

SENTECH: Um combinador de análise técnica e de sentimento para o mercado de ações

SENTECH: A stock market sentiment and technical analysis combiner

RESUMO

Isabela Nunes Caetano
isabelac@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, PR, Brasil

Giovani Volnei Meinerz
giovanimainerz@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, PR, Brasil

O mercado financeiro apresenta extrema importância para as relações econômicas modernas. Além de se consolidar com o passar do tempo, cresce o número de pessoas que se tornam investidores. Para tanto, é necessário estratégias que auxiliem o processo de tomada de decisões. Para resolver este problema da previsão do comportamento das ações, encontram-se propostas que se separam por séculos. A primeira, baseada em indicadores matemáticos e a segunda baseada nos sentimentos identificados em texto. A fusão das técnicas já foi explorada em panorama internacional, entretanto, o mesmo não acontece em âmbito nacional. Visto isso, este trabalho apresenta o SENTECH, um modelo preditivo que combina as duas abordagens técnicas citadas, desenvolvido com objetivo de identificar a correlação entre o valor predito e o preço de fechamento das ações listadas na bolsa de valores brasileira. Como resultado da tarefa de classificação de dados, tem-se resultados de correlação moderados, com alta significância (95%), para duas empresas da Carteira Teórica do índice IBOVESPA.

PALAVRAS-CHAVE: Bolsa de valores. Twitter. Inteligência Artificial. Aprendizado do Computador. Correlação.

ABSTRACT

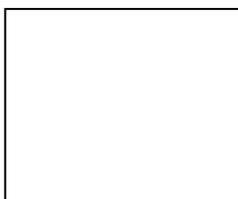
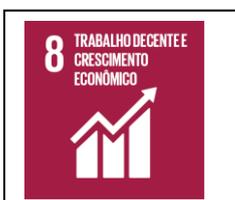
THE FINANCIAL MARKET IS EXTREMELY IMPORTANT FOR MODERN ECONOMIC RELATIONS. IN ADDITION TO CONSOLIDATING ITSELF OVER TIME, THE NUMBER OF PEOPLE WHO BECOME INVESTORS GROWS. FOR THIS PURPOSE, STRATEGIES ARE NEEDED TO ASSIST THE DECISION-MAKING PROCESS. TO SOLVE THIS PROBLEM OF FORECASTING THE BEHAVIOR OF ACTIONS, THERE ARE PROPOSALS THAT ARE SEPARATED FOR CENTURIES. THE FIRST BASED ON MATHEMATICAL INDICATORS AND THE SECOND BASED ON THE FEELINGS IDENTIFIED IN TEXT. THE FUSION OF TECHNIQUES HAS ALREADY BEEN EXPLORED ON THE INTERNATIONAL SCENE, HOWEVER, THE SAME DOES NOT HAPPEN AT THE NATIONAL LEVEL. HAVING SAID THAT, THIS PAPER PRESENTS THE SENTECH, A PREDICTIVE MODEL COMBINING THE TWO TECHNIQUES MENTIONED, DEVELOPED WITH THE OBJECTIVE OF IDENTIFYING THE CORRELATION BETWEEN THE PREDICTED VALUE AND THE CLOSING PRICE OF THE STOCKS LISTED ON THE BRAZILIAN STOCK EXCHANGE. AS A RESULT OF THE DATA CLASSIFICATION TASK, THERE ARE MODERATE CORRELATION RESULTS, WITH HIGH SIGNIFICANCE (95%), FOR TWO COMPANIES LISTED IN THE IBOVESPA INDEX.

KEYWORDS: Stock markets. Twitter. Artificial Intelligence. Machine Learning. Correlation.

Recebido: 19 ago. 2020.

Aprovado: 01 out. 2020.

Direito autoral: Este trabalho está licenciado sob os termos da Licença Creative Commons-Atribuição 4.0 Internacional.



INTRODUÇÃO

A expansão da digitalização acelerou a crescente esfera de dados global atingindo a produção de Zettabytes (ZB) a partir de 2010, já estimada para 50 ZB no ano corrente (REINSEL, GANTZ e RYDNING, 2018).

Consequência de tamanho volume será devido a uma maior interação dos usuários, com foco para as redes sociais, cenário dos dados que descrevem os comportamentos do público, já estimado em 140 milhões (STATISTA, 2020).

