**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, pois naquela época as negociações ainda aconteciam por viva voz, as ofertas feitas pelo auto falante, e os negócios fechados ali mesmo, boca-a-boca.

Apenas na década de 90 que a negociação eletrônica foi implementada na Bovespa, e apenas em 2005 que as negociações via pregão viva voz de ações foram encerradas e quatro anos depois de contratos derivativos como conta (historiaBolsa). Um processo lento e demorado que levou vários anos, (DEMARCO, 2016) fala que os papéis eram negociados no “viva a voz” e no sistema eletrônico ao mesmo tempo, e apenas em 2005 o pregão por viva voz foi desativado e deixou de existir.

**Com esse avanço da tecnologia e a popularização da função de *Trader,* um novo mecanismo para investimento tem ganhado força e relevância, os robôs investidores, eles são algoritmos automatizados que usam de várias áreas de conhecimento, seja da Inteligência Artificial e modelos econométricos ou análise de sentimentos para prever o valor dos ativos de renda variável em uma bolsa de valores.**

A medida que a tecnologia passou a ser mais frequente no mundo dos ativos, começaram a surgir mecanismos automatizados para investimentos, ou seja, algoritmos que usam a Inteligência Artificial (**livroInteligenciaArtificialMiguel**) para fazer previsões do preço das ações, existem vários termos para se referir a este tipo de tecnologia, um deles é o High Frequency Trading (HFT).

HFT, que em português significa negociação de alta frequência, é o uso de negociação algorítmica para operações de mercado como commodities e ativos financeiros no geral, já a negociação algorítmica é o uso de algoritmos computacionais (citado anteriormente como robô investidor) para realização dos processos de negociação de forma automática e independente (AldridgeHFT).

Partindo para exemplos reais, o HFT provou ser muito forte na bolsa de valores dos E.U.A, (AldridgeHFT) fala que ele chegou a representar 60% das transações na mesma. O HFT também tem seu espaço na bolsa de valores do Brasil, segundo dados da (<https://ri.b3.com.br/ptb/998/DFBVMF2011.pdf>) ele representou 10,3% do volume total negociado no terceiro e no quarto trimestre de 2011, isso também é influenciado pelos programas de incentivo oferecidos pela própria B3, como a redução de tarifas baseada na quantidade média de contratos negociados (ADV – Avarage Daily Volume)**(**[**http://www.b3.com.br/pt\_br/produtos-e-servicos/tarifas/listados-a-vista-e-derivativos/programas-de-incentivo/tarifas-de-programa-hft/**](http://www.b3.com.br/pt_br/produtos-e-servicos/tarifas/listados-a-vista-e-derivativos/programas-de-incentivo/tarifas-de-programa-hft/)**).**

(falar do aumento de robôs, mostrar o exemplo do japão)

<https://www.livemint.com/Companies/ptTJe5RODwPRVqlEjBpu7M/Humans-lose-out-as-robots-take-Tokyo-Stock-Exchange.html>

NAKAMURA, Y.; HASEGAWA, T. Humans Lose Out as Robots Take Tokyo Stock Exchange. 2015. Disponível em: . Acesso em: 13/10/2016. Citado na página 11.

O termo High Frequency Trading (HFT) é usado pela B3 em suas demonstrações financeiras (<https://ri.b3.com.br/ptb/998/DFBVMF2011.pdf>) e por **AldridgeHFT** para se referir a todo tipo de negociação feita por algoritmos, já no mundo acadêmico (**estrategiaInvestimentoRedeNueralPreditor)** se refere a estes algoritmos como robô investidor e estratégia automatizada, além disso em trabalhos como (**EURUSDwithIA**), (**SVMinFinancialTimeSeriesStockIndexFuture**) e (**ForeignExhangeSVM**) tratam apenas como modelos para previsão de séries temporais, sem usar algum termo especifico. Neste trabalho, para fins de simplicidade, será usado a terminologia robô investidor para se referir a todo tipo de negociação automatizada ou qualquer modelo computacional criado para atuar no mercado financeiro.

Antecipar o comportamento da bolsa de valores é algo essencial para a obtenção de lucros, como o mercado reflete os acontecimentos políticos e sociais e o sentimento coletivo em relação a esses acontecimentos de uma parcela da população nas redes sociais.

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

* 1. **Justificativa / Motivação**

**Falar oq motiva**, que varias pessoas estão tentando ingressar na bolsa, o pq os robôs são importantes, quantos robores atuam no brasil,

Se as corretoras falam de robô

Se a b3 fala de robôs, e se ela facilita de alguma forma

Falar que uma boa quantidade de trabalho fala sobre isso

Nos trabalhos de fulano.... já tratam disso, eles tratam de tais técnicas para resolver esses problemas, etc

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.

**(oMercadoDeCapitais)** fala que os investidores tem avançado nas transformações sociais e enriquecimento das famílias, passando a ter maiores excedentes de consumos. A possibilidade de investir no mercado de capitais também é uma chance que as pessoas possuem de mudar suas vidas, e até mesmo de subir suas classes sociais.

A possibilidade de mudança de vida oferecida por este mercado se reflete nos números, com mais pessoas ingressando nesse estilo de vida segundo relatório da **(**[**http://www.b3.com.br/pt\_br/market-data-e-indices/servicos-de-dados/market-data/consultas/mercado-a-vista/historico-pessoas-fisicas/**](http://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/consultas/mercado-a-vista/historico-pessoas-fisicas/)**)** o número de pessoas físicas que possuíam ações na BOVESPA em 2018 era de 813.291 mil, em 2019 este valor saltou para 1.681.033 milhões, e até abril de 2020 este número já era de 2.385.290 milhões, percebe-se que o mercado acionário no Brasil está em grande evolução e expansão, abrindo uma infinidade de possibilidades para as mais diversas áreas de estudo

Esse ambiente leva à um maior interesse de pesquisadores para trabalharem na área, porém já existem muitas abordagens de trabalhos estrangeiros sobre o mercado financeiro que utilizam as mais diversas técnicas. O trabalho de \citeonline{ForeignExhangeSVM} por exemplo, usa a técnica Máquina de Vetores de Suporte, voltada para classificação, para prever a direção da taxa de câmbio do Euro (EUR) em relação a outras 3 moedas.

