UNIVERSIDADE FEEVALE

FREDERICO AUGUSTO SCHMITZ

##### *MACHINE LEARNING* NA IDENTIFICAÇÃO DO SENTIMENTO ASSOCIADO AOS *TWEETS* E PREÇO DE AÇÕES

(Título Provisório)

###### Anteprojeto de Trabalho de Conclusão

Novo Hamburgo

2021

FREDERICO AUGUSTO SCHMITZ

##### *MACHINE LEARNING* NA IDENTIFICAÇÃO DO SENTIMENTO ASSOCIADO AOS *TWEETS* E PREÇO DE AÇÕES

(Título Provisório)

Anteprojeto de Trabalho de Conclusão de Curso, apresentado como requisito parcial

à obtenção do grau de Bacharel em

Sistemas de Informação pela

Universidade Feevale

Orientador: Juliano Varella de Carvalho

Novo Hamburgo

2021

# RESUMO

O número de novos investidores cadastrados na bolsa de valores brasileira quase dobrou em 2020, em relação ao ano anterior. A maioria destes novos investidores tem como principal motivação buscar conhecimento em novas formas de investimento bem como uma maior rentabilidade. Pesquisas mostram que as redes sociais e plataformas online são a principal fonte utilizada por estes investidores para aprendizagem e busca de informações. O número crescente de investidores e transações executadas criam um alto volume de dados que escondem padrões e correlações, fazendo o mercado financeiro um candidato relevante à execução de técnicas de *Machine Learning* sobre *Big Data*. A análise de sentimentos é o campo responsável por analisar opiniões e sentimentos das pessoas em relação a produtos e serviços, o rápido crescimento desta área coincide com o crescimento das redes sociais e a grande quantidade de opiniões geradas por seus usuários. Pelos motivos e informações apresentadas anteriormente, este trabalho propõe o desenvolvimento de um protótipo para extração de *tweets* e dados da bolsa de valores brasileira com o objetivo de aplicar técnicas de *Machine Learning* a fim identificar relações entre o sentimento associado aos *tweets* e o preço de ações.

Palavras-chave: *Machine Learning*.Análise de sentimentos. Bolsa de Valores. Ações. *Tweets.*

SUMÁRIO

MOTIVAÇÃO ...........................................................................................................................5

OBJETIVOS ..............................................................................................................................8

METODOLOGIA ......................................................................................................................9

CRONOGRAMA ....................................................................................................................10

BIBLIOGRAFIA ....................................................................................................................11

#  MOTIVAÇÃO

O número de pessoas cadastradas no Brasil Bolsa Balcão (B3), a bolsa de valores oficial do Brasil, quase dobrou em 2020 em relação a 2019. De acordo com os dados divulgados pela B3, a maior motivação destes novos investidores foi a busca de aprendizado em novas modalidades de investimentos, bem como a procura por uma maior rentabilidade (BRASIL BOLSA BALCÃO, 2020).

A B3 relata que o perfil do novo investidor é jovem, e utiliza principalmente as redes sociais e outras plataformas *online* para aprender e se informar sobre investimentos. Cerca de 60% destes novos investidores seguem ou consultam influenciadores digitais para obter informações (BRASIL BOLSA BALCÃO, 2020).

A pesquisa da B3 também mostra que o produto mais conhecido do investidor são as ações. Cerca de 72% dos entrevistados possuem ações atualmente, e 43% adquiriram ações como primeiro produto de investimento. Além disso, a maior diversificação de ativos é observada em contas com menores volumes financeiros, que formam a maioria na bolsa e possuem ao menos cinco papéis diferentes (BRASIL BOLSA BALCÃO, 2020).

A principal definição de *Big Data* está diretamente relacionada às características das informações processadas, e são elas: volume, variedade e velocidade (GANDOMI; HAIDER, 2015). O alto volume de transações faz o mercado financeiro ser um candidato relevante para aplicações de técnicas de *Big Data*. Diversas estratégias de compra e venda de ativos são desenvolvidas, utilizando como base as informações e padrões analisados a partir do resultado deste grande volume de dados. (TRELEWICZ, 2017). De acordo com reportagem publicada pelo Estadão, em 2020 a B3 atingiu um volume histórico de 5,13 bilhões de dólares movimentados por dia (APUD, 2020).

