	38.19%	45.69%
Regressão Logística	75.91%	73.30%	67.81%	78.10%
Random Forest	72.30%	67.70%	78.17%	71.30%
SVM	78.21%	64.35%	64.30%	70.19%
KNeighbors	71.19%	61.53%	75.10%	71.10%

Fonte: Elaborado pelo autor.

Verifica-se pela Tabela 13, que o modelo RNNTW-TC apresentou melhor desempenho em quatro quesitos analisados (Acurácia, Precisão, Revocação e F1). O quesito Acurácia diferente do Período 1, nos sequenciamentos de 15 e 30 minutos teve o valor acima dos demais, apesar de próximo. Ao comparar com a rede LSTM sem dados do Twitter, podemos ver que os dados do Twitter conseguem uma leve melhora sobre os resultados obtidos em todos os critérios.

6.2.2 Avaliação financeira para o Período 2

Os resultados da avaliação financeira para o Período 2 demonstra a comparação entre os três sequenciamentos usados neste trabalho. Além disso é feito a comparação com a estratégia *buy and hold*, comprar e segurar. É feito a demonstração primeiramente dos resultados referentes a pontos ganhos por contrato do mini-índice e na sequência a demonstração dos valores referentes a cada ponto movimentado, cada ponto movimentado do contrato do mini-índice equivale a R$ 0,20 (vinte centavos de reais).

Passamos a apresentar os resultados da avaliação financeira para o Período 2:

Para a análise financeira para o Período 2, a Tabela 14 apresenta a quantidade de pontos movimentados, tanto para cima quanto para baixo, para cada um dos sequenciamentos e para a estratégia de *buy and hold* para todos os dias presentes neste período.

Tabela 14 - Demonstração de pontos movimentados no Período 2.

Data	Buy/Hold	5min	15min	30min
15-Mar-21	720	1480	600	400
16-Mar-21	-1560	2170	500	200
17-Mar-21	2700	3420	1200	800
18-Mar-21	-1135	2365	2100	2000
19-Mar-21	660	2250	1800	1300
	1385	**11685**	**6200**	**4700**

Fonte: Elaborado pelo autor.

Na análise financeira para o Período 2, o modelo RNNTW-TC obteve valores maiores relacionados a comparação com a estratégia *buy and hold*, comprar e segurar, em todos os sequenciamentos gerados pelo modelo.

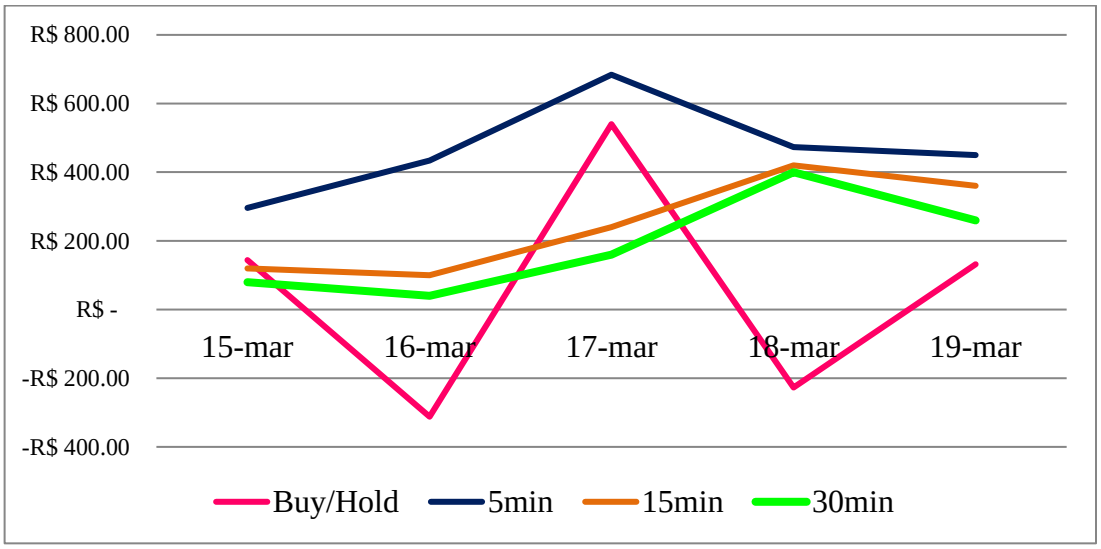
Tabela 15 - Demonstração do valor financeiro diário no Período 2.

Data	Buy/Hold	5min	15min	30min
15-Mar-21	R$ 144.00	R$ 296.00	R$ 120.00	R$ 80.00
16-Mar-21	-R$ 312.00	R$ 434.00	R$ 100.00	R$ 40.00
17-Mar-21	R$ 540.00	R$ 684.00	R$ 240.00	R$ 160.00
18-Mar-21	-R$ 227.00	R$ 473.00	R$ 420.00	R$ 400.00
19-Mar-21	R$ 132.00	R$ 450.00	R$ 360.00	R$ 260.00
	R$ 277.00	**R$ 2,337.00**	**R$ 1,240.00**	**R$ 940.00**

Fonte: Elaborado pelo autor.

Na comparação entre os sequenciamentos, assim como na avaliação estatística do período é possível notar que os valores para o sequenciamento de 5 minutos são melhores em relação aos outros sequenciamentos com tempo maior, em seguida temos o sequenciamento de 15 minutos e de 30 minutos nessa ordem. Abaixo é demonstrado através do gráfico em linhas que apresenta uma visão na comparação entre os dias presentes neste período.

Figura 24 - Demonstração da rentabilidade gerada para os dias do Período 2.



Fonte: Elaborado pelo autor.

6.3 RESULTADOS NO PERÍODO 3

Esse período vai do dia 29/03/2021 ao 01/04/2021, totalizando quatro (4) dias úteis, no qual coletamos 130.486 *tweets* que foram agrupados com os dados do mini-índice. Quanto ao período da análise dos *tweets*, agrupamos por períodos de 5, 15 e 30 minutos, conforme sequenciamento de movimentação do ativo mini-índice. Nesse período tivemos uma geração de 428 *candles* de fechamento para o sequenciamento de 5 minutos, 144 *candles* de fechamento para o sequenciamento de 15 minutos e 72 *candles* de fechamento para o sequenciamento de 30 minutos.

Foi dividido a análise em resultados estatísticos (acurácia, precisão, revocação e F1) em comparação com outros algoritmos e em resultados financeiros, vejamos:

6.3.1 Resultados Estatísticos no Período 3

Da análise dos gráficos estatísticos e financeiros que elaboramos no desenvolvimento da nossa pesquisa, e que abaixo passamos a apresentar de modo detalhado, entendemos ser oportuno apresentar de antemão os principais resultados que constatamos para o período, deixando desde já evidente que o método RNNTW-TC tem performance com valores em destaque comparada aos outros algoritmos de classificação.

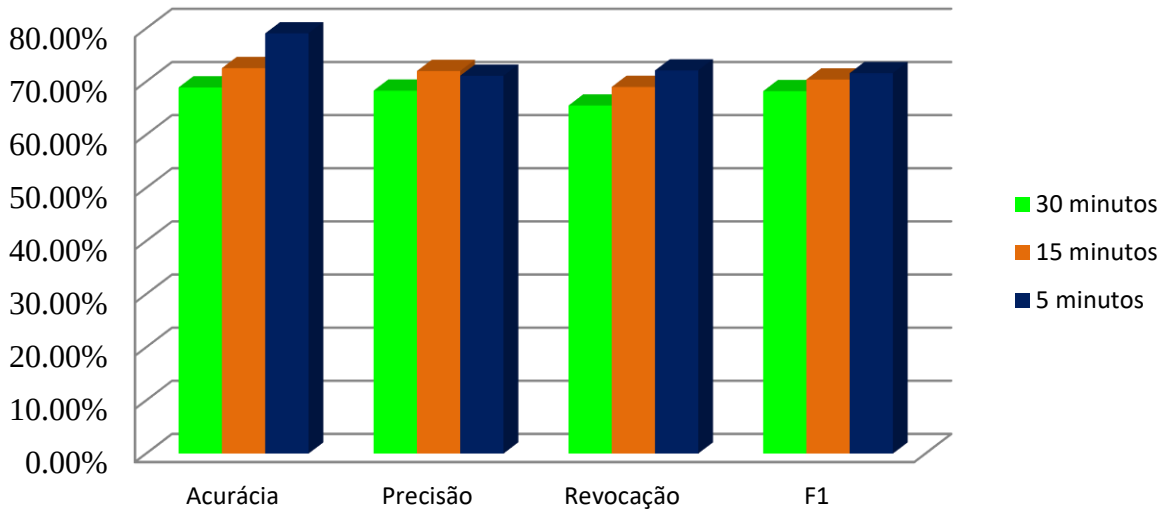
Abaixo apresentamos tabela e gráfico com os dados mais relevantes dos minutos agrupados apenas do modelo RNNTW-TC para esse Período 3.

Tabela 16 - Métricas de avaliação estatística o Período 3.

Sequenciamento	Acurácia	Precisão	Revocação	F1
30 minutos	68.90%	68.30%	65.50%	68.20%
15 minutos	72.54%	72.00%	69.00%	70.40%
5 minutos	79.09%	71.10%	72.10%	71.60%

Fonte: Elaborado pelo autor.

Figura 25 - Resultado estatístico em barras no Período 3.



Fonte: Elaborado pelo autor.

Conforme desenvolveremos com melhor atenção quando da análise final, valido neste momento apontar que quando o agrupamento dos *tweets* é feito por período de 5 minutos os resultados estatísticos são melhores e conforme o período do agrupamento aumenta os

resultados pioram. Neste período também é possível checar através dos resultados que na comparação entre os sequenciamentos o sequenciamento de 5 minutos se sobre sai sobre os demais em todas as métricas, abaixo segue resultados estatísticos de forma linear.

