**1 Introdução**

O desejo de ganhar dinheiro em bolsas de valores não é algo novo, porém com o avanço da tecnologia isso se tornou cada vez mais acessível para pessoas comuns, mas nem sempre comprar e vender ações foi algo simples. Demarco (2016) conta que, por causa da unificação das 7 bolsas de valores na bolsa de São Paulo no ano de 1999, os operadores se concentraram em um canto só, causando muito tumulto e desentendimentos, visto que naquela época as negociações ainda aconteciam por viva voz, ou seja, as negociações ocorriam tudo pessoalmente, as ofertas feitas pelo auto falante, e os negócios fechados ali mesmo, boca-a-boca.

A automação eletrônica no mercado de capitais como um todo, teve início somente a partir da década de 70 e, no Brasil, somente a partir da década de 90 (construcaoRoboInvestMestrado). A evolução aconteceu aos poucos, os papéis eram negociados no “viva voz” e no sistema eletrônico ao mesmo tempo, e apenas em 2005 o pregão por viva voz foi desativado e deixou de existir, segundo (DEMARCO, 2016).

Com esse avanço da tecnologia e a popularização do função de *Trader,* um novo mecanismo para investimento tem ganhado força e relevância, os *Expert Advisors* (robôs investidores) , eles são algoritmos que usam de várias áreas de conhecimento, seja da Inteligência Artificial (IA) ou de modelos econométricos, para tentar prever o valor dos ativos de renda variável de determinada bolsa de valores, como a B3 do Brasil, por exemplo.

Predizer o comportamento da bolsa de valores é um diferencial estratégico que pode representar ganhos significativos (LIMA, 2016). É pensando nisso que este trabalho foi desenvolvido, com o intuito de antecipar os acontecimentos da bolsa de valores do Brasil, e auxiliar o *Trader* na compra de Mini Contratos Dólar Futuro (WDO), utilizando da análise de séries temporais nos dados históricos do dólar e análise de sentimentos no Twitter (livro/autor), afim de achar uma relação entre o sentimento coletivo e o comportamento da moeda brasileira em relação ao da moeda americana.

* 1. **Justificativa**

Falar dos problemas de pesquisa aqui

não esquecer de falar da HME, com o objetivo de refutá-la, pois ela quem é a maior oposição a esta área de pesquisa.

Como já falado anteriormente, a evolução tecnológica e relacionado as negociações de ativos tanto na compra, quanto na venda, aconteceram entre o final do século passado e começo dos anos 2000, a presença da computação neste meio é ainda mais recente. A popularização do uso de algoritmos para negociação foi um evento ainda mais recente, pois somente nos últimos anos é que corretoras de valores começaram a disponibilizar plataformas Home Brokers para seus clientes (construcaoRoboInvestMestrado).

Comprar por um preço e vender por outro mais alto, este é o princípio básico no mundo das ações. Mas a escolha de um ativo seguro e lucrativo nunca é uma tarefa simples. Na maioria das vezes, a emoção é que toma a decisão pelo investidor (estrategiaAnaliseFundamentalisticaTecnica). Isso acaba por refletir em más escolhas e resulta em prejuízo financeiro, pois o usuário perde a capacidade de se antecipar aos outros Trades, e garantir a melhor oferta para si, caso ele possuísse uma ferramenta computacional para auxilia-lo, talvez o cenário fosse diferente, é aqui que entra a motivação para o uso de algoritmos inteligentes . De acordo com (mestradoAnaliseSentimentosSVMRedesSociais) um dos principais motivos é a vantagens que uma ferramenta de predição com grandes índices de acerto representaria para o seu detentor ou manipulador frente aos demais negociadores.

Já existem pesquisas e estudos nessa área, em relação a análise de séries temporais a tese de mestrado (construcaoRoboInvestMestrado) mostra duas frentes, os preditores econométricos e as redes neurais para a construção do seu robô investidor, obtendo bons resultados e lucro no final, agora sobre a análise de sentimentos, pode-se citar a tese de mestrado do (mestradoAnaliseSentimentosSVMRedesSociais), ele fez um trabalho muito parecido com o que será proposto neste, explorando duas frentes: Mineração de Opinião e análise de séries financeiras. Ele realiza a coleta de dados no Twitter, e tenta estabelecer uma relação entre o sentimento coletivo e o comportamento financeiro da empresa Petrobrás (PETR4), no segundo ponto é utilizado a Máquina de Vetor de Suporte (SVM) para saber se o valor de fechamento do ativo no dia atual será de alta ou baixa, em relação ao dia anterior.