Nesse passo, um dos métodos para extrair informação é a **Análise de Sentimentos (AS)**, que tem como objetivo classificar a emoção e o humor denotados em texto. A referida técnica tem sido amplamente utilizada como estratégia para descrever o comportamento do mercado de ações. Isso, com maior assertividade no curto prazo e, possivelmente, menor assertividade no longo prazo (SHAH, ISAH e ZULKERNINE, 2019). Assim, BOLLEN, MAO e ZENG (2011) foram os precursores na utilização do Twitter como fonte de dados para a AS.

Em outro viés, para a mesma finalidade de acompanhamento do mercado financeiro, tem-se uma abordagem atemporal, a **Análise Técnica (AT)**, comumente adotada por investidores e que trata de dados quantitativos, a fim de buscar padrões recorrentes e previsíveis nos preços das ações. Ademais, a filosofia técnica, pressupõe que o preço é influenciado pela psicologia do investidor tanto quanto pelas variáveis financeiras (DAMODARAN, 2010).

Na literatura, ATTANASIO et al. (2019) se destacou ao explorar, em conjunto, as estratégias de AT e AS para identificação da reversão de tendência, que obtiveram maior sucesso do que em relação as estratégias analisadas individualmente. O seu trabalho se assimila com o executado por PICASSO et al. (2019), que se diferencia pelo intuito de prever o preço de fechamento das ações. No entanto, apesar dos trabalhos citados explorarem a combinação das técnicas, não há registros desta aplicação no âmbito nacional.

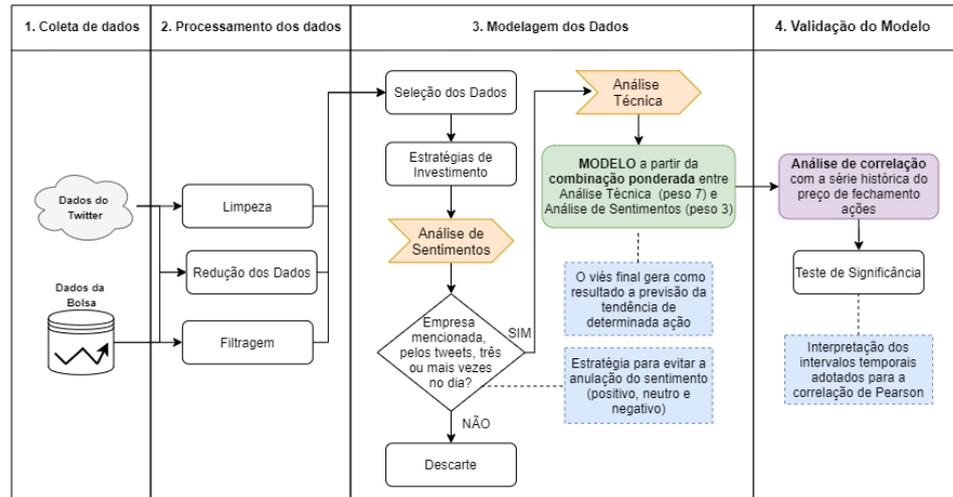
Assim sendo, este trabalho de pesquisa tem como objetivo desenvolver um modelo preditivo que combina as análises técnica e de sentimento, visando **identificar a correlação** entre o valor predito com o preço de fechamento das ações listadas na **bolsa de valores brasileira**.

MATERIAL E MÉTODOS

Definido o objetivo, levanta-se a hipótese da relação existente entre as variáveis, objeto de estudo deste trabalho. Isto posto, optou-se por realizar a validação da hipótese do ponto de vista da metodologia de um projeto fundamentado na ciência de dados, ilustrado na Figura 1, que tem como principais etapas para o desenvolvimento (*pipeline*) de um modelo: coleta, processamento e modelagem de dados e, por fim, a validação do modelo.

As etapas que constituem o *pipeline* do modelo, requerem a realização de tarefas específicas. Para tanto, escolheu-se a linguagem Python, que possui bibliotecas que dão subsídio para que os requisitos necessários fossem alcançáveis, tais como: Pandas, para manipulação e análise de dados; TA-Lib, para cálculo dos indicadores técnicos; SciPy, para análises estatísticas e a Matplotlib, para visualização dos dados por meio de gráficos.

Figura 1 - Fluxograma de desenvolvimento do *pipeline* do modelo



Fonte: Autoria própria (2020).

