UNIVERSIdade FEEVALE

FREDERICO AUGUSTO SCHMITZ

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

Novo Hamburgo

2021

frederico augusto schMITZ

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

Trabalho de Conclusão de Curso

apresentado como requisito parcial

à obtenção do grau de Bacharel em

Sistemas de Informação pela

Universidade Feevale

Orientador: Prof. Dr. Juliano Varella de Carvalho

Novo Hamburgo

2021

Agradecimentos

Gostaria de agradecer a todos os que, de alguma maneira, contribuíram para a realização desse trabalho de conclusão, em especial:

Aos meus pais, Acir e Lucia por toda dedicação e suporte durante esta jornada, além de todo o apoio incondicional.

Aos meus irmãos Felipe e Fabricio, pela amizade e companheirismo.

Ao meu orientador Juliano, pelas sugestões, paciência e apoio durante a execução deste trabalho.

E a todos que de alguma forma tenham contribuído para a execução deste trabalho.

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 propôs a utilização de técnicas de *Machine Learning* com o objetivo de analisar se o sentimento expressado pelos investidores em *tweets* tem correlações com os movimentos do mercado financeiro. Para se chegar neste objetivo realizou-se a coleta de um grande volume de *tweets* através da aplicação de técnicas de *data mining.* Estes *tweets* foram então analisados e processados por meio da aplicação de algoritmos de Processamento de Linguagem Natural, visando filtrar apenas *tweets* que estejam relacionados com investidores. O sentimento destes *tweets* foi analisado com o auxílio da ferramenta *Google Cloud Natural Language*, e então validado através da comparação de seu resultado com um evento externo. Por fim um modelo de predição foi construído com a utilização do algoritmo de classificação *Support Vector Machine.* Os resultados apresentados por este modelo vão de encontro com a Hipótese do Mercado Eficiente. De acordo com ela, o preço de um ativo reflete de forma eficiente toda e qualquer informação já publicada, sendo assim o preço do amanhã não pode ser previsto utilizando informações do passado.

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

Abstract

The number of new investors registered on the Brazilian stock exchange almost doubled in 2020, in comparison to the previous year. Most of these new investors have as their main motivation to acquire knowledge in new ways of investment and are looking for better profitability. Research shows that social networks and online platforms are the main sources used by these investors to learn and seek information. The growing number of investors and executed transactions create a high volume of data that hides patterns and correlations, making the financial market a relevant candidate to apply Machine Learning and Big Data techniques. Sentiment analysis is the field responsible for analyzing people's opinions and feelings about products and services, the rapid growth of this area coincides with the growth of social networks and the large number of opinions generated by its users. For the reasons and information presented above, this work proposes the use of Machine Learning techniques to analyze whether the sentiment expressed by investors in tweets is correlated with the movements of the financial market. To reach this objective, a large volume of tweets was collected through the application of data mining techniques. These tweets were then analyzed and processed by Natural Language Processing algorithms, aiming to filter only tweets that are related to investors. The sentiment of these tweets was analyzed with the help of the Google Cloud Natural Language tool, and then validated by comparing their result with an external event. Finally, a prediction model was built using the Support Vector Machine classification algorithm. The results presented by this model are in line with the Efficient Market Hypothesis. According to it, the price of an asset efficiently reflects all information already published, so the tomorrow price of assets cannot be predicted using information from the past.

Palavras-chave: Machine Learning. Sentiment Analysis. Stock Exchange. Stocks. Tweets.

Lista de Figuras

[Figura 1 - Resultado dos artigos encontrados nas bases 19](#_Toc87311415)

[Figura 2 - Progressão da Revisão Sistemática 20](#_Toc87311416)

[Figura 3 - Comparativo entre os modelos utilizados por Bouktif, Fiaz e Awad (2019) 24](#_Toc87311417)

[Figura 4 - Comparativo dos modelos de predição utilizados por Pasupulety et al. (2019) 24](#_Toc87311418)

[Figura 5 - Comparação dos algoritmos utilizados em Bing, Chan e Ou (2014) 25](#_Toc87311419)

[Figura 6 - Comunidade de usuários do setor financeiro 31](#_Toc87311420)

[Figura 7 - Representação gráfica da metodologia 34](#_Toc87311421)

[Figura 8 - Comunidade de usuários do Twitter 37](#_Toc87311422)

[Figura 9 - Processo de extração de usuários e seguidores 39](#_Toc87311423)

[Figura 10 – Processo de extração do *tweets*. 41](#_Toc87311424)

[Figura 11 - Distribuição do tweets coletados e filtrados 42](#_Toc87311425)

[Figura 12 - Estrutura de um *tweet* 43](#_Toc87311426)

[Figura 13 - Estrutura do registro de *tweet* processado e importado 44](#_Toc87311427)

[Figura 14 - Exemplo da técnica de *Tokenization* 46](#_Toc87311428)

[Figura 15 - Exemplo de aplicação da técnica de *Named Entity Recognition* 47](#_Toc87311429)

[Figura 16 - Nuvem de palavras das entidades filtradas 48](#_Toc87311430)

[Figura 17 - Nuvem de palavras dos substantivos filtrados 48](#_Toc87311431)

[Figura 18 - Palavras e expressões utilizadas para limpeza de *tweets*. 50](#_Toc87311432)

[Figura 19 - Nuvem de palavras dos termos utilizados como critério de exclusão 51](#_Toc87311433)

[Figura 20 - Progresso de filtragem dos *tweets* 52](#_Toc87311434)

[Figura 21 - Distribuição final dos *tweets* coletados ao longo do tempo. 52](#_Toc87311435)

[Figura 22 - Hierarquia do aprendizado de máquina 55](#_Toc87311436)

[Figura 23 - Tabela de preços do *Cloud Natural Language* 60](#_Toc87311437)

[Figura 24 - Interpretação de valores do *Cloud Natural Language* 61](#_Toc87311438)

[Figura 25 - Histograma de distribuição do sentimento nos *tweets*. 62](#_Toc87311439)

[Figura 26 - Histograma de distribuição do sentimento nos tweets do trabalho de Yang, Mo e Liu (2015) 63](#_Toc87311440)

[Figura 27 - Análise de sentimentos do trabalho de Bollen, Mao e Zeng (2011) 64](#_Toc87311441)

[Figura 28 - Histograma de distribuição do sentimento nos *tweets* entre 16/03/2020 e 26/03/2020. 64](#_Toc87311442)

[Figura 29 - Média diária dos valores de sentimento de março de 2019. 65](#_Toc87311443)

[Figura 30 - Média diária dos valores de sentimento de março de 2020 65](#_Toc87311444)

[Figura 31 - Média diária dos valores de sentimento de março de 2021. 66](#_Toc87311445)

[Figura 32 - Exemplo de documento do *dataset* final 68](#_Toc87311446)

[Figura 33 - Resultado dos modelos de predição do trabalho de Bouktif, Fiaz e Awad (2019). 71](#_Toc87311447)

Lista de QUADROS

[Quadro 1 - Quadro de artigos selecionados para análise e discussão. 20](#_Toc87311448)

[Quadro 2 - Técnicas de ML utilizadas 21](#_Toc87311449)

[Quadro 3 - Algoritmos de ML utilizados 21](#_Toc87311450)

[Quadro 4 - Elemento em que a predição foi realizada 22](#_Toc87311451)

[Quadro 5 - Critérios utilizados para seleção das ações 22](#_Toc87311452)

[Quadro 6 - Fonte de dados utilizadas para análise de sentimentos 22](#_Toc87311453)

[Quadro 7 - Atributos utilizados na construção do modelo 22](#_Toc87311454)

[Quadro 8 - Métricas utilizadas na validação 23](#_Toc87311455)

[Quadro 9 - Operações de normalização dos *tweets* coletados 44](#_Toc87311456)

[Quadro 10 - Métricas de avaliação dos modelos utilizando o índice de sentimento 70](#_Toc87311457)

[Quadro 11 - Métricas de avaliação dos modelos desconsiderando o índice de sentimento 70](#_Toc87311458)

Lista de Abreviaturas e Siglas

|  |  |
| --- | --- |
| API | Application Programming Interface |
| B3 | Brasil Bolsa Balcão |
| CSV | Comma Separated Values |
| GPOMS | Google Profile of Mood States |
| LSTM | Long Short-Term Memory |
| ML | Machine Learning |
| MAE | Mean Absolute Error |
| MAPE | Mean Absolute Percentage Error |
| PLN | Processamento de Linguagem Natural |
| R² | R-Squared |
| RBF | Radial Basis Function |
| RMSE | Root Mean Square Error |
| SVM | Support Vector Machine |
| VXEWZ | CBOE Brazil Etf Volatility |

Sumário

[1 Introdução 11](#_Toc87311459)

[2 REVISÃO SISTEMÁTICA 14](#_Toc87311460)

[2.1 PROTOCOLO DA REVISÃO SISTEMÁTICA 14](#_Toc87311461)

[2.1.1 O Protocolo 14](#_Toc87311462)

[2.2 DESENVOLVIMENTO DA REVISÃO SISTEMÁTICA 18](#_Toc87311463)

[2.2.1 Pesquisa nas bases 18](#_Toc87311464)

[2.2.2 Fases de seleção 19](#_Toc87311465)

[2.2.3 Discussão dos resultados 20](#_Toc87311466)

[2.2.4 Análise dos artigos 23](#_Toc87311467)

[2.2.5 Análise crítica 28](#_Toc87311468)

[2.2.6 Pesquisa complementar 29](#_Toc87311469)

[2.2.7 Considerações finais 32](#_Toc87311470)

[3 AQUISIÇÃO e processamento DOS DADOS 34](#_Toc87311471)

[3.1 ARQUITETURA 34](#_Toc87311472)

[3.2 INTRODUÇÃO 35](#_Toc87311473)

[3.3 DEFINIÇÃO DOS USUÁRIOS 35](#_Toc87311474)

[3.4 COLETA DE DADOS - TWITTER 37](#_Toc87311475)

[3.4.1 Coletor de dados 38](#_Toc87311476)

[3.4.2 Análise e processamento dos dados 41](#_Toc87311477)

[3.4.3 Processamento textual dos *tweets* 43](#_Toc87311478)

[3.5 PROCESSAMENTO DOS TWEETS COLETADOS 45](#_Toc87311479)

[3.5.1 Construção do *corpus* de palavras-chaves 45](#_Toc87311480)

[3.5.2 Filtragem dos *tweets* coletados 49](#_Toc87311481)

[3.5.3 Análise dos *tweets* filtrados 51](#_Toc87311482)

[3.5.4 Limitações 53](#_Toc87311483)

[4 APLICAÇÃO DE *MACHINE LEARNING* NA IDENTIFICAÇÃO DO SENTIMENTO ASSOCIADO AOS *TWEETS* E PREÇO DE AÇÕES 54](#_Toc87311484)

[4.1 MACHINE LEARNING 54](#_Toc87311485)

[4.1.1 Aprendizagem Supervisionada 56](#_Toc87311486)

[4.1.2 Aprendizagem não supervisionada 56](#_Toc87311487)

[4.2 AQUISIÇÃO DOS DADOS FINANCEIROS 57](#_Toc87311488)

[4.3 APLICAÇÃO DE ANÁLISE DE SENTIMENTOS 58](#_Toc87311489)

[4.3.1 *Google Cloud Natural Language* 59](#_Toc87311490)

[4.3.2 Janela de análise 61](#_Toc87311491)

[4.3.3 Resultados 61](#_Toc87311492)

[4.4 MODELO DE PREDIÇÃO 66](#_Toc87311493)

[4.4.1 Seleção dos Atributos 67](#_Toc87311494)

[4.4.2 Pré-processamento dos dados 68](#_Toc87311495)

[4.4.3 Construção do modelo de predição 68](#_Toc87311496)

[4.4.4 Resultados e análise 70](#_Toc87311497)

[5 CONCLUSÃO 72](#_Toc87311498)

[Referências Bibliográficas 75](#_Toc87311499)

# Introdução

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, 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) e pode-se citar as redes sociais como uma das principais fontes de dados. Entre as principais podemos nomear: *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 americano.

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.

Este trabalho está dividido em cinco capítulos. No primeiro capítulo é apresentada uma introdução sobre o tema do trabalho. No segundo capítulo é apresentada uma revisão sistemática sobre o tema proposto e a discussão dos seus resultados. No terceiro capítulo é apresentado o procedimento de coleta de dados utilizado neste trabalho. O quarto capítulo aborda a utilização de técnicas de *Machine Learning* sobre os dados coletados. Por fim no quinto capítulo são apresentadas as conclusões deste trabalho.

# REVISÃO SISTEMÁTICA

Como forma de auxiliar a execução deste trabalho, optou-se por realizar uma revisão sistemática. Segundo Kitchenham (2007), uma revisão sistemática é uma forma de sintetizar, analisar e avaliar o conhecimento disponível sobre um determinado assunto ou área de pesquisa. A autora também acrescenta que um dos motivos para realização de uma revisão sistemática: a criação de uma base para realização de pesquisas e trabalhos posteriores.

Esta revisão sistemática tem como objetivo identificar na literatura as principais técnicas e o estado da arte na aplicação de *Machine Learning* na predição de bolsa de valores, levando em consideração o sentimento dos investidores, bem como auxiliar no levantamento de referencial teórico para a execução deste Trabalho de Conclusão de Curso.

## PROTOCOLO DA REVISÃO SISTEMÁTICA

O protocolo desta revisão sistemática foi desenvolvido com base no protocolo utilizado por Medeiros (2016).

### O Protocolo

O planejamento desta revisão sistemática obedece a seguinte estrutura:

**a)** Título

Revisão sistemática de técnicas de *Machine Learning* aplicadas à predição de bolsa de valores.

**b)** Resumo

Esta revisão sistemática foi elaborada com o objetivo de auxiliar e buscar referencial teórico para a execução deste trabalho de conclusão. Para isto, os trabalhos foram selecionados e analisados com o objetivo de verificar as principais técnicas e algoritmos de ML na predição de bolsa de valores.

**c)** Objetivo

O principal objetivo desta revisão sistemática é verificar na literatura atual as principais abordagens e técnicas de ML aplicadas na predição de bolsa de valores.

Para responder às questões de pesquisa desta revisão optou-se por utilizar a metodologia PICOC (população, intervenção, comparação, resultados e contexto) proposta por Kitchenham (2007), que se encontra descrita na seção seguinte.

#### Formulação da pesquisa

A formulação da pesquisa tem a seguinte estrutura:

**a)** Foco da questão:

Esta revisão sistemática tem como objetivo encontrar em artigos e trabalhos o estado da arte em relação a aplicação de técnicas ML na predição de bolsa de valores e análise de sentimento dos investidores.

**b)** Questões de interesse:

* Quais as técnicas e algoritmos de ML utilizados?
* Quais os atributos utilizados na elaboração do modelo de predição?
* Quais as métricas utilizadas para a validação dos modelos de predição?
* Como a análise de sentimento dos investidores foi aplicada no modelo de predição?

**c)** Palavras-chave:

*Machine Learning*. Análise preditiva. Análise de sentimentos. Bolsa de valores.

**d)** Intervenção:

Identificar técnicas de análise de predição e análise de sentimentos aplicadas na bolsa de valores.

**e)** Controle:

Não há.

**f)** Efeito:

Identificar oportunidades de pesquisa na área de ML com aplicação de predição na bolsa de valores.

**g)** Medida de resultado:

A partir da resposta às questões de interesse, espera-se que a revisão possa produzir embasamento e referencial teórico para auxiliar na execução do Trabalho de Conclusão de Curso, além de indicar algoritmos, atributos e métricas de validação a serem utilizados.

**h)** População de interesse:

Pesquisadores, professores, desenvolvedores e investidores de forma geral.

**i)** Aplicação:

Esta revisão sistemática irá gerar uma base de conhecimento que poderá ser utilizada posteriormente na execução de novas pesquisas e trabalhos.

**j)** Desenho do experimento:

Não será realizado.

**k)** Financiamento:

Sem financiamento.

#### Seleção das bases de dados

**a)** Definição dos critérios para seleção das fontes de dados:

As fontes de dados foram selecionadas com o auxílio do orientador desta pesquisa, bem como a sugestão de Kitchenham (2007) sobre as fontes de dados mais relevantes para execução de pesquisas relacionadas à área de ciências da computação. Foram selecionadas as bases *IEEE Xplore* e *ACM Digital Library*. A *IEEE Xplore* é um repositório de publicações científicas e hospeda mais de 5 milhões de publicações de autoria própria e de parceiros, ela abrange as áreas de engenharia elétrica, ciências da computação e eletrônica (IEEE Xplore, 2021). *A ACM Digital Library* possui o maior acervo digital de publicações da área de computação e tecnologia da informação, contando com diversos periódicos e aproximadamente 2,9 milhões de publicações (ACM Digital Library, 2021).

**b)** Idiomas das fontes de dados:

Serão consideradas as publicações que estiverem nos idiomas inglês e português.