O campo de *Big Data* cresce a cada ano, a estimativa para 2020 era que o volume global de dados atingisse cerca de 35 Zettabytes(YE; LI, 2017). As operações envolvendo *Big Data* podem ser resumidas da seguinte forma: coletar dados de diversas fontes, analisar os dados coletados para identificar correlações e realizar predições utilizando as correlações encontradas (LEHIKOINEN; KOISTINEN, 2014). Diversas empresas incorporam a utilização de *Big Data* em seus negócios para auxiliar na tomada de decisões, e seu uso é aplicado em campos como marketing, finanças, política e ciências sociais (GANDOMI; HAIDER, 2015).

A utilização de *Big Data* envolve o processamento de diferentes tipos de dados (estruturados, semi-estruturados e não estruturados), sendo as redes sociais uma das principais fontes de dados. Entre as principais redes sociais é possível citar: *YouTube, Twitter, Instagram, Facebook* (YAQOOB et al., 2016). De acordo com a *Hootsuite,* em 2021 o número de usuários de redes sociais chegou a 4.2 bilhões. O *Twitter,* por exemplo,é uma rede social do tipo de *microblog*, responsável por 353 milhões destes usuários (HOOTSUITE, 2021).

De acordo com Mitchell (1997) a área de *Machine Learning* (ML) tem como objetivo a construção de programas que aprendam e se aperfeiçoem automaticamente com a experiência adquirida durante sua execução. *Text Mining* é uma subárea de *Big Data* e ML*,* responsável pela extração de informações adquiridas a partir de textos, estes dados normalmente não estão estruturados, sendo necessária a aplicação de técnicas de análise estatística e ML, a fim de gerar informação útil para análise humana. Um outro exemplo da utilização de ML é a predição da bolsa de valores, através de informações extraídas de notícias do mercado financeiro (GANDOMI; HAIDER, 2015).

A análise de sentimentos é o campo responsável por analisar opiniões, sentimentos e emoções das pessoas em relação a produtos, serviços, acontecimentos e outros tópicos. O rápido crescimento dessa área coincide com o crescimento das redes sociais e a grande quantidade de opiniões e dados gerados por seus usuários. Apesar da grande quantidade de informações disponíveis, o usuário comum possui dificuldade para processar e extrair informações relevantes, por isso a necessidade de aplicações automatizadas de análise de sentimentos (LIU, 2012).

A correlação entre a atividade de usuários na internet e a predição de tendências é um assunto que já atraiu diversos pesquisadores, conforme Zhang, Fuehres e Gloor (2011). Em seu trabalho, os autores realizaram uma pesquisa para encontrar uma relação entre os índices de ações *DOW, NASDAQ* e *S&P 500,* e o humor negativo dos usuários no *Twitter*. Os pesquisadores observaram que conforme os usuários expressam mais palavras de medo, ansiedade e esperança, o índice de ações tende a cair no dia seguinte. Para validar seu trabalho, os autores compararam seus resultados com o índice *VIX,* que mede a volatilidade do mercado de ações.

Uma pesquisa desenvolvida por Bing, Chan e Ou (2014), buscou responder às seguintes questões: O sentimento público pode ser analisado para prever movimentos de uma ação? O preço de uma ação ou setor pode ser mais previsível que outros? Para responder às perguntas anteriores, os pesquisadores mineraram dados do *Twitter,* buscando menções de empresas específicas e para cada empresa foi aplicada uma estratégia própria, buscando *tweets* que não mencionam apenas a empresa diretamente, mas também seus produtos e serviços. Para cada *tweet* foi definido um peso para identificar o sentimento em relação a empresa. Em seu experimento, o algoritmo proposto pelos pesquisadores atingiu uma precisão de 81,45% na estimativa das ações.