6.3.1.1 Sequenciamento 5 minutos no Período 3

Quanto ao sequenciamento de 5 minutos, no qual tivemos 535 *candles* de pontos de fechamentos. Passamos a apresentar os resultados estatísticos em comparação com outros algoritmos de classificação. O modelo RNNTW-TC, que usa o modelo desenvolvido com uma rede LSTM acrescida aos dados da rede social Twitter, foi comparado também com a execução dos dados com a rede LSTM sem os dados do Twitter (LSTM na tabela), e nessa comparação o modelo apresentou uma superioridade também.

Tabela 17 - Avaliação estatística no sequenciamento de 5 minutos para o Período 3.

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1**
RNNTW-TC	**79.09%**	**71.10%**	**72.10%**	**71.60%**
LSTM	75.38%	67.39%	68.39%	67.89%
Regressão Logística	76.11%	56.60%	63.30%	59.60%
Random Forest	75.80%	58.10%	59.01%	51.20%
SVM	79.50%	51.70%	51.84%	54.51%
KNeighbors	78.89%	54.40%	58.17%	57.77%

Fonte: Elaborado pelo autor.

Verifica-se pela Tabela 17, que o modelo RNNTW-TC apresentou melhor desempenho nos quatro quesitos analisados (Acurácia, Precisão, Revocação e F1), e fazendo uma análise geral podemos constatar que o desempenho do sistema por nos desenvolvido em alguns casos tem resultado com Precisão, Revocação e F1 bem superior ao dos outros algoritmos.

6.3.1.2 Sequenciamento 15 minutos no Período 3

Quanto ao sequenciamento de 15 minutos, no qual tivemos 180 *candles* (pontos de fechamentos). Apresentaremos os resultados estatísticos em comparação com outros algoritmos de classificação:

Tabela 18 - Avaliação estatística no sequenciamento de 30 minutos para o Período 3.

Modelo	Acurácia	Precisão	Revocação	F1
RNNTW-TC	**72.54%**	**72.00%**	**69.00%**	**70.40%**
LSTM	71.31%	70.77%	67.77%	69.17%
Regressão Logística	68.49%	57.38%	61.94%	59.40%
Random Forest	65.14%	57.16%	60.55%	58.75%
SVM	67.50%	55.86%	61.63%	58.19%
KNeighbors	68.89%	55.60%	68.99%	55.12%

Fonte: Elaborado pelo autor.

É possível verificar pela Tabela 18, que o modelo RNNTW-TC apresentou melhor desempenho nas quatro métricas analisadas. Destaque para ao F1 que obteve valores bem maiores que os demais algoritmos. Ao comparar com a rede LSTM sem dados do Twitter, podemos ver que os dados do Twitter conseguem uma leve melhora sobre os resultados obtidos em todos os critérios.

6.3.1.3 Sequenciamento 30 minutos no Período 3

Quanto ao sequenciamento de 30 minutos, no qual tivemos 90 *candles* de pontos de fechamentos. Apresentaremos os resultados estatísticos em comparação com outros algoritmos de classificação:

Tabela 19 - Avaliação estatística no sequenciamento de 30 minutos para o Período 3.

Modelo	Acurácia	Precisão	Revocação	F1
RNNTW-TC	**68.90%**	**68.30%**	**65.50%**	**68.20%**
LSTM	68.12%	67.52%	64.72%	67.42%
Regressão Logística	71.08%	61.90%	68.00%	69.13%
Random Forest	70.73%	58.29%	63.30%	60.61%
SVM	71.62%	59.60%	57.77%	55.62%
KNeighbors	72.60%	51.29%	67.83%	69.21%

Fonte: Elaborado pelo autor.

É possível verificar pela Tabela 19, que o modelo RNNTW-TC apresentou melhor desempenho em três quesitos analisados (Precisão, Revocação e F1). Para o quesito Acurácia o valor foi menor para todas as comparações, mas teve um valor bem próximo aos outros

algoritmos. Ao comparar com a rede LSTM sem dados do Twitter, podemos ver que os dados do Twitter conseguem uma leve melhora sobre os resultados obtidos em todos os critérios.

6.3.2 Avaliação financeira para o Período 3

Os resultados da avaliação financeira para o Período 3 demonstra a comparação entre os três sequenciamentos usados neste trabalho. Além disso é feito a comparação com a estratégia *buy and hold*, comprar e segurar. É feito a demonstração primeiramente dos resultados referentes a pontos ganhos por contrato do mini-índice e na sequência a demonstração dos valores referentes a cada ponto movimentado, cada ponto movimentado do contrato do mini-índice equivale a R$ 0,20 (vinte centavos de reais).

Passamos a apresentar os resultados da avaliação financeira para o Período 3:

Tabela 20 - Demonstração de pontos movimentados no Período 3.

Data	Buy/Hold	5min	15min	30min
29-Mar-21	1100	1375	280	1065
30-Mar-21	1475	2300	2265	670
31-Mar-21	260	1770	1450	940
01-Apr-21	-2395	300	200	80
	440	**5745**	**4195**	**2755**

Fonte: Elaborado pelo autor.

É apresentado a quantidade de pontos movimentados na Tabela 20, sendo para cada um dos sequenciamentos e para a estratégia de "*buy and hold*" para todos os dias presentes neste período.

Na análise financeira para o Período 3, o modelo RNNTW-TC obteve valores maiores relacionados a comparação com a estratégia *buy and hold*, comprar e segurar, em todos os sequenciamentos gerados pelo modelo.

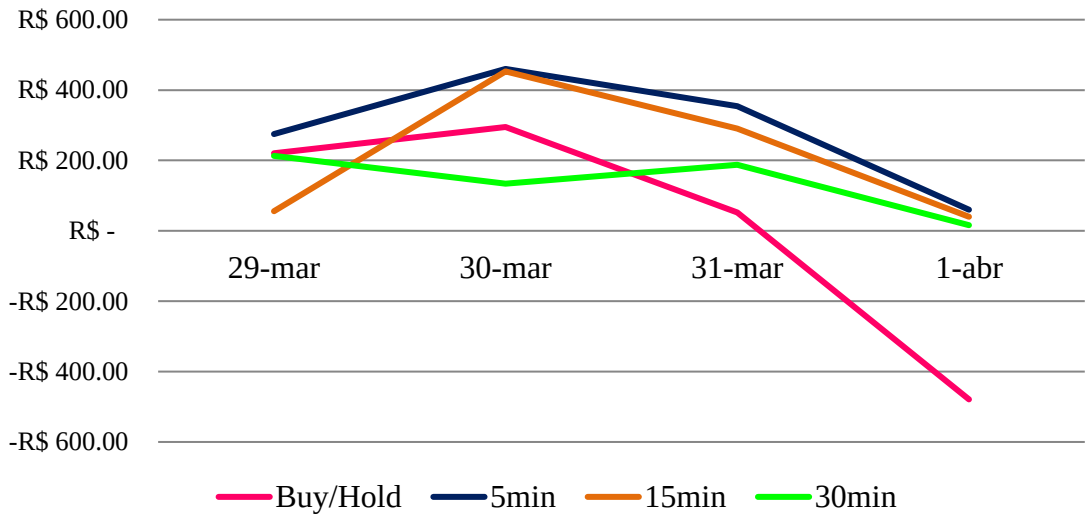
Na comparação entre os sequenciamentos, assim como na avaliação estatística deste período é possível notar que os valores para o sequenciamento de 5 minutos são melhores em relação aos outros sequenciamentos com tempo maior, em seguida temos o sequenciamento de 15 minutos e de 30 minutos nessa ordem.

Tabela 21 – Demonstração do valor financeiro diário no Período 3.

Data	Buy/Hold	5min	15min	30min
29-Mar-21	R$ 220.00	R$ 275.00	R$ 56.00	R$ 213.00
30-Mar-21	R$ 295.00	R$ 460.00	R$ 453.00	R$ 134.00
31-Mar-21	R$ 52.00	R$ 354.00	R$ 290.00	R$ 188.00
01-Apr-21	-R$ 479.00	R$ 60.00	R$ 40.00	R$ 16.00
	R$ 88.00	**R$ 1,149.00**	**R$ 839.00**	**R$ 551.00**

Fonte: Elaborado pelo autor.

Figura 26 - Demonstração da rentabilidade gerada para os dias do Período 3.



Fonte: Elaborado pelo autor.

6.4 RESULTADOS NO PERÍODO 4

Esse período vai do dia 05/04/2021 ao 09/04/2021, totalizando 5 dias úteis, no qual coletamos 114.251 *tweets* que foram agrupados com os dados do mini-índice. Quanto ao período da análise dos *tweets*, agrupamos por períodos de 5, 15 e 30 minutos, conforme sequenciamento de movimentação do ativo mini-índice. Nesse período tivemos uma geração de 535 *candles* de fechamento para o sequenciamento de 5 minutos, 180 *candles* de fechamento para o sequenciamento de 15 minutos e 90 *candles* de fechamento para o sequenciamento de 30 minutos.

Foi dividido a análise em resultados estatísticos (acurácia, precisão, revocação e F1) em comparação com outros algoritmos e em resultados financeiros, vejamos:

6.4.1 Resultados Estatísticos no Período 4

Da análise dos gráficos estatísticos e financeiros que elaboramos no desenvolvimento da nossa pesquisa, e que abaixo passamos a apresentar de modo detalhado, entendemos ser oportuno apresentar de primeiramente os principais resultados que constatamos, deixando desde já evidente que o método RNNTW-TC tem performance com valores em destaque comparada aos outros algoritmos de classificação.