**c)** *String* de busca:

Seguindo a recomendação de Kitchenham (2007), a *string* de busca foi desenvolvida levando em consideração as questões de interesse levantadas anteriormente.

*(("machine learning" OR “predictive analysis” OR “predictive analytics”) AND ("sentiment analysis" OR “opinion mining”) AND ("stock market" OR "stocks") AND (“social media” OR “twitter” OR “microblogging” OR “microblog”))*

**d)** Artigos de controle:

O trabalho (BING, CHAN e OU, 2014) foi encontrado em uma pesquisa pregressa realizada nas mesmas bases definidas neste protocolo. Este trabalho foi selecionado como artigo de controle pela quantidade relevante de citações.

#### Seleção dos estudos

**i.** Critérios para inclusão e exclusão dos resultados:

**a)** O ano de publicação do trabalho deve ser maior ou igual a 2018.

**b)** O artigo deve ter sido publicado em um *journal* ou periódico de congresso científico.

**c)** O artigo deve estar relacionado com todas as questões de interesse levantadas anteriormente.

**d)** O trabalho deve ter aplicado *Machine Learning* na predição de bolsa de valores, levando em consideração o sentimento dos investidores.

**e)** O trabalho deve ter informado ao menos uma técnica/algoritmo utilizado.

**ii.** Procedimentos para seleção dos trabalhos:

Primeiramente os trabalhos serão obtidos com a execução da *string* de busca nas bases de dados definidas. Para a etapa de execução e análise, os trabalhos serão exportados das bases e importados na ferramenta *StArt*, criada para auxiliar pesquisadores na execução de revisões sistemáticas. Esta ferramenta auxilia o pesquisador na execução das etapas da pesquisa, além de fornecer estatísticas e dados sobre ela (FABBRI et al., 2016). Após, serão executadas as etapas de seleção e extração dos critérios de qualidade dos trabalhos. As etapas estão descritas da seguinte maneira:

**iii.** Etapas para seleção dos trabalhos:

Etapa 1 – Leitura do título, palavras chaves e resumo

Etapa 2 – Leitura da introdução, metodologia e conclusão.

Etapa 3 – Leitura completa do artigo.

**iv.** Critérios de qualidade dos trabalhos:

* Quais as técnicas de ML utilizadas?
* Quais os algoritmos utilizados?
* Quais os atributos utilizados na construção do modelo?
* Quais as métricas utilizadas na validação do modelo?
* A predição foi aplicada sobre uma ação, setor de ações ou índice de ações?
* Qual o critério utilizado para selecionar uma ação, setor ou índice de ações?
* Qual a fonte de dados utilizada para extrair o sentimento dos investidores?

## DESENVOLVIMENTO DA REVISÃO SISTEMÁTICA

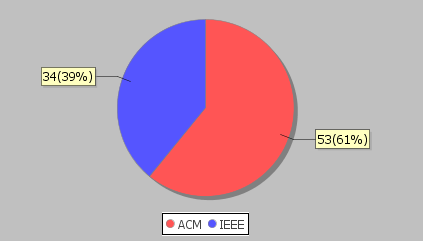
Após a elaboração do protocolo da revisão sistemática, iniciou-se a fase de desenvolvimento. Para auxiliar na execução da revisão sistemática optou-se por utilizar o software *StArt*. Essa ferramenta permite a execução da revisão de forma organizada e sistematizada, diminuindo assim o viés do pesquisador e possíveis erros em sua condução. Ela abrange todas as etapas da pesquisa, desde a criação do protocolo até a seleção e avaliação dos estudos, além disso é possível gerar gráficos e visualizar informações sobre o andamento do trabalho (FABBRI et al., 2016).

### Pesquisa nas bases

A pesquisa foi executada no dia 02 de maio de 2021, utilizando a *string* de busca definida no protocolo. Ambas as bases de estudo fornecem ferramentas para aplicação de filtros avançados e utilizando essa ferramenta foi possível aplicar os critérios de inclusão relacionados ao ano e tipo de publicação diretamente na pesquisa.

Na base de dados *ACM Digital Library* foram encontrados 53 trabalhos e na base de dados *IEEE Xplore* foram encontrados 34 trabalhos. Posteriormente os trabalhos foram exportados para o formato *BibTex* e então importados na ferramenta *StArt*. A Figura 1 ilustra os resultados de ambas as bases. Juntamente com os trabalhos encontrados, foi incluído o artigo de controle, que estava presente em ambas as bases.

Figura 1 - Resultado dos artigos encontrados nas bases



Fonte: elaborado pelo autor

### Fases de seleção

Após a pesquisa dos trabalhos se iniciou a fase de seleção, onde foram avaliados os critérios de inclusão e exclusão, além da etapa de extração dos critérios de qualidade. Segue abaixo detalhamento das fases de seleção.

#### Primeira etapa de seleção

A verificação de artigos duplicados foi executada com auxílio da ferramenta *StArt* e nenhum trabalho foi removido por este critério. Em seguida iniciou-se a leitura do título, palavras-chave e *abstract* dos trabalhos selecionados. Nesta etapa, 62 trabalhos foram rejeitados, sendo a maioria rejeitado por serem trabalhos não relacionados com predição de bolsa de valores, sendo assim não relacionados com as questões de interesse desta pesquisa. No total 25 trabalhos foram aprovados para a etapa seguinte.

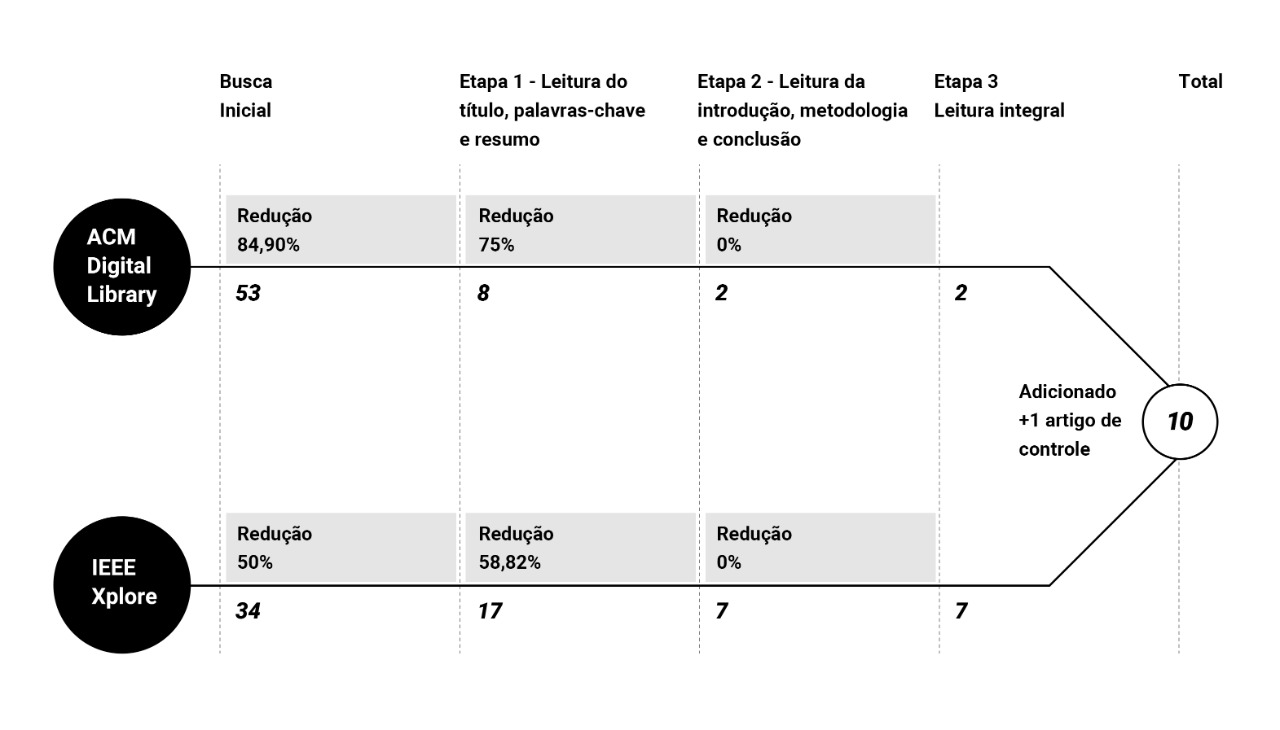
#### Segunda etapa de seleção

Nesta etapa foi realizada a leitura da introdução, metodologia e conclusão dos trabalhos selecionados. Novamente os critérios de inclusão e exclusão foram validados para verificar se os trabalhos possuem relação com as questões de interesse levantadas anteriormente. O principal objetivo desta etapa foi filtrar trabalhos que tivessem informado ao menos uma técnica ou algoritmo de ML. Nesta fase foram rejeitados 16 trabalhos, sendo 9 aprovados para a próxima fase. Um dos trabalhos foi rejeitado pois não estava disponível para leitura integral.

#### Terceira etapa de seleção

Na última etapa de seleção os artigos foram lidos de forma integral, nenhum trabalho foi removido nesta etapa. Durante esta etapa também foi realizada a extração dos critérios de qualidade definidos anteriormente, o artigo de controle foi adicionado neste momento, totalizando 10 trabalhos para realizar a extração dos dados. As informações extraídas foram catalogadas na ferramenta *StArt*. Ao final iniciou-se a análise dos resultados obtidos com a extração dos dados. A Figura 2 ilustra o progresso do processo de seleção da revisão sistemática.

Figura 2 - Progressão da Revisão Sistemática



Fonte: elaborado pelo autor

### Discussão dos resultados

Após a leitura e extração das informações dos artigos lidos de forma integral, deu-se início a análise dos resultados. Os resultados são debatidos nesta seção, através das perguntas definidas no protocolo da revisão sistemática. Abaixo, no Quadro 1, encontram-se os artigos que foram analisados de forma integral.

Quadro 1 - Quadro de artigos selecionados para análise e discussão.

|  |  |
| --- | --- |
| **Ano** | **Título** |
| 2014 | Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements |
| 2018 | Machine Learning Approach to Extracting Emotions Information from Open Source Data for Relative Forecasting of Stock Prices |
| 2018 | Stock Market Analysis Using Social Networks |
| 2019 | Prediction of Stock Prices with Sentiment Fusion and SVM Granger Causality |
| 2019 | Stock Market Prediction Using Machine Learning Techniques |
| 2019 | Stock Market Movement Prediction using Disparate Text Features with Machine Learning |
| 2019 | Predicting Stock Prices using Ensemble Learning and Sentiment Analysis |
| 2019 | The Influence of Tweets and News on the Brazilian Stock Market through Sentiment Analysis |
| 2020 | SENN: Stock Ensemble-based Neural Network for Stock Market Prediction using Historical Stock Data and Sentiment Analysis |
| 2021 | Predicting Stock Closing Price After COVID-19 Based on Sentiment Analysis and LSTM |

Fonte: elaborado pelo autor

#### Perguntas respondidas

Nesta seção se encontram os Quadros 2, 3, 4, 5, 6, 7 e 8 que respondem às perguntas em relação aos critérios de qualidade dos trabalhos.

1. Quais as técnicas de ML utilizadas?

Quadro 2 - Técnicas de ML utilizadas

|  |  |
| --- | --- |
| **Técnica** | **Artigos** |
| Classification | 6 |
| Ensemble | 5 |
| Recurrent Neural Network | 3 |
| Perceptron | 1 |

Fonte: elaborado pelo autor

2. Quais os algoritmos utilizados?

Quadro 3 - Algoritmos de ML utilizados

|  |  |
| --- | --- |
| **Algoritmo** | **Artigos** |
| Support Vector Machine (SVM) | 6 |
| Random Forest | 3 |
| LSTM | 3 |
| Logistic Regression | 2 |
| Naive Bayes | 1 |
| Naive Bayes Bernoulli | 1 |
| C4.5 | 1 |
| Multilayer Perceptron | 1 |
| Extremely Randomized Tree | 1 |
| XGBoost | 1 |
| Stacking | 1 |
| Algoritmo próprio | 1 |

Fonte: elaborado pelo autor

3. A predição foi aplicada sobre uma ação, setor de ações ou índice de ações?

Quadro 4 - Elemento em que a predição foi realizada

|  |  |
| --- | --- |
| **Resposta** | **Artigos** |
| Ação | 7 |
| Índice | 2 |
| Setores | 1 |

Fonte: elaborado pelo autor

4. Qual o critério utilizado para selecionar uma ação, setor ou índice de ações?

Quadro 5 - Critérios utilizados para seleção das ações

|  |  |
| --- | --- |
| **Critério** | **Artigos** |
| Não informado | 7 |
| Volume | 2 |
| Volatilidade | 1 |

Fonte: elaborado pelo autor

5. Qual a fonte de dados utilizada para extrair o sentimento dos investidores?

Quadro 6 - Fonte de dados utilizadas para análise de sentimentos

|  |  |
| --- | --- |
| **Fonte** | **Artigos** |
| Twitter | 9 |
| StockTwits | 2 |
| Notícias | 2 |
| Search Engine Query | 1 |

Fonte: elaborado pelo autor

6. Quais os atributos utilizados na construção do modelo?

Quadro 7 - Atributos utilizados na construção do modelo

|  |  |
| --- | --- |
| **Atributo** | **Artigos** |
| Data | 10 |
| Preço de fechamento | 10 |
| Índice de sentimento | 10 |
| Preço de abertura | 6 |
| Preço mínimo do dia | 5 |
| Preço máximo do dia | 5 |
| Volume | 5 |
| *21-day rate of change* | 1 |
| *Stochastic oscillator* | 1 |
| *Channel commodity index* | 1 |
| *Moving average convergence divergence* | 1 |

Fonte: elaborado pelo autor

7. Quais as métricas utilizadas na validação do modelo?

Quadro 8 - Métricas utilizadas na validação

|  |  |
| --- | --- |
| **Métrica** | **Artigos** |
| Accuracy | 5 |
| Mean Absolute Percentage Error | 4 |
| F1-Score | 3 |
| Root Mean Square Error | 3 |
| R-Squared | 2 |
| Precision | 2 |
| Recall | 2 |
| K-fold cross validation | 2 |
| Mean Absolute Error | 1 |

Fonte: elaborado pelo autor

### Análise dos artigos

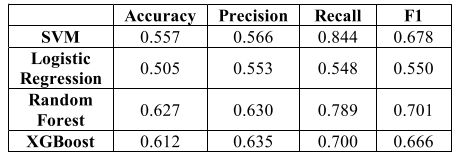
A partir da análise dos critérios de qualidade e das perguntas respondidas na seção anterior, foi realizada uma discussão detalhada sobre os artigos. A discussão será apresentada nesta seção.

#### Algoritmos e técnicas

Conforme critério de inclusão especificado no protocolo, todos os trabalhos informaram ao menos um algoritmo/técnica de ML aplicado na execução dos mesmos. Em relação a técnica escolhida, houve uma preferência por *Classification*, aplicada em 6 trabalhos (BOUKTIF, FIAZ e AWAD, 2019) (PASUPULETY, ANEES, *et al.*, 2019) (PARK, MA e LEUNG, 2019) (BATHIA, HAGRAS e LEPLEY, 2018) (BING, CHAN e OU, 2014) (LI, YANG, *et al.*, 2018). Quanto aos algoritmos, o mais utilizado foi SVM, sendo aplicado em todos os 6 trabalhos que utilizaram a técnica de *Classification*, correspondendo assim a aproximadamente 55% dos trabalhos analisados. Na amostragem geral, 5 trabalhos (BOUKTIF, FIAZ e AWAD, 2019) (PASUPULETY, ANEES, *et al.*, 2019) (BATHIA, HAGRAS e LEPLEY, 2018) (BING, CHAN e OU, 2014) (LI, YANG, *et al.*, 2018) aplicaram mais de um algoritmo/técnica.

Os autores Bouktif, Fiaz e Awad (2019) implementaram os algoritmos SVM, *Logistic Regression, Random Forest e XGBoost*. Conforme a Figura 3, os algoritmos *Random Forest* e *XGBoost*, ambos do tipo *Ensemble,* tiveram uma acurácia superior aos do tipo *Classification* e, entre eles, *Random Forest* obteve a melhor acurácia.

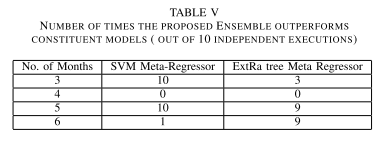
Figura 3 - Comparativo entre os modelos utilizados por Bouktif, Fiaz e Awad (2019)



Fonte: Imagem adaptada de (BOUKTIF, FIAZ e AWAD, 2019)

Pasupulety et al*.* (2019) também realizou comparações utilizando mais de um algoritmo em seu trabalho. Os autores inicialmente desenvolveram dois modelos utilizando os algoritmos SVM e *Extremely Randomized Tree*, além de utilizarem um algoritmo de *Stacking* para combinar ambos os modelos desenvolvidos anteriormente. De acordo com os autores (Figura 4), em um *dataset* com dados de 5 meses, o algoritmo de *Stacking* obteve melhor desempenho, e nos outros cenários, SVM geralmente obteve a melhor *perfomance*.