A principal barreira encontrada é o fato de que o mercado financeiro é muito volátil, e muito sensível a mudanças. (Forecasting500StocksIndexFuturesHybridSystem) fala que muitos fatores influenciam o mercado financeiro, incluindo eventos políticos, condições econômicas gerais e expectativas dos investidores. Por esse fator de imprevisibilidade, que a Hipótese do Mercado Eficiente (HME) é tão forte, proposto por (FAMA, 1970), onde o mesmo defende que os preços dos ativos refletem as informações disponíveis no mercado, onde os indivíduos nela são racionais e buscam o lucro baseado em informações disponíveis igualmente a todos, assim o mercado tende a se manter em equilíbrio, pois sempre existirá compradores e vendedores. No trabalho de (hipoteseDeMercadoEficiente), onde é mostrado a evolução do estudo e testes da HME, fala que o mercado é considerado eficiente se refletisse rapidamente qualquer informação nos preços dos ativos, impossibilitando ganhos anormais, pois de acordo com a própria HME ações sempre são negociadas a um preço justo, impedindo a venda delas por preços inflacionados ou a compra em um estado super desvalorizado. Com todos esses pontos apresentados, e considerando que todos eles sejam sempre uma verdade, seria impossível criar mecanismos automatizados que encontrassem padrões nas séries financeiras, onde analisando dados do passado, seria possível antecipar os acontecimentos do futuro.

O primeiro objetivo é refutar a Hipótese do Mercado Eficiente, criando um mecanismo que antecipe os acontecimentos do mercado financeiro e que sirva para auxiliar o usuário a minimizar as incertezas em relação a imprevisibilidade do mesmo. O segundo objetivo é mostrar que é possível obter bons resultados na compra de ativos, usando Expert Advisors como conselheiros, ou seja, desenvolver um Robô que justifique o *Trader* a colocá-lo como sendo uma de suas principais fonte de pesquisa, na hora de movimentar um contrato de mini dólar.

Modelos que buscam modelar com rigor estatístico a volatilidade dos retornos levando em conta características importantes como assimetria e heterocedasticidade condicional só foram introduzidos na década de 80. (construcaoRoboInvestMestrado)

Com muita pesquisa(muita aonde? Como? Citar vários artigos seguidos aqui pra provar oq eu estou falando) foi observado que na literatura existem muitos estudos sobre a previsão de séries temporais, principalmente voltada ao mercado financeiro, também foi constatado bastante material sobre análise de sentimentos envolvendo o Twitter, também voltado ao mercado financeiro, como, por exemplo, o preço dos ativos de empresas como a Petrobras (citar os artigos que falam e usam os ativos da Petrobras como dados históricos), mas acima de tudo notou-se bastante artigo sobre a previsão da taxa de câmbio de países emergentes em relação a grandes potências (citar os artigos que falam disso, aqueles em inglês, eu acho). Porém não foi encontrado no levantamento bibliográfico feito até então, um método que usa as duas áreas do conhecimento, ou até mesmo desenvolvido um robô investidor onde ele utilizasse destas duas frentes de pesquisa com a finalidade de ter uma base maior, para auxiliar o investidor na hora da compra e da venda de um ativo.

* 1. **Problema de pesquisa**

(apresentar o problema de pesquisa, falando da não-linearidade dos dados, e de quantas variantes existem que podem influenciar no preço do dólar, )

A literatura relacionada à previsão de séries temporais tem registrado, desde a década de 90, importantes avanços relacionados à incorporação de novas metodologias que tentam determinar padrões de relacionamentos presentes nos dados do mercado financeiro

(Podemos prever a taxa de cambio brasileira? Evidência empírica utilizando inteligência computacional e modelos econométricos)

**1.3 Objetivo Geral e Específico**

Objetivo Geral

Objetivos Especificos

Previsão de séries temporais usando dados históricos

Análise de sentimentos no twitter sobre fatos políticos para previsão do preço do dólar

Construir um Expert Advisor para operação swing trade com mini-dolar

Esta pesquisa tem como principal objetivo o desenvolvimento de um Robô-Investidor para prever a tendência do mercado financeiro, ou seja, a movimentação de preços do mesmo, e assim auxiliar na compra e venda de Mini Contratos Dólar Futuro (WDO). Será utilizado três áreas de pesquisa: A análise de dados históricos do preço do dólar, mediante *machine learning,* análise de sentimentos em dados extraídos da rede social *Twitter* e do estudo dos dados estatísticos que as séries temporais das duas primeiras partes irão gerar.