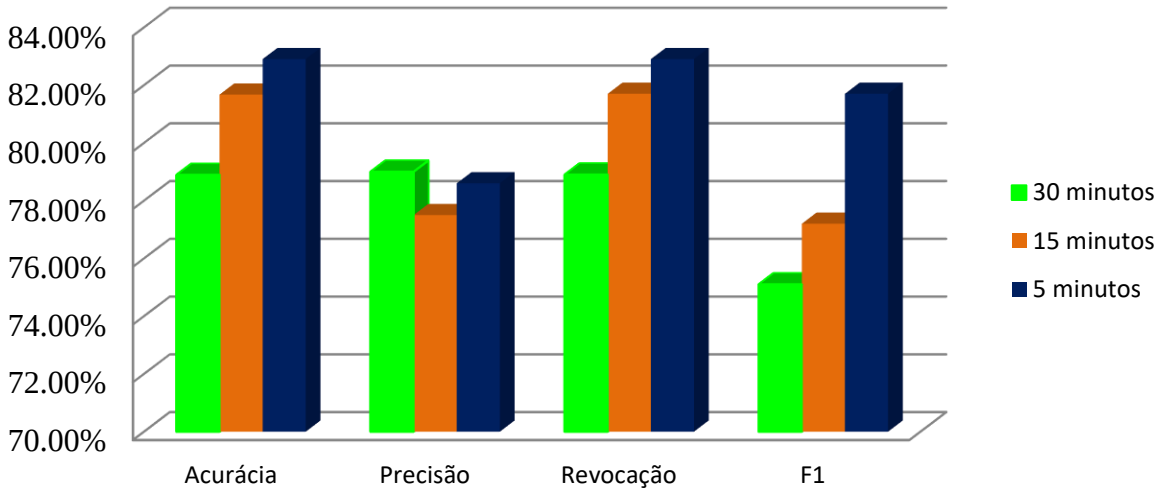
Abaixo apresentamos tabela e gráfico com os dados mais relevantes dos minutos agrupados apenas do modelo RNNTW-TC para esse Período 4.

Tabela 22 - Métricas de avaliação estatística o Período 4.

Sequenciamento	Acurácia	Precisão	Revocação	F1
30 minutos	78.89%	79.00%	78.90%	75.10%
15 minutos	81.67%	77.50%	81.70%	77.20%
5 minutos	82.90%	78.60%	82.90%	81.70%

Fonte: Elaborado pelo autor.

Figura 27 - Resultado estatístico em barras no Período 4.



Fonte: Elaborado pelo autor.

Conforme desenvolveremos com melhor atenção quando da análise final, valido neste momento apontar que quando o agrupamento dos *tweets* é feito por período de 5 minutos os resultados estatísticos são melhores e conforme o período do agrupamento aumenta os

resultados pioram, assim como em todos outros períodos. Abaixo segue gráfico linear com a comparação de sequenciamentos entre os resultados estatísticos do modelo RNNTW-TC.

6.4.1.1 Sequenciamento 5 minutos no Período 4

Quanto ao sequenciamento de 5 minutos, no qual tivemos 535 *candles* de pontos de fechamentos. Passaremos a apresentar os resultados estatísticos em comparação com outros algoritmos de classificação. O modelo RNNTW-TC, que usa o modelo desenvolvido com uma rede LSTM acrescida aos dados da rede social Twitter, foi comparado também com a execução dos dados com a rede LSTM sem os dados do Twitter (LSTM na tabela), e nessa comparação o modelo apresentou uma superioridade também.

Tabela 23 - Avaliação estatística no sequenciamento de 5 minutos para o Período 4.

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1**
RNNTW-TC	**82.90%**	**78.60%**	**82.90%**	**81.70%**
LSTM	79.43%	75.13%	79.43%	78.23%
Regressão Logística	81.70%	59.90%	63.33%	61.52%
Random Forest	71.90%	62.23%	71.50%	59.60%
SVM	59.80%	85.32%	63.30%	59.90%
KNeighbors	78.70%	59.90%	61.50%	63.30%

Fonte: Elaborado pelo autor.

Verifica-se pela Tabela 23, que o modelo RNNTW-TC apresentou melhor desempenho nos quatro quesitos analisados (Acurácia, Precisão, Recall e F1), e fazendo uma análise geral podemos constatar que o desempenho do sistema desenvolvido em alguns casos tem resultado com Precisão, Revocação e F1 bem superior ao dos outros algoritmos (exemplo: quesitos Precisão, Revocação e F1).

6.4.1.2 Sequenciamento 15 minutos no Período 4

Quanto ao sequenciamento de 15 minutos, no qual tivemos 180 *candles* de pontos de fechamentos. Apresentaremos os resultados estatísticos em comparação com outros algoritmos de classificação:

Tabela 24 - Avaliação estatística no sequenciamento de 15 minutos para o Período 4

Modelo	Acurácia	Precisão	Revocação	F1
RNNTW-TC	**81.67%**	**77.50%**	**81.70%**	**77.20%**
LSTM	80.24%	76.07%	80.27%	75.77%
Regressão Logística	76.11%	60.55%	63.33%	61.80%
Random Forest	78.33%	68.77%	59.01%	68.19%
SVM	76.60%	64.46%	69.34%	64.53%
KNeighbors	76.11%	68.99%	69.90%	67.83%

Fonte: Elaborado pelo autor.

Verifica-se pela Tabela 24, que o modelo RNNTW-TC apresentou melhor desempenho nas quatro métricas analisadas. Destaque para ao F1 que obteve valores bem maiores que os demais algoritmos. Ao comparar com a rede LSTM sem dados do Twitter, podemos ver que os dados do Twitter conseguem uma leve melhora sobre os resultados obtidos em todos os critérios.

6.4.1.3 Sequenciamento 30 minutos no Período 4

Quanto ao sequenciamento de 30 minutos, no qual tivemos 90 *candles* de pontos de fechamentos. Apresentaremos os resultados estatísticos em comparação com outros algoritmos de classificação:

É possível verificar pela Tabela 25, que o modelo RNNTW-TC apresentou melhor desempenho nas quatro métricas analisadas (Precisão, Revocação e F1). Diferentemente de alguns períodos como o Período 1 e 2 neste Período não tivemos resultados menores para a métrica Acurácia para o modelo proposto. Ao comparar com a rede LSTM sem dados do Twitter, podemos ver que os dados do Twitter conseguem uma leve melhora sobre os resultados obtidos em todos os critérios.

Tabela 25 - Avaliação estatística no sequenciamento de 30 minutos para o Período 4.

Modelo	Acurácia	Precisão	Revocação	F1
RNNTW-TC	**78.89%**	**79.00%**	**78.90%**	**75.10%**
LSTM	78.24%	78.39%	78.22%	74.45%
Regressão Logística	74.40%	59.63%	63.33%	61.30%
Random Forest	78.80%	64.70%	66.60%	65.60%
SVM	73.30%	59.60%	61.37%	63.30%
KNeighbors	72.20%	60.55%	73.75%	50.28%

Fonte: Elaborado pelo autor.

6.4.2 Avaliação financeira para o Período 4

Os resultados da avaliação financeira para o Período 4 demonstra a comparação entre os três sequenciamentos usados neste trabalho. Além disso é feito a comparação com a estratégia *buy and hold*, comprar e segurar. É feito a demonstração primeiramente dos resultados referentes a pontos ganhos por contrato do mini-índice e na sequência a demonstração dos valores referentes a cada ponto movimentado, cada ponto movimentado do contrato do mini-índice equivale a R$ 0,20 (vinte centavos de reais).

Passamos a apresentar os resultados da avaliação financeira para o Período 4:

Tabela 26 - Demonstração de pontos movimentados no Período 4.

Data	Buy/Hold	5min	15min	30min
05-Apr-21	1700	1960	300	200
06-Apr-21	330	1830	1500	1000
07-Apr-21	625	2420	1700	800
08-Apr-21	-205	615	350	300
09-Apr-21	-240	2460	1650	1200
	2210	**9285**	**5500**	**3500**

Fonte: Elaborado pelo autor.

Na Tabela 26, é demonstrado a quantidade de pontos movimentados, tanto para cima quanto para baixo, para cada um dos sequenciamentos e para a estratégia de *buy and hold* para todos os dias presentes neste período.

Na análise financeira para o Período 4, assim como para todos os outros períodos avaliados, o modelo RNNTW-TC obteve valores maiores relacionados a comparação com a estratégia *buy and hold*, comprar e segurar, em todos os sequenciamentos gerados pelo modelo.

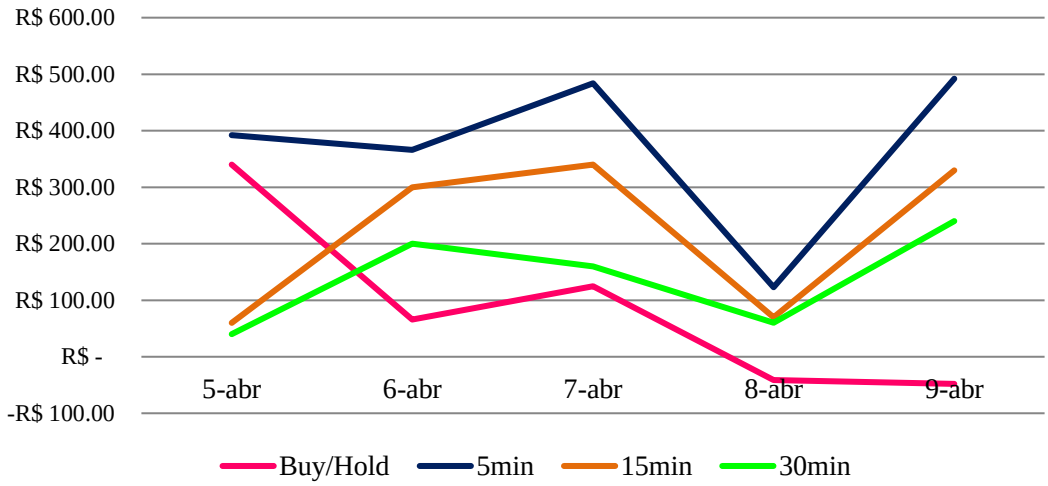
Tabela 27 - Demonstração do valor financeiro diário no Período 4.