Figura 4 - Comparativo dos modelos de predição utilizados por Pasupulety et al. (2019)



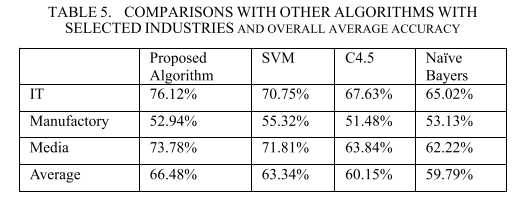
Fonte: Imagem adaptada de Pasupulety et al*.* (2019)

O trabalho de (BATHIA, HAGRAS e LEPLEY, 2018) também utilizou mais de um algoritmo na construção de seus modelos de predição, os autores utilizaram *Random Forest,* *Logistic Regression* e SVM, porém não detalharam qual algoritmo teve a melhor precisão, apenas realizaram comparações para verificar se a adição de um índice de emoções aumentava a precisão dos modelos. Os autores concluíram que em alguns ativos financeiros é possível melhorar a predição utilizando um índice de emoções dos investidores.

Os autores Li et al*.* (2018), utilizam diferentes algoritmos para as etapas de seu trabalho. Na primeira etapa foi realizada a análise de sentimentos utilizando os algoritmos *Naive Bayes* *Bernoulli* e SVM, o algoritmo SVM obteve o melhor desempenho com uma taxa de acuracia de 85,34%. Na segunda etapa foram utilizados os algoritmos *Random Forest* e SVM para predição de bolsa de valores utilizando apenas as séries históricas das ações, nesta etapa o algoritmo SVM obteve melhor desempenho em 3 cenários diferentes, com taxa de acuracia de aproximandente 54%. Ná ultima etapa para realizar a predição com o índice de sentimento dos investidores, apenas o algoritmo SVM foi utilizado. Nesta etapa a taxa de acuracia foi de 56,5%

Em outro trabalho, os autores Bing, Chan e Ou (2014), propuseram um algoritmo próprio, além de avaliar o desempenho dos algoritmos SVM, C4.5 e *Naive Bayes*. O algoritmo desenvolvido pelos autores leva em consideração as características dos produtos e serviços ofertados pelas empresas, para então calcular o índice de sentimento. Cada empresa possui um número de *features* relacionadas e organizadas em uma estrutura de multiplos níveis. Conforme a Figura 6, o algoritmo proposto pelos autores obteve a melhor precisão.

Figura 5 - Comparação dos algoritmos utilizados em Bing, Chan e Ou (2014)



Fonte: Imagem adaptada de Bing, Chan e Ou (2014)

#### Escolha das ações, setores ou índice de ações

Todos os trabalhos informaram ao menos uma ação, setor ou índice de ações em que foi aplicada a predição, porém apenas 3 trabalhos (OWEN e OKTARIANI, 2020) (PASUPULETY, ANEES, *et al.*, 2019) (BATHIA, HAGRAS e LEPLEY, 2018) informaram os critérios utilizados para a escolha. Os autores justificaram a escolha dos critérios relacionando a volatilidade e o volume de negociações com uma possível maior quantidade de dados para realizar a extração dos sentimentos dos investidores.

Notou-se uma preferência pelo setor de tecnologia, apesar de não justificarem a escolha, 6 trabalhos (LI, YANG, *et al.*, 2018) (BATHIA, HAGRAS e LEPLEY, 2018) (CHOU, PARK e CHOU, 2021) (PARK, MA e LEUNG, 2019) (PASUPULETY, ANEES, *et al.*, 2019) (BOUKTIF, FIAZ e AWAD, 2019) utilizaram ações deste setor. Apenas Bing, Chan e Ou (2014) realizaram a predição em setores de ações, e constaram que os algoritmos utilizados tiveram uma melhor acurácia no setor de tecnologia.

#### Fonte de dados utilizada para análise dos sentimentos dos investidores

O *Twitter* foi a fonte de dados mais citada, sendo utilizada em 9 trabalhos (BOUKTIF, FIAZ e AWAD, 2019) (PASUPULETY, ANEES, *et al.*, 2019) (PARK, MA e LEUNG, 2019) (SIRIMEVAN, MAMALGAHA, *et al.*, 2019) (CHOU, PARK e CHOU, 2021) (BATHIA, HAGRAS e LEPLEY, 2018) (BING, CHAN e OU, 2014) (CAROSIA, COELHO e SILVA, 2019) (LI, YANG, *et al.*, 2018). Destes, 3 trabalhos (CAROSIA, COELHO e SILVA, 2019) (CHOU, PARK e CHOU, 2021) (SIRIMEVAN, MAMALGAHA, *et al.*, 2019) utilizaram mais de uma fonte. Dois trabalhos (SIRIMEVAN, MAMALGAHA, *et al.*, 2019) (LI, YANG, *et al.*, 2018) citaram a utilização de *crawlers* e outras bibliotecas para contornar as limitações de requisições impostas na API oficial do *Twitter.*

Os autores Carosia, Coelho e Silva (2019) analisaram a literatura e identificaram que a maioria dos trabalhos utilizam notícias ou *tweets* para construção de modelos de predição com análise de sentimentos, então em seu trabalho propuseram um comparativo entre ambas as fontes. De acordo com os autores, ao utilizar os atributos de preço de abertura e volume de transações, notícias apresentam uma melhor acurácia para realizar a predição do dia seguinte, enquanto ao utilizar o preço de fechamento, *tweets* tem uma acurácia superior.

Os pesquisadores Chou, Park e Chou (2021), coletaram informações de duas fontes, *Twitter* e *StockTwits*. *StockTwits* é uma plataforma de *microblogging* similar ao *Twitter*, porém focada apenas em comentários relacionados ao mercado financeiro. As publicações do *StockTwits* podem ser identificadas pelo usuário com um sentimento, indicando se a postagem tem carater positivo ou negativo. Os autores utilizaram as postagens identificadas do *StockTwits* para criar um modelo de classificação de sentimentos*,* este modelo foi então aplicado para nomear o sentimento das postagens do *Twitter*.

#### Atributos utilizados na construção do modelo

De forma geral, os trabalhos utilizaram atributos relacionados com o preço e movimento de mercado das ações. O preço de fechamento e a data foram considerados em todos os trabalhos, enquanto 6 trabalhos (LI, YANG, *et al.*, 2018) (CAROSIA, COELHO e SILVA, 2019) (CHOU, PARK e CHOU, 2021) (PASUPULETY, ANEES, *et al.*, 2019) (OWEN e OKTARIANI, 2020) (BOUKTIF, FIAZ e AWAD, 2019) também utilizaram o preço de abertura, 5 trabalhos (BOUKTIF, FIAZ e AWAD, 2019) (OWEN e OKTARIANI, 2020) (CHOU, PARK e CHOU, 2021) (LI, YANG, *et al.*, 2018) (PASUPULETY, ANEES, *et al.*, 2019) utilizaram o volume de transações, preço máximo durante o dia e preço mínimo durante o dia. Estes atributos estão nas informações que são normalmente divulgadas periodicamente pelas bolsas de valores, e estão disponíveis de forma pública.

Um trabalho (PASUPULETY, ANEES, *et al.*, 2019) além de utilizar as informações básicas, adicionou 4 indicadores técnicos (*21-day rate of change, Stochastic oscillator, Channel commodity index, Moving average convergence divergence*), que são obtidos através de cálculos envolvendo os outros atributos. Os autores utilizaram o algoritmo *Extremely Randomized Tree* para identificar quais os atributos que tiveram o maior peso na predição do preço das ações, estes atributos foram utilizados na criação do modelo final.

Todos os trabalhos analisados implementaram de alguma forma um índice de sentimento que é calculado de forma diária ou por hora, 5 destes trabalhos (BOUKTIF, FIAZ e AWAD, 2019) (OWEN e OKTARIANI, 2020) (PASUPULETY, ANEES, *et al.*, 2019) (PARK, MA e LEUNG, 2019) (CAROSIA, COELHO e SILVA, 2019), com o objetivo de verificar o impacto dos sentimentos ao longo do tempo, aplicaram atrasos na relação entre o sentimento e o preço da ação. Um trabalho (PARK, MA e LEUNG, 2019) utilizou a quantidade de *retweets* e seguidores dos autores da postagem para adicionar peso na hora de definir índice de sentimento do *tweet*.

#### Métricas utilizadas na validação do modelo

Quanto as validações, foi observado a utilização de diversas métricas de validação. A acurácia foi a validação mais utilizada, sendo mencionada em 5 trabalhos (BOUKTIF, FIAZ e AWAD, 2019) (BATHIA, HAGRAS e LEPLEY, 2018) (BING, CHAN e OU, 2014) (CAROSIA, COELHO e SILVA, 2019) (LI, YANG, *et al.*, 2018). Seis trabalhos (BOUKTIF, FIAZ e AWAD, 2019) (OWEN e OKTARIANI, 2020) (PASUPULETY, ANEES, *et al.*, 2019) (CHOU, PARK e CHOU, 2021) (BING, CHAN e OU, 2014) (CAROSIA, COELHO e SILVA, 2019) utilizaram mais de uma métrica de validação.

Entretanto notou-se uma preferência por métricas utilizadas em modelos envolvendo séries temporais, que coincidem com a predição de bolsa de valores. A métrica MAPE foi a segunda mais utilizada, sendo citada em 4 trabalhos (OWEN e OKTARIANI, 2020) (PARK, MA e LEUNG, 2019) (CHOU, PARK e CHOU, 2021) (BING, CHAN e OU, 2014). Já a métrica RMSE foi citada em 3 trabalhos (PASUPULETY, ANEES, *et al.*, 2019) (SIRIMEVAN, MAMALGAHA, *et al.*, 2019) (CHOU, PARK e CHOU, 2021)

### Análise crítica

Ao executar esta pesquisa notou-se que o número de trabalhos subiu conforme o passar dos anos, mostrando um interesse crescente pelo assunto. Os pesquisadores de 4 trabalhos (PASUPULETY, ANEES, *et al.*, 2019) (CAROSIA, COELHO e SILVA, 2019) (BING, CHAN e OU, 2014) (LI, YANG, *et al.*, 2018) citaram como uma das motivações para a execução de seus trabalhos, a hipotese do mercado eficiente, que indica que o mercado financeiro não pode ser previsto através do estudo e o mesmo é afetado por diversos fatores externos (como por exemplo, o sentimento dos investidores).

Como ponto de início verificou-se que todos os trabalhos iniciaram de forma semelhante, utilizando técnicas de *data mining* para obtenção de dados que expressem a opinião dos investidores, estes dados então foram pré-processados e um sentimento positivo ou negativo foi extraído. Como resultado desta etapa, todos os trabalhos criaram um índice de sentimentos temporal, para que possa então ser levado em consideração no modelo de predição.

Quanto a construção dos modelos de análise de sentimentos, apenas um trabalho (CAROSIA, COELHO e SILVA, 2019) utilizou um *dataset* elaborado e revisado por profissionais do setor financeiro. Um trabalho (CHOU, PARK e CHOU, 2021) criou o modelo de análise de sentimentos a partir de dados com a opinião já rotulada, estes dados foram extraídos da rede social *StockTwits*, que é uma rede social específica para publicação de comentários relacionados ao mercado financeiro.

De forma geral os modelos de predição foram construídos de forma parecida, utilizando informações relacionadas ao preço histórico das ações e o índice de sentimentos criado na etapa anterior. Um trabalho (PASUPULETY, ANEES, *et al.*, 2019) explorou a utilização de indicadores técnicos, porém não obteve melhora significativa em seu modelo de predição. Os trabalhos (BATHIA, HAGRAS e LEPLEY, 2018) (CHOU, PARK e CHOU, 2021) avaliaram a utilização do índice de sentimentos contra modelos contendo apenas os preços históricos.

Bathia, Hagras e Lepley (2018) realizaram testes em 5 cenários com ações diferentes para avaliar o impacto da utilização da análise de sentimentos na predição Em um cenário houve um aumento significativo de aproximadamente 26% na acurácia, enquanto nos demais o impacto foi menor, variando entre 2,07% a 7,93%. Os autores avaliam estes resultados menos significativos como uma possível consequência da limitação de informações extraídas para criação do índice de sentimentos.

Chou, Park e Chou (2021) também avaliaram o impacto do índice de sentimentos no modelo de predição. Os autores criaram 3 modelos utilizando o algoritmo *Long short-term memory* (LSTM), o primeiro modelo com apenas os preços históricos, o segundo modelo utilizando o índice de sentimentos, e o terceiro modelo utilizando o índice de sentimentos com peso e a técnica de *attention mechanism*. O terceiro modelo obteve os melhores resultados, de acordo com os autores, utilizando o algoritmo LSTM foi possível identificar que o sentimento de um dia esta altamente correlacionado com o dia seguinte.

Houve uma utilização mista de métricas para validar os modelos, porém notou-se que as métricas mais citadas são comumente utilizadas em problemas envolvendo séries temporais. Chou, Park e Chou (2021) fizeram o uso das métricas MAE, RMSE, MAPE e R² no modelo de predição de bolsa de valores, optando por utilizar apenas F1-Score na análise de sentimentos. Em outro trabalho, os autores Owen e Oktariani (2020) utilizaram *k-fold cross-validation* no modelo de análise de sentimentos e MAPE no modelo de predição de bolsa de valores.

Quanto a conclusão, os trabalhos no geral apresentaram resultados positivos dentro de suas propostas, indicando uma correlação entre o sentimento dos investidores e o preço das ações. Apenas um trabalho (PASUPULETY, ANEES, *et al.*, 2019) citou ter verificado um impacto insignificante na utilização da análise de sentimentos.

### Pesquisa complementar

Como forma de complementar lacunas encontradas durante a revisão sistemática, dois trabalhos foram revisados para serem utilizados no auxílio do desenvolvimento da metodologia. O trabalho de (BOLLEN, MAO e ZENG, 2011) foi escolhido pelo alto número de citações na revisão sistemática, sendo considerado por alguns autores (LI, YANG, *et al.*, 2018) como sendo um dos primeiros trabalhos significativos a avaliar a utilização de *Machine Learning* e *tweets* na predição do mercado financeiro.

Durante a avaliação dos resultados da revisão sistemática, verificou-se que a maioria dos trabalhos utilizou uma ferramenta própria do *Twitter* para filtrar *tweets* que façam menção a empresas negociadas na bolsa de valores, porém a ferramenta está disponível apenas para empresas presentes na bolsa de valores dos Estados Unidos. Sendo assim, se fez necessário uma pesquisa adicional para desenvolver uma estratégia de busca de t*weets* que possa ser replicada neste trabalho. O trabalho de (YANG, MO e LIU, 2015) propos uma metodologia para criação de comunidades de usuários do setor financeiro no *Twitter,* que será discutida neste capítulo.

#### Bollen, Mao e Zeng (2011)

O trabalho de Bollen, Mao e Zeng (2011) aborda a relação entre o humor dos usuários e sua relação com indicadores econômicos. O trabalho investiga se a análise do humor das massas, expressada através de postagens diárias na rede social *Twitter,* está correlacionada ao índice *Dow Jones Industrial Average* da bolsa de valores americana.

Os pesquisadores coletaram aproximadamente 10 milhões de *tweets* em um período de 10 meses no ano de 2009. Os *tweets* foram agrupados por dia e filtrados para conter apenas expressões que indiquem algum estado de humor e/ou emoção do usuário, como por exemplo: *“i feel”, “i am feeling”, “i dont feel”* e etc. Os resultados foram pré-processados para remoção de *stop words* e pontuação.

A partir dos *tweets* filtrados, os pesquisadores iniciaram a etapa de extração dos sentimentos dos usuários. A análise de sentimentos foi realizada através da construção de um índice diário, com o auxílio de duas ferramentas: *OpinionFinder* e *Google Profile of Mood States* (GPOMS). A ferramenta OpinionFinder utiliza um dicionário léxico para avaliar a subjetividade (positivo ou negativo) das expressões, está ferramenta já foi utilizada em outras coleções de *tweets* e obteve resultados satisfatórios. Diferente de outras ferramentas que realizam análise de sentimentos de forma binária, isto é, positivo e negativo, GPOMS realiza a análise de sentimentos através de uma matriz de 6 dimensões com diferentes emoções: *calm, alert, sure, vital, kind* e *happy.* Como forma de validar os resultados da análise de sentimentos, um período de testes de dois meses foi avaliado, este período foi selecionado especificamente por envolver dois grandes eventos que costumam gerar diversas opiniões dos usuários: a eleição presidencial e o dia de ações de graça americano.