Ye e Li realizaram uma revisão da literatura (2017) para analisar o estado atual de *Big Data* em pesquisas acadêmicas aplicadas ao mercado de capitais. Os pesquisadores indicam que a maioria das informações estão concentradas em redes sociais e motores de busca, sendo essas as principais fontes de dados utilizadas em pesquisas na área. Baseando-se em suas estruturas, eles classificaram as fontes em três tipos: fórum, *microblog* e motor de busca. Em relação aos indicadores construídos com a utilização de *Big Data* foi possível notar impacto no volume de ações negociadas, retornos em curto prazo e no comportamento do investidor. Os autores concluem que apesar das pesquisas terem produzido resultados, ainda há espaço para melhorias nos métodos de análise e extração de dados.

Pelos motivos e informações apresentadas acima, este trabalho propõe o desenvolvimento de um protótipo para extração de *tweets* e dados da Bovespa com o objetivo de aplicar técnicas de *Machine Learning* a fim identificar relações entre o sentimento associado aos *tweets* e o preço de ações.

OBJETIVOS

Objetivo geral:

 Aplicar técnicas de *Machine Learning* em *tweets* relacionados a empresas listadas na bolsa de valores brasileira, procurando identificar relações entre o sentimento associado aos *tweets* e o preço da ação.

Objetivos específicos:

* Pesquisar trabalhos correlatos.
* Identificar algoritmos de *Machine Learning (ML)* a seremutilizados.
* Estabelecer critérios para escolha de ações e/ou setores.
* Desenvolver um protótipo para importação de dados do Bovespa.
* Estabelecer critérios para seleção de contas no *Twitter,* de onde serão extraídos os *tweets.*
* Desenvolver um protótipo para importação de *tweets*.
* Estabelecer critérios para filtrar *tweets*.
* Extrair um índice de sentimento diário sobre uma ação e/ou setor.
* Aplicar ML sobre os *tweets* de ações e/ou setores.
* Desenvolver uma interface gráfica para exibição dos dados.

# METODOLOGIA

Figura 1: Classificação da Pesquisa

Fonte: Adaptado de Prodanov e Freitas (2013, P.51)

Conforme a Figura 1, neste trabalho será desenvolvida uma pesquisa de natureza aplicada. De acordo com Prodanov e Freitas (2013), a pesquisa aplicada envolve a geração de conhecimento através de aplicação prática. Para atender este requisito, serão estudadas técnicas de *Machine Learning* para aplicação em *tweets* relacionados a empresas listadas na bolsa de valores brasileira.

Esta pesquisa tem como objetivo explicativo, pois através da aplicação de *Machine Learning* procura identificar uma relação entre os preços de ações e o sentimento extraído de *tweets.*

Serão aplicados neste estudo os procedimentos técnicos de pesquisa bibliográfica e pesquisa experimental. A pesquisa bibliográfica terá como objetivo levantar o referencial teórico e analisar técnicas e algoritmos de *Machine Learning* aplicados em trabalhos similares, a fim de auxiliar na execução prática deste trabalho.

Será desenvolvido um protótipo para extração dos *tweets,* das cotações da bolsa de valores e para visualização dos resultados. As cotações históricas da bolsa de valores estão disponíveis em forma de arquivo de texto, que serão processados e inseridos em um banco de dados. Os *tweets* podem ser extraídos através de sua *API* oficial, ou com a utilização de bibliotecas de *scraping*. Por fim, antes da aplicação de *Machine Learning* serão definidos critérios para seleção de ações e *tweets*.

A pesquisa terá uma abordagem quantitativa, pois através do modelo gerado pela aplicação de *Machine Learning,* é possível obter uma precisão do algoritmo utilizado*.* Para auxiliar esta análise, um índice de sentimento diário poderá ser criado e utilizado para comparação com as oscilações de preços das ações. Por fim, o trabalho busca responder a seguinte questão: é possível identificar uma relação entre os sentimentos expressados em *tweets* e oscilações no preço de ações no curto prazo?