Data	Buy/Hold	5min	15min	30min
05-Apr-21	R$ 340.00	R$ 392.00	R$ 60.00	R$ 40.00
06-Apr-21	R$ 66.00	R$ 366.00	R$ 300.00	R$ 200.00
07-Apr-21	R$ 125.00	R$ 484.00	R$ 340.00	R$ 160.00
08-Apr-21	-R$ 41.00	R$ 123.00	R$ 70.00	R$ 60.00
09-Apr-21	-R$ 48.00	R$ 492.00	R$ 330.00	R$ 240.00
	R$ 442.00	**R$ 1,857.00**	**R$ 1,100.00**	**R$ 700.00**

Fonte: Elaborado pelo autor.

Assim como nos outros períodos também, na comparação entre os sequenciamentos, assim como na avaliação estatística deste período é possível notar que os valores para o sequenciamento de 5 minutos são melhores em relação aos outros sequenciamentos com tempo maior, em seguida temos o sequenciamento de 15 minutos e de 30 minutos nessa ordem.

Tabela 28 - Demonstração da rentabilidade gerada para os dias do Período 4.



Fonte: Elaborado pelo autor.

6.5 RESULTADOS GERAIS

Para apresentação a demonstração dos resultados gerais foi feito divisão em duas partes para melhor apresentação, a primeira parte se refere à demonstração do resultado estatístico geral, a segunda parte ao resultado financeiro bruto e por fim o resultado financeiro líquido. A seguir apresentaremos esses resultados:

6.5.1 Resultado Estatístico Geral

Com o intuito de analisar melhor os resultados de forma total e melhor apresentação destes resultados, foi feito a média de todos os resultados e comparações gerados neste trabalho para os quatro períodos. Na Tabela 29 é apresentado o resultado estatístico geral para o modelo RNNTW-TC.

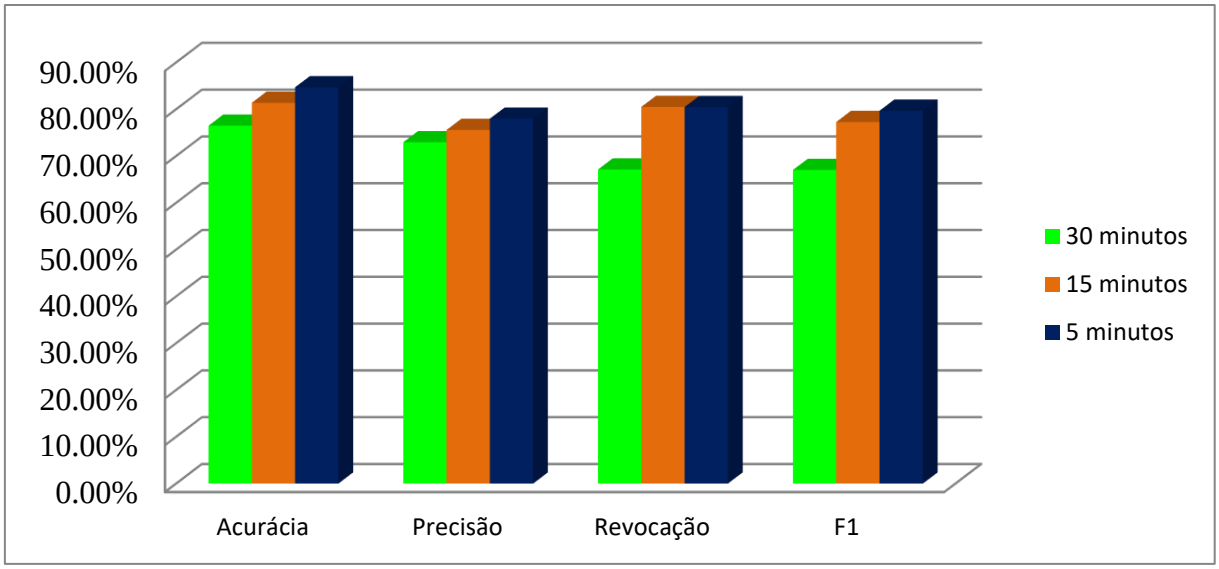
Tabela 29 - Métricas de avaliação estatística geral.

Sequenciamento	Acurácia	Precisão	Revocação	F1
30 minutos	76.49%	72.95%	67.13%	67.03%
15 minutos	81.38%	75.58%	80.50%	77.28%
5 minutos	84.69%	77.98%	80.45%	79.69%

Fonte: Elaborado pelo autor.

Em seguida é apresentado a Figura 28, no qual podemos verificar a comparação entre os sequenciamentos analisados neste trabalho. Conforme todos os períodos analisados em separado o sequenciamento de 5 minutos obteve resultados melhores em comparação com os outros sequenciamentos analisados, seguido pelo sequenciamento de 15 e 30 minutos, nessa ordem.

Figura 28 - Resultado estatístico em barras no geral.



Fonte: Elaborado pelo autor.

Na comparação feita entre o modelo e outros algoritmos é possível ver que o modelo fica acima dos outros algoritmos em todos os sequenciamentos analisados neste trabalho. As Tabelas 30, 31 e 32 demonstram a média de todos os quatro períodos analisados, conforme já descrito acima. Assim como nos resultados apresentados para cada um dos períodos o sequenciamento de 5 minutos é superior aos demais, e com as tabelas apresentadas com os valores com média calculada fica notável essa superioridade.

É possível notar que o modelo RNNTW-TC, que usa o modelo desenvolvido com uma rede LSTM acrescida aos dados da rede social Twitter, quando comparado com a execução dos

dados com a rede LSTM sem os dados do Twitter (LSTM na tabela), apresenta resultado superior em todas as quatro métricas. Isso demonstra que os dados do Twitter podem melhorar o desempenho de um modelo.

Tabela 30 - Avaliação estatística no sequenciamento de 5 minutos no geral.

Modelo	Acurácia	Precisão	Revocação	F1
RNNTW-TC	**84.69%**	**77.98%**	**80.45%**	**79.69%**
LSTM	80.90%	74.19%	76.49%	75.90%
Regressão Logística	83.22%	64.53%	65.76%	63.57%
Random Forest	77.89%	65.06%	65.73%	62.97%
SVM	78.75%	69.32%	62.58%	64.79%
KNeighbors	83.54%	65.43%	63.70%	65.90%

Fonte: Elaborado pelo autor.

Tabela 31 - Avaliação estatística no sequenciamento de 15 minutos no geral.

Modelo	Acurácia	Precisão	Revocação	F1
RNNTW-TC	**81.38%**	**75.58%**	**80.50%**	**77.28%**
LSTM	80.14%	74.08%	78.90%	75.78%
Regressão Logística	79.46%	64.06%	67.06%	65.85%
Random Forest	78.95%	66.65%	66.68%	66.94%
SVM	78.94%	63.56%	70.53%	67.33%
KNeighbors	78.59%	66.05%	70.54%	65.92%

Fonte: Elaborado pelo autor.

Tabela 32 - Avaliação estatística no sequenciamento de 30 minutos no geral.

Modelo	Acurácia	Precisão	Revocação	F1
RNNTW-TC	**76.49%**	**72.95%**	**67.13%**	**67.03%**
LSTM	75.77%	72.24%	66.39%	66.30%
Regressão Logística	76.40%	65.78%	66.72%	70.11%
Random Forest	76.70%	64.77%	68.84%	67.54%
SVM	77.02%	61.19%	63.64%	65.08%
KNeighbors	75.43%	60.44%	73.90%	66.31%

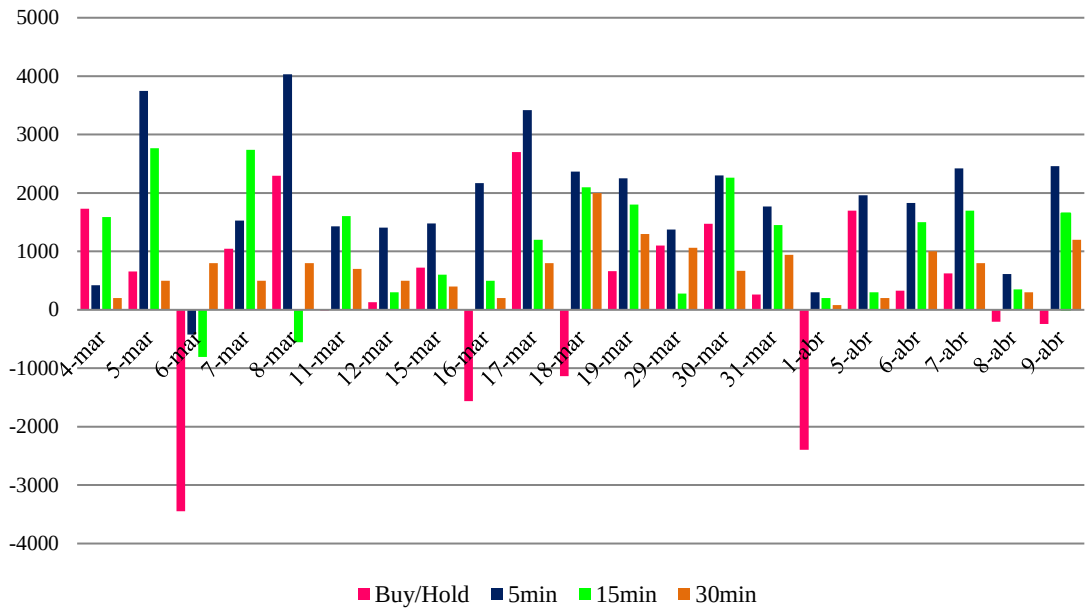
Fonte: Elaborado pelo autor.

6.5.2 Resultado Financeiro Bruto

O resultado da avaliação financeira bruta é apresentado considerando todos os dias de simulação analisados neste trabalho. Através da Tabela 33, os resultados são apresentados realizando a comparação entre os sequenciamentos e da estratégia *buy and hold*.