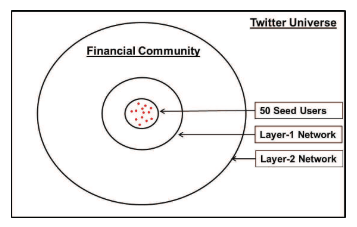
Para avaliar a influência dos sentimentos, os autores utilizaram o teste de causalidade de Granger, para verificar se a série temporal do humor dos usuários poderia ser utilizada na predição do índice *Dow Jones Industrial Average.* Os pesquisadores chegaram à conclusão de que mudanças no humor dos usuários podem estar correlacionadas com movimentos na bolsa de valores. O teste de Granger mostrou que de todas as emoções analisadas pela ferramenta GPOMS, a emoção que representa a “calma” dos usuários se mostrou mais preditiva, já na análise do *OpinionFinder*, as polaridades positivas e negativas obtiveram resultados similares. Entre as possibilidades de continuação de seu trabalho, os autores sugeriram a avaliação de apenas um subconjunto de usuários do Twitter.

#### YANG, MO e LIU (2015)

O trabalho de Yang, Mo e Liu (2015) analisa o efeito e o poder preditivo de uma comunidade de usuários no mercado de ações. Os autores sugerem que o movimento do mercado financeiro está relacionado ao humor dos usuários, porém nem todo usuário possui o mesmo poder de influência. Para isto o trabalho buscou identificar e classificar uma comunidade de usuários relacionados ao setor financeiro, identificando os principais influenciadores desta comunidade para verificar se eles possuem um maior poder de predição.

Os autores iniciaram o trabalho definindo a raiz da comunidade de usuários, que será composta por pessoas que são consideradas referência no setor de investimentos. No total foram selecionados 50 usuários através da análise de notícias e artigos que compilavam as contas mais influentes no mercado americano, estes usuários foram classificados como a raiz da rede. A partir destes usuários, a comunidade foi expandida em mais duas camadas, utilizando os perfis dos seguidores de cada camada para criar a próxima. A Figura 6 ilustra este processo.

Figura 6 - Comunidade de usuários do setor financeiro



Fonte: Imagem adaptada de Yang, Mo e Liu (2015).

Como forma de otimizar os resultados da comunidade criada e extrair apenas os usuários mais relevantes, alguns filtros foram aplicados. Primeiro, foram selecionados apenas os usuários que utilizam o idioma inglês e estejam localizados nos Estados Unidos ou Reino Unido. O segundo critério para filtragem foi localizar apenas usuários que tenham interesse no setor financeiro, para isto um compilado de palavras chaves relacionadas ao setor financeiro foi criado, através da análise das mensagens das 50 contas principais da comunidade. Estas palavras chaves foram agrupadas em tópicos e então utilizadas para filtrar os usuários das camadas externas em que as 20 mensagens mais recentes não contenham nenhum destes tópicos. No total aproximadamente 154 mil usuários foram selecionados.

Para verificar os principais mais influentes desta comunidade, os autores utilizaram uma representação em grafos e então utilizaram medidas de centralidade para identificar os nodos mais influentes desta comunidade. No total foram selecionados 2500 dos usuários mais influentes e a partir das postagens destes usuários, foi criado um índice diário de análise de sentimentos, a centralidade do usuário foi utilizada como peso no cálculo da polaridade do sentimento. Por fim os autores criaram um modelo de regressão para verificar o peso do sentimento na predição da bolsa de valores.

Durante a análise dos resultados os pesquisadores verificaram que entre as medidas de centralidade utilizadas para selecionar os usuários, a medida de centralidade de intermediação obteve o melhor resultado no modelo de predição. Também foram realizados testes para verificar o impacto da quantidade de usuários na predição, verificou-se o intervalo de 100 até 2500 dos usuários mais influentes, os melhores resultados foram obtidos utilizando os 200 usuários mais influentes. Por fim os autores concluíram que é possível verificar uma forte relação entre os movimentos do mercado financeiro e uma comunidade de usuários relacionada. Também foi possível verificar que o sentimento negativo desta comunidade possui uma maior taxa de acerto na predição.

### Considerações finais

Foi realizada uma revisão sistemática buscando encontrar na literatura os algoritmos e técnicas de ML mais utilizados na predição de bolsa de valores com aplicação de análise de sentimento em relação às opiniões dos investidores. No total foram encontrados 87 trabalhos nas bases *ACM Digital Library* e *IEEE Xplore*, após as fases de validação dos requisitos de inclusão e exclusão, 10 trabalhos foram lidos e analisados de forma integral.

Os trabalhos foram analisados com o objetivo de verificar o estado da arte em relação à utilização de algoritmos e técnicas de ML na predição de bolsa de valores, bem como extração de outras conhecimentos que serão úteis para a realização deste trabalho de conclusão. O algoritmo *Support Vector Machine* foi o mais utilizado e *Classification* foi a técnica mais citada. Os atributos utilizados na construção do modelo foram analisados e constatou-se que todos os trabalhos, além de utilizar o preço histórico das ações, criaram um índice temporal com o sentimento dos investidores. Quanto as métricas de validação, a acurácia dos modelos foi a mais citada, seguida por métricas comumente utilizadas em problemas envolvendo séries temporais, como por exemplo *Mean Absolute Percentage Error.*

Após a execução da revisão sistemática verificou-se uma lacuna em relação a uma estratégia de busca de *tweets* que possa ser replicada neste trabalho, portanto, optou-se por realizar uma pesquisa complementar. Além disso, esta pesquisa avaliou de forma individual um trabalho com alto número de citações entre os trabalhos encontrados na revisão sistemática.

Com a execução desta revisão sistemática foi possível verificar o estado da arte em relação a utilização de ML na predição de bolsa de valores, o conhecimento adquirido será utilizado para nortear a execução deste trabalho de conclusão.

# AQUISIÇÃO e processamento DOS DADOS

Neste capítulo será detalhada a metodologia aplicada na extração e processamento de dados, a arquitetura proposta, e as ferramentas utilizadas nesta etapa. A metodologia foi construída com o apoio das informações coletadas durante a execução da revisão sistemática.

## ARQUITETURA

O objetivo deste trabalho é a aplicação de técnicas de *Machine Learning* (ML) 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. Para se chegar neste objetivo, optou-se por replicar em parte a metodologia utilizada no trabalho de Yang, Mo e Liu (2015), que consiste em primeiro construir uma comunidade de usuários que tenham relação com o domínio financeiro para então construir um modelo de predição utilizando os *tweets* dos usuários desta comunidade. A Figura 7, representa este processo de forma geral e simplificada.

Figura 7 - Representação gráfica da metodologia

Diagrama

Descrição gerada automaticamente

Fonte: elaborado pelo autor.

O ambiente utilizado para desenvolvimento e testes possui a seguinte configuração: sistema operacional Windows 10, processador AMD Ryzen 5 2600 @ 3.4GHz e 16,0 GB de memória RAM. Quanto aos softwares utilizados, as ferramentas e scripts foram desenvolvidas nas linguagens de programação C#, utilizando a versão 4.7.2 do .NET Framework, e na linguagem de programação Python 3.9.6. Os dados coletados foram armazenados e manipulados em um banco de dados MongoDB 5.0.3.

## INTRODUÇÃO

O trabalho de Bollen, Mao e Zeng (2011) avaliou a utilização do humor do Twitter como forma de avaliar se o sentimento das massas está correlacionado com as movimentações do mercado financeiro. Os autores obtiveram conclusões positivas quanto as suas propostas, identificando uma correlação entre os *tweets* e o mercado financeiro. Bollen, Mao e Zeng (2011) também sugeriram que o idioma e a localização dos usuários poderia influenciar nos resultados, além de propor que a identificação e análise de uma comunidade, ou segmento especifico de usuários poderia melhorar os resultados.

Yang, Mo e Liu (2015) testaram esta hipótese em seu trabalho. Os autores propuseram a criação de uma comunidade de usuários que estejam relacionados ou tenham interesse no setor financeiro, para então extrair apenas *tweets* dos perfis identificados. Os autores chegaram a conclusão que há uma forte correlação entre o humor dos usuários desta comunidade e o preço de ativos do mercado financeiro, além de identificar que o humor negativo tem um peso maior na predição e na volatilidade dos preços.

Baseando-se nas informações apresentadas, este trabalho de conclusão buscou restringir a busca por *tweets* de usuários que tenham alguma relação com o setor financeiro. Para atingir este objetivo, foi desenvolvida uma estratégia de busca baseando-se na metodologia apresentada pelo trabalho de Yang, Mo e Liu (2015), onde uma rede de usuários chaves foi identificada.

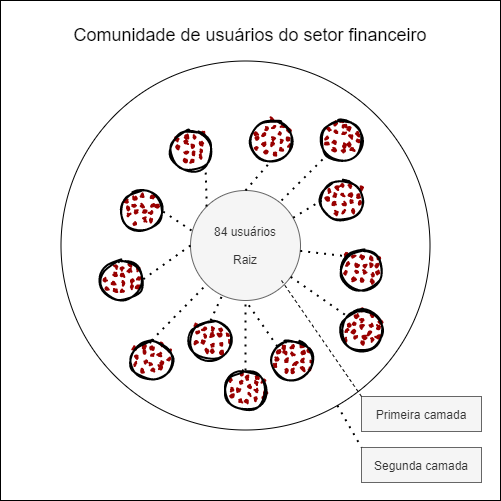
## DEFINIÇÃO DOS USUÁRIOS

A base da comunidade de usuários foi construída através da identificação de influenciadores, especialistas, e outras contas relacionadas ao setor financeiro. Para a construção e identificação destes usuários foi utilizado como auxílio um relatório publicado pela Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais (ANBIMA, 2021), onde foram identificados e analisados os perfis dos principais influenciadores brasileiros. Este relatório foi elaborado a partir da análise de dados coletados de 266 influenciadores nas redes sociais Twitter, Facebook, Instagram e Youtube. No total foram analisadas 160.312 publicações. De acordo com este relatório, o Twitter foi a rede social mais utilizada pelos influenciadores, e apresentou o maior engajamento entre os usuários. Este relatório também destaca e classifica alguns dos perfis dos influenciadores encontrados:

* Produtores de conteúdo: são usuários que divulgam e criam conteúdos planejados sobre investimentos, além de fazer análises e tratar de temas diversos relacionados a economia e educação financeira.
* Analistas financeiros: são profissionais do setor financeiro, que trabalham produzindo relatórios, análises técnicas e indicações de investimentos.
* *Traders*: são influenciadores que utilizam as redes sociais para compartilhar sua rotina profissional e fazer análise sobre transações do mercado financeiro.
* Casas de análise: são empresas independentes que produzem conteúdo e relatórios pagos, além de oferecerem consultoria para investidores.
* Especialistas: são profissionais ligados à área de economia, costumam utilizar seu conhecimento técnico para comentar sobre assuntos ligados a investimentos e economia.

Através da leitura e análise deste relatório, foram encontrados 62 perfis de influenciadores com contas no Twitter, sendo posteriormente adicionados mais 22 perfis de forma manual, totalizando 84 contas. Os perfis adicionados de forma manual foram selecionados por estarem relacionados com o setor financeiro e dentre estes perfis estão contas oficiais de organizações como o perfil oficial da B3, do Banco Central Brasileiro e corretoras de valores. Estas 84 contas formam a primeira camada da comunidade de usuários do setor financeiro, e a partir delas serão selecionadas as contas que irão compor a segunda camada de usuários. A Figura 8 ilustra o universo dos usuários. A segunda camada de usuários será construída através da obtenção dos seguidores das contas que foram a raiz da comunidade.

Figura 8 - Comunidade de usuários do Twitter



Fonte: elaborado pelo autor

## COLETA DE DADOS - TWITTER

Criado em 2006, o Twitter é uma rede social do tipo *microblogging* que se caracteriza pela publicação de mensagens curtas, de até 280 caracteres, chamadas de *“tweets”.* Assim como outras redes sociais, o Twitter permite a formação de comunidades de usuários por meio das conexões criadas pelo mecanismo de seguidores. Um usuário pode definir qual o tipo de conteúdo ele deseja consumir, utilizando estas conexões.

O Twitter foi selecionado como plataforma para coleta de dados pelo grande volume de dados não padronizados produzidos diariamente, sendo um excelente candidato para a aplicações de *Big Data,* além de atrair a atenção de pesquisadores de diversos campos(BING, CHAN e OU, 2014). De acordo com a ANBIMA (2021), o Twitter foi a rede social mais utilizada no Brasil por influenciadores do setor financeiro, produzindo também maior engajamento entre seus seguidores.

O Twitter possui uma *Application Programming Interface* (API) oficial que permite o acesso a diversos dados, entre eles estatísticas e informações sobre as contas dos usuários e suas mensagens publicadas. Para acessar estes dados é necessário solicitar acesso de desenvolvedor através da criação de uma conta na plataforma. Apesar da disponibilidade destes dados via API, os autores em Sirimevan et al. (2019), sugeriram que a coleta retroativa de dados pode ser uma tarefa desafiadora devido as limitações impostas na API oficial, por isso sugeriram avaliar a utilização de *crawlers* para executar esta tarefa.

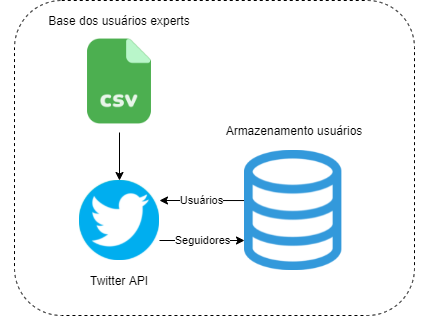
Seguindo esta recomendação, avaliou-se a utilização do pacote Python *twint*[[1]](#footnote-1), que é uma ferramenta de *crawler* para coletar dados do Twitter, sem a necessidade de autenticação ou utilização da API oficial. Os dados retornados por esta ferramenta correspondem aos mesmos campos retornados nas consultas da API oficial. Após testes e consultas na documentação da ferramenta, verificou-se que algumas funcionalidades da mesma foram comprometidas após atualizações do Twitter em 2021. No momento da execução deste trabalho, não foi possível utilizar as funções de consulta à lista de seguidores dos usuários.

Sendo assim optou-se por uma utilização mista da API oficial e do *crawler*. As informações, estatísticas e listagem dos seguidores das contas foram consultadas via API oficial e os *tweets* postados pelas contas foram obtidos pelo *crawler*. Quanto à obtenção dos *tweets* postados pelos usuários em sua linha do tempo, o *crawler* possui a mesma limitação imposta pelo Twitter em sua API: é possível listar apenas as últimas 3200 postagens de um usuário em sua linha tempo, também não sendo possível aplicar filtros nesta pesquisa. Para a utilização da API do Twitter, optou-se pelo pacote Python *tweepy*[[2]](#footnote-2), que possui implementadas todas as requisições definidas pela documentação da API.

### Coletor de dados

Devido às limitações descritas anteriormente, a tarefa de coleta de dados foi executada em diversas etapas para contemplar os passos necessários de acordo com a utilização da API e do *crawler.* A primeira etapa de coleta foi realizada com o objetivo de processar os perfis encontrados na etapa de definição dos usuários base da comunidade, e para isto foi criada uma coleção no banco de dados, com campos equivalentes aos atributos do objeto de perfil de usuário descrito na documentação oficial da API do Twitter. Além destes campos foi adicionado um atributo para identificar se o usuário é do nível base ou secundário na comunidade de usuários. Os usuários iniciais então foram importados através de um arquivo de texto com uma lista dos identificadores das contas encontradas na etapa de definição dos usuários, para cada usuário foi feita uma consulta na API para buscar as informações de sua conta.

Figura 9 - Processo de extração de usuários e seguidores



Fonte: elaborado pelo autor.

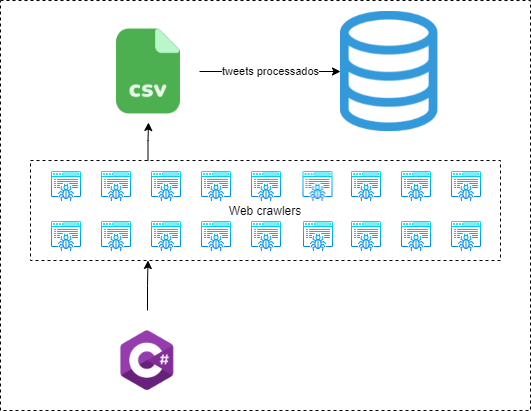
A segunda etapa teve como objetivo obter a lista de seguidores dos perfis coletados anteriormente. Estes seguidores formaram a segunda camada da comunidade de usuários. Para realizar esta tarefa foi desenvolvido um *script* em Python responsável por acessar a API do Twitter. Entretanto ao executar esta tarefa encontrou-se duas limitações impostas pela API: o método responsável por listar os usuários retorna apenas a informação do identificador númerico do usuário seguidor e a segunda limitação diz a respeito ao número de dados retornados e da quantidade de requisições. O método de listagem de seguidores retorna no máximo 5000 registros por requisição e permite apenas 15 requisições em intervalos de 15 minutos, totalizando assim a possibilidade de se obter 75000 seguidores a cada 15 minutos. A Figura 9 ilustra o processo inicial de obtenção dos usuários e seus seguidores.