A etapa inicial do *pipeline*, a **coleta de dados**, consiste em mapear o escopo do trabalho. Para tanto, foi utilizado um conjunto refinado de *tweets* emitidos por veículos de comunicação especializados no mercado de ações brasileiro, em um intervalo temporal que vai de 12/11/2015 à 16/05/2019. O processo para a identificação dos perfis válidos foi realizado manualmente por MELO e MEINERZ (2019), bem como a rotulação da polaridade do sentimento presente em uma mensagem de texto, classificado em: positivo (1), neutro (0) ou negativo (-1).

Para a segunda base de dados, referente a **série histórica diária das ações** das empresas que fazem parte da Carteira Teórica do IBOVESPA, utilizou-se como fonte de dados a Brasil, Bolsa, Balcão (B3), no intervalo temporal correspondente ao da base de dados do Twitter.

Em seguida, para o refino dos dados, cerne do desenvolvimento do modelo, são necessárias tarefas de **filtragem e limpeza**, típicas da etapa de **processamento de dados**. Assim, os dados do Twitter foram adaptados para análise diária, enquanto que do *dataset* da B3 foram filtrados os registros do segmento ações, a vista e fracionário, equivalente a 36,4% dos registros da base original.

Finalizada a etapa de processamento, inicia-se a da **modelagem de dados**, que tem como objetivo principal a construção do modelo preditivo. A primeira tarefa empregada foi a **seleção dos dados**, com a decisão de escolha pelas empresas que seriam exploradas. Para isso, criaram-se variáveis explanatórias, de agrupamento dos sentimentos, para responder à seguinte questão: quais empresas são mencionadas, nos *tweets*, três ou mais vezes em um dia?

A limitação do volume válido (mínimo de três *tweets*), mencionado na questão, se deu a fim de evitar a anulação entre os sentimentos positivo e negativo, o que também garante que haja uma maior convicção do sentimento, por parte do mercado, para o próximo pregão.

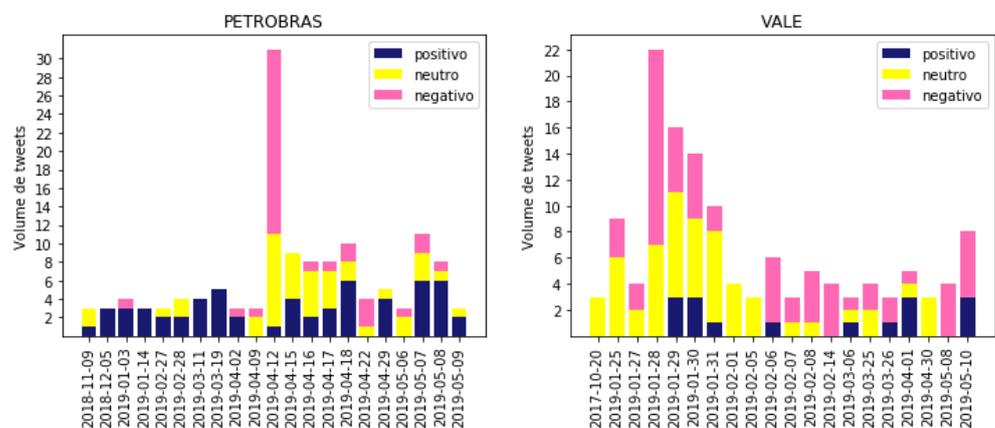
A resposta manteve, então, somente 35 empresas, correspondente a 56,45% do total de 62 empresas que compunham a carteira teórica do índice Ibovespa considerado em maio de 2019. Ainda assim, as tarefas seguintes prosseguiram somente com duas empresas que, em algum dia, apresentaram um notório volume de dados, sendo este o caso da PETROBRAS e VALE.

O volume dos *tweets* delimitam o intervalo temporal de análise, onde para todas as empresas observa-se um volume notório e válido a partir de 2019, conforme ilustra a Figura 2. Entretanto, a frente, são analisadas as ações das empresas no mês em que há maior quantidade de postagens, com destaque para o mês de abril, para a PETROBRAS, e janeiro, para a VALE.

Sabendo que uma empresa listada em uma bolsa de valores pode ter mais do que um tipo de ação sendo negociada, decidiu-se selecionar aquelas com maior liquidez, ou seja, as mais negociadas. São elas:

- a) PETR4 - Petróleo Brasileiro S.A;
- b) VALE3 - Vale S.A.

Figura 2 - Sentimentos rotulados para as empresas analisadas (PETROBRAS E VALE)



Fonte: Autoria própria (2020).