Outro trabalho como (preverTaxadeCambioBrasileira) usa as Redes Neurais Artificiais e métodos tradicionais, os preditores econométricos, para previsão da taxa de câmbio brasileira, e faz comparações dos resultados das técnicas em questão. Enquanto (estrategiaInvestimentoRedeNueralPreditor) desenvolveu um robô investidor que combina as técnicas de redes neurais artificiais e preditores econométricos para criar um mecanismo de geração de sinais de compra e venda para determinados ativos.

**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), ele defende que os preços dos ativos refletem as informações disponíveis no mercado e 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, é mencionado 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.**

**Isto também motiva este trabalho, pois uma divergência entre autores, onde alguns se apoiam na HME e outros criam mecanismos para atuar no mercado (já citados anteriormente) e se opondo a HME, é uma ótima oportunidade para novas pesquisas cientificas. Uma nova abordagem de previsão do preço da moeda local em relação ao dólar pode servir de referência aos trabalhos que vierem depois desse, pois qualquer contribuição para esta área de pesquisa com tais divergências é algo válido.**

No lado dos autores que criaram métodos para atuar no mercado financeiro, temos o (**ForeignExhangeSVM**) por exemplo, ele usa a técnica Máquina de Vetores de Suporte, voltada para classificação, para prever a direção da taxa de câmbio do Euro (EUR) em relação a outras 3 moedas.

Outras pesquisas como (...) usam as Redes Neurais Artificiais para contratos futuros e taxa de câmbio, outras pesquisas como (..) usam métodos tradicionais, os preditores econométricos, porém foi visto nestes mesmos trabalhos que eles são inferiores as técnicas de Aprendizagem de Máquina.

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. **(estas informações irão nos trabalhos correlatos, tentar reecrever para apenas citar, assim como ficou no paragrafo a seguir**)

**Outros trabalhos como (construcaoRoboInvestMestrado, citar as outras teses e outros trabalhos aqui) demostraram que é possível obter bons resultados usando tanto a Inteligência Artificial quanto a Análise de Sentimento para prever o comportamento de séries financeiras, indo no caminho oposto do que defende a HME. Logo esta pesquisa também procura obter tal feito, visto que o mercado financeiro é algo tão complexo e cheio de incertezas e qualquer contribuição para minimizar esses problemas é bem-vinda.**

**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.**

Qualquer contribuição é algo valido

**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.**

* 1. **Problema de pesquisa**

(muitos indicadores, difícil escolher, como gerenciar o dinheiro) referenciar um livro do Al brooks

Fazer a pergunta pra trabalhar com mini dólar

Se vai ser swing trade ou day trade

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

**Como já falado anteriormente, a introdução tecnológica relacionado as negociações de ativos, aconteceram entre o final do século passado e começo dos anos 2000, a presença da computação neste meio também é algo novo, em relação ao trabalhos estrangeiros (Forecasting500StocksIndexFuturesHybridSystem) foi o mais antigo achado, e data da década de 90, outro trabalho antigo foi (FinancialTimeSeriesForecastingSVM) de 2003.**

A previsão de séries temporais financeiras é considerada uma das aplicações mais desafiadoras da previsão moderna de séries temporais (**SVMinFinancialTimeSeriesStockIndexFuture**). Além do fato do mercado ser muito volátil e muito sensível a fatores externos, como eventos políticos e condições econômicas, fatores do cotidiano não são o único problema para a previsão do mercado financeiro.

\cite{TradingOnTheEdge} se refere as séries temporais financeiras como inerentemente barulhentas, não estacionárias e deterministicamente caóticas. \citeonline{SVMinFinancialTimeSeriesStockIndexFuture} explica os conceitos anteriores como: Elas são não-estacionária pois a distribuição das séries temporais está mudando ao longo do tempo, e deterministicamente caótica porque elas são aleatórias a curto prazo, mas determinísticas a longo prazo.

**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).**

Muitas vezes emoções como medo, ganancia e raiva acabam por diminuir a capacidade de decisão do investidor. (Trading:Atitude mental) explica que a atitude mental com que nos posicionamos num *trade,* condiciona fortemente os ganhos potenciais que venhamos a obter, e que o impulso é um dos maiores inimigos de um *trader*.

Esse fato acaba por refletir em más escolhas que resultam em prejuízo financeiro, pois o indivíduo perde a capacidade de se antecipar aos outros Investidores, e garantir a melhor oferta para si. Caso ele possuísse uma ferramenta computacional para auxiliá-lo, como por exemplo um robô investidor, talvez o cenário e o resultado final fossem diferentes.

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.

Existe, no mundo acadêmico, a Hipótese do Mercado Eficiente (HME), proposto por (FAMA, 1970), ele defende que os preços dos ativos refletem as informações disponíveis no mercado e 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, é mencionado 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, a partir de dados passados, padrões nas séries financeiras para antecipar os acontecimentos do futuro. E isso é apenas uma das barreiras que existe em relação a essa área de pesquisa e também ao próprio mercado financeiro como um todo. Falta de conhecimento técnico ou até mesmo fatores psicológicos fazem com que muitas pessoas tenham prejuízos financeiros, ao invés de lucros.

**Isto também motiva este trabalho, pois uma divergência entre autores, onde alguns se apoiam na HME e outros criam mecanismos para atuar no mercado (já citados anteriormente) e se opondo a HME, é uma ótima oportunidade para novas pesquisas cientificas. Uma nova abordagem de previsão do preço da moeda local em relação ao dólar pode servir de referência aos trabalhos que vierem depois desse, pois qualquer contribuição para esta área de pesquisa com tais divergências é algo válido.**

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)**

Como prever as séries temporais financeiras usando técnicas de aprendizagem de máquina e análise de sentimentos, considerando que elas são não-estacionárias e aleatórias a curto prazo?

O problema considerado neste trabalho é expresso na seguinte questão: Como fazer a previsão de séries temporais financeiras do contrato futuro de dólar(WDO), unindo as técnicas de aprendizagem de máquina para análise dos dados históricos do preço do dólar com a análise de sentimentos para os dados recolhidos da rede social Twitter?

**1.3 Objetivos**

**Objetivo Geral**

O objetivo principal deste trabalho é a criação de uma ferramenta para auxiliar investidores na compra e venda de um contrato de mini-dólar. Através das técnicas de Aprendizagem de Máquina e Análise de Sentimentos em conjunto. Com a ideia de dar um amparo maior para o usuário, onde ele poderá ter informações estatísticas de duas fontes distintas: das séries históricas e os sentimentos dos textos das pessoas do Twitter..