Como forma de contornar estas limitações, a coleta de dados foi realizada gradualmente entre os dias 24/08/2021 e 02/09/2021, e os identificadores númericos do seguidores foram inseridos de forma avulsa, sem as informações adicionais das contas. A biblioteca *tweepy* possui implementação própria para realizar a espera entre os intervalos de 15 minutos das requisições, não sendo necessário alteração no script em relação a este aspecto. Quanto a limitação anterior de obter apenas o identificador númerico da conta, para utilização do *crawler* de *tweets* é necessário informar o *username* da conta que será coletada os *tweets.* Como forma de obter o restante das informações destes usuários encontrados anteriormente, foi desenvolvido um *script* Python responsável por acessar a API do Twitter para buscar os demais dados, como por exemplo a quantidade de *tweets* e *username* dos usuários.

A terceira etapa de coleta foi responsável por buscar os *tweets* de todas as contas definidas nas etapas anteriores, ao todo foram encontrados 4.444.899 usuários únicos entre os seguidores de todas as contas. Para esta etapa optou-se pela utilização do *crawler*, que apesar de possuir a mesma limitação de retornar apenas os últimos 3.200 *tweets* da linha do tempo do usuário, o mesmo não possui limites quanto ao número de requisições. De acordo com a documentação disponível na página da biblioteca *twint*, o *crawler* realiza requisições em um serviço *web* do Twitter, que é destinado para acesso e utilização por dispositivos mobiles com limitações. A aquisição dos dados pelo *crawler* ocorre de forma prográmatica, simulando a atividade de um usuário humano visualizando a linha do tempo de um usuário, levando em torno de 30 segundos para listagem completa dos últimos 3200 *tweets* de um usuário. Os dados obtidos pelo *crawler* são escritos em um arquivo de texto no formato *comma separated values* (CSV*),* que é salvo em disco*.*

Devido a grande quantidade de contas a serem verificadas, e as limitações de tempo deste trabalho, optou-se pela utilização de múltiplos *crawlers* de forma simultânea. Para isto decidiu-se por uma solução mista desenvolvida nas linguagens Python e C#. O *script* Python é responsável unicamente pela execução da biblioteca do *crawler,* recebendo o nome do usuário a ser pesquisado através de argumentos de linha de comando, e a solução desenvolvida em C# é responsável por gerenciar a execução das tarefas de forma simultânea, utilizando o recurso de paralelismo presentes na linguagem. Durante testes verificou-se que o número máximo de *crawlers* ativos é em torno de 30, ao testar um número maior o Twitter passa a bloquear as conexões.

Figura 10 – Processo de extração do *tweets*.



### Análise e processamento dos dados

A coleta de dados iniciou na data de 02 de setembro de 2021 e foi finalizada na data de 11 de outubro de 2021. A partir das contas dos usuários da primeira camada da rede do domínio financeiro, foram identificados 4.444.899 seguidores únicos. destes usuários, 3.373.805 possuem uma contagem de *tweets* maior que zero. Devido as restrições de tempo e limitações impostas na coleta dos dados, ao final do período de extração foi extrair os *tweets* de apenas 808.938 usuários*.* Ao todos, estes usuários foram responsáveis pela publicação de 1.526.462.923 *tweets,* resultando em um volume de dados de 768 *gigabytes* obtidos através das requisições realizadas ao Twitter durante as tarefas de *crawling*. A Figura 10 demonstra o processo executado para obtenção dos *tweets*.

Os arquivos com as informações obtidas pelo *crawler* estão no formato CSV, e possuem diversos campos que não são uteis para a execução deste trabalho, como forma de facilitar a manipulação e otimizar o espaço em disco ocupado, optou-se por realizar um pré-processamento com o objetivo de eliminar estas informações desnecessárias, além da aplicação de um filtro de data e de idioma. Apenas os seguintes campos foram mantidos: *id, date, time, timezone, user\_id, username, tweet, language.*

O filtro de data foi aplicado para considerar apenas *tweets* enviados a partir da data de 01 de janeiro de 2019, que corresponde ao período que será utilizado na construção do modelo de predição. O filtro de idioma foi aplicado após uma análise visual de uma amostra dos dados coletados. Nesta análise foi possível verificar que o Twitter não atribui um idioma a *tweets* com frases que não possuem sentido, e, desta forma, estes *tweets* foram excluídos. Por fim com a aplicação dos filtros e da remoção das informações necessárias, foram excluídos 553.398.628 *tweets*, restando 973.064.295 *tweets* que agora correspondem a um volume de 155 *gigabytes*. A Figura 11 demonstra a distribuição destes *tweets* em relação a data de publicação dos mesmos.

Figura 11 - Distribuição do tweets coletados e filtrados

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Fonte: elaborado pelo autor

Ao analisar a distribuição dos dados coletados é possível verificar que a maior concentração de dados está no ano de 2021, isto pode estar relacionado com a limitação na extração de dados dos perfis do Twitter, onde só foi possível obter no máximo os últimos 3200 *tweets* postado. Sendo assim a concentração destes dados corresponde ao nível da atividade dos usuários. Também é possível notar um grande salto no mês de março de 2020, correspondente ao aumento no uso de redes sociais durante a pandemia de Covid-19 que também foi noticiado por alguns meios de comunicação (STATISTA, 2021).

### Processamento textual dos *tweets*

Como as mensagens trocadas em redes sociais não seguem um padrão estruturado e normalmente são escritas em linguagem informal, os *tweets* podem conter diversos problemas que podem afetar de alguma forma o resultado da aplicação de algoritmos de *machine learning.* Desta forma é necessário preparar estes dados para processamento (BOUKTIF, FIAZ e AWAD, 2019).

De acordo com Alves (2015), o pré processamento de textos extraídos de redes sociais é uma tarefa comum visando a execução de análise de sentimentos para detectar corretamente o humor do público. O texto de redes sociais do tipo de *microblogging* pode conter diversos elementos únicos, como *links* para reportagens, imagens, menções e citações a outros usuários, e conteúdos não relacionados ao campo de pesquisa. A Figura 12 ilustra a estrutura de um *tweet*.

Figura 12 - Estrutura de um *tweet*

Interface gráfica do usuário, Texto, Aplicativo, chat ou mensagem de texto

Descrição gerada automaticamente

Fonte: Imagem adaptada de Alves (2015).

Ao analisar os trabalhos avaliados durante a revisão, foi possível verificar que alguns autores detalharam as tarefas realizadas para o pré-processamento de *tweets*. O trabalho de (LI, *et al.*, 2018) removeu *links* e *stop-words* dos *tweets.* Os autores em (PASUPULETY, *et al.*, 2019) utilizaram expressões regulares para remover espaços e pontuação, e uma biblioteca de terceiros para remover *emoticons, hashtags e* menções a usuários.

Com base nos resultados encontrados durante a revisão sistemática e uma análise visual de uma amostra dos *tweets* coletados, o Quadro 9 foi elaborado detalhando quais operações foram realizadas para processar os *tweets* coletados:

Quadro 9 - Operações de normalização dos *tweets* coletados

|  |
| --- |
| **Operação** |
| Remoção do termo “RT” que é encontrado no início das publicações de *retweet*. |
| Remoção de *emoticons* e caracteres especiais utilizando expressões regulares. |
| Remoção de endereços *web* presentes nas mensagens utilizando expressões regulares. |
| Remoção dos termos de *hashtags,* que são caracterizados pelo formato “#texto” utilizando expressões regulares. Alguns usuários fazem menções a ações utilizando *hashtags* (exemplo: “#PETR4”), nestes casos foi removido apenas o símbolo “#”, mantendo a referência à ação intacta |
| Remoção de menções a usuários, que são caracterizadas pelo formato “@usuario”. |
| Remoção de múltiplos espaços em brancos. |

Fonte: elaborado pelo autor.

Os *tweets* processados foram inseridos em um banco de dados MongoDB visando a facilitação das próximas atividades. Também foi mantido um campo com o registro do texto não tratado, caso fosse necessária uma nova análise em alguma etapa posterior. Além das operações detalhadas anteriormente, foi utilizada uma expressão regular para criação de um novo campo contendo apenas os códigos de ações mencionados na mensagem, caso haja algum. Os códigos das ações na bolsa brasileira seguem um formato fixo de 4 letras + 1 ou 2 números. A Figura 13 ilustra o documento final de um *tweet* processado e inserido no banco de dados.

Figura 13 - Estrutura do registro de *tweet* processado e importado



Fonte: elaborado pelo autor

## PROCESSAMENTO DOS TWEETS COLETADOS

Seguindo a metodologia apresentada por Yang, Mo e Liu (2015), após a definição dos usuários e coleta de seus *tweets*, é necessário filtrar quais usuários realmente estão interessados por temas relacionados ao setor financeiro. De acordo com os autores ao eliminar usuários não relevantes, é possível diminuir o efeito de ruídos na etapa de predição utilizando o sentimento dos usuários como atributo.

Para realizar este filtro, os autores propuseram analisar os *tweets* enviados pelas contas dos usuários considerados *experts*, que compõem a base da camada de usuários. Esta análise tem como objetivo criar um *corpus* de palavras-chaves relacionadas a tópicos financeiros, para então aplicar um filtro nos usuários da segunda camada, onde caso nenhuma das últimas 20 mensagens enviadas por um destes usuários contenha algum termo presente no *corpus* de palavras-chaves, o usuário e seus *tweets* são então descartados.

Os autores Yang, Mo e Liu (2015) utilizaram técnicas de processamento de linguagem natural para construir seu *corpus* de palavras-chaves. Os *tweets* foram processados e as *stop words* foram removidas, após então os substantivos foram extraídos e classificados de acordo com sua frequência. Por fim os autores adicionaram no *corpus* o nome e o código da ação de todas as empresas listadas em três bolsas de valores americanas.

### Construção do *corpus* de palavras-chaves

Com base na metodologia proposta pelos autores Yang, Mo e Liu (2015), foi criado um *corpus* de palavras-chaves contendo os termos mais frequentes encontrados nos *tweets* dos usuários da primeira camada. Para realização das atividades de Processamento de Linguagem Natural (PLN), foi utilizada a biblioteca Python s*paCy*[[3]](#footnote-3).

De acordo com a documentação do site oficial, s*paCy* é uma biblioteca *open-source* para aplicação de PLN, atingindo o estado da arte na implementação de diversas tarefas, entre elas: *Part-of-Speech Tagging,* *Named Entity Recognition* e *Tokenization*. Algumas das tarefas são executadas com a utilização de modelos de *deep learning* já treinados, fornecendo suporte para diversos idiomas, entre eles, o português. O modelo no idioma português foi construído utilizando como base o trabalho desenvolvido por Rademaker *et al.* (2017), no qual foi criado um *corpus* com as regras de gramática do idioma português. Este modelo foi então treinado em uma base de noticias contendo aproximadamente 500 mil registros.

Como forma de facilitar a manipulação destes dados, uma nova coleção no banco de dados MongoDB foi criada, contendo apenas os *tweets* dos usuários da primeira camada, que somam no total 216.497 registros. Para execução das tarefas de processamento textual foi desenvolvido um *script* em Python*,* na qual todos os registros contidos nesta nova coleção foram processados. Na primeira etapa foi realizada a extração dos substantivos presentes nos *tweets*, utilizando as técnicas de *Tokenization* e *Part-of-Speech Tagging.* Na tarefa de *Tokenization* o texto dos *tweets* é dividido e segmentado em palavras, símbolos e pontuação, a Figura 14 ilustra esse processo.

Figura 14 - Exemplo da técnica de *Tokenization*

Linha do tempo

Descrição gerada automaticamente

Adaptado de spaCy (2021)

Já na tarefa de *Part-of-Speech Tagging,* o texto segmentado é então classificado de acordo com o vocabulário e as regras gramaticais do idioma, deste modo é possível diferenciar e extrair os substantivos das sentenças. Ao todo foram extraídos uma lista de 75.259 substantivos únicos, ordenados de acordo com a frequência dos termos.

Já na segunda etapa foi executa a tarefa de *Named Entity Recognition,* que tem como objetivo identificar e classificar entidades mencionadas em um texto não classificado. O *spaCy* realiza esta atividade através da aplicação de técnicas de estatística utilizando um modelo já treinado com exemplos, além disso é levado em consideração regras gramaticais e contexto do texto. A Figura 15 demonstra um exemplo da aplicação desta técnica. No total foram identificadas 59.911 entidades únicas, que foram ordenadas de acordo com sua frequência.

Figura 15 - Exemplo de aplicação da técnica de *Named Entity Recognition*

Tabela

Descrição gerada automaticamente

Os autores Yang, Mo e Liu (2015) utilizaram apenas a extração de substantivos para criação de seu dicionário de palavras-chaves e não especificaram nenhum critério para definir quais termos fazem parte do dominio financeiro. Desta forma neste trabalho optou-se por realizar uma verificação e classificação manual.

Ao analisar as entidades extraídas, verificou-se que há 55.021 registros com uma frequência menor que cinco, e ao analisar de forma visual estes termos é possível verificar que em sua maioria são registros que podem ser considerados como “falso-positivo” devido a natureza informal do texto dos *tweets,* onde por exemplo textos ou palavras escritas em caixa alta foram identificadas como uma entidade. Deste modo optou-se por analisar e filtrar apenas as entidades que possuem ao menos uma frequência de 100 registros, correspondendo a 319 registros.

Ao todo foram filtradas 120 entidades que em sua maioria correspondem ao nome e código de ações de empresas listadas na bolsa de valores e nomes e siglas de entidades relacionadas com o setor financeiro. Entre os registros removidos estão entidades e nomes de empresas não relacionadas ao setor financeiro, como menções a redes sociais, organizações mundiais e etc.

Entretanto ao realizar um teste filtrando uma amostragem de aproximadamente um milhão de *tweets,* utilizando este dicionário de entidades*,* verificou-se alguns falsos positivos onde o nome da entidade também corresponde a uma palavra, como por exemplo nos casos da empresa Vale e da sigla IR que corresponde ao imposto de renda. O trabalho de Silva (2015) obteve a mesma constatação, e optou por remover este tipo de registro em sua lista de palavras utilizadas na busca de *tweets.* A mesma ação foi tomada neste trabalho, resultando em um número final de 116 entidades. A Figura 16 demonstra os termos e suas frequências em uma exibição de nuvem de palavras.

Figura 16 - Nuvem de palavras das entidades filtradas

Texto

Descrição gerada automaticamente com confiança média

Fonte: elaborado pelo autor.

Quanto a análise dos substantivos, utilizou-se os mesmos critérios definidos anteriormente, foram analisados apenas substantivos com uma frequência maior que 100, correspondendo no total 2445 registros. Ao todo foram selecionados 113 substantivos que em sua maioria correspondem a termos e jargões utilizados por pessoas relacionadas ao mercado financeiro. As entidades encontradas na etapa anterior foram desconsideradas nesta lista. A Figura 17 demonstra a nuvem de palavras correspondentes aos itens filtrados.

Figura 17 - Nuvem de palavras dos substantivos filtrados

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente com confiança média

Fonte: elaborado pelo autor

### Filtragem dos *tweets* coletados

Nesta etapa foi realizada a filtragem dos 973.064.295 *tweets* adquiridos nas etapas anteriores. Na metodologia aplicada por Yang, Mo e Liu (2015) os *tweets* foram analisados por usuário, e foram considerados todos os *tweets* de usuários em que ao menos uma de suas 20 últimas mensagens contenha algum dos termos identificados nas palavras-chaves anteriormente.

Devido ao grande volume de dados obtidos, neste trabalho optou-se por modificar este processo, todos os *tweets* coletados foram analisados e filtrados individualmente utilizando o dicionário de palavras-chaves construído previamente. Desta forma busca-se reduzir uma maior quantidade de *tweets* que não tenham relação ao contexto financeiro, evitando assim situações em que por exemplo um usuário possa ter feito apenas uma ou poucas publicações em relação ao dominio financeiro, e suas demais postagens são relacionadas a outros assuntos.

Para realizar esta tarefa foi desenvolvido um *script* Python, responsável por analisar todos os registros de *tweets* inseridos no banco de dados. A lista de palavras-chaves compostas pelas entidades e substantivos, criada anteriormente, foi salva em um arquivo e carregada na memória. A etapa de comparação de *string* e verificação das palavras-chaves foi realizada com o auxílio da biblioteca *spaCy* e das técnicas *Tokenization* e *Part-of-Speech Tagging.* O texto dos *tweets* foi segmentado e classificado, e apenas os *tokens* correspondentes a objetos e substantivos foram comparados com a lista de palavras-chaves.

Ao todo foram foram selecionados 18.641.168 *tweets* que possuem ao menos uma palavra-chave presente em seu texto. Ao analisar de forma visual uma amostra de aproximadamente 100 mil registros, foi possível observar a realidade dos dados coletados. Boa parte dos *tweets* não segue um padrão estruturado, contém erros de ortografia, mensagens de *spam* e, além disso, foi possível perceber que há diversos *tweets* sem relação ao mercado financeiro. Ao analisar estes *tweets* e comparar com a lista de palavras-chaves elaboradas anteriormente, pode-se perceber que algumas palavras como por exemplo “bolsa”, “investir”, “mercado”, “alta”, “queda” podem ser empregadas em outras situações, tendo assim seu sentido dependente do contexto envolvido.