Com isso pode-se listar os seguintes objetivos específicos:

Fazer o levantamento bibliográfico, e analisar os trabalhos relacionados disponíveis;

Observar quais as técnicas mais usadas nestas três áreas de pesquisa para o problema de previsão de preços dos ativos de uma bolsa de valores;

Escolher qual a técnica de aprendizagem de máquina para ser usada na análise dos dados históricos;

Determinar qual é o melhor modelo econométrico para a análise de sentimentos no *Twitter*;

Selecionar qual os princípios estatísticos para o estudo dos gráficos gerados pelas bases de dados em questão;

Executar e validar o robô com dados reais do mercado financeiro.

**1.4 Metodologia**

Esta pesquisa vai abranger algumas áreas, dentre elas a Mineração de Dados no Twitter, também conhecido como Mineração de Opinião, da análise de séries temporais, ou análise de séries financeiras para este caso, e análise de dados estatísticos.

A primeira parte consiste em fazer duas coletas de dados:

a primeira são os dados históricos do dólar, para isto foi escolhido a plataforma Metatrader 5, nela há possibilidade de exportar os dados que serão usados para análise;

a segunda coleta de dados será feita na plataforma Twitter, foi usado a API própria do Twitter para isso, nela serão tirados os textos que falam sobre o dólar, como postagens de usuários, postagem das páginas de jornais online, entre outros. Todas as pesquisas serão em português, pois o objetivo é a operação dólar em relação ao real.

A segunda parte é o tratamento dos dados, em relação as séries financeiras, é necessário, antes de aplicá-los aos algoritmos, a retirada dos ruídos, dos outliers e outras possíveis inconsistências. Para os textos alguns tratamentos são necessários de serem feitos, por exemplo em um texto publicado no Twitter existem muitas palavras desnecessárias, e que só atrapalhariam o desempenho dos algoritmos, principalmente se for uma publicação de um usuário, onde contem coisas como como gírias, abreviações, caracteres especiais, é preciso tirá-las e deixar apenas o grupo de palavras principais, que atribuem um significado à publicação.

Para a parte de análise de dados, é necessário escolher uma técnica para determinada atividade, em relação a previsão de séries temporais, foi escolhido a Máquina de Vetor de Suportes

Em mineração de opinião, depois de se extrair o conteúdo da rede social em questão, é necessário um analisador léxico para a classificação dos textos

(falar da api do google translate, para traduzir os textos em portugues)

(falar da técnicas escolhidas, analisador léxico e da SVM)

(falar sobre o tipo de ativo que será investido, o pq dessa escolha)

(falar sobre a predição de series temporais)

(falar do fator emoção e de como isso prejudica na maximização dos ganhos)

(falar sobre a analise de sentimentos no twitter)

1. **Fundamentação Teórica**
   1. O mercado financeiro
   2. Support Vector Machine– SVM
   3. Redes Sociais ????
   4. Mineração de Dados
   5. Análise de Sentimentos de Processamento de Linguagem Natural

**2.2 Aprendizagem Supervisionada**

Na aprendizagem supervisionada, o objetivo é aprender um mapeamento a partir da entrada para uma saída onde os valores corretos são fornecidos por um supervisor (**alppaydin\_machinelearning\_2010**), isso quer dizer que o algoritmo é “treinado” sobre um conjunto de dados rotulados (conhecidos como dados de treinamento), e depois receber novos dados sem rótulos e tomar decisões sobre eles, ou seja, atribui-los aos rótulos previamente existentes dentro do programa.