**Objetivos Especificos** (não é um passo a passo, objetivo especifico são componentes do objetivo geral, objetivo geral destrinchado, explicado de forma mais sucinta)

1 – Fazer o levantamento bibliográfico em cada uma das áreas já especificadas (Aprendizagem de Máquina e Análise de Sentimentos no Mercado Financeiro).

2– Recolher a série histórica do período de 2019 da plataforma MetaTrader5.

3 – Recolher os textos da rede social Twitter utilizando a palavra “dólar” como filtro, e apenas na região do Brasil com textos em português.

4– Tratar das duas bases de dados, para retirada de ruídos como outliers das séries históricas e gírias ou abreviações dos textos.

5 Desenvolver um algoritmo com AM para análise das séries históricas tratadas.

6 Desenvolver um algoritmo com análise de sentimentos para os textos minerados e tratados

7- Unificar os dois algoritmos e construir um Robô Investidor para auxiliar nas operações de swing trade com mini contratos de dólar (WDO) na bolsa de valores do Brasil.

-------------------------------------------------------------------------------

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 Organização do Trabalho**

A introdução deste trabalho apresentou: o contexto, motivação, definição do problema, objetivos, metodologia e contribuições dessa pesquisa. Os capítulos restantes são organizados da seguinte forma:

\begin{itemize}

\item No \autoref{chap:conceitos}, \textbf{Fundamentos Teóricos}, são apresentados os conceitos abordados neste trabalho, especificamente: XXX, YYY, e ZZZ.

%

\item No \autoref{chap:trabcorrelatos}, \textbf{Trabalhos Correlatos}, são analisados os trabalhos correlatos a solução proposta.

%

\item No \autoref{chap:solucaoproposta}, \textbf{Método da Solução Proposta}, é descrito as etapas de execução do método da solução proposta para XXXX.

%

\item No \autoref{chap:planavaexp}, \textbf{Planejamento para Avaliação Experimental}, é apresentado o planejamento e projeto para execução da avaliação da solução proposta.

%

\item No \autoref{chap:cronograma}, \textbf{Cronograma de execução para o TCC $2$}, é descrito o cronograma proposto para as próximas etapas do trabalho de conclusão de curso, bem como, a descrição das atividades propostas.

%

\item E por fim no \autoref{chap:consideparciais}, \textbf{Considerações Parciais}, apresenta-se as considerações parciais e análise das atividades já desenvolvidas.

**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
      1. Bolsa de Valores do Brasil
      2. Day trade e Swing Trade
      3. Análise Gráfica
      4. Indicadores Técnicos de Análise Gráfica
      5. Contratos de Dólar
   2. Aprendizagem de Máquina
      1. Séries Temporais
      2. Algoritmos de Classificação e Regressão
      3. Algoritmo Support Vector Machines – SVM
      4. Métricas de Desempenho
   3. Análise de Sentimentos
      1. Processamento de Linguagem Natural
      2. Data Scraping
      3. Técnicas de Pré-processamento e Texto

* 1. Ferramentas de Suporte a Machine Learning
     1. Scikit Learn
     2. Natural Language Toolkit – NLTK
     3. Análose de Dados com Pandas
  2. Análise de Sentimentos de Processamento de Linguagem Natural

**2.1 O mercado financeiro**

O mercado financeiro é o local, o ambiente onde ocorre a compra e venda de valores imobiliários (Ações, títulos de Renda Fixa, Fundos, etc), do câmbio de moedas estrangeiras e de mercadorias como minérios e produtos agrícolas. Esse mercado se divide em quatro outros mercados: Cambio, Monetário, Crédito e o de capitais. (<https://blog.easynvest.com.br/mercado-financeiro-e-de-capitais/>).

**i –Bolsa de Valores do Brasil**

**O** Mercado de Capitais é um sistema de distribuição de valores mobiliários, com o objetivo de prover liquidez aos vários títulos existentes, de forma a melhorar o fluxo de capitais entre os agentes econômicos. (<https://www.infomoney.com.br/colunistas/economia-com-renata-barreto/voce-sabe-a-importancia-do-mercado-de-capitais/#:~:text=Com%20o%20mercado%20de%20capitais,barata%20do%20que%20por%20empréstimos>.). Esse mercado é formado por instituições financeiras, corretoras de valores e a própria bolsa de valores, que por sua vez são negociados: Ações, Debêntures, Títulos Privados, Mercado Futuro e Renda Fixa (<https://blog.easynvest.com.br/mercado-financeiro-e-de-capitais/>).

A bolsa de valores é o local do mercado de capitais onde se negociam ações de empresas com capital aberto, seja privada ou pública, entre outros valores imobiliários. Segunda a própria (infomoney) uma ação é a menor fração do capital social de uma empresa, quando o investidor a obtém, se torna acionista desta empresa.

Existe algumas formas de lucrar com isso, através do recebimento de dividendos, que são parte dos lucros da empresa em questão dividia com os acionistas (https://www.infomoney.com.br/guias/proventos/), com a valorização daquele ativo que o investidor comprou no passado (se ele tiver valorizado). A segunda opção abre um espaço para os chamados especuladores, eles são pessoas que apostam na alta ou baixa de determinado ativo para lucrar em um curto ou médio espaço de tempo nos mais diversos ativos presentes na Bolsa de Valores. (https://www.infomoney.com.br/guias/mercado-futuro/),

**ii – Day Trade e Swing Trade**

Dentre as mais diversas modalidades para se especular no mercado de capitais, existem duas que estão se tornando cada vez mais famosas, Day Trade e Swing Trade. Uma operação se caracteriza como day trade se ela é aberta e finalizada no mesmo dia, já para operações que finalizam no dia seguinte, alguns dias ou semanas depois se caracterizam como swing trade.

Especuladores que realizam day trade, acreditam que várias operações mais rápidas em um curto período de tempo podem trazer lucros maiores, mesmo que isso leve a maiores exposições ao risco. Já os que realizam swing trade acreditam em menos operações em períodos determinado de tempo, com menor exposição ao risco (por realizar menos operações).

Toda estratégia tem suas vantagens e desvantagens, o Swing trade, por exemplo, por ser operações pautada vários dias, acaba por permitir que a pessoa tenha mais tempo para as suas tomadas de decisões, faça menos operações (com isso pague menos taxas de corretagem) dentro de um período de tempo determinado. Outra vantagem desta estratégia é em relação aos impostos, existe isenção de imposto de renda nos ganhos líquidos se as movimentações financeiras ficarem abaixo de 20 mil reais no mês em especifico.