O trabalho de Alves (2015) também realizou uma coleta de *tweets* do mercado financeiro*,* porém utilizando uma estratégia de coleta diferente da aplicada por este trabalho. Como forma de melhorar seu resultado, a autora construiu uma lista com as palavras e expressões mais citadas e que claramente não tenham a ver com o dominio financeiro. Os *tweets* coletados foram então filtrados utilizando estas listas. A Figura 18 ilustra a lista construída pela autora.

Figura 18 - Palavras e expressões utilizadas para limpeza de *tweets*.

Texto, Carta

Descrição gerada automaticamente

Fonte: adaptado de (ALVES, 2015).

Como forma de melhorar os resultados dos *tweets* filtrados anteriormente, optou-se por aplicar uma segunda etapa de filtragem, utilizando a mesma estratégia descrita anteriormente pela autora. Para realizar esta etapa de filtragem e construir a lista de termos a serem utilizados como critério de exclusão, foi utilizada a mesma estratégia empregada na etapa de construção da lista de palavras-chaves deste trabalho. O *script* Python responsável por executar as técnicas de *Part-of-Speech Tagging,* *Named Entity Recognition* e *Tokenization* foi aplicado nesta coleção de *tweets* filtrados. Ao todo foram obtidos 6.399.064 termos e entidades, destes 22.026 registros possuem uma frequência maior que 100. Optou-se por analisar de forma manual, em ordem descrecente até ser obtido uma quantidade de 100 termos para realizar o filtro de exclusão, devido a limitações de tempo deste trabalho. Ao todo foram removidos 948.642 *tweets* que se encaixaram no filtro de exclusão. A Figura 19 demonstra a nuvem de palavras com estes termos.

Figura 19 - Nuvem de palavras dos termos utilizados como critério de exclusão

Texto

Descrição gerada automaticamente

Fonte: elaborado pelo autor.

### Análise dos *tweets* filtrados

Ao todo foram obtidos inicialmente a quantidade de 1.526.462.923 *tweets*, que com a aplicação do primeiro filtro de período e de idioma foram reduzidos para 973.064.295 *tweets*. Com a aplicação da primeira etapa de filtragem, onde os *tweets* foram filtrados de acordo com a presença de termos que remetem ao setor financeiro, a quantidade total foi reduzida para 18.641.168 *tweets.* Por fim com a última etapa de filtragem, onde foram removidos *tweets* que contenham termos que claramente não fazem parte do setor financeiro, mais 948.642 *tweets* foram removidos, restando a quantidade final de 17.692.526, apresentando uma redução de aproximadamente 98,84% em relação ao número inicial de *tweets*. A Figura 20 representa o progresso das etapas de filtragem. A Figura 21 representa a distribuição final destes *tweets* ao longo do tempo.

Figura 20 - Progresso de filtragem dos *tweets*

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Fonte: elaborado pelo autor.

Figura 21 - Distribuição final dos *tweets* coletados ao longo do tempo.

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Fonte: elaborado pelo autor

Ao analisar o conteúdo destes *tweets*, 139.989 mensagens fazem menções diretas a códigos de ações de empresas listadas na bolsa de valores brasileira. A empresa Petrobras foi a mais citada nestas mensagens com 23.348 menções divididas entre as ações PETR3 e PETR4, a segunda empresa mais citada foi a VALE S.A através da ação VALE3, com 7.485 menções. Atualmente, no mês de outubro de 2021, estas duas empresas possuem o maior peso na composição do índice Ibovespa e ambas normalmente estão entre os maiores volumes de negociações durante os pregões da bolsa, atraindo assim o interesse dos investidores.

Ao analisar as mensagens buscando pelo nome da empresa, 214.968 mencionaram diretamente a Petrobras. Quanto a Vale S.A não foi possível fazer essa mensuração visto a possibilidade da utilização da palavra “vale” como verbo ou como substantivo, dependendo assim do contexto para fazer esta análise. Para efeitos de comparação foi analisado também as menções da empresa Banco do Brasil, a qual foi citada 21.658 vezes pelo nome e 4.056 vezes pelo código da ação. Apesar da quantidade significativa de menções a uma empresa listada na bolsa de valores, não necessariamente o público pode estar comentando sobre sua ação, visto que muitas empresas listadas prestam diversos tipos de serviços, sendo assim podem atrair o interesse geral da população. Tal constatação também foi percebida na análise dos dados coletados pelo trabalho de Alves (2015).

### Limitações

Quanto as limitações encontradas durante a execução desta etapa do trabalho, a estratégia inicial de busca dos *tweets* previa a pesquisa de empresas diretamente pelo seu código de ação, visto que a maioria dos trabalhos analisados durante a revisão sistemática fizeram uso desta mesma estratégia. Porém, o recurso de *cashtag* (quando uma ação é citada com o prefixo de cifrão, exemplo: $AAPL) disponibilizado pelo Twitter apenas oferece suporte a empresas listadas nas bolsas norte americanas.

Em relação ao aspecto da qualidade dos dados obtidos, o Twitter impõe uma limitação de listar apenas os últimos 3200 *tweets* de um usuário. Sendo assim o período de dados coletados não foi uniforme e sim dependente da atividade e quantidade de *tweets* publicados pelas contas analisadas. Já nas etapas de filtragem não foi possível encontrar uma estratégia de seleção de termos já analisada e validada por outros pesquisadores. Desta forma, foi realizada de forma manual, havendo viés do pesquisador na seleção dos termos.

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

Neste capítulo serão aplicadas técnicas de ML em *tweets* relacionados ao índice e empresas listadas na bolsa de valores brasileira, procurando identificar relações entre o sentimento associado aos *tweets* e o preço da ação. A execução desta etapa se deu com o auxílio do referencial teórico analisado durante a execução da revisão sistemática executada anteriormente.

## MACHINE LEARNING

Segundo Mitchell (2006), o principal objetivo de pesquisa da área de *Machine Learning* (ML) é a construção de algoritmos e programas de computador que possam se aperfeiçoar automaticamente com experiencias e conhecimentos adquiridos durante execuções passadas. A construção destes algoritmos se dá através de uma combinação multidisciplinar de conceitos de diversas áreas como: estatística e probabilidade, inteligência artificial, ciências cognitivas e outras. Com aprimoramento do aprendizado de máquina pode-se abrir novos caminhos e percepções de como a utilização da computação pode melhorar o aprendizado humano (MITCHELL 2006).

Um dos principais usos de ML é a resolução de forma eficiente de problemas que normalmente são resolvidos com facilidade por seres humanos. Anteriormente estes problemas eram resolvidos com a codificação de um programa de computador com diversos condicionais e regras lógicas, que por sua vez eram criadas através da aquisição de conhecimento específicos. Com a utilização de ML é possível desenvolver uma hipótese ou função, que é capaz de resolver estes problemas através do conhecimento adquirido com execuções passadas. Um exemplo é a descoberta de uma hipótese que leva a criação de regras que definem quais clientes devem receber um material de propaganda específico, com base nos dados de suas compras passadas (FACELI, *et al.*, 2011).

De acordo com Alpaydin (2010), com a utilização de MLpode-se resolver problemas computacionais para os quais não há um algoritmo específico, ou então evitar cenários onde é necessário a utilização de múltiplos algoritmos para condições especificas. Uma das principais tarefas de ML é a extração e reconhecimento de padrões em dados, que serão utilizados para a tomada de decisões inteligentes (HAN; KAMBER; PEI, 2012). Este processo ocorre através do processamento de grandes quantidades de dados, onde é possível rotular uma amostra deles, e com a utilização de técnicas de ML o programa é capaz de extrair o algoritmo necessário para a classificação e extração do conhecimento destes dados (ALPAYDIN, 2010).

Conforme Mitchell (2006), a definição de um problema de ML inicia através da definição de uma tarefa ou objetivo a ser cumprido, necessita de métricas para avaliar o seu desempenho e por fim uma fonte de dados que possa ser utilizada como forma de treinamento e geração de experiência. Um dos principais desafios a ser superado na resolução de um problema de ML é na obtenção de uma fonte de experiências de qualidade, para isto é necessário atentar a alguns atributos como uma quantidade adequada de dados, a possibilidade de extrair métricas de desempenho destes dados e uma distribuição diversificada que represente situações que serão encontradas ao longo da execução do programa desenvolvido (MITCHELL, 2006).

A utilização de ML permite providenciar soluções para uma gama de problemas encontrados no mundo real, como a predição do mercado financeiro e de fraudes em sistemas bancários, a otimização de processos industriais e o auxílio no diagnóstico de doenças. Para se chegar nestes resultados é necessário avaliar e definir quais tipos de algoritmos e técnicas de ML serão empregadas na execução, e o resultado será um modelo estatístico que pode ser tanto preditivo, para realizar predições futuras através da análise de dados do passado, quanto um modelo descritivo, utilizado na extração de conhecimento a partir de uma amostra de dados (ALPAYDIN, 2010).

De acordo com Faceli et al. (2011), as tarefas, algoritmos e técnicas de ML podem ser caracterizadas pela seguinte hierarquia, representada na Figura 22.

Figura 22 - Hierarquia do aprendizado de máquina

Diagrama

Descrição gerada automaticamente

Fonte: adaptado de Faceli et al. (2011)

### Aprendizagem Supervisionada

A aprendizagem supervisionada recebe este nome pois o aprendizado de máquina se dá através da utilização de um conjunto de dados já classificado anteriormente por um “supervisor externo”. A resolução de problemas de ML utilizando aprendizagem supervisionada se resume a utilização dos algoritmos de classificação e regressão, a escolha destes algoritmos se dá de acordo com o objetivo que se busca atingir na análise dos dados (HAN; KAMBER; PEI, 2012).

A utilização deste paradigma de aprendizado é utilizada para criação de modelos de análise preditiva, que por sua vez tem como objetivo realizar predições de valores e testar hipóteses com base nos atributos de entrada. Para realizar esta tarefa, a fonte de dados utilizada deve possuir atributos de entrada e saída. Com o conhecimento dos valores de entrada e saída do conjunto de dados de treinamento, é possível avaliar a capacidade do modelo de realizar predições e obter valores de saída para novos exemplos (FACELI, *et al.*, 2011).

### Aprendizagem não supervisionada

A aprendizagem não supervisionada é responsável pela criação de modelos descritivos, que tem como objetivo explorar e extrair conhecimento de um conjunto de dados, buscando informações e padrões escondidos. Este paradigma recebe este nome pois a fonte de dados utilizada na tarefa não possui um atributo de saída definido, apenas atributos de entrada. Estes modelos normalmente são utilizados em atividades de agrupamento de dados, que tem como objetivo encontrar objetos semelhantes em um conjunto de dados e em atividades de associação, que buscam encontrar atributos relacionados (FACELI, *et al.*, 2011).

## AQUISIÇÃO DOS DADOS FINANCEIROS

Esta etapa iniciou-se com a aquisição das séries históricas dos ativos negociados na Bovespa, estes dados estão disponíveis de forma pública e são disponibilizados para download no site[[4]](#footnote-4) oficial da B3. Estes dados são disponibilizados em arquivos de texto com formatação própria, a própria B3 disponibiliza um manual de como realizar a leitura destes dados. Conforme a B3 (2021) informa, as seguintes informações estão presentes nestes arquivos:

Esse produto contempla as principais informações dos ativos, como: nome e código da empresa, código da ação, código ISIN, tipo de mercado (a vista, termo, opções), especificação (ON/PN), preços (anterior, abertura, mínimo, médio, máximo, fechamento), quantidade de negócios e volume negociado com o papel, dentre outros dados disponíveis.

Escolheu-se utilizar extrair o período de negociações de 01/01/2019 até 30/09/2021. Os arquivos referentes a estes períodos foram processados através da execução de um script Python, onde foram coletadas as informações de negociações do ativo BOVA11, que é um ETF utilizado para representar o índice Bovespa, e das empresas Vale S.A e Petrobrás. Além do índice, optou-se verificar de forma individual estas duas empresas pelos seguintes motivos: A Petrobrás é uma empresa pública e estatal, sendo assim está mais suscetível a opinião pública, e atualmente ela é a segunda empresa mais importante quanto ao seu peso na composição do índice Ibovespa. Já a Vale S.A foi selecionada por ser a empresa de maior peso no cálculo da composição do índice Ibovespa. Atualmente no mês de outubro de 2021 a Vale S.A corresponde a 12,26% do índice e a Petrobrás a 11,5%.

## APLICAÇÃO DE ANÁLISE DE SENTIMENTOS

Nesta etapa busca-se obter o sentimento geral da multidão através da aplicação de técnicas e algoritmos de análise de sentimentos nos *tweets* coletados durante a etapa anterior, com o objetivo de verificar sua relação com o desempenho do índice e de empresas presentes na bolsa de valores. Visto que não é o objeto de estudo deste trabalho o desenvolvimento de uma ferramenta para análise de sentimentos, optou-se por realizar um levantamento e utilizar uma ferramenta ou estratégia já analisada e validada por outros pesquisadores e que atendesse aos requisitos deste trabalho: ser capaz de processar textos na língua portuguesa e em textos de redes sociais.

De acordo com Liu (2012), a análise de sentimentos ou mineração de opiniões é o campo de estudo que analisa os sentimentos, opiniões e emoções de indivíduos em relação a produtos, organizações e acontecimentos. Este processo de análise ocorre através da utilização do processamento de linguagem natural e linguística computacional para analisar e identificar opiniões, subjetividade e emoções em textos e sentenças.

Apesar do processamento de linguagem ser um campo de pesquisa que levanta grande interesse da comunidade acadêmica, foi apenas a partir dos anos 2000 que a análise de sentimentos começou a atrair o interesse tanto desta comunidade, quanto da indústria. Isto se deve ao advento e a utilização em massa das redes sociais, e o grande volume de dados produzidos por elas (LIU, 2012).

O rápido crescimento da análise de sentimento se deve também ao seu impacto em diversas outras áreas, sendo considerada uma informação essencial em diversos tipos de negócios, auxiliando na tomada de decisões em empresas do setor da saúde, análise do mercado financeiro e em agências governamentais. As opiniões também são um grande influenciador do comportamento humano, sendo comum pessoas realizarem pesquisas para verificar outras opiniões antes de realizar uma compra, escolher um candidato político, e até em tomada de decisões pessoais (LIU, 2012).

De acordo com os autores Kansaon, Brandão e Pinto (2018) e Pereira (2020) há diversas pesquisas e ferramentas desenvolvidas visando o idioma inglês, enquanto pesquisas envolvendo o idioma português ainda se encontram em estágios iniciais. De acordo com Kansaon, Brandão e Pinto (2018) a falta de estudos no idioma português se deve pela falta de bases de dados e poucas ontologias disponíveis no português brasileiro para realização de análises de texto. A grande quantidade de regras gramaticais, variação dos tempos e formas verbais dificultam a criação de dicionários léxicos no idioma Português (JUNQUEIRA e FERNANDES, 2018).

Pereira (2018) também acrescenta que boa parte dos recursos disponíveis em português foram criados através da tradução de base de dados desenvolvidas originalmente no idioma inglês. De acordo com Pereira (2018), o desempenho destes recursos é limitado, sendo que, em sua maioria, ao traduzir o texto para o idioma Inglês e utilizar a ferramenta originalmente desenvolvida para este idioma, é possível obter melhores resultados. Apesar disto, a tradução dos textos para inglês também não é uma prática recomendada, pois ao utilizar meios automáticos de tradução pode haver perda de sentido em algumas sentenças (PEREIRA, 2018).

Visto as dificuldades relatadas pelos autores acima, neste trabalho decidiu-se explorar uma nova abordagem, que é a utilização de uma ferramenta paga, visto que boa parte destas ferramentas oferece alguma forma de créditos para teste ou para estudantes. Até a data de execução deste trabalho, não foi possível encontrar publicações mencionando a utilização destas ferramentas para análise de sentimentos no idioma português brasileiro.

Optou-se por utilizar a solução *Google Cloud Natural Language* visto que a mesma oferece uma quantidade significativa de créditos para teste, sendo possível analisar uma boa amostra dos *tweets* coletados. Outras soluções como *Amazon Comprehend, Azure Machine Learning* e *IBM Watson Natural Language Understanding* também foram avaliadas, porém as mesmas oferecem uma quantidade relativamente menor de recursos gratuitos.

### *Google Cloud Natural Language*

A plataforma *cloud* do Google oferece diversas soluções no âmbito de ML e análises de PLN. É possível analisar textos de forma inteligente, utilizar modelos de ML já treinados ou criar modelos personalizados. A solução *API Natural Language* é responsável por realizar as tarefas de processamento de texto, sendo possível obter diversos *insights,* entre eles a extração de sentimentos no português brasileiro, atendendo assim os requisitos deste trabalho. Conforme documentação oficial, esta solução possui os seguintes recursos:

Os avançados modelos pré-treinados da API Natural Language permitem que os desenvolvedores apliquem facilmente o processamento de linguagem natural (NLU) aos aplicativos com recursos como análise de sentimento, análise de entidade, análise de sentimento de entidade, classificação de conteúdo e análise sintática.