O livro (**alppaydin\_machinelearning\_2010**) usa uma situação com carros para exemplificar a aprendizagem supervisionada, o exemplo é o seguinte: O objetivo é aprender uma classe *C* de uma família de carros. Tem-se um conjunto de carros e um grupo de pessoas a qual será mostrado estes veículos, as pessoas olham os carros e os rotulam, os que eles consideram ‘carros de família’ são classificados como *exemplos positivos*, e os outros como *exemplos negativos.* Aprender uma classe é achar uma descrição que se repete em todos os e*xemplos positivos,* ou seja, uma ou várias característicasem comum entre eles, e que não aparece nos *exemplos negativos.* Agora é o momento da predição, pegando um carro nunca antes visto e comparando com a descrição já memorizada, será dito se ele é um carro de família ou não.

**2.2.1 Algoritmos para classificação**

Algoritmos usados para classificação também são do tipo supervisionado, o próprio exemplo anteriormente mostrado é um problema de classificação, ou seja, classificar é o ato de atribuir rótulos aos novos dados recebidos baseado em dados rotulados previamente analisados. Algoritmos de classificação normalmente trabalha com saídas binárias, ou seja, 0 (zero) para ‘não’ e 1 (um) para sim.

O exemplo que o livro (**alppaydin\_machinelearning\_2010)** mostra é uma situação com bancos, onde eles emprestam dinheiro com juros (crédito). Para um banco realizar um empréstimo, ele precisa avaliar os riscos e a probabilidade do cliente não o pagar, para isso ele toma informações sobre o cliente para determinar sua capacidade financeira, entre esses dados estão: renda, poupança, garantias, profissão, idade, histórico financeiro, passado entre outros.

O exemplo acima é um problema de classificação com duas classes existentes: clientes de *baixo risco* e *alto risco.* As informações dos clientes (renda, poupança...) são as entradas do algoritmo, e a saída do mesmo será os clientes rotulados em uma das duas classes.

**2.2.1.1 Classificador Naive Bayes**

Um exemplo de algoritmo para classificação é o Naive Bayes, normalmente é usado em classificar textos, como definir se um e-mail é um *spam* ou não, classificar emoções em um texto como positiva ou negativa, entre outros.

Segundo o livro (**alppaydin\_machinelearning\_2010)** as entradas do Naive Bayes são independentes, poque ele ignora possíveis dependências ou correlações entre as variáveis de entrada, além disso ele reduz um problema multivariado (várias variáveis) a grupo de problemas univariados (de uma variável). Por exemplo em uma situação para classificar frutas, o algoritmo recebe uma maçã, mas os valores de entrada recebidos são a cor vermelha, formato arredondado e superfície lisa, o Naive Bayes simplesmente irá ignorar qualquer correção entre essas características e realizar seus cálculos.

(**alppaydin\_machinelearning\_2010)** mostra a equação do Naive Bayes como:

P(C|x) = P(C)p(x|C)

P(x)

Onde x é a variável de entrada e *C* uma variável multinominal que toma um dos estados K para o código de classe.

**2.2.1.2 Decision Tree**

18.3.1 Representação da árvore de decisão Livro inteligência artificial

**2.2.1.3 Support Vector Machine (SVM)**

**(**kernel machine**)**

Criada por (VAPNIK at all, 2003), foi desenvolvido originalmente para problemas de classificação, maximizando a margem entre os dados de treinamento e o hiperplano criado, com o intuito de classificar dados linearmente separáveis, porém (**precificacaoDeOpcoesPorAcoesTeseMestrado**) conta que em 1995 foi introduzido para a SVM variáveis de folga, com isso se tornou possível encontrar um hiperplano de separação com margens flexíveis que minimiza os erros em casos de dados não linearmente separáveis

**(alppaydin\_machinelearning\_2010)** fala que nos últimos anos algoritmos baseados em kernel, como a SVM, se tornaram muito populares, pois através do uso de funções de kernel podem ser adaptados à vários tipos de problemas, como na bioinformática e no Processamento de Linguagem Natural (NLP)**.** Eletambém explica que a SVM é um modelo linear, porém em casos de problemas não lineares pode-se mapear o problema para um novo espaço, fazendo uma transformação não linear usando funções básicas escolhidas previamente, então usa-se um modelo linear nesse novo espaço.