(https://www.infomoney.com.br/onde-investir/day-trade-ou-swing-trade-os-dois/#:~:text=O%20swing%20trade%20se%20diferencia,do%20tempo%2C%20com%20menos%20operações.)

(<https://www.infomoney.com.br/minhas-financas/como-declarar-swing-trade-no-imposto-de-renda/>)

**IV – Análise Gráfica**

Para trabalhar com operações no mercado financeiro, existem diferentes tipos de abordagens, analise fundamentalista, que se baseia em fundamentos da empresa em questão, análise gráfica, ou técnica que usa técnicas relacionadas ao movimento do preço no gráfico para para encontrar padrões que se repetem de forma recorrente. Neste momento o foco será na análise gráfica.

De acordo com o (Elder) os primeiros grafistas dos EUA surgiram na virada do século XIX para o XX, como o próprio Charles Dow, o responsável pela famosa teoria de Dow. Porém (Elder) defina o período de 1930, logo após a crise de 1929, como a era de ouro dos gráficos, figuras importantes como Elliott (criador das ondas de Elliott), Rhea, e Gann publicaram suas pesquisas durante este período.

Por fim (Elder) fala que a maioria dos analistas técnicos usam gráficos de barras, eles possuem o preço de fechamento, o preço máximo e mínimo, volume e contratos em abertos. Esses dados são muito parecidos com a própria base de dados usada por esta pesquisa, quem contém os preços de abertura, fechamento, máxima e mínima e o volume negociado.

Historia da analise gráfica do livro do Elder, , posso falar do livro da de do mark douglas

Agora eu falo da relação com a minha pesquisa, falo do que o Elder disse, da análise gráfica ser usado muito em gráfico de barras, preço máximo, mínimo. Etc

**V – Indicadores Técnicos da Análise Gráfica**

Dentro da análise gráfica/técnica, existe algo chamado de indicadores técnicos, eles são ferramentas baseadas em cálculos matemáticos sobre a variação do preço e/ou do volume negociado. (Elder) mostra uma divisão em três grupos dos indicadores, são eles: rastreadores de tendência, osciladores e mistos.

Rastreadores de tendência são muito úteis para preços em alta ou em baixa. Osciladores funcionam bem para mercados laterais, consolidados, já os mistos são muito bons para leitura da psicologia de massa.

Ai eu falo do do trab (trab2 eu acho) que usou em um dos modelos indicadores da análise técnica (ou deixo pro próximo capitulo)

**VI - Contratos de Dólar**

Um dos ativos existentes na Bolsa de Valores e no Mercado de Capitais são os Contratos Futuros. Eles representam o compromisso de comprar ou de vender um certa quantia de um determinado ativo em uma data futura por um preço determinado previamente (<https://www.infomoney.com.br/guias/mercado-futuro/>). Eles são derivativos, ou seja, possuem o preço derivado de outro ativo.

O contrato futuro de mini dólar (WDO) por sua vez tem o seu valor derivado do preço de mercado do próprio dólar no mercado a vista, os mini contratos acabam sendo vantajosos que os contratos cheios, pois seus valores para negociações são consideravelmente mais baixos, o WDO em especifico tem um valor mínimo 5 vezes menor que o contrato cheio de dólar, sendo isso uma vantagem para investidores e especuladores com capital menor. A estratégia varia de pessoa pra pessoa, algumas quando especulam preferem encerrar as suas operações no mesmo dia, outros preferem carregar suas posições por mais de um dia

**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, 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

Uma árvore de decisão representa uma função que toma como entrada um vetor de valores de atributos e retorna uma “decisão” — um valor de saída único. Os valores de entrada e saída podem ser discretos ou contínuos **(Inteligencia-Artificial-3a-Ed-Russell-Stuart-Norvig-Peter-pdf).** Para as situações em que a entrada tem valores discretos e a saída tem duas opções: verdadeiro (positivo) ou falso (negativo), é considerado um algoritmo de classificação. Ainda nesse livro é falado que uma árvore de decisão alcança sua decisão em uma sequência de testes, onde cada nó corresponde a um teste de um dos atributos de entrada, se verdade tal comparação ou pergunta, por exemplo, tal resultado vai para um nó, caso não irá para outro nó, e assim por diante, os nós das folhas da árvore representam o valor que é retornado pela função

**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, isso os ajuda a generalizar bem.

Em várias situações quando o hiperplano é traçado, os dados dos dois lados podem ficar muito próximos dos separadores traçados, neste caso a SVM tenta minimizar a perda de generalização, em vez de minimizar a perda empírica, escolhendo o separador mais distante possível dos pontos (cada ponto no plano é um dado), ele é chamado de separador de margem máxima. **(Inteligencia-Artificial-3a-Ed-Russell-Stuart-Norvig-Peter-pdf).**

**(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.

**?? (Colocar uma imagem do livro, citando a fonte do mesmo, que mostra o gráfico de uma classificação) ??**

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, a saída do algoritmo (a resposta para o problema) será 0 (zero) para mulher ou 1 (um) para homem.

**2.2.2 Regressão Linear**

6 REGRESSÃO E CLASSIFICAÇÃO COM MODELOS LINEARES

Pags 9, 34, 73, 348

Nonlinear Regression

Problemas cujo a saída é um número, são problemas de regressão, e além disso assim como a classificação a Regressão Linear também é um problema de aprendizagem supervisionada, como explica (**alppaydin\_machinelearning\_2010)**, ou seja, a entrada são certos valores representados como *X*, e a saída é representada como *Y*. Em uma situação onde usa-se da Regressão Linear para resolver determinado problema, é necessário plotar os dados em um gráfico, com o objetivo de criar uma relação entre *x* e *y,* onde *x* é o valor (ou os valores) de entrada, e *y* o valor de saída, a resposta do algoritmo *,* ou seja, um função, onde a variável *x* (ou as variáveis *x*´s) é a variável independente, e a variável *y*  é a variável dependente, depois disso uma reta é plotada com o objetivo de ser o mais próximo possível desses dados.

**?? (Colocar uma imagem do livro, citando a fonte do mesmo, que mostra o gráfico de uma regressão) ??**

Para explicar melhor (**alppaydin\_machinelearning\_2010)** aplica a Regressão Linear em uma situação onde deseja-se ter um sistema para predizer o preço de um carro usado, com as entradas sendo os atributos dele: marca, ano, potência do motor, quilometragem, e qualquer outros atributos que possam afetar o seu valor final. A saída é o valor do carro.