Esta solução é acessível através da utilização de uma *API Rest,* sendo necessário realizar um cadastro prévio para obtenção das credenciais de acesso. Como forma de degustação dos seus serviços, o Google oferece um crédito de 300 dólares para ser utilizado por novas contas de usuários. A cobrança pela utilização deste recurso é realizada por unidades de processamento utilizado. Cada documento enviado é considerado como uma unidade, e caso o documento ultrapasse o limite estabelecido de 1.000 caracteres, serão cobradas unidades extras. A Figura 23 demonstra a tabela de preços e quantidades.

Figura 23 - Tabela de preços do *Cloud Natural Language*

Tabela

Descrição gerada automaticamente

Fonte: adaptado de Google (2021).

Visto que cada *tweet* publicado contém no máximo 280 caracteres, eles serão contabilizados como um documento que equivale a uma unidade. Sendo assim este trabalho estará limitado ao processamento de 300.000 *tweets*.

De acordo com a documentação oficial, a análise de sentimentos busca determinar a atitude expressa no texto enviado como positiva, neutra ou negativa. O sentimento é obtido através de dois valores numéricos, retornados nos campos *score* e *magnitude*. O atributo *score* pode variar de -1.0 (negativo) a 1.0 (positivo), e corresponde ao sentimento geral do texto enviado.

Já o campo *magnitude* identifica a carga emocional no texto enviado, com valores a partir de 0.0, sem limite máximo. O valor deste campo é formado a partir da análise de blocos dentro do texto enviado, sendo assim textos maiores podem atingir valores maiores de magnitude. Com a utilização do campo *magnitude* é possível diferenciar emoções mistas de neutras, por exemplo um texto contendo um misto de blocos positivos e negativos pode ser classificado com sentimento neutro, porém com uma alta carga de magnitude. A Figura 24 demonstra uma sugestão de interpretação destes valores.

Figura 24 - Interpretação de valores do *Cloud Natural Language*

Tabela

Descrição gerada automaticamente

Fonte: adaptado de Google (2021).

### Janela de análise

Conforme as limitações de análise de sentimentos da plataforma selecionada, será possível realizar uma análise parcial em relação a quantidade de *tweets* coletados nas etapas anteriores. Sendo assim, com a ajuda do orientador deste trabalho, optou-se por reduzir o escopo de análise deste trabalho para o seguinte formato: serão analisados três períodos de um mês dos anos de 2019, 2020 e 2021.

Escolheu-se o mês de março de cada ano, em função do início da pandemia de Covid-19 no Brasil (março de 2020). A ideia é encontrar padrões ao comparar o mesmo período antes e depois. Para cada mês escolhido serão sorteados um máximo de 100.000 *tweets* de forma aleatória, ignorando *tweets* publicados em finais de semana devido a nestes dias não haver pregão na bolsa de valores.

### Resultados

Foi desenvolvido um *script* Python para realização desta tarefa. Na primeira etapa o *script* foi responsável por selecionar de forma aleatória quais *tweets* foram analisados; para realização desta etapa foi utilizada a função *sample* do banco de dados MongoDB, na qual são retornados uma quantidade de documentos, selecionados de forma aleatória pelo próprio banco de dados. A quantidade diária de documentos foi definida levando em consideração o número de dias úteis do mês, variando de aproximadamente 4500 a 5000 registros. Esta função foi executada para cada dia útil do mês de março dos anos de 2019, 2020 e 2021. Visto a não uniformidade dos dados, em alguns dias não foi possível obter a quantidade máxima de documentos definida, resultando assim numa coleta total de 291.000 *tweets* a serem analisados.

Já na segunda etapa, foram realizadas as consultas à API do *Google Cloud Natural Language*. Estas requisições foram realizadas com auxílio da biblioteca oficial de Python, disponibilizada pelo próprio Google. Esta etapa demorou em torno de 9h, visto que esta API possui um limite de 600 requisições por minuto/800.000 requisições diárias. Os resultados destas consultas foram salvos em um novo campo no documento dos *tweets* armazenados no banco de dados.

Figura 25 - Histograma de distribuição do sentimento nos *tweets*.

Gráfico, Gráfico de barras, Histograma

Descrição gerada automaticamente

Fonte: elaborado pelo autor.

No total foram analisados os sentimentos de 291.000 *tweets*, sendo que 72.334 destes *tweets* obtiveram um *score* de sentimento 0.0, sendo assim considerados como neutro, 141.131 obtiveram um *score* maior que zero, sendo classificados como positivos, e 77.353 tiveram um valor *score* menor que zero, sendo considerados como negativos. De acordo com Liu (2012), o sentimento neutro pode representar uma falta de opinião, sendo ignorado por muitos pesquisadores em seus trabalhos. Sendo assim optou-se pela mesma decisão na execução deste trabalho. Ao eliminar os *tweets* considerados como neutros, é obtida a quantidade final de 218.666 *tweets*. A Figura 25 demonstra o histograma com a distribuição dos mesmos.

A maior concentração destes *tweets* se encontra na faixa de -0,2 a 0.2. O trabalho de Yang, Mo e Liu (2015) obteve resultados similares em sua análise de sentimentos, os *tweets* classificados pelos autores se encontraram em uma faixa de concentração similar a deste trabalho. A Figura 26 demonstra os resultados de seu trabalho.

Figura 26 - Histograma de distribuição do sentimento nos tweets do trabalho de Yang, Mo e Liu (2015)

Gráfico, Histograma

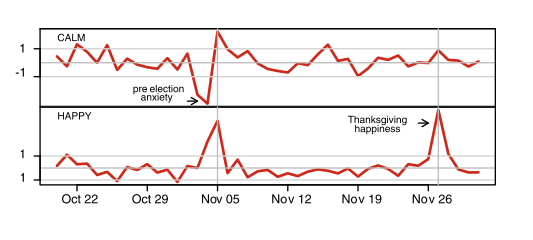
Descrição gerada automaticamente

Fonte: adaptado de (YANG, MO e LIU, 2015)

No trabalho de Bollen, Mao e Zeng (2011), os autores realizaram a coleta de *tweets* utilizando como texto de busca expressões que remetem a como o usuário está se sentindo, como por exemplo: “eu sinto que ...” e “estou me sentindo...”. Para realizar a análise de sentimento destes *tweets* os autores utilizaram a ferramenta *Google Profile of Mood States* (GPOMS), na qual se utiliza um dicionário léxico para extrair o humor do texto, que é representado em seis dimensões: *Calm, Alert, Sure, Vital, Kind and Happy.*

Como forma de validar seus resultados, os autores selecionaram alguns períodos de tempo no qual se esperava uma alta concentração de um humor em especifico, como datas festivas, feriados e eleições. Em um de seus experimentos, os autores validaram o período de feriado de ação de graças americano, onde se esperava encontrar uma grande concentração de *tweets* com a emoção *Happy*, que representa “felicidade”. Em outro experimento os autores validaram a emoção “*Calm*”, que representa “calma”, no período de eleições e esperavam encontrar uma grande concentração negativa. Em ambos os experimentos, os resultados foram validados com sucesso, a Figura 27 demonstra o resultado deste trabalho.

Figura 27 - Análise de sentimentos do trabalho de Bollen, Mao e Zeng (2011)



Fonte: adaptado de (BOLLEN, MAO e ZENG, 2011)

Optou-se por repetir este experimento neste trabalho, através da análise de um período no mês de março de 2020, considerado como início da pandemia de Covid-19 no Brasil, onde a bolsa de valores também bateu recordes negativos. Sendo assim é esperado uma maior concentração de valores de sentimento negativo.

Escolheu-se então avaliar o período do dia 16 até o dia 26, sendo que no dia 23 de março, o índice Bovespa teve seu menor valor no ano de 2020. Diferente do trabalho de Bollen, Mao e Zeng (2011), não foi possível observar o mesmo resultado neste experimento, a distribuição dos sentimentos neste período se manteve similar ao comparar com o resultado anterior, considerando todos os *tweets* coletados (Figura 23). A Figura 28 demonstra estes resultados.

Figura 28 - Histograma de distribuição do sentimento nos *tweets* entre 16/03/2020 e 26/03/2020.

Gráfico, Histograma

Descrição gerada automaticamente

Fonte: elaborado pelo autor.

Visto que esta validação apenas considerou os *tweets* de forma quantitativa, optou-se por realizar uma nova análise considerando os valores de sentimento atribuídos aos *tweets*. Nesta análise foi calculada a média diária, de cada mês, dos valores de *score* obtidos na análise de sentimento dos *tweets.* As Figuras 29, 30 e 31 trazem o resultado mensal destas análises.

Figura 29 - Média diária dos valores de sentimento de março de 2019.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Fonte: elaborado pelo autor.

Figura 30 - Média diária dos valores de sentimento de março de 2020

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Fonte: elaborado pelo autor

Figura 31 - Média diária dos valores de sentimento de março de 2021.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Fonte: elaborado pelo autor.

Ao analisar estes dados ainda podemos verificar que há uma concentração maior de *tweets* positivos de forma geral, porém ao comparar o ano de 2020 com 2019 e 2021 podemos validar através das médias que o sentimento deste ano foi no geral mais negativo que nos outros dois anos. Ao comparar o ano de 2021 com 2019, antes da pandemia de Covid-19, podemos observar que apesar de haver uma melhora no sentimento geral, em relação a 2020 (auge da pandemia), o mesmo ainda se encontra em níveis mais baixos do que antes da pandemia.

Sendo assim podemos considerar que a hipótese de um evento externo, neste caso a pandemia de Covid-19, pode ser considerada como parâmetro para avaliar os dados coletados durante a análise de sentimentos.

## MODELO DE PREDIÇÃO

Nesta etapa foi realizada a construção de um modelo de predição, optou-se por selecionar um classificador *Support Vector Machine* (SVM), que foi a técnica utilizada com maior frequência nos trabalhos identificados durante a execução da revisão sistemática. Visto que esta técnica é aplicada em problemas de classificação binária, este trabalho terá como objetivo verificar se o preço da ação irá subir ou descer, sem avaliar o percentual ou valor de queda/subida. Para execução desta etapa, optou-se pela utilização do pacote *Python scikit-learn[[5]](#footnote-5),* o qual já possui diversas implementações de algoritmos e técnicas de ML, em seu estado da arte.

### Seleção dos Atributos

A seleção dos atributos utilizados foi realizada com base nos resultados encontrados durante a execução da sistemática. Foram definidos os seguintes atributos: “preço de fechamento”, “índice de sentimento”, “preço de abertura”, “índice de volatilidade”.

Os atributos “preço de fechamento” e “preço de abertura”, estão presentes nos dados financeiros coletados do site oficial da B3. Já o atributo “índice de volatilidade” será referenciado pelo medidor CBOE Brazil Etf Volatility (VXEWZ), que é um índice criado para calcular a volatilidade do índice Bovespa de forma anualizada. Apesar deste atributo não ter sido mencionado em nenhum trabalho durante a revisão sistemática, os trabalhos de (BOLLEN, MAO e ZENG, 2011) e (YANG, MO e LIU, 2015) verificaram que existe uma correlação entre a predição de bolsa de valores e o índice de volatilidade VIX, utilizado para verificar a volatilidade do mercado norte americano. O atributo “índice de sentimento” foi obtido na etapa anterior, ao realizar o cálculo das médias diárias da análise de sentimento dos *tweets*. Por fim foi criado um último atributo, chamado de “movimento”. Este atributo será o alvo do modelo de predição, e seus valores foram preenchidos com “1”, caso o preço de fechamento seja maior que o preço de abertura, ou “0” caso o preço de fechamento seja menor que o preço de abertura.

Visto que estas informações se encontram em três fontes de dados separadas, como forma de facilitar a manipulação destes dados optou-se pela criação de um novo *dataset* contendo as informações necessárias de forma centralizada. Este *dataset* foi armazenado em uma nova coleção no banco de dados MongoDB, a Figura 32 demonstra um documento com as informações importadas. Os campos “data\_pregao” e “ticker” foram mantidos para separação dos dados por empresas e período.

Figura 32 - Exemplo de documento do *dataset* final

Texto

Descrição gerada automaticamente

Fonte: elaborado pelo autor.

### Pré-processamento dos dados

Visto que algoritmos de SVM trabalham com a criação de hiperplanos e distância entre pontos, é necessário a aplicação de técnicas de normalização e escalonamento de dados, evitando assim que o modelo de predição realize inferências incorretas em relação ao peso dos valores (FACELI, *et al.*, 2011).

Optou-se pela realização do teste de Shapiro-Wilk para verificar a normalidade das distribuições dos dados dos atributos selecionados, apenas a coluna de dados “índice\_sentimento” apresentou um valor p > 0.05. Sendo assim neste atributo optou-se pela normalização *Z-Score*, aplicada através da classe *StandardScaler* do pacote *scikit-learn.* Os demais atributos foram normalizados com a técnica *MinMax,* aplicada pela classe *MinMaxScaler* do pacote *scikit-learn.* De acordo com os autores Han, Kamber e Pei (2012), ao aplicar uma normalização *Z-Score* é possível diminuir o impacto de *outliers* em distribuições normais. Já uma normalização *MinMax* preserva a escala entre os valores (HAN, KAMBER, PEI, 2020). Entre os trabalhos analisados durante a revisão sistemática, apenas o trabalho de Owen e Oktariani (2020) mencionou explicitamente qual a técnica utilizada durante a normalização nos dados, o mesmo aplicou a técnica de *MinMax* em todos os atributos coletados do mercado financeiro.

### Construção do modelo de predição

Inicialmente previu-se a construção de modelos de predição para cada ativo financeiro (PETR4, VALE3 e BOVA11) e mês (março/2019, março/2020 e março/2021) coletados. Visto que devido as limitações na quantidade de *tweets* analisados durante a etapa de análise de sentimentos, foram considerados apenas os dias úteis em que houve pregão na bolsa de valores. Sendo assim cada período possui uma quantidade pequena de dados, de uma média de 20 dias. Durante a execução inicial, estes dados foram verificados novamente através da visualização do *dataframe* gerado pela biblioteca *pandas,* e foi possível notar um desbalanceamento nos dados de 2020, onde por exemplo, para o código de ação PETR4, há apenas 6 registros onde o preço de fechamento foi maior que o preço de abertura.

Levando em consideração este desbalanceamento, optou-se por alterar a estratégia de criação dos modelos, desconsiderando o período de negociação. Desta forma, os modelos foram construídos utilizando os dados de todos os meses, ignorando o filtro de data no *dataset*. Assim é possível diminuir o víes do modelo de predição devido a pouca quantidade de dados.

Visto que normalmente a predição de bolsa de valores é tratada como um problema de série temporal, alguns autores como (BOUKTIF, FIAZ e AWAD, 2019) realizaram testes utilizando um *delay* para verificar o impacto da predição em um número de dias posteriores. Ao desconsiderar as datas neste trabalho, se cria uma limitação para realização destes experimentos. Após excluir o filtro de data, o novo *dataset* que será utilizado para construção dos modelos de predição possui 62 registros para cada ativo financeiro.

A construção do modelo de predição iniciou-se com a separação da classe alvo de predição dos atributos descritivos. A classe alvo de predição foi identificada pela variável “y” e os atributos pela variável “X”. Os autores Pang-Ning, Steinbach e Kumar (2005) sugerem que a técnica de validação de modelo *10-fold cross-validation* fornece uma melhor métrica de estimação de modelos em relação a utilização de uma divisão aleatória entre conjunto de treinamento e testes. O trabalho de Heckler (2018) sumariza a aplicação desta técnica da seguinte maneira:

Na técnica k-fold cross validation, os dados são divididos em k subconjuntos aleatórios com tamanho aproximado. Após isso, são realizados treinamento e teste k vezes. A cada vez, um subconjunto é utilizado para teste e os demais para treinamento. Desta forma, os subconjuntos são utilizados o mesmo número de vezes para treinamento e uma vez para teste. Um dos objetivos da divisão é manter a distribuição de classes do conjunto de dados original em cada subconjunto. Em geral, 10-fold cross validation é recomendada para estimar a acurácia de um modelo devido à baixa variação.

A biblioteca *scikit-learn* fornece meios de aplicar está técnica através da utilização da classe *KFold*. O parâmetro “*shuffle”* foi utilizado para realizar a divisão destes subconjuntos de forma aleatória, e com o objetivo de manter a replicabilidade dos testes, o parâmetro “*random\_state*” foi utilizado. Por fim, um classificador SVM foi instanciado e optou-se pela utilização do *kernel Radial basis function* (RBF). De acordo com Faceli *et al.* (2011), *o kernel* é uma fórmula matemática que define como será construida a linha do hiperplano. De acordo com os autores, o *kernel* RBF é uma boa escolha inicial.