Para a tarefa de **análise de sentimento**, sabe-se que o sentimento coletado pelo conjunto de *tweets* em um dia t representará a tendência, o que chamamos de viés, para o próximo pregão ($t + 1$) – dia em que ocorre negociação das ações. Para este fim, calculou-se a força do sentimento diário (F) a partir da divisão entre a soma (S) dos sentimentos classificados em positivo (1), neutro (0) e negativo (-1) e a quantidade total de *tweets* (T), conforme a Equação 1:

$$F = \frac{S}{T} \tag{1}$$

Os valores emitidos para os vieses da AS foram mapeados, em uma escala de -1 a 1, de forma análoga a variação de tendência nos níveis de venda forte, venda, neutro, compra e compra forte, conforme a Tabela 1.

Tabela 1 - Escala de tendência para a polaridade dos sentimentos

Força do sentimento (F)	Viés de tendência
-1,0 à -0,6	VENDA FORTE
-0,59 à -0,2	VENDA
-0,19 à +0,19	NEUTRO
+0,2 à +0,59	COMPRA
+0,6 à +1,0	COMPRA FORTE

Fonte: Autoria própria (2020).

Para a modelagem do ponto de vista da AT, tem-se uma **estratégia de investimento** que combina grupos de indicadores técnicos. Sendo assim, o primeiro indicador escolhido foi o das **Médias Móveis Exponenciais (MME)**, que levam em consideração a comparação entre duas médias de períodos (n) diferentes, sendo uma de curto prazo (5 dias) e outra de prazo maior (20 dias), atribuindo maior peso aos dados mais recentes (MURPHY, 1999), demonstrado na Equação 2:

$$MME_{(atual)} = Preço_{(atual)} - MME_{(anterior)} * \frac{2}{n+1} + MME_{(anterior)} \quad (2)$$

Uma aplicação mais ampla das médias móveis é o indicador **Convergência/Divergência das Médias Móveis (MACD)**, reconhecido por MURPHY (1999) como um híbrido ao combinar indicadores de oscilação com o de médias móveis. Este, compõe-se de duas linhas que sinalizam o início ou fim de uma tendência: a linha MACD, onde o valor da MME mais curta (12 dias) é subtraído pelo valor da MME mais longa (26 dias), e a linha de Sinal, composta pela linha MACD ajustada por uma MME de 9 dias.

Por último, buscou-se um indicador capaz de realizar a confirmação das tendências encontradas. A decisão foi pelo **Volume**. Ele se refere a quantidade de dinheiro sendo negociado, representando a força do mercado e, ainda, retratando a espinha dorsal da AT - oferta/demanda. E para o viés final da AT, a combinação dos indicadores técnicos ficou definida conforme a Equação 3:

$$Viés_{AT} = (0.5 * MME) + (0.3 * MACD) + (0.2 * Volume) \quad (3)$$

Assim como realizado na tarefa de identificação do viés da AS, o conjunto dos indicadores técnicos foram mapeados para a escala de -1 a 1, como pode ser observado na Tabela 2.

Tabela 2 - Escala de tendência para a AT

MME	MACD	VOLUME	ESCALA	F
COMPRA	COMPRA	SIM	+1	COMPRA FORTE
COMPRA	COMPRA	NÃO	+0,5	COMPRA
COMPRA	VENDA	SIM	+0,5	COMPRA
COMPRA	VENDA	NÃO	+0,5	COMPRA
VENDA	COMPRA	SIM	-0,5	VENDA
VENDA	COMPRA	NÃO	-0,5	VENDA
VENDA	VENDA	NÃO	-0,5	VENDA
VENDA	VENDA	SIM	-1	VENDA FORTE

Fonte: Autoria própria (2020).

Cumpridas as técnicas, a fase de Modelagem dos Dados chega à última tarefa de implementação: a **combinação ponderada entre o viés de ambas técnicas** para a consolidação do modelo preditivo, nomeado como SENTECH (Combinador de Análise Técnica e Sentimento), que atribui maior peso para a AT (0,7), visto que a AS (0,3) apresenta maior sensibilidade em seu viés.