Nessa situação *X* representa os atributos e *Y* o preço do carro, o programa treina uma função para os dados coletados para aprender Y como um a função de X,

Y = wx + w0 ( f(x) = wx + w0 )

onde *w* e *w0*  são os coeficientes dos valores reais, atuando como pesos da função.

**(Inteligencia-Artificial-3a-Ed-Russell-Stuart-Norvig-Peter-pdf)** para explicar a Regressão Linear, ele a separa em: regressão com função linear univariada e com função linear multivariada. A univariada é resumida pela formula mostrada acima, com uma variável independente, já a multivariada é feita por um somatório e cada exemplo x é um vetor de *n* elementos, uma função com várias variáveis independentes.

**2.2.2.1 Support Vector Regression (SVR)**

Kernel Machines for Regression no livro alppaydin\_machinelearning\_2010

Pags 9, 34, 73, 348

Nonlinear Regression

No tópico dos algoritmos de classificação foi mencionada, descrita e exemplificada a SVM, porém existe uma extensão dela que é usada em problemas de regressão, essa extensão é a SVR. (**precificacaoDeOpcoesPorAcoesTeseMestrado**) menciona que no ano de 1995 a SVM 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.

A ideia é a mesma da SVM, uma linha é traçada entre os dados, porém a linha pode assumir várias formas dependendo do kernel usado, como **(alppaydin\_machinelearning\_2010)** demostra graficamente, ele pode ser em formato de parábola, para o kernel quadrático, ou uma reta para o caso do kernel linear, por exemplo. O intuito na SVR é, quando o hiperplano é traçado, garantindo o máximo de pontos possíveis (cada ponto são os dados de entrada do algoritmo) dentro das margens traçadas paralelo ao hiperplano, mantendo essas mesmas margens o mais próximo possível da linha original sem sacrificar muitos pontos de dentro delas, o *x* é a variável de entrada e o  *y* a variável de saída, o segundo sendo a resposta do algoritmo. Em **(alppaydin\_machinelearning\_2010)** é explicado que existem 3 situações, a primeira é o dado dentro da margem, a segunda é o dado em cima das margens, e a terceira é ele do lado de fora, neste último caso o mesmo é chamado de *outlier.*

Um exemplo para descrever a aplicabilidade esse algoritmo é uma situação onde deseja-se prever a renda mensal de uma família, existem certos programas sociais onde é necessário saber de dados desta natureza para identificar quais os grupos que se encaixam em determinada faixa de renda para poder receberem tal auxilio. Os dados de entrada poderiam ser: se a família possui uma geladeira em casa, se possuem televisão a cabo, água encanada, se os filhos (caso possuam algum) estudam em escola pública ou privada, entre outras informações consideradas relevantes. Uma linha é traçada o mais próximo possível dos pontos plotados no gráfico, essas informações são os dados de entrada *x,* e o *y* será a saída, a renda média da família, a resposta do algoritmo sempre será representada por um valor real.

**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 Inteligência artificial

linguística computacional ou processamento de linguagem natural.

**PLI**

1. Método da solução proposta

Neste capítulo será apresentado a solução proposta por esta pesquisa, utilizando ferramentas para auxiliar neste processo, como uma *big Picture****.*** Para mostrar o fluxo da solução, será usado um diagrama de fluxo. As ferramentas para o desenvolvimento deste trabalho também serão mostradas e explicadas aqui. Por fim algumas considerações finais sobre a construção da solução proposta, pontos positivos e limitações.

**4.1 Arquitetura**

A arquitetura será explicada usando a figura ..., ela fornece uma visão geral da solução proposta por esta pesquisa. Um robô investidor para auxiliar o usuário na tomada de decisão na hora de comprar ou vender um ativo. Os próximos parágrafos serão destinados para a explicação do funcionamento do software baseado na figura ...

O software se inicia com a ordem do usuário, a máquina dele irá mandar o comando solicitando os dados para o software, que é ilustrado pela imagem do robô, para obter a resposta se ele deve comprar ou vender, este por sua vez coleta os dados da rede social (os textos e publicações referente ao dólar) e da plataforma de negociação (os dados históricos do período selecionado do ativo). Logo após o processamento dos dados, o robô envia a resposta, o sinal de compra ou venda, para o computador do usuário.

A base de dados, o histórico de preços é fornecido pela plataforma de negociação chamada Metatrader (referência dela aqui). Essa plataforma fornece um período histórico de até 5 anos. Porém, seguindo o trabalho do trab4 usarem um período de ..... Esses dados são analisados por um classificador SVM, onde ele divide e classifica os dados em A (preço vai subir) ou B (preço vai descer), usando vetores de suportes, como é mostrado no gráfico central da figura.

Sobre a parte de análise de sentimentos, perceba que logo em cima existe uma imagem de uma publicação da rede social sobre dólar. No texto, de publicações coletadas semanalmente, existe uma palavra circulada, e a negatividade dela em 80%. Isso significa que o algoritmo analisou esta frase e atribuiu um valor de sentimento para o verbo (aumentou) que faz referência ao substantivo (dólar). Este valor de sentimento é passado como parâmetro para uma SVM, que por sua vez irá classificar a direção do preço do ativo e retornar o resultado para a máquina do usuário.

A figura .., contém dois diagramas de fluxo, o primeiro ilustra o toda a execução do modelo, com as entrada ,saídas e resultados gerados por ele, com a sua construção. O Segundo diagrama mostra a execução do software pelo usuário, o passo a passo para a utilização do mesmo.

O diagrama A está dividido em dois fluxos de dados, o primeiro com a construção da análise das séries históricas iniciado na etapa A, fornecida pela própria plataforma Metatrader () como já mencionado anteriormente. Na etapa B acontece o pré-processamento dos dados, onde será feita a limpeza do mesmo, com a retirada de outliers, dados ausentes, repetidos ou duplicados, entre outras irregularidades e ruídos, como é exibido nas etapas B2 e B3.