Um bom exemplo para explicar esses conceitos é uma situação em que se precise classificar o gênero de uma pessoa, o algoritmo recebe os dados de entrada daquele individuo (altura, peso, entre outros) e precisa determinar se é um homem ou uma mulher, com isso um hiperplano será criado para separar os dados em duas classes (problema de duas classes), para então rotulá-los em uma dessas duas.

**2.2.2 Algoritmos para regressão**

**2.2.2.1 Support Vector Regression (SVR)**

Kernel Machines for Regression no livro alppaydin\_machinelearning\_2010

**2.4 Métricas**

**2.2 Support Vector Machine – SVM**

Criada por (VAPNIK at all, 2003), foi desenvolvido originalmente para problemas de classificação, maximizando a margem entre os dados de treinamento e o hiperplano criado, com o intuito de classificar dados linearmente separáveis, porém (**precificacaoDeOpcoesPorAcoesTeseMestrado**) conta que em 1995 foi introduzido para a SVM variáveis de folga, com isso se tornou possível encontrar um hiperplano de separação com margens flexíveis que minimiza os erros em casos de dados não linearmente separáveis.

Agora para mostrar a aplicação da SVM em séries financeiras, temos o trabalho do (**FinancialTimeSeriesForecastingSVM**), ele a usou para prever mudanças nos preços do índice das ações da Coréia do Sul (KOSPI), como é um algoritmo de classificação, o uso foi da seguinte forma, o autor fez uma previsão do valor do dia seguinte em relação ao dia atual, ou seja, 0 (zero) se o índice do próximo dia é menor que o do dia atual e 1 (um) se o índice do dia seguinte for maior que o do dia atual. Foram usados dados de 2928 dias, entre os anos de 1989 e 1998. Ele observado que a SVM é bastante sensível aos parâmetros de limite superior C e a função de kernel, e que a performance da previsão é dependente desses dois valores. Para os resultados foi comparado a SVM com as Redes Neurais Artificiais (RNA´s) e ao Raciocínio Baseado em Casos (CBR), a SVM foi superior aos outros dois por se sair melhor na tarefa de generalização, se mostrando uma ótima alternativa para a previsão de séries financeiras.

(botar no rodapé o livro sobre machine learning para mostrar um canto que fala sobre a SVM)

**2.2.1 Support Vector Regression – SVR**

A SVM não serve apenas para problemas de classificação, (**precificacaoDeOpcoesPorAcoesTeseMestrado**) menciona que no ano de 1995 ela é introduzida aos casos de regressão através do acréscimo de uma função de perda, passando a ser chamada de *Support Vector Regression* quando aplicada para esse tipo de problema.

Para mostrar a aplicabilidade desse algoritmo no mercado de ações, temos a tese de mestrado da (**precificacaoDeOpcoesPorAcoesTeseMestrado**) onde é usado a SVR para determinar preços teóricos de contratos de opções, neste trabalho foi analisado os dados históricos referentes às opções de compra americana sobre a ação da Petrobras (PETR4) negociada na BOVESPA no período entre novembro de 2008 até maio de 2009. Foram escolhidos 11 parâmetros de entrada para o algoritmo, entre elas o preço do exercício da opção, o tempo em dias até o vencimento da opção, taxa de juros livre de risco (a taxa SELIC), entre outros. Foram criados 84 modelos da SVR com distinções entre as variáveis de entrada, valor constante de regularização C, valor da margem ε, tipo de função kernel e respectivos parâmetros, para então escolher o modelo com melhor desempenho, o interessante é que o modelo m14 foi o escolhido, sendo que ele exerceu a média da quantidade de vetores de suporte, porém isso não afetou a sua capacidade de generalização. Para os resultados foram comparados os preços teóricos obtidos pelas SVR´s com os preços reais na BOVESPA. A SVR se mostrou superior ao Black & Scholes (B&S) (um modelo matemático já existente na análise no mercado e ações) em quase todas as previsões.