A princípio é feito a demonstração primeiramente dos resultados referentes a pontos ganhos por contrato do mini-índice e na sequência a demonstração dos valores referentes a cada ponto movimentado, cada ponto do contrato do mini-índice equivale a R$ 0,20 (vinte centavos de reais), conforme já informado.

Figura 29 - Demonstração de quantidade de pontos diários gerados.



Fonte: Elaborado pelo autor.

A Tabela 33, tabela de demonstração de quantidade de pontos movimentados diariamente, nos serviu para realização de uma comparação para todos os dias, demonstrado na Figura 29, no qual é possível reforçar a os resultados já apresentados. É possível notar a superação dos ganhos com o sequenciamento de 5 minutos em relação aos demais sequenciamentos. Ainda, é possível observar que na maioria dos dias esse sequenciamento tem resultados melhores que os demais.

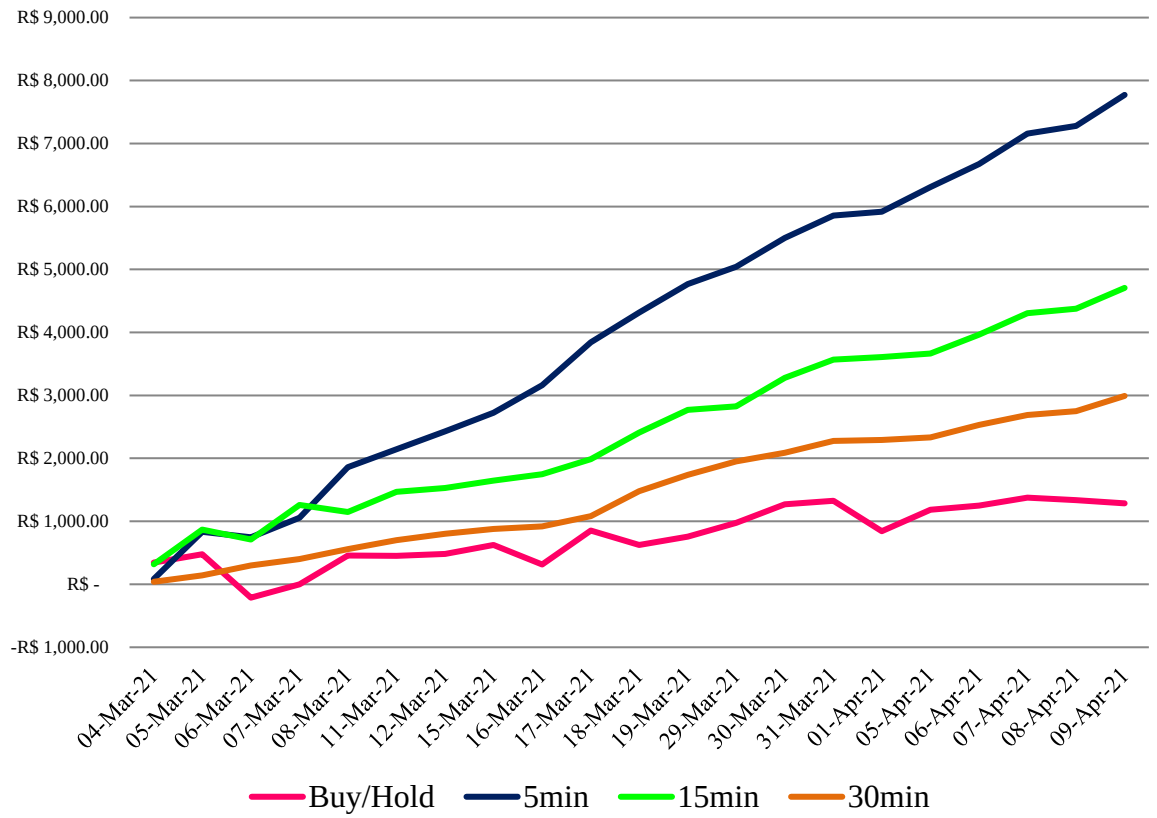
Tabela 33 - Demonstração de quantidade de pontos diários gerados.

Data	Buy/Hold	5min	15min	30min
4-Mar-21	1730	420	1590	200
5-Mar-21	655	3750	2765	500
6-Mar-21	-3445	-425	-805	800
7-Mar-21	1045	1530	2740	500
8-Mar-21	2295	4030	-555	800
11-Mar-21	-10	1430	1605	700
12-Mar-21	130	1410	300	500
15-Mar-21	720	1480	600	400
16-Mar-21	-1560	2170	500	200
17-Mar-21	2700	3420	1200	800
18-Mar-21	-1135	2365	2100	2000
19-Mar-21	660	2250	1800	1300
29-Mar-21	1100	1375	280	1065
30-Mar-21	1475	2300	2265	670
31-Mar-21	260	1770	1450	940
1-Apr-21	-2395	300	200	80
5-Apr-21	1700	1960	300	200
6-Apr-21	330	1830	1500	1000
7-Apr-21	625	2420	1700	800
8-Apr-21	-205	615	350	300
9-Apr-21	-240	2460	1650	1200
Total	6435	38860	23535	14955

Fonte: Elaborado pelo autor.

Com os pontos convertidos em valor financeiro, conforme valor de R$ 0,20 (vinte centavos de reais) para cada ponto movimentado, é apresentado na Figura 30 o valor acumulado para todos os dias usados nas simulações feitas neste estudo. Na Figura 30 apresentamos o resultado referente a comparação entre as sequencias analisadas e sua comparação com a estratégia *buy and hold*.

Figura 30 - Resultado financeiro acumulado bruto.



Fonte: Elaborado pelo autor.

As operações de compra e venda em todo período avaliado geraram no acumulado 38.860 pontos para o sequenciamento de 5 minutos, 23.535 pontos para o sequenciamento de 15 minutos e 14.955 pontos para o sequenciamento de 30 minutos.

Esses pontos monetizados, conforme valor de R$ 0,20 (vinte centavos de reais) geram a rentabilidade bruta de R$ 7.772,00 para o sequenciamento de 5 minutos, R$ 4.707 para o sequenciamento de 15 minutos e R$ 2.991,00 para o sequenciamento de 30 minutos, valores apresentados na Figura 30. Para a estratégia *buy and hold* a pontuação agregada para todo o período avaliado foi de 6.435 pontos, R$ 1.287,00 no que diz a valores brutos.

Os valores apresentados se referem ao valor bruto gerado pelo modelo, já na seção 6.5.3 será apresentado o resultado financeiro líquido, no qual é feito a subtração dos descontos para possibilitar a visão do valor líquido.

Vale ressaltar que esse resultado está sendo calculado para a negociação de um contrato do ativo mini-índice, podendo ser multiplicado, caso houvesse mais de um contrato na simulação, pelo número de contratos negociados, aumentando a possibilidade de ganhos usando

o modelo proposto. Por fim é possível afirmar que os resultados gerados pelo modelo proposto neste trabalho são satisfatórios.

6.5.3 Resultado Financeiro Líquido

É importante informar que os resultados apresentados na seção 6.5.2 estão considerando apenas os valores brutos gerados pelo modelo. No entanto é necessário considerar que esses valores gerados precisam ser subtraídos pelos débitos referentes ao Imposto de Renda (IR) sobre o lucro gerado, corretagem cobrado pela intermediadora das operações, gastos com os emolumentos da Bolsa, no qual se resumem aos custos e valores cobrados pela própria Bolsa referente aos serviços prestados por ela.

Todos esses valores são gerenciados pela B3 (Bolsa Brasil Balcão), empresa que atua desde 2017 é a junção entre antiga BM&F Bovespa (Bolsa de Valores, Mercadorias e Futuros de São Paulo) e a Cetip (Central de Custódia e de Liquidação Financeira de Títulos) (B3, 2021).

No que diz respeito a operações de *Day Trade*, usadas neste trabalho, a alíquota referente ao Imposto de Renda (IR) é de 20% sobre o lucro gerado, esse valor é pago ao final de todo mês.

Já a corretagem, se refere à taxa cobrada pela corretora (intermediadora) sobre o valor de cada operação, podendo ser fixo ou de forma variável dependendo da corretora. Essa taxa não será considerada neste trabalho, pois muitas das corretoras oferecem planos ou até mesmo o serviço de corretagem gratuito.

Diferente da taxa de corretagem, os emolumentos, também chamado de taxa de negociação, são taxas e custos operacionais que são cobradas diretamente pela B3. Esses valores são cobrados para cada liquidação feita (compra ou venda) e isso acaba tendo um impacto neste trabalho, pois trabalhamos com operações *day trade* isso gera uma quantidade de operações feitas maior do que operações comuns. O valor médio para os emolumentos considerado neste trabalho é de R$0,31 para cada operação realizada (B3, 2021).

Diante da necessidade acima apresentada, no qual é necessário entender o que nos sobraria de líquido para as simulações feitas neste trabalho, foi desenvolvida a seguinte fórmula para cálculo do valor líquido:

$$vRLiq = vRBruto - ((vAliqIR * vRBruto) - (vTxEmol * qOper)) \quad (20)$$

Sendo que:

$vAliqIR$= Valor da alíquota do IR, 20% para operações Day Trade.

$vRBruto$= Valor do rendimento bruto gerado pelo modelo.

$vTxEmol$ = Valor médio da taxa de emolumento.

$qOper$ = Quantidade de operações feitas.

O modelo executa uma média de 40 operações para o sequenciamento de 5 minutos, 14 operações para o sequenciamento de 15 minutos e para o período de 30 minutos uma média de 4 operações diárias. A partir disso foi feito o cálculo do total de emolumentos gastos nos 21 dias úteis que foram usados nesta pesquisa. Vale ressaltar que para as operações com menos período, exemplo sequenciamento de 5 minutos, o número de operações é maior, consequentemente o número de operações também.