As etapas C1, C2 e W, são da base de dados final, onde receberá os dados históricos tratados e também os textos tratados. A etapas D1 e D2 se referem a divisão da base de dados, 94,7% para treino e 5,3% para teste, seguindo o que foi feito no (trab4). As etapas E1 e E2 são sobre o próprio algoritmo, um classificador SVM com um kernel do tipo radial, também seguindo o que foi proposto por (trab4). Logo depois, na etapa F os resultados da predição de todo o modelo, e as etapas G1 e G2 divide os resultados em dois, um com os sentimentos e o outro sem os sentimentos.

Para o segundo fluxo, onde é explicado a construção do modelo para análise dos sentimentos dos textos, tem-se a etapa H, onde é recebido os textos das publicações, a fonte dos dados é a própria rede social. Na etapa I acontece o pré-processamento dos dados. Nas etapas J1 e J2 acontecem a tradução dos textos, que estão em português, para o inglês., utilizando a ferramenta Google translate API (referencia) K1 e K2 mostram um exemplo de um texto passando por este processo de tradução.

A etapa L trata da divisão dos textos em tokens, M1 e M2 explicam esse processo com um exemplo, poise os algoritmos para análise de sentimentos analisam melhor palavra por palavra. A etapa N é justamente a própria análise de sentimentos, onde é atribuído um dos três valores, positivo negativo ou neutro, como é mostrado em O1 e O2, Para atribuição de sentimento aos termos / aos textos, foi usado o recurso lexical chamado. SentiWordNet (), o mesmo usado pelo trab2. Ele classifica os sentimentos das palavras em 3 tipos: negatividade, objetividade e positividade.

. Com os resultados é passado para a etapa W, onde tem-se a base de dados final, junto do histórico de preços, o resto do processo já foi explicado anteriormente.

Para atribuição de sentimento aos termos / aos textos, foi usado o recurso lexical chamado. SentiWordNet (), o mesmo usado pelo trab2. Ele classifica os sentimentos das palavras em 3 tipos: negatividade, objetividade e positividade.

* 1. **Ferramentas e Implementações**

Para a implementação da solução proposta será usado a linguagem de programação Python (v3.6), junto com as bibliotecas de software Pandas (v1.1.5) e o Scikit-learn (v0.22.2). O pandas será utilizado para leitura e escrita na base de dados importada na extensão *.csv. O* Scikit-learn será usado para os algoritmos de aprendizagem de máquina, como a própria Máquina de Vetor de Suportes e a validação cruzada.

* 1. **Considerações adicionais**

Uma limitação desta pesquisa é a velocidade de mineração dos dados, o acesso com a plataforma Metatrader demora um tempo significativo, por esta razão não é possível criar um modelo de previsão que atue para especular o preço do mesmo dia. A segunda limitação é que nosso modelo se restringe apenas a orientar o investidor, ele não faz operações de forma automática, porque para isso precisaria criar uma ligação direta com a bolsa de valores (B3) ou alguma corretora que atue nela com o algoritmo criado.

1. Trabalhos Correlatos

**Trabalho 1 - Forecasting Stock Market Movement Direction Using**

**Sentiment Analysis and Support Vector Machine**

No trabalho proposto por (..), é desenvolvido um método que integra a análise de sentimentos em um sistema que usa aprendizagem de máqina baseado em Máquina de Vetor de Suportes (SVM). O intuito é prever a direção do movimento do SSE 50 Index, o índice da Bolsa de Valores de Xangai, e auxiliar o investidor na hora da tomada de decisão. Com a ideia antecipar a tendência do mercado de ações, foi implementado uma SVM com fivefold cross validation. Para elimitar o viés de antecipação, uma abordagem com rolling window é utilizada. Depois é implementado uma estratégia de stop-loss, para reduzir e limitar as perdas e por consequência potencializar os ganhos dos investidores.

(aquela imagem gigante)

Sobre o processo dos sentimentos, as primeira etapa é a contrução de um Web Crawler, a sua função é baixar de forma automática os documentos textuais direcionados e armazená-los em um database. A própria ferramenta gerencia a fila de URLs, decidindo a prioridade e descartando os endereços duplicados. Por fim, depois das páginas da web terem sido recolhidas, os seus conteúdos recolhidos são: a notícia textual com a data dos sites e o outro é o URL contido na página. O diagrama mostrando este processo está logo abaixo.

(Print da figura 2)

Step 2 Daily sentiment: A sentence-based sentim, HowNet and Chinese Sentiment Analysis Ontology Base

A ideia é analisar as frases para extrair a idéia de todo o documento, em vez de uma palavra só. Com isso um documento é dividido em frases, depois as sentenças são divididas em palavras, por fim é contado o número de palavras positivas e negativas para ser atribuida um valor específico de sentimento e decidida a polaridade de cada frase, os dicionários usados para isso foram: HowNet and Chinese Sentiment Analysis Ontology Base.

Foram feitos e experimentos (Experiment 1 e 2), o experimento 1 foi usado os dados históricos, que tem preço de abertura, fechamento, máxima e mínima do dia, volume do número de negociações, volume em RMB, mudança em RMB, e mudança de porcentagem. Experimento 2 foi uma combinação do experimento 1 com os sentimentos. A performace foi calculada com com uma fórmula de Acurácia.

c foi adaptada pra treinar uma SVM. Experimento 1 os parâmetros da SVM foram: função de kernel RBF, C=256, Y=0,9942, pro Experimento 2: função de kernel RBF, C=181.0193, y=0,005524. A ideia do cross validation não é ser usado pra previsão, mas sim pra selecionar a melhor função de kernel e os melhores parâmetros, la no artigo ele explica o motivo.

SVM com rolling Windows, a ideia é selecionar a melhor janela, usa-se N dias anteriores pra prever a direção do movimento dos próximos dias. Os parâmetros para os experimentos 1 e 2 são os mesmos do anterior. Repita em loop o procedimento alterando o valor de N até que a SVM atinja a maior precisão com os parâmetros a função de kernel que já temos selecionados. No experimento 1 a janela continua ideal é 68 com precisão de 71,33%, para o experimento 2 é 76, com precisão de 89,93%. É evidente que a adição dos sentimentos no experimento 2 elevou bastante os resultados.

logistic regression (LR) com fivefold cross validation é usada pra reexaminar as conclusões. Validação cruzada também é aplicada pra prever a direção do movimento do SSE 50 index. Experimento 1 teve 70,96% de acurácia. GD com 600 interações é implementado para convergir o resultado (?). No experimento 2 a acurácia foi de 86,69%. GD com 1000 interações é implementado para convergir o resultado.

Resultado: support vector machine (SVM) com fivefold cross validation foi o melhor modelo desses testados.