Com o modelo desenvolvido, inicia-se a etapa de **Validação do Modelo**, que consiste da aplicação e análise de testes estatísticos, com vistas a apoiar o processo de inferência. Portanto, aplicou-se a **análise por correlação**, exploração descritiva que mede se há, e em que grau, dependência entre duas variáveis, embora não

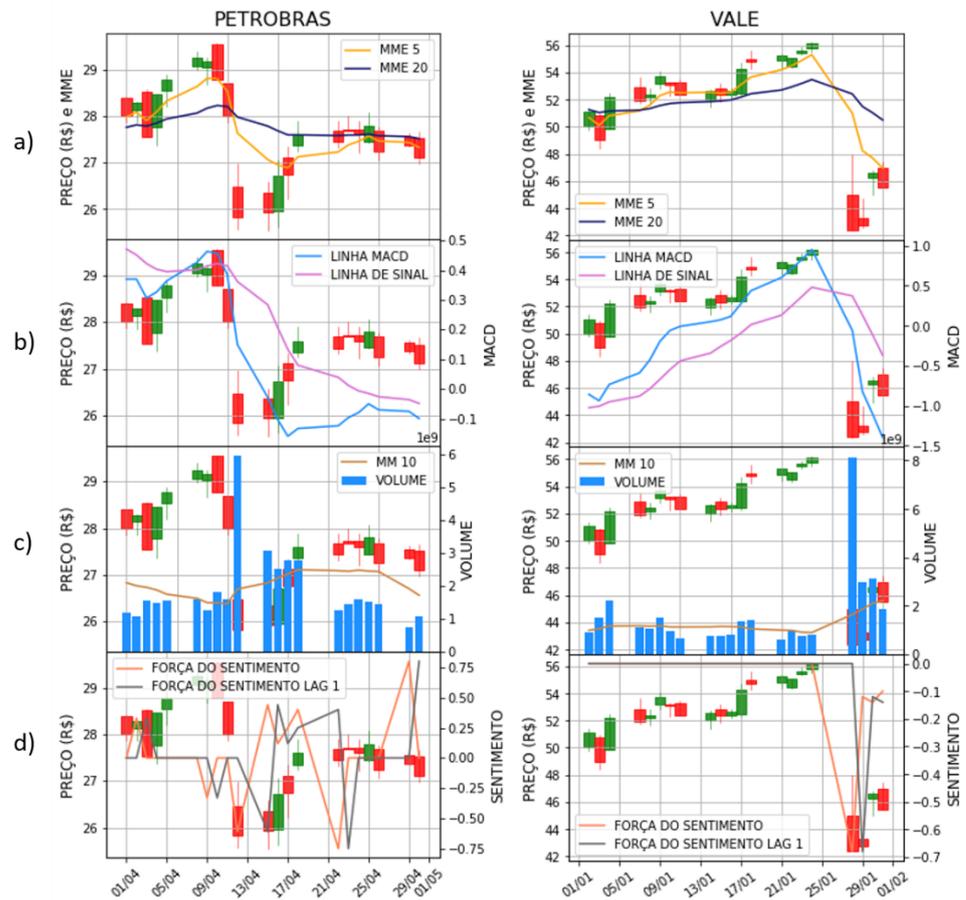
implique em causalidade (MOORE e KIRKLAND, 2007). Empregou-se o **Coefficiente de Correlação de Pearson (ρ)**, alcançável pela divisão entre a covariância (*cov*) e o desvio padrão (σ), conforme a Equação 4. Logo, quanto mais próximo de 1 ou -1 está o valor (ρ), mais forte é a relação entre as variáveis.

$$\rho_{x,y} = \frac{cov_{x,y}}{\sigma_x \sigma_y} \quad (4)$$

RESULTADOS E DISCUSSÃO

A Figura 3 apresenta a aplicação dos indicadores técnicos em comparação com o Gráfico de Candles, uma das principais ferramentas da AT, com informações de preço de fechamento, abertura, mínimo e máximo.

Figura 3 - AT e AS aplicadas para PETROBRAS e VALE



Fonte: Autoria própria (2020).

Sendo assim, é possível notar o comportamento dos indicadores técnicos, que acompanham a movimentação real do preço das ações analisadas. Além do mais, nos momentos em que há inversão de tendência, nota-se o cruzamento das linhas que compõe a MME (a) e o MACD (b), além do Volume (c) acima da média. Já para o indicador que representa a força do sentimento (d), nota-se um atraso de um dia (*lag 1*) em relação ao comportamento traçado pelos *candles*.

Já na etapa Validação do Modelo, a hipótese foi validada para a correlação entre o viés da combinação dos indicadores e o preço de fechamento das ações. Assim, pelo teste de significância, obteve-se valores de correlação válidos com um nível de confiança de 95% ao adotar a probabilidade de erro de 5%, conforme elenca a Tabela 3, com os valores de p-valor menores que o α adotado (0,05), sendo, portanto, significativos.