É fato que o número de operações para o sequenciamento de 5 minutos é maior pois esse sequenciamento nos propõe oportunidade de mais negócios devido a ter mais movimentação.

Ainda, foi feito o cálculo do valor do IR sobre o lucro bruto de 20% para as operações *day trade* e de 15 % para operações com mais de um dia de duração, no caso da operação *buy and hold*. Ao final feito a aplicação da fórmula acima para cálculo do resultado líquido.

Diante disso temos os seguintes resultados conforme tabela abaixo.

Tabela 34 - Resultado Financeiro Líquido.

	Buy/Hold	**5min**	**15min**	**30min**
Variação Bruta	R$ 1.287,00	R$ 7.772,00	R$ 4.707,00	R$ 2.991,00
Emolumentos	R$ -	R$ 272,80	R$ 95,48	R$ 27,28
Imp. Renda	R$ 193,05	R$ 1.554,40	R$ 941,40	R$ 598,20
Variação Líquida	R$ 1.093,95	R$ 5.944,80	R$ 3.670,12	R$ 2.365,52
Valor Líquido	**R$ 23.360,95**	**R$ 28.211,80**	**R$ 25.937,12**	**R$ 24.632,52**

Fonte: Elaborado pelo autor.

Conforme apresentado na tabela acima, tivemos para a estratégia *buy and hold* uma variação bruta de R$ 1.287,00, conforme mencionado também na seção 6.5.2. Não foi considerado emolumentos, pois essa estratégia só teríamos uma compra e uma venda, então acaba sendo um valor que não impactaria no resultado. Já para o valor de IR, essa estratégia teve que realizar o recolhimento de R$ 193,05. Por fim obteve uma variação líquida de R$

1.093,95, gerando o acumulado de R$ 23.360,95. Esse valor líquido é equivalente à 4,913% do valor inicial investido.

Já para a estratégia com sequenciamento de 5 minutos, tivemos uma variação bruta de R$ 7.772,00, mencionado na seção 6.5.2. O valor de emolumentos para essa estratégia foi de R$ 272,80, sendo o maior valor para tal entre as possibilidades apresentadas neste trabalho. O IR foi igual à R$ 1.554,40 para esta estratégia. Obtendo ao final uma variação líquida de R$ 5.994,80 e valor líquido de R$ 23.360,95. Esse valor líquido é equivalente à 26,698% do valor inicial investido.

Na estratégia de 15 minutos, o valor da variação bruta é de R$ 4.707,00, emolumentos de R$ 95,48, IR sobre o lucro de R$ 941,40. Com isso tivemos uma variação líquida de R$ 3.670,12 totalizando o valor líquido de R$ 25.937,12, o que equivale à 16,482% do valor inicial investido.

Por fim a estratégia de 30 minutos, gerou a variação bruta é de R$ 2.991,00, emolumentos de R$ 27,28, menor valor para esse débito de todos os sequenciamentos devido ao número menor de operações realizadas, IR sobre o lucro de R$ 598,20. Isso gerou uma variação líquida de R$ 2.365,52 totalizando o valor líquido de R$ 24.632,52, equivalente à 10,623% do valor inicial investido.

7. CONCLUSÕES

Quando nos propusemos a desenvolver o estudo que ora apresentamos, havia como objetivo responder as indagações apresentadas no item 1.2 e acreditamos que conseguimos atingir nosso objetivo, pois durante a pesquisa, fizemos as análises propostas, coletando mensagens da rede social Twitter, agrupadas em períodos de 5, 15 e 30 minutos conforme a movimentação do ativo mini-índice e agora passamos a apresentar os resultados obtidos.

No que se refere à análise financeira considerando um determinado período e um determinado ativo, constatamos que os resultados financeiros são positivos e superiores às propostas conservadoras de investimentos *(buy and hold)*, pois apresentam ganhos líquidos no período avaliado de R$ 5.944,80 para sequenciamento de 5 minutos, R$ 3.670,12 para o sequenciamento de 15 minutos e R$ 2.365,52 para o sequenciamento de 30 minutos. De acordo com as análises que realizamos, concluímos que os valores são superiores ao valor gerado pela estratégia conservadora *(buy and hold)*, que foi R$ 1.093,95 no mesmo período. Essa variação líquida é de 26,698% para o sequenciamento de 5 minutos, 16,482% para o sequenciamento de 15 minutos, 10,623% para o sequenciamento de 30 minutos e de 4,913% para a estratégia *buy and hold.*

Comparando o modelo desenvolvido com outros algoritmos já existentes, e analisando as métricas estáticas (acurácia, precisão, revocação e F1 *Score*), bem como as mensagens por agrupamento de 5, 15 e 30 minutos, apuramos que a acurácia gerada foi de 84,69% para o sequenciamento de 5 minutos, 81,38% para o sequenciamento de 15 minutos e 76,49% para o sequenciamento de 30 minutos, no modelo proposto, valores acima dos valores gerados por outros algoritmos usados nas comparações. Considerando a precisão temos os seguintes resultados de 77,98% para o sequenciamento de 5 minutos, 75,98% para o sequenciamento de 15 minutos e 72,95% para o sequenciamento de 30 minutos. Referente a revocação temos os seguintes resultados de 80,45% para o sequenciamento de 5 minutos, 80,50% para o sequenciamento de 15 minutos e 67,13% para o sequenciamento de 30 minutos. Por fim, quanto ao F1 *Score*, os valores foram de 79,69% para o sequenciamento de 5 minutos, 77,28% para o sequenciamento de 15 minutos e 67,03% para o sequenciamento de 30 minutos. Vale ressaltar que o modelo teve melhores resultados estatísticos em comparação com os outros algoritmos já existentes no mercado e que foram utilizados como paradigmas para a nossa pesquisa.

Comparando o modelo desenvolvido com a rede LSTM, apenas com os dados sequenciais do ativo e sem dados do Twitter, foi possível notar que os dados do Twitter podem trazer uma leve melhora nos resultados estatísticos. O percentual de melhora é baixo, no

entanto, é notável a melhora nos resultados ao usar os dados da rede social informada em todas as métricas estáticas (acurácia, precisão, revocação e F1 *Score*).

No que se refere a análise acerca das mensagens da rede social Twitter influenciarem operações de compra e venda do ativo mini-índice, chegamos à conclusão de que as mensagens podem sinalizar a trajetória que o ativo realizará, pois há relação entre as postagens e as negociações referentes ao ativo analisado.

Ao analisarmos a questão referente a interferência no resultado de acordo com o período em que houve a postagem no Twitter, ou seja, com relação ao agrupamento das mensagens em 5, 15 e 30 minutos é possível verificar se há maior influência quando as mensagens são postadas próximas ao fechamento do mini-índice, concluímos que quanto mais próximo do horário do fechamento das negociações maior é a influência das mensagens postadas no Twitter com as tomadas de decisões de compra e venda do ativo mini-índice.

Outrossim, valida a afirmação no sentido de que todas as análises que realizamos para o desenvolvimento do nosso trabalho, tanto quanto aos resultados referentes ao agrupamento das mensagens do Twitter, bem como a análise dos índices estatísticos o desempenho do modelo que desenvolvemos é superior em todos os aspectos analisados aos algoritmos que elegemos para fazer as comparações.

Foi apurado que, quando as mensagens são postadas próximas ao período de fechamento das negociações do mini-índice, os retornos financeiros e estatísticos são melhores. Apuramos ainda que no agrupamento dos minutos em menor período (5 minutos) obtivemos resultados melhores do que dos agrupamentos dos períodos maiores (15 e 30 minutos).

Por fim, valida a afirmativa que os resultados do modelo que propomos, foram superiores aos resultados dos algoritmos analisados, independentemente das análises comparativas (análise dos sequenciamentos dos agrupamentos, análise estatística e análise financeira).

Como trabalhos futuros desejamos testar o modelo com uma quantidade maior de dados obtidos do Twitter e de outras fontes, tal como jornais, de forma a comprovar a hipótese de que uma melhor qualidade dos dados pode fazer com que o modelo consiga dar melhores resultados quanto a hora de comprar ou vender um ativo. Com o aumento de fontes de textos podemos realizar operações em outros ativos da bolsa de valores, uma vez que pode poderemos ampliar a quantidade de textos específicos. Ainda como trabalhos futuros, há desejo em realizar uma plataforma em tempo real para realização das operações com o modelo proposto.

REFERÊNCIAS

ALZAZAH, Faten; CHENG, Xiaochun. **Recent Advances in Stock Market Prediction Using Text Mining: A Surve**y. In: [s.l.: s.n.], jun. 2020. DOI: 10.5772/intechopen.92253.

AMIDI, Afshine; AMIDI, Shervine. **Recurrent Neural Networks cheatsheet**. Distribuido digitalmente. [S.l.], nov. 2018. Disponível em: <https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks>. Acesso em: 4 ago. 2020.

ANUP, Pokhrel et al. **Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis**. 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), p. 128–132, 2018.

ARANGO, H. G. **Bioestatística: Teórica e Computacional**. Rio de Janeiro: Guanabara Koogan, 2001, 440p.

ATASHIAN, Gasia; HRACHYA, Khachatryan. **Sentiment Analysis To Predict Global Cryptocurrency Trends.** Mai. 2018. Tese (Doutorado). DOI: 10.13140/RG.2.2.24311.32163.

B3, Inc: **Tarifas de Ibovespa e Índice Brasil 50**. Disponível em: http://www.b3.com.br/pt_br/produtos-e-servicos/tarifas/listados-a-vista-e-derivativos/renda-variavel/tarifas-de-ibovespa-e-indice-brasil-50/futuros-e-estruturadas/. Acesso em: 4 abr. 2021.

BAKER, Malcolm; WURGLER, Jeffrey. **Investor Sentiment in the Stock Market. The Journal of Economic Perspectives**, American Economic Association, v. 21, n. 2, p. 129–151, 2007. ISSN 08953309. Disponível em: <http://www.jstor.org/stable/30033721>.