Para o DataSet o autor escolheu explorar o mercado da China, mais especificamente o índice da Bolsa de Valores de Xangai, o SSE 50 Index. Os dados recolhidos em relação aos preços foram: máxima e mínima do dia, preço de abertura e de fechamento, volume negociado, volume em, RMB mudança em RMB e mudança em porcentagem. Os dados foram retirados do Wind Economic Database. Os dados textuais foram retirados dos foruns Sina stock e Eastmoney stock. O período dos dados vai de 17 de junho de 2014, até 7 de junho de 2016, somando um total de 485 dias de negociação.

A tabela IV mostra os resultados com o incremento do stop loss.

Sobre os resultados.... Acuracia de 89,93 porcento, com aumento de 18,6 porcento depois de introduzir os sentimentos.

Depois dos resultados ele avalia se os resultados sao bons para o o investidor, no item D. Investiment Performace:

AL é uma medida usada pra calculo dos pontos de estoque, por exemplo se um ativo é comprado pelo preço de 100 e vendido por 150, temos 50 pontos de lucro, logo o valor de Al é 50 pontos de estoque. Um calculo de drawdrown (MDD) também é feito, MDD é o declínio, o rebaixamento máximo de uma série de um pico até um vale durante um período de tempo determinado. Expected maximum drawdown (EDD) é uma expectativa da média de perdas máximas.

Resultados dos experimentos 1 e 2 são usados para calcular o Al e o MDD. Na tabela abaixo já fica evidente a diferença, o experimento 2 obteve mais que o dobro de pontos em Al do que o experimento 1. Os resultados de MDD e EDD também foram melhores no experimento 2. Com isso percebe-se que a adição dos sentimentos promove um resultado consideravelmente melhor, com lucros maiores e riscos menores.

Por fim, abaixo tem o item do experimento 2 com ‘strategy’, aqui o trabalho implementou ao experimento 2 uma estratégia de ‘stop-loss’ com o intuito de limitar as perdas. O limite de perdas foi definido em 95 pontos de estoque. Ou seja, se as perdas, no dia de negociação em questão, atingirem 95 pontos, a negociação aberta será fechada e finalizada. Na tabela é mostrado que essa estratégia melhorou os resultados, tanto dos pontos em Al quanto da redução dos valores de MDD e EMD. Pois em sequencias grande de perdas o stop-loss corta essa sequência, e menos perdas significa um lucro líquido maior.

**{Trabalho 2 - Sentiment analysis on social media for stock movement prediction}**

**Introdução**

No trabalho proposto por ... o objetivo foi criar um modelo para previsão do preço de ações baseado nos sentimentos coletados dos textos. O diferencial deste trabalho é que em vez de analisar apenas -os sentimentos gerais, aqui é coletado também os sentimentos dos temas específicos de cada uma das 18 empresas analisadas neste trabalho, mostrando resultados superiores do modelo contendo os sentimentos, em relação ao modelo que analise apenas os preços históricos.

**Descrição das bases de dados**

Duas bases de dados foram utilizadas, do período de 23 de julho de 2012 até 19 de julho de 2013, a seguir a descrição detalhada de cada uma. A primeira é o histórico de preços de 18 ativos, onde foram coletados do Yahoo Finance, e as informações de cada um são: preço de abertura, máxima, mínima, fechamento e fechamento ajustado. Abaixo a tabela com a lista dos 18 ativos coletados.

(print da imagem da tabela 1)

A segunda é o conjunto de dados de informações de humor. Onde foram coletados 18 quadros de mensagens das 18 ações, a fonte foi o Yahoo Finance Message Board. A vantagem destes quadros é que os usuários normalmente discutem notícias e previsões sobre as ações em questão, além disso um ponto importantíssimo obervado pelo autor é que em 15,6 porcento das mensagens coletadas possuem marcações de sentimentos nelas, são 5 tipos de marcações existen tes: Compra forte, Compra, Espera,Venda e Venda Forte.

As bases de dados foram divididas da seguinte forma: dos 249 dias de transações recolhidas, 171 foram destinadas para treino e 71 para testes. A métrica usada para avaliação do método proposto foi a Accuracy. A fóruma dela está mostrada logo abaixo.

(print da imagem da Acurracy)

Abaixo duas imagens, a primeira com exemplo de texto, nota-se que essa mensagem é uma resposta à outra mensagem, além disso no canto inferior direito existe a marcação com o tipo de sentimento, feito pelo próprio autor do texto. A segunda imagem é a tabela com o número de mensagens sobre cada empresa no período do ano coletado,

(print da figura 1 e da tabela 2)

**Descrição dos modelos**

(tabela 3 mostrando os modelos)

O primeiro, chamado de Price Only, usa apenas o histórico de preços para treino e previsão dos preços dos ativos, segundo o autor o objetivo deste modelo, além de tentar descobrir padrões nos preços passados dos ativos, é usá-lo como base para comparação com os outros modelos que terão o aditivo dos sentimentos, e saber se houve ou não melhora nos resultados.

O segundo é o Human Sentiment, aqui é incorporado a análise do histórico de preços com os sentimentos, porém foram usados apenas os sentimentos anotados pelos próprios usuários, que corresponde, segundo a descrição da base de dados, a 15.6 porcento do total das mensagens. O resto da base de dados não foi usada para este modelo.

O terceiro modelo foi o Sentiment Classification, neste além do histórico de preços, foram usados os textos colhidos, porém a outra parte de 84,4 porcento que não possuem os sentimentos anotados pelos usuários. Para ser possível usá-las, o trabalho desenvolveu um modelo de classificação, usando SVM, para extrair os sentimentos das mensagens.

Ele foi treinado com os textos que já possuem os sentimentos anotados (os 15,6 porcento mencionados anteriormente), depois disso o modelo foi usado para classificar a parte da base de dados sem as marcações de sentimentos, em cinco classes diferentes. São elas: Compra Forte, Compra, Espera, Venda, e Venda Forte. Os stop words foram removidos dos textos, então todas as palavras passaram pelo processo de lemantização atravez da ferramenta Stanford CoreNLP. A ferramenta TF-IDF foi utilizada como ponderador de recursos.

Para o quarto modelo tem-se o LDA-based method, aqui cada mensagem é considerada como uma mistura de tópicos ocultos, os documentos são representados como misturas aleatórias sobre tópicos latentes, onde cada tópico é caracterizado por uma distrivuição de palavras, Latent Dirichlet Allocation (LDA), um modelo estatisticos generativo, foi escolhido nesta pesquisa para descobrir estes tópicos.