Tabela 3 - Resultado da correlação para a PETROBRAS e VALE

Empresa	Intervalo	Correlação (ρ)	p-valor
PETROBRAS	1 mês 01/04/2019 à 31/04/2019	0,57 MODERADA	0,0073
	1 ano 01/05/2018 à 31/05/2019	0,24 DESPREZÍVEL	5,29e-05
VALE	1 mês 01/01/2019 à 31/01/2019	0,65 MODERADA	0,0013
	1 ano 01/05/2018 à 31/05/2019	0,34 FRACA	9,76e-09

Fonte: Autoria própria (2020).

Ainda, para ambas ações foi encontrado um baixo nível de correlação ao considerar a janela de análise de um ano, o que ainda pode-se justificar pelo baixo volume de tweets encontrados no longo prazo. A justificativa foi analisada, na Tabela 4, para a empresa VALE, aumentando a janela de análise para dois meses, intervalo no qual houve uma maior constância de postagens referentes a empresa.

Tabela 4 – Correlação para a VALE em janela de dois meses

Empresa	Intervalo	Correlação (ρ)	p-valor
VALE	2 meses 01/01/2019 à 31/02/2019	0,7 FORTE	4.30e-07

Fonte: Autoria própria (2020).

CONCLUSÃO

Voltando à questão colocada no início deste estudo, agora é possível afirmar, com 95% de confiança, que o resultado predito pelo SENTECH, modelo desenvolvido a partir da combinação do viés das abordagens de análise técnica e análise de sentimentos, está correlacionado, para o curto prazo, com o preço de fechamento das ações.

A pesquisa apresenta como principais benefícios propiciados pelo modelo a quantificação da correlação entre o resultado do modelo e a série histórica dos preços de fechamento das ações e a alta significância dos testes de correlação.

Em consequência, a aplicação do modelo reforça a importância do mercado financeiro para as relações modernas, unindo o elo do conservadorismo da análise técnica e o contemporâneo da análise de sentimentos.

Para a continuidade deste trabalho, destaca-se como principal recomendação, caracterizada como uma extensão natural do SENTECH, a validação do modelo por meio do *backtesting*, a fim de avaliar o percentual de lucro obtido a partir das decisões tomadas baseadas na previsão realizada, produto deste trabalho.

REFERÊNCIAS

- ATTANASIO, G. et al. Combining News Sentiment and Technical Analysis to Predict Stock Trend Reversal. In: **2019 International Conference on Data Mining Workshops (ICDMW)**, Beijing, 2019. Disponível em: <https://ieeexplore.ieee.org/document/8955650>. Acesso em: 23 abr. 2020.
- BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. **Journal of Computational Science**, v. 2, n. 1, p. 1-8, 2011. Disponível em: <http://www.sciencedirect.com/science/article/pii/S187775031100007X>. Acesso em: 23 mai. 2020.
- DAMODARAN, A. **Avaliação de investimentos: Ferramentas e Técnicas para Determinação do Valor de Qualquer Ativo**. 2. ed. Rio de Janeiro: Qualitymark, 2010.
- MELO, J. G.; MEINERZ, G. V. Análise de sentimentos de textos voltados ao mercado de ações. In: **XXIV Seminário de Iniciação Científica e Tecnológica**, Pato Branco, 2019. Disponível em: <https://eventos.utfpr.edu.br/sicite/sicite2019/paper/viewFile/4739/1331>. Acesso em: 20 mar. 2020.
- MOORE, D. S.; KIRKLAND, S. **The basic practice of statistics**. New York: WH Freeman, 2007.
- MURPHY, J. **Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications**. New York Institute of Finance, 1999.
- PICASSO, A. et al. Technical analysis and sentiment embeddings for market trend. **Expert Systems with Applications**, v. 135, p. 60-70, 2019. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0957417419304142?via%3Dihub>. Acesso em: 23 abr. 2020.
- REINSEL, D.; GANTZ, J.; RYDNING, J. The digitization of the world from edge to core. **Framingham: International Data Corporation**, 2018. Disponível em: <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>. Acesso em: 01 jun. 2020.
- SHAH, D.; ISAH, H.; ZULKERNINE, F. Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. **International Journal of Financial Studies**, Kingston, v. 7, n. 2, p. 26, 2019. Disponível em: <https://www.mdpi.com/2227-7072/7/2/26>. Acesso em: 23 mar. 2020.

STATISTA. **Number of social network users in Brazil from 2017 to 2023 (in millions)**. 2020. Disponível em:

<https://www.statista.com/statistics/278408/number-of-social-network-users-in-brazil/>. Acesso em: 01 abr. 2020.