BENEVENUTO, Fabrício; RIBEIRO, Filipe; ARAÚJO, Matheus. **Métodos para Análise de Sentimentos em mídias sociais.** Minicurso em Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia), 2015.

BENGIO, Yoshua et al. **A Neural Probabilistic Language Model**. J. Mach. Learn. Res., JMLR.org, v. 3, null, p. 1137–1155, mar. 2003. ISSN 1532-4435.

BITTENCOURT, Jairo Alano de et al. **Análise da relação entre o perfil de investidor a realidade do mercado de renda fixa e variável e a teoria da aversão à perda**. Revista Razão Contábil E Finanças, v. 9, 2018.

BM&FBOVESPA. **Por dentro da BM&FBOVESPA**. fev-2017 Disponível em: http://bvmf.bmfbovespa.com.br/pt-br/download/LivroPQO.pdf. Acesso em: 4 abr. 2021

CAMPOLINA, Paulo Azevedo Meijon; BATISTA, Lucas S. Batista. **Uma estratégia automatizada de day-trade por meio de comite de indicadores técnicos**. 2019.

CAROSIA, Arthur E.; COELHO, Guilherme P.; SILVA, Ana E. A. da. **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. Rio de Janeiro, Brazil: Association for Computing Machinery, 2019. (WebMedia '19), p. 385–392. ISBN 9781450367639. Disponível em: <https://doi.org/10.1145/3323503.3349564>.

CAUX, Marcelo de; BERNARDINI, Flavia; VITERBO, Jose. **Short-Term Forecasting in Bitcoin Time Series Using LSTM and GRU RNNs**. In: ANAIS do VIII Symposium on Knowledge Discovery, Mining and Learning. Evento Online: SBC, 2020. P. 97–104. DOI: 10.5753/kdmile.2020.11964.Disponível em: https://sol.sbc.org.br/index.php/kdmile/article/view/11964.

CHEN, Qian; ZHANG, Wenyu; LOU, Yu. **Forecasting Stock Prices Using a Hybrid Deep Learning Model Integrating Attention Mechanism, Multi-Layer Perceptron, and Bidirectional Long-Short Term Memory Neural Network**. IEEE Access, v. 8, p. 117365–117376, 2020.

CHOLLET, François et al. Keras. [S.l.: s.n.], 2015. https://keras.io.

CHOWDHURY, Gobinda G. **Natural language processing**. Annual Review of Information Science and Technology, v. 37, n. 1, p. 51–89, 2003. DOI: 10.1002/aris.1440370103. Eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/aris.1440370103. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440370103>.

COINANALYSIS; **Predict cryptocurrency prices based on news and historical price data**. 2018. Distribuido digitalmente. [S.l.], jan. 2018. Disponível em: <http://tiny.cc/urivmz>. Acesso em: 24 set. 2020.

COVINGTON, M.; NUTE, D.; VELLINO, André. **Prolog Programming in Depth**. In: COVINGTON, Michael; BARKER, Ken; SZPAKOWICZ, Stan. Natural Language Processing for Prolog Programmers. Prentice Hall, 1994.

CVM, Comissão de Valores Mobiliários. **Análise de Investimentos (Histórico, Principais Ferramentas e Mudanças Conceituais para o Futuro).** 4. ed. Rio de Janeiro, RJ, Brasil: [s.n.], 2017. Disponível em: <https://www.investidor.gov.br/portaldoinvestidor/export/sites/portaldoinvestidor/publicacao/Livro/livro_TOP_analise_investimentos.pdf>.

CVM, Comissão de Valores Mobiliários. **Mercado de valores mobiliários brasileiro.** 4. ed. Rio de Janeiro, RJ, Brasil: [s.n.], 2019. Disponível em: <https://www.investidor.gov.br/portaldoinvestidor/export/sites/portaldoinvestidor/publicacao/Livro/Livro-IBRI-CVM.pdf>.

DALE, Robert; MOISL, Hermann; SOMERS, Harold (Ed.). **Handbook of Natural Language Processing.** [S.l.]: Marcel Dekker, 2000. ISBN 0824790006.

DAMODARAN, Aswath. **Avaliação de investimentos: ferramentas e técnicas para determinação de valor de qualquer ativo**. Rio de Janeiro, RJ, Brasil: Qualitymark, 2015.

DAULTANI, D. **Stock predictions through news sentiment analysis**. Distribuido digitalmente. [S.l.], 2017. Disponível em: <http://tiny.cc/urivmz>. Acesso em: 24 set. 2020.

DAUTEL, Alexander et al. **Forex exchange rate forecasting using deep recurrent neural networks**. Digital Finance, mar. 2020. DOI: 10.1007/s42521-020-00019-x.

DAVID, Easley; JON, Kleinberg. **Networks, Crowds, and Markets: Reasoning About a Highly Connected World.** USA: Cambridge University Press, 2010. ISBN 0521195330.

DIETTERICH, Thomas G. **Machine Learning for Sequential Data: A Review**. In: [s.l.: s.n.], 2002. P. 227–246. Disponível em: <http://www.springerlink.com/content/av8l8hjl6yc2ya3m>.

DRUCKER, H.; BURGES, C. J.; KAUFMAN, L.; SMOLA, A.; VAPNIK, V. **Support vector regression machines. Advances in neural information processing systems,** Morgan Kaufmann Publishers, p. 155–161, 1997.

FELDMAN, Ronen; SANGER, James. **The text mining handbook: Advanced approaches in analyzing unstructured data**. [S.l.: s.n.], jan. 2007.

FERNEDA, Edberto. **Redes neurais e sua aplicação em sistemas de recuperação de informação**. pt. Ciência da Informação, scielo, v. 35, p. 25–30, abr. 2006. ISSN 0100-1965. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652006000100003&nrm=iso>.

FERREIRA, A. B. de H. **Aurélio: o dicionário da língua portugues**a. Curitiba, PR, Brasil: Positivo, 2008.

FERREIRA, Rodrigo S. et al. **Data Science in Financial Markets: Characterization and Analysis of Stocktwits**. In: PROCEEDINGS of the 25th Brazillian Symposium on Multimedia and the Web. Rio de Janeiro, Brazil: Association for Computing Machinery, 2019. (WebMedia '19), p. 393–400. Disponível em: <https://doi.org/10.1145/3323503.3360298>.

GITMAN, Laurence. **Princípios de Administração Financeira**. Porto Alegre, RS, Brasil: Bookman, 2001.

GOLDBERG, Yoav; HIRST, Graeme. **Neural Network Methods in Natural Language Processing.** [S.l.]: Morgan & Claypool Publishers, 2017. ISBN 1627052984.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep learning**. [S.l.]: MIT press, 2016.

HACHICHA, Ahmed; HACHICHA, Fatma. **Analysis of the bitcoin stock market indexes using comparative study of two models SV with MCMC algorithm**. Review of Quantitative Finance and Accounting, jul. 2020. DOI: 10.1007/s11156-020-00905-w.

HAYKIN, Simon. **Neural Networks and Learning Machines, 3ª edição**. Prentice Hall, 2008.

HARTMANN, Nathan et al. **Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks**, ago. 2017.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. **Long Short-Term Memory. Neural Computation**, v. 9, n. 8, p. 1735–1780, 1997.

JAIN, Arti et al. **Forecasting Price of Cryptocurrencies Using Tweets Sentiment Analysis**. In: p. 1–7. DOI: 10.1109/IC3.2018.8530659.

JIANG, Weiwei. **Applications of deep learning in stock market prediction: recent progress**. [S.l.: s.n.], 2020. arXiv: 2003.01859 [q-fin.ST].

JOACHIMS, Thorsten. **A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization.** In: FISHER, Douglas H. (Ed.). Proceedings of ICML-97, 14th International Conference on Machine Learning. Nashville, US: Morgan Kaufmann Publishers, San Francisco, US, 1997. P. 143–151. Disponível em: <http://citeseer.ist.psu.edu/54920.html>.

JOSHI, Kalyani; N, Bharathi; RAO, Jyothi. **Stock Trend Prediction Using News Sentiment Analysis.** International Journal of Computer Science and Information Technology, v. 8, p. 67–76, jun. 2016. DOI: 10.5121/ijcsit.2016.8306.

KIM, Young Bin et al. **Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies**. PLOS ONE, v. 11, e0161197, ago. 2016. DOI: 10.1371/journal.pone.0161197.

KRAAIJEVELD, Olivier; DE SMEDT, Johannes. **The predictive power of public Twitter sentiment for forecasting cryptocurrency prices**. Journal of International Financial Markets, Institutions and Money, v. 65, p. 101188, 2020. ISSN 1042-4431. DOI: https://doi.org/10.1016/j.intfin.2020.101188. Disponível em: <http://www.sciencedirect.com/science/article/pii/S104244312030072X>.

LI, Quanzhi; SHAH, Sameena. **Learning Stock Market Sentiment Lexicon and Sentiment-Oriented Word Vector from StockTwits**. In: PROCEEDINGS of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Vancouver, Canada: Association for Computational Linguistics, ago. 2017. P. 301–310. DOI: 10.18653/v1/K17-1031. Disponível em: <https://www.aclweb.org/anthology/K17-1031>.

LI, Tianyu Ray et al. **Sentiment-Based Prediction of Alternative Cryptocurrency Price Fluctuations Using Gradient Boosting Tree Model.** Frontiers in Physics, v. 7, p. 98, 2019. ISSN 2296-424X. DOI: 10.3389/fphy.2019.00098.

LI, Xinyi et al. DP-LSTM: **Differential Privacy-inspired LSTM for Stock Prediction Using Financial News.** [S.l.: s.n.], 2019. arXiv: 1912.10806 [q-fin.ST].

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

MACHADO, Kascilene Gonçalves. **Análise Comparativa dos Retornos Efetuados e Estimados de Ações de Empresas Brasileiras.** Revista Ciência Administrativas, v. 26, 2020.