### Resultados e análise

Ao todo foram criados seis modelos de predição utilizando um classificador SVM, um modelo para cada ação analisada, considerando ou não o índice de sentimento criado anteriormente. Como forma de avaliar os resultados, optou-se pela extração das métricas de acurácia, precisão, sensibilidade e F1-Score. O Quadro 10 demonstra os valores levando em consideração o índice de sentimento e o Quadro 11 demonstra os valores desconsiderando este atributo.

Quadro 10 - Métricas de avaliação dos modelos utilizando o índice de sentimento

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ativo** | **Acurácia** | **Precisão** | **Sensibilidade** | **F1-Score** |
| **BOVA11** | 54,76% | 54,83% | 55,83% | 52,90% |
| **PETR4** | 53,10% | 51,00% | 55,67% | 51,98% |
| **VALE3** | 48,24% | 51,17% | 58,00% | 46,93% |

Fonte: elaborado pelo autor.

Quadro 11 - Métricas de avaliação dos modelos desconsiderando o índice de sentimento

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ativo** | **Acurácia** | **Precisão** | **Sensibilidade** | **F1-Score** |
| **BOVA11** | 42,14% | 45,12% | 57,83% | 47,10% |
| **PETR4** | 58,26% | 60,67% | 60,83% | 56,33% |
| **VALE3** | 49,52% | 47,50% | 61,83% | 51,06% |

Fonte: elaborado pelo autor.

Os autores Bouktif, Fiaz e Awad (2019) citam que a tarefa de realizar predições do mercado financeiro é desafiadora devido aos preços serem afetados por diversos fatores externos. De acordo com os autores Carosia, Coelho e Silva (2019), apesar de haver um grande interesse dos pesquisadores em verificar a viabilidade da previsão do comportamento financeiro, atualmente a Hipótese do Mercado Eficente é a mais aceita. De acordo com ela, o preço de um ativo reflete de forma eficiente toda e qualquer informação já publicada, sendo assim o preço do amanhã irá levar em consideração apenas as informações publicadas amanhã, desconsiderando o passado (BING, CHAN e OU, 2014). Visto que a publicação de notícias é imprevisível, se torna difícil realizar predições com uma acurácia maior que 50% (BOLLEN, MAO e ZENG, 2011).

Ao levar em consideração os resultados acima, os modelos criados obtiveram resultados de acurácia geral levemente superiores ou inferiores a 50%, ao encontro da Hipótese do Mercado Eficiente. Se compararmos os modelos que utilizaram o índice de sentimento contra os modelos que desconsideraram este atributo, apenas o ativo BOVA11 apresentou uma melhora significativa em sua predição. Visto que o BOVA11 representa o índice completo do Ibovespa e a coleta de *tweets* foi realizada de forma geral, é possível que o mesmo seja mais sensível a opinião pública do que empresas em específico. O trabalho de Yang, Mo e Liu (2015) avaliou esta possibilidade ao testar ações de diferentes setores e foi possível perceber diferenças de desempenho em seus modelos de predição. De forma geral ao comparar os resultados deste trabalho com os trabalhos verificados durante a revisão sistemática, a precisão dos modelos é similar. Os trabalhos de Bathia, Hagras e Lepley (2018) e Bouktif, Fiaz e Awad (2019) obtiveram valores de acurácia similares a este trabalho. A Figura 33 demonstra o resultado obtido no trabalho de Bouktif, Fiaz e Awad (2019).

Figura 33 - Resultado dos modelos de predição do trabalho de Bouktif, Fiaz e Awad (2019).

Tabela

Descrição gerada automaticamente

Fonte: elaborado pelo autor.

Outro aspecto que podemos considerar nesta análise, é a possibilidade de que os *tweets* utilizados na construção do índice de sentimento sejam apenas reações a informações atuais. Uma forma de testar esta hipótese seria utilizar o índice de sentimento com um atraso em relação aos valores do mercado financeiro, porém este trabalho ficou limitado no número de dados obtidos para a predição. Sendo assim não foi possível testar esta hipótese.

De forma geral a etapa de aplicação de técnicas de ML no *dataset* construído apresentou limitações quanto a exploração de possibilidades, visto a baixa quantidade de registros para se trabalhar. Quanto a não utilização dos atributos de data, o problema analisado deixou de ser do tipo de série temporal, sendo assim não foi possível testar algumas hipóteses em relação a utilização de outros atributos do mercado financeiro. O trabalho de (OWEN e OKTARIANI, 2020), por exemplo avaliou a utilização de métricas utilizadas por profissionais do setor de análise técnica.

# CONCLUSÃO

O presente trabalho buscou avaliar a influência dos sentimentos dos investidores, expressos em *tweets*, nos preços de ativos da bolsa de valores brasileira, a Bovespa. Para isto foram empregadas técnicas de análise de sentimentos e técnicas de análise preditiva, por meio da utilização de algoritmos de *Machine Learning*.

Inicialmente foi executada uma revisão sistemática, com um escopo reduzido, para verificar o estado da arte em relação ao tema proposto neste trabalho, além de produzir referencial teórico para execução deste trabalho. A revisão sistemática teve como principal objetivo verificar as principais técnicas e algoritmos de *Machine Learning* utilizados. Também foi possível extrair outras informações relevantes que foram utilizadas na construção da metodologia deste trabalho, como por exemplo quais atributos foram selecionados na construção do modelo de predição, como os modelos fizeram uso da análise de sentimento dos investidores, além das métricas observadas na validação do modelo.

Apesar da revisão sistemática ter fornecido informações importantes para a execução deste trabalho, posteriormente foram identificadas limitações quanto a coleta dos dados necessários. Os trabalhos analisados utilizaram recursos próprios do Twitter para pesquisas de *tweets* relacionados à ativos financeiros, porém este recurso se encontra disponível apenas para ativos negociados no mercado norte americano. Para contornar esta limitação foi necessária a execução de uma pesquisa bibliográfica complementar para avaliar novas estratégias.

Com o resultado desta pesquisa complementar, foi possível iniciar a busca e coleta dos dados necessários. A estratégia de busca destes dados foi construída com base em uma metodologia já validada e aplicada no trabalho de (YANG, MO e LIU, 2015), no qual envolveu a identificação de uma comunidade de usuários com interesse no setor financeiro. Esta comunidade foi construída através da identificação de contas do Twitter de usuários considerados *experts* e seus seguidores. Seguindo esta metodologia foi possível realizar a coleta de uma grande quantidade de dados que foram analisados posteriormente.

Para a aplicação da etapa de análise de sentimentos foi necessário realizar o pré-processamento dos dados coletados, onde foi necessário a aplicação de diversas técnicas de processamento de linguagem natural. Através destas técnicas foi possível identificar entidades e substantivos relacionados ao mercado financeiro, presentes nos *tweets* enviados pelos usuários. Com a identificação destes termos foi possível desenvolver estratégias para filtragem dos dados coletados, onde procurou-se excluir *tweets* que não possuíam relação ao mercado financeiro, visando assim obter um resultado mais fiel a realidade durante a etapa de análise de sentimentos.

Durante a execução da etapa de análise de sentimentos dos *tweets* coletados verificou-se novamente limitações. Como o objetivo deste trabalho não era o desenvolvimento de ferramentas para aplicação de análise de sentimentos, optou-se por utilizar uma ferramenta já desenvolvida e testada, visto que parte dos trabalhos analisados durante a revisão sistemática utilizaram esta mesma estratégia. Entretanto não foi possível encontrar uma ferramenta disponível no idioma português brasileiro. Desta maneira foi proposta a utilização de uma ferramenta paga, com a utilização de créditos grátis para novos usuários. Optou-se pela utilização da ferramenta *Google Cloud Natural Language.* Com a utilização desta ferramenta a quantidade de *tweets* analisados foi limitada, sendo analisados no total 291.000 registros.

Com a limitação de análise dos *tweets*, optou-se por avaliar o mesmo período em três anos diferentes. Os resultados encontrados durante a execução da análise de sentimentos foram interessantes, foi possível avaliar o desempenho da análise utilizando como métrica um acontecimento externo, que neste caso foi a pandemia de Covid-19. Os *tweets* postados neste período foram claramente menos positivos ao realizar comparações com o mesmo período do ano anterior e posterior.

Por fim foram aplicadas técnicas de construção de modelos preditivos utilizando o algoritmo classificador SVM. A escolha por este algortimo se deu pelos resultados da análise sistemática, o mesmo foi o mais utilizado nos trabalhos analisados. A aplicação da análise de sentimento no modelo se deu através da criação de um índice diário com a média dos resultados obtidos na etapa de classificar o sentimento dos *tweets*. Todos os trabalhos analisados durante a revisão sistemática utilizaram esta mesma estratégia, construindo um índice de sentimentos.

Devido a limitação encontrada anteriormente, onde só foi possível utilizar uma pequena quantidade dos *tweets* coletados, a quantidade dados utilizados para construção do modelo de predição foi baixa, limitando assim uma análise mais aprofundada do impacto do sentimento relacionado ao *tweets* nos preços de ativos do mercado financeiro. A precisão dos modelos avaliados não foi satisfatória e de certa forma corrobora com a Hipótese do Mercado Eficiente, de que os preços dos ativos sempre refletem todas as informações já disponíveis, não sendo possível realizar predições utilizando dados passados.

Através da execução deste trabalho foi possível verificar que o campo estudado vem gradativamente recebendo mais atenção com o passar dos anos. Sendo assim há diversas possibilidades para trabalhos futuros. Em relação a este trabalho, o mesmo poderia ser continuado através da aplicação de análise de sentimentos em uma quantidade maior de *tweets*, obtendo-se assim uma maior quantidade de dados para se utilizar durante as etapas de aplicação de algoritmos de *Machine Learning*. O trabalho dos autores Souza, Nogueira e Lotufo (2020) produziu um modelo BERT pré-treinado no idioma português, que pode ser utilizado para diversas atividades de processamento de linguagem natural, como extração de *features* e análise de sentimentos.

Quanto aos algoritmos e técnicas utilizadas na construção dos modelos de predição, pode-se explorar novas opções como forma de realizar comparações com os modelos de SVM construídos neste trabalho. Visto que o problema de predição de bolsa de valores é tratado como um problema de série temporal, algoritmos de redes neurais recorrentes se mostram um candidato interessante neste aspecto. Também é possível explorar a aplicação de atributos financeiros de análise técnica na construção dos modelos de predição e comparar o poder preditivo de diferentes ações e setores

Em relação ao *dataset* construído durante a execução deste trabalho, não foi possível encontrar uma plataforma para compartilhá-lo, devido ao seu grande volume de dados. Sendo assim futuramente pode-se buscar um local e uma estratégia para divulgá-lo com a comunidade cientifica.

Referências Bibliográficas

ALVES, D. S. **Uso de técnicas de Computação Social para tomada de decisão**. UNIVERSIDADE DE BRASÍLIA. Brasília. 2015.

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

BALCÃO, B. B. B3 divulga estudo sobre os 2 milhões de investidores que entraram na bolsa entre 2019 e 2020. **B3**, 2020. Disponivel em: <http://www.b3.com.br/pt\_br/noticias/investidores.htm>. Acesso em: 07 Março 2021.

BATHIA, A.; HAGRAS, H.; LEPLEY, J. J. **Machine Learning Approach to Extracting Emotions Information from Open Source Data for Relative Forecasting of Stock Prices**. 2018 10th Computer Science and Electronic Engineering (CEEC). [S.l.]: IEEE. 2018.

BING, L.; CHAN, K. C. C.; OU, C. 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. 232-239.

BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. **Journal of Computational Science**, 2011. p1-8.

BOUKTIF, S.; FIAZ, A.; AWAD, M. Stock Market Movement Prediction using Disparate Text Features with Machine Learning. **2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)**, Outubro 2019.

CAROSIA, A.; COELHO, G. P.; SILVA, A. E. A. D. S. The Influence of Tweets and News on the Brazilian Stock Market through Sentiment Analysis. **Proceedings of the 25th Brazillian Symposium on Multimedia and the Web**, 2019.

CHOU, C.; PARK, J.; CHOU, E. Predicting Stock Closing Price After COVID-19 Based on Sentiment Analysis and LSTM. **2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)**, Março 2021.

FACELI, K. et al. **Inteligência Artificial:** Uma Abordagem de Aprendizado de Máquina. 1ª. ed. [S.l.]: LTC - Livros Técnicos e Científicos Editora Ltla., 2011.

GANDOMI, A.; HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. **International Journal of Information Management, 35**, Abril 2015. p137-144.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining:** Concepts and Techniques. [S.l.]: Morgan Kaufmann, 2012.

HECKLER, W. F. Análise Preditiva Sobre Pacientes do "Projeto de Extensão Reabilitação Pulmonar" da Universidade Feevale, Novo Hamburgo, 2018.

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

IEEEXPLORE. About IEEE Xplore. **IEEEXplore**. Disponivel em: <https://ieeexplore.ieee.org/Xplorehelp/overview-of-ieee-xplore/about-ieee-xplore/>. Acesso em: 17 Abril 2021.

JUNQUEIRA, K. T. C.; FERNANDES, A. Análise de Sentimento em Redes Sociais no Idioma Português. **IX Computer on the Beach** , 2018.

KANSAON, D. P.; BRANDÃO, M. A.; PINTO, S. A. D. P. P. Análise de Sentimentos em ​Tweets​ em Português Brasileiro. **BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM)**, Porto Alegre, 2018.

KITCHENHAM, B. **Guidelines for performing Systematic Literature Reviews in Software Engineering**. University of Durham. Durham, UK. 2007.

LEHIKOINEN, J.; KOISTINEN, V. In big data we trust? **Interactions**, v. 21, n. 5, p. 38-41, Setembro 2014.

LI, M. et al. Stock market analysis using social networks. **ACSW '18: Proceedings of the Australasian Computer Science Week Multiconference**, 2018. p1-10.

LIBRARY, A. D. An Archive of Original Research. **ACM Digital Library**. Disponivel em: <https://www.acm.org/publications/digital-library>. Acesso em: 17 Abril 2021.

LIU, B. **Sentiment Analysis and Opinion Mining**. [S.l.]: Morgan & Claypool Publishers, 2012.

MEDEIROS, E. R. D. **Revisão Sistemática Sobre os Dispositivos Vestíveis na Área da Saúde**. Universidade Feevale. Novo Hamburgo. 2016.

MITCHELL, T. **Machine Learning**. 1ª. ed. Nova Iorque, EUA: McGraw-Hill, Inc., 1997.

OWEN, L.; OKTARIANI, F. SENN: Stock Ensemble-based Neural Network for Stock Market Prediction using Historical Stock Data and Sentiment Analysis. **2020 International Conference on Data Science and Its Applications (ICoDSA)**, Agosto 2020.

PANG-NING, T.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. 1ª. ed. [S.l.]: Pearson Education, 2005.

PARK, J.; MA, K.; LEUNG, H. Prediction of Stock Prices with Sentiment Fusion and SVM Granger Causality. **2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)**, Agosto 2019.

PASUPULETY, U. et al. Predicting Stock Prices using Ensemble Learning and Sentiment Analysis. **2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)**, Junho 2019.

PEREIRA, D. A. A survey of sentiment analysis in the Portuguese language. **Artifcial Intelligence Review**, 2020.

RADEMAKER, A. et al. Universal Dependencies for Portuguese. **Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)**, p. 197-206, Setembro 2017.

SIRIMEVAN, N. et al. Stock Market Prediction Using Machine Learning Techniques. **2019 International Conference on Advancements in Computing (ICAC)**, Maio 2019.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. **9th Brazilian Conference, BRACIS 2020**, Outubro 2020.

STATISTA. **Statista**, 2021. Disponivel em: <https://www.statista.com/topics/7863/social-media-use-during-coronavirus-covid-19-worldwide/#dossierKeyfigures>. Acesso em: 15 Outubro 2021.

TRELEWICZ, J. Q. Big Data and Big Money: The Role of Data in the Financial Sector. **IT Professional**, v. 19, n. 3, p. 8-10, Junho 2017.

YANG, S.; MO, S.; LIU, A. Twitter financial community sentiment and its predictive relationship to stock market movement. **Quantitative Finance**, 2015. 1637-1656.

YAQOOB, I. et al. Big Data: From Beginning to Future. **International Journal of Information Management**, Dezembro 2016. p1231–1247.

YE, M.; LI, G. Internet big data and capital markets: a literature review. **Financial Innovation**, Março 2017. 6.

ZHANG, X.; FUEHRES, H.; GLOOR, P. A. Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear”. **Procedia - Social and Behavioral Sciences**, 2011. p55-62.

1. https://github.com/twintproject/twint [↑](#footnote-ref-1)
2. https://github.com/tweepy/tweepy [↑](#footnote-ref-2)
3. https://spacy.io/ [↑](#footnote-ref-3)
4. https://www.b3.com.br/pt\_br/market-data-e-indices/servicos-de-dados/market-data/historico/mercado-a-vista/cotacoes-historicas/ [↑](#footnote-ref-4)
5. https://scikit-learn.org/stable/ [↑](#footnote-ref-5)