A autora do trabalho anterior também escreveu um artigo (SVRvsRNAPrecificacaoPrecoOpcoes), onde ela comparou o desempenho da SVR com as RNA´s. Para a base de dados também foi usado os dados históricos referentes às opções de compra americana sobre a ação da Petrobras (PETR4) no mesmo período, mantidos os mesmos dados de treinamento e de teste para os dois algoritmos. Para a escolha dos modelos foi usado o melhor modelo da SVR do trabalho anterior, ajustada da seguinte forma, as variáveis de entrada: preço da ação, preço de exercício e tempo em dias até o vencimento da opção; já os parâmetros: constante de regularização C=100, margem ε =0,05, função kernel polinomial não homogêneo com grau p=5 e constante k=1; a variável de saída foi o preço teórico da opção. Agora para a RNA foram criados 450 modelos que tinham como padrão a arquitetura MLP com uma camada escondida e o algoritmo de treinamento *backpropagation,* o número de neurônios variava entre 1 e 10, além de outros ajustes específicos, as variáveis de entrada e saída foram as mesmas da SVR, com isso foi escolhido, entre esses 450, o modelo que melhor captou o comportamento das opções. Os resultados mostraram a superioridade da SVR em relação a RNA em 12 das 15 séries avaliadas, mostrando sua capacidade de acompanhar a alta oscilação do mercado.

**2.2 Análise de Sentimentos**

. Processamento de linguagem natural no livro Inteligencia artificial

Estrutura do TCC

<https://comoelaborarumtcc.net/estrutura-fundamental-do-tcc/>

**Referências Bibliográficas**

ABHYANKAR, A.; COPELAND, L.; WONG, W. Nonlinear dynamics in real-time equity markets indices: evidence from United Kingdom. The Economic Journal, v. 105, n. 431, p. 864-880, 1995.

DEMARCO, A. E. A gritaria acabou: Do pregão viva-voz à negociação eletrônica. 2016. Disponível em: <http://vemprabolsa.com.br/2016/03/18/gritaria-acabou-do-pregao-viva-voz-negociacao-eletronica/>. Acesso em: 01/02/2020. Citado na página ?.

LIMA, Milson Louseiro. Um modelo para predição de bolsa de valores baseado em mineração de opinião. 2016. 113 f. Dissertação (Mestrado em Engenharia) - Universidade Federal do Maranhão, São Luís, 2016.

FAMA, E. F. Efficient capital markets: A review of theory and empirical work\*. The journal of Finance, Wiley Online Library, v. 25, n. 2, p. 383–417, 1970

BOSER, B. E.; GUYON. I. M.; VAPNIK. V. N. A Training Algorithm for Optimal Margin Classifiers. In: ANNUAL WORKSHOP ON COMPUTATIONAL LEARNING THEORY, 5.,1992, Pittsburgh. Proceedings of the fifth annual workshop on Computational learning theory. New York: ACM, 1992. p. 144-152

Existem diversas estratégias para se operar no mercado financeiro e uma delas é a chamada arbitragem. Segundo Siegel (2009), a arbitragem consiste em avaliar as discrepâncias temporárias nos preços de séries de bens que sejam idênticos (como é o caso do dólar cheio com o mini dólar) ou bens que estejam altamente correlacionados, como papéis da mesma empresa (PETR3 e PETR4, por exemplo). Caso o valor de um bem caia ou aumente suficientemente em relação ao seu par entende-se que uma oportunidade de operação no mesmo sentido exista no ativo que se está operando.

#Percentage of Change in Direction (POCID)

Buy and Hold

**Uma justificativa para usar um modelo linear (a SVM) para tratar de dados não lineares**

SVM uses linear model to implement nonlinear class boundaries through some nonlinear mapping the input vectors x into the high-dimensional feature space. A linear model constructed in the new space can represent a nonlinear decision boundary in the original space.

**Sobre a capacidade de generalização de um SVM / SVR**

Além do cálculo do EQRM, foi levado em consideração o número de vetores suporte gerados por cada modelo de precificação. Pois, um número pequeno de vetores suportes representa uma alta capacidade de generalizar os dados, enquanto o contrário representa uma generalização pobre, ou seja, o overffiting

O Means Square erro é usado nos trabalhos da

Beltrami

No trabalho do criador da SVM

**Uma idéia**

Ainda, sugere-se como alternativa ao uso da ferramenta computacional LINGO a resolução do problema quadrático do SVR por meio do algoritmo Sequential Minimal Optimization (SMO), o qual vem sendo amplamente empregado na resolução de problemas do Support Vector Machine.

metradoASEconometricosAutomaticoNaoSupervisionado compara o léxico que ele criou com o NRC Hashtag e o Sentiment140