MEDEIROS, Murilo; BORGES, Vinicius. **Tweet Sentiment Analysis Regarding the Brazilian Stock Market**. In: DOI: 10.5753/brasnam.2019.6550.

MEHTAB, Sidra; SEN, Jaydip. **A Robust Predictive Model for Stock Price Prediction Using Deep Learning and Natural Language Processing.** [S.l.: s.n.], 2019. arXiv: 1912.07700 [q-fin.ST].

MERN, John; ANDERSON, S.; POOTHOKARAN, John. **Using Bitcoin Ledger Network Data to Predict the Price of Bitcoin**.

MIKOLOV, Tomas; LE, Quoc V.; SUTSKEVER, Ilya. **Exploiting Similarities among Languages for Machine Translation.** [S.l.: s.n.], 2013. arXiv: 1309.4168 [cs.CL].

MIKOLOV, Tomas et al. **Efficient Estimation of Word Representations in Vector Space**. Proceedings of Workshop at ICLR, v. 2013, jan. 2013.

MITTAL, Anshul. **Stock Prediction Using Twitter Sentiment Analysis**.

NELSON, David; PEREIRA, Adriano; OLIVEIRA, Renato de. **Stock market's price movement prediction with LSTM neural networks**. In: p. 1419–1426. DOI: 10.1109/IJCNN.2017.7966019.

NGUYEN, Thien Hai; SHIRAI, Kiyoaki; VELCIN, Julien. **Sentiment Analysis on Social Media for Stock Movement Prediction**. Expert Syst. Appl., Pergamon Press, Inc., USA, v. 42, n. 24, p. 9603–9611, dez. 2016. ISSN 0957-4174. DOI: 10.1016/j.eswa.2015.07.052. Disponível em: <https://doi.org/10.1016/j.eswa.2015.07.052>.

NTI, Isaac kofi; ADEKOYA, Adebayo; WEYORI, Benjamin. **Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence From Ghana**. Applied Computer Systems, v. 25, p. 33–42, jun. 2020. DOI: 10.2478/acss-2020-0004.

OLAH, Christopher. **Understanding LSTM Networks**. [S.l.: s.n.], 2015. Disponível em: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>.

PENG, Yangtuo; JIANG, Hui. **Leverage Financial News to Predict Stock Price Movements Using Word Embeddings and Deep Neural Network**s. In: p. 374–379. DOI: 10.18653/v1/N16-1041.

PEREIRA, Alexandre André Santos; SILVA COELHO, Fernando Miguel Teixeira da; SILVA MONTEIRO, Jean Carlos da. **O Twitter no webjornalismo: os impactos da cibercultura e da mobilidade digital na narrativa jornalística**. Anais do XXI Congresso de Ciências da Comunicação na Região Nordeste - Intercom Nordeste, 2019.

PRING, Martin J. **Technical Analysis Explained: The Successful Investor's Guide to Spotting Investment Trends and Turning Points**. [S.l.]: McGraw-Hill Education, 2002. ISBN 9780071381932. Disponível em: <https://books.google.com.br/books?id=ng-4a53H87gC>.

POWERS, David MW. **Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation**. arXiv preprint arXiv:2010.16061, 2020

QIU, Jiayu; WANG, Bin; ZHOU, Changjun. **Forecasting stock prices with long-short term memory neural network based on attention mechanism.** PLOS ONE, Public Library of Science, v. 15, n. 1, p. 1–15, jan. 2020. DOI: 10.1371/journal.pone.0227222. Disponível em: <https://doi.org/10.1371/journal.pone.0227222>.

RADFORD, Alec et al. **Language Models are Unsupervised Multitask Learners.** Relatório Técnico. São Francisco: OpenAI, 2019.

RAJU, S M; TARIF, Ali Mohammad. **Real-Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis**. [S.l.: s.n.], 2020. arXiv: 2006.14473 [q-fin.ST].

REILLY, Frank K.; BROWN, Keith C. **Investment Analysis and Portfolio Management.** [S.l.]: Dryden Press, 1997. (Dryden Press series in finance).

RHEA, R. **The Dow Theory**. [S.l.]: Igal Meirovich, 2013. ISBN 9781607966289. Disponível em: <https://books.google.com.br/books?id=FO65ngEACAAJ>.

RONG, Xin. **Word2vec Parameter Learning Explained**. [S.l.: s.n.], 2014.
RUCHIGA, Mariana Antunes. **Comunicação E Mídias Sociais: Estratégias De Personalização E Humanização De Marca No Twitter**. COMUNICOLOGIA, v. 12, p. 88–109, 2019.

RUDGE, Luiz Fernando; CAVALCANTE, Francisco. **Mercado de Capitais Comissão Nacional.** 3. ed. Belo Horizonte, MG, Brasil: CNBV, 1996.

RUSSELL, Stuart J.; NORVIG, Peter. **Artificial Intelligence: A Modern Approach, 3ª edição**. Upper Saddle River: Prentice Hall, 2010.

SAK, Hasim; SENIOR, Andrew; BEAUFAYS, Françoise. **Long short-term memory recurrent neural network architectures for large scale acoustic modeling**. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, p. 338–342, jan. 2014.

SANTAELLA, L.; LEMOS, R. **Redes sociais digitais: a cognição conectiva do Twitter.** [S.l.]: Paulus, 2010. (Comunicação (Paulus (Firm : Brazil)))). ISBN 9788534932394. Disponível em: <https://books.google.com.br/books?id=xa9BYgEACAAJ>.

SATTAROV, Otabek et al. **Recommending Cryptocurrency Trading Points with Deep Reinforcement Learning Approach**. Applied Sciences, v. 10, p. 1506, fev. 2020. DOI: 10.3390/app10041506.

SIEGEL, Jeremy. **Investindo em Ações no Longo Prazo**. 5. ed. New York, US: Bookman, 2015.

SILVER, David et al. "**A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play**". Science, v. 362, n. 6419, pp. 1140–1144, 2018.

SOCIAL & HOOTSUITE, **We are. Digital in 2020.** Distribuído digitalmente. [S.l.], 2020. Disponível em: <https://wearesocial.com/digital-2020>. Acesso em: 4 out. 2020.

SOUZA, Dyliane Mourí Silva De; LUCENA, Wenner Glaucio Lopes; QUEIROZ, Dimas Barrêto De. **O Efeito do Sentimento do Investidor Expresso via Twitter sobre o Comportamento do Mercado Acionário Brasileiro Durante o Período Eleitoral**. XVIII Congresso USP de Iniciação Cientifica em Contabilidade, 2019.

TAVARES, Cristiano Viana Cavalcanti Castellão; SAMPAIO, Valdeci Cira Filgueira. **O poder da influência das redes sociais na decisão de compra do consumidor universitário da cidade de Juazeiro do Norte-CE**. Semana Acadêmica = Revista Científica, v. 01, 2017.

TELLES, André. **A revolução das mídias sociais: Cases, Conceitos, Dicas e Ferramentas.** São Paulo, SP, Brasil: M. Books do Brasil Editora Ltda., 2011.

TENSORFLOW. **Word2Vec tutorial**. Distribuído digitalmente. [S.l.], 2016. Disponível em: <https://www.tensorflow.org/tutorials/word2vec>. Acesso em: 4 out. 2020.

TURNER, Zane; LABILLE, Kevin; GAUCH, Susan. **Lexicon-Based Sentiment Analysis for Stock Movement Prediction**. v. 14, n. 5, p. 185–7, 2020.

URAS, Nicola et al. **Forecasting Bitcoin closing price series using linear regression and neural networks models.** [S.l.: s.n.], 2020.

USP, Universidade de São Paulo. **Repositório de Word Embeddings do NILC.** Distribuído digitalmente. [S.l.], 2017. Disponível em: <https://www.tensorflow.org/tutorials/word2vec>. Acesso em: 4 out. 2020.

VARGAS, Manuel et al. **Deep Leaming for Stock Market Prediction Using Technical Indicators and Financial News Articles.** In: p. 1–8. DOI: 10.1109/IJCNN.2018.8489208.

VELAY, Marc; DANIEL, Fabrice. **Using NLP on news headlines to predict index trends.** ArXiv, abs/1806.09533, 2018.

WELLMAN, Barry. For a Social Network Analysis of Computer Networks: A Sociological Perspective on Collaborative Work and Virtual Community. In: PROCEEDINGS of the 1996 ACM SIGCPR/SIGMIS Conference on Computer Personnel Research. Denver, Colorado, USA: Association for Computing Machinery, 1996. P. 1–11. Disponível em: <https://doi.org/10.1145/238857.238860>.

WITTGENSTEIN, Ludwig. **Philosophical Investigations**. Oxford: Basil Blackwell, 1953. ISBN 0631119000.

XING, Frank Z.; CAMBRIA, Erik; WELSCH, Roy E. **Natural Language Based Financial Forecasting: A Survey.** Artif. Intell. Rev., Kluwer Academic Publishers, USA, v. 50, n. 1, p. 49–73, jun. 2018. ISSN 0269-2821. DOI: 10.1007/s10462-017-9588-9. Disponível em: <https://doi.org/10.1007/s10462-017-9588-9>.

XU, Yumo; COHEN, Shay. **Stock Movement Prediction from Tweets and Historical Prices.** In: p. 1970–79. DOI: 10.18653/v1/P18-1183.

YACIM, Joseph; BOSHOFF, Douw. **Impact of Artificial Neural Networks Training Algorithms on Accurate Prediction of Property Values**. Journal of Real Estate Research, v. 40, p. 375–418, nov. 2018.

YU, Pengfei; YAN, Xuesong. **Stock price prediction based on deep neural networks.** Neural Computing and Applications, v. 32, mar. 2020. DOI: 10.1007/s00521-019-04212-x.

ZOEN, Joshua et al. **BERT-based Financial Sentiment Index and LSTM-based Stock Return Predictability.** [S.l.: s.n.], 2019. arXiv: 1906.09024.