# CRONOGRAMA

Trabalho de Conclusão I

|  |  |
| --- | --- |
| Etapa  | Meses |
| Mar | Abr | Mai | Junho |
| Anteprojeto |  |  |  |  |
| Pesquisa bibliográfica sobre aplicação de *Machine Learning* em ações |  |  |  |  |
| Estudar algoritmos de *Machine Learning* e análise de sentimentos |  |  |  |  |
| Definição de critérios para seleção de ações e *tweets* |  |  |  |  |
| Desenvolvimento TC I |  |  |  |  |

Trabalho de Conclusão II

|  |  |
| --- | --- |
| Etapa  | Meses |
| Ago | Set | Out | Nov |
| Protótipo para importação de ações e preços históricos |  |  |  |  |
| Protótipo para importação de *tweets* |  |  |  |  |
| Aplicar *Machine Learning* |  |  |  |  |
| Analisar resultados |  |  |  |  |
| Protótipo para visualização dos dados |  |  |  |  |
| Desenvolvimento TC II |  |  |  |  |

# BIBLIOGRAFIA

APUD, Mateus. **Com atual volume negociado, B3 seria a 7ª ação mais negociada nos EUA.** Estadão, São Paulo, julho 2020. Disponível em: <https://einvestidor.estadao.com.br/mercado/com-atual-volume-negociado-b3-seria-a-7a-acao-mais-negociada-nos-eua/>. Acesso em 28 mar. 2021.

BING, Li; CHAN, Keith C. C.; OU, Carol, **Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements**, 2014 IEEE 11th International Conference on e-Business Engineering, Guangzhou, China, 2014, pp. 232-239. Disponível em: <<https://ieeexplore.ieee.org/document/6982085>>. Acesso em: 05 mar. 2021.

BRASIL BOLSA BALCÃO. **B3 divulga estudo sobre os 2 milhões de investidores que entraram na bolsa entre 2019 e 2020**, São Paulo, SP, 2020. Disponível em: <http://www.b3.com.br/pt\_br/noticias/investidores.htm>. Acesso em: 07 mar. 2021.

GANDOMI, Amir.; HAIDER, Murtaza. **Beyond the hype: Big data concepts, methods, and analytics**. International Journal of Information Management. 35. p. 137-144. 2015.

Disponível em: <https://www.sciencedirect.com/science/article/pii/S0268401214001066>. Acesso em 21 mar. 2021

HOOTSUITE. **15.5 Users Join Social Every Second (and Other Key Stats to Know)**. Disponível em: <https://blog.hootsuite.com/simon-kemp-social-media/>. Acesso em: 21 mar. 2021.

LEHIKOINEN, Juha; KOISTINEN, Ville. **In big data we trust?**, Interactions, 21(5), p. 38–41. 2014. Disponível em: <<https://dl.acm.org/doi/10.1145/2641398>>. Acesso em: 14 mar. 2021

LIU, Bing. **Sentiment Analysis and Opinion Mining**. Synthesis Lectures on Human Language Technologies. 5. 2012.

MITCHELL, Tom. **Machine Learning (1st. ed.).** McGraw-Hill, Inc**.** 1997.

TRELEWICZ, Jennifer. Q., **Big Data and Big Money: The Role of Data in the Financial Sector**. IT Professional, vol. 19, no. 3, pp. 8-10, 2017. Disponível em: <https://ieeexplore.ieee.org/document/7945155>. Acesso em 24 mar. 2021.

YAQOOB, Ibrar; HASHEM, Ibrahim; GANI, Abdullah; MOKHTAR, Salimah; AHMED, Ejaz; ANUAR, Nor; VASILAKOS, Athanasios. **Big Data: From Beginning to Future**. International Journal of Information Management. 2016. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0268401216304753>. Acesso em 21 mar. 2021

YE, Minjian; LI, Guangzhong. **Internet big data and capital markets: a literature review.** Financial Innovation. 2017. Disponível em: <<https://www.researchgate.net/publication/316335254_Internet_big_data_and_capital_markets_a_literature_review>>. Acesso em: 13 mar. 2021

ZHANG, Xue; FUEHRES, Hauke; GLOOR, Peter. **Predicting Stock Market Indicators Through Twitter – “I Hope it is Not as Bad as I Fear**. Procedia - Social and Behavioral Sciences. 2011. p. 55–62. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1877042811023895>>. Acesso em: 08 mar. 2021