As stop words foram removidas, depois é feito o processo de lemantização atravéz da ferramenta StanfordCoreN. a LDA foi trainada no conjunto de treino, a ferramenta Gibbs Sampling também é utilizada para inferir os tópicos através de 1000 intereações

No quinto modelos temos o JST-based method, cada mensagem é uma mistura de tópicos e sentimentos, a idéia dele é extrair tópicos e sentimentos simultaneamente. Diferente da LDA, onde existia apenas uma distribuição de tópico específico de documento para cada documento, no JST cada documento é associado com S rótulos de sentimentos, Cada um dos rótulos de sentimento está associado a uma distribuição de tópico específica do documento com o mesmo número de tópicos. Uma palavra no documento é extraído da distribuição sobre as palavras definidas pelo tópico e etiqueta de sentimento. A imagem abaixo mostra um exemplo de um texto com tópicos e sentimentos.

(print da figura 4)

O sexto modelo Aspect-based sentiment foi o que obteve o melhor resultado. Aqui foi adicionado à SVM que utiliza os preços históricos, os sentimentos extraídos de cada mensagem do DataSet formado.

**O sexto e último modelo é o Aspect-based sentiment. aqui as misturas de sentimentos e tópicos não estão ocultas**. Cada mensagem é representada como uma lista de tópicos e seus valores de sentimento correspondentes. No Aspect-based sentiment o tópico é substantivo da frase. A Imagem de um exemplo de texto abaixo exemplifica isso, 'profit' (cuja a tradução é lucro) é o tópico, e 'up' (cujo nesta frase tem sentido de irá aumentar) é o sentimento positivo.

(print da figura 4)

Depois da divisão das mensagens em frases, a ferramenta Stanford CoreNLP é usada para fazer a marcação de partes do discurso em corpus e para lemantização de cada palavra em cada frase. Os substantivos consecutivos foram extraídos como tópicos das frases, e tópicos que apareceram menos de 10 vezes, formam descartados. Com a lista de tópicos formadas, e os valores dos sentimentos de cada frase são extraídos. Neste processo as palavras que contém opiniões de cada frase analisada são extraídas usando um recurso léxico chamado SentiWordNet, ele serve para classificar as palavras atribuindo-lhes uma nota para o grau de positividade, negatividade e objetividade. Por fim essas pontuações obtidas de cada tópico naquele data de transação ‘t’ foram usadas como dados de entrada para o modelo de SVM.

A base de dados foi dividida da seguinte forma: dos 249 dias de transações recolhidas, 171 foram destinadas para treino e 71 para testes. A métrica usada para avaliação do méotdo proposto foi a Accuracy.A fóruma dela está mostrada logo abaixo.

(print da imagem da Acurracy)

**Sobre os resultados**

O modelo Aspect-based sentiment obteve o melhor resultado médio para as 18 ações, com 54,41 %. Para algumas ações em especifico ele foi ainda maior, como por exemplo 71,05 % na AMZN e 64,47 % para a DELL. Os resultados também mostraram a evolução dos modelos quando adicionados nele a análise de sentimentos, o Aspect-based sentiment obteve 2,07 % médio a mais que o Price Only (algoritmo mostrado da figura 3 que usa apenas os preços históricos para análise).

( aqui um print da tabela 6)

**Trabalho 3 - Social Media Aided Sentiment Analysis for**

**Stock Prediction}**

(...) propõe um método que emprega medidas combinadas de Indicadores da Análise Técnica (TI), Indicadores Fundamentais (FI) e Indicadores de Sentimentos (SI). São usados para previsão do preço das ações de 5 empresas distintas. Com o intuito de auxiliar o investidor na compra e na venda de determinado ativo da bolsa de valores. O algoritmo escolhido foi a Máquina de Vetor de Suportes (SVM). No geral foram implementados vários algoritmos de SVM, e para os vários modelos desses algoritmos, foram também usados vários tipos de indicadores, que serviram como entrada para cada SVM.

Os indicadores de TI foram RSI, Money Flow Index (MFI), Exponential Moving Average (EMA), Stochastic Oscillator (SO) e Moving Average Convergence and Divergence (MACD), neste caso esses dados foram obtidos dos preços históricos dos ativos. Os do FI foram: Preço do ouro (GP), Taxa de Câmbio (ER) e o Preço do pretóleo bruto (COP). Por fim para o SI foi criado uma tabela com várias palavras chaves que possuem relação com o mercado financeiro, e foram classificadas como positiva ou negativas, alguns exemplos de palavras chave são: touro, urso, perigo, baixa, dentre outras. A seguir a imagem ilustrando o diagrama do algoritmo.

(A imagem com o diagrama do algoritmo)

O Dataset utilizado referente ao histórico do preço das ações foi retirado das bolsas de valores National Stock Exchange (NSE) e Bombay Stock Exchange (BSE), e do Yahoo Finance. O Dataset referente aos textos que contém os sentimentos das pessoas foi retirado do Money Control e do Twitter. As empresas escolhidas foram: Infosys, Wipro, Hindustan Computer Limited (HCL), Tata Consultancy Services (TCS) e a Tech Mahindra, junto com os índices de ações SENSEX e Nifty da BSE e o índice BSE. Por último o preço do ouro (GP), taxa de câmbio (ER) e o preço do petróleo bruto (COP) também foram coletados para o dataset.

Logo depois, na parte de pré-processamento, os dados recolhidos eles são tratados, como remoção de palavras indesejadas, informações sem relevância para a opinião expressa no texto, dentre outros. Para o histórico de preços, foi usado as informações de: Máxima do preço, Mínima, Abertura, Fechamento, e o Volume. O preço de fechamento foi escolhido para ser o valor previsto pelos algoritmos.

Sobre as regras, se o modelo prever que a tendência do dia (t+1) for de alta, então invista no ativo, se a tendência for de baixa, então venda, e por último se não houver tendência prevista, não venda.

Os resultados dos modelos de SVM foram calculados usando as métricas: Acurracy, Sensitivity, Specificity, RMSE e MAPE. No geral o modelo que combina os 3 tipos de indicadores (CI) se saiu melhor do que todos os outros que usavam indicadores individualmente (TI, SI e FI) ou uma combinação de dois deles (TI\&SI e FI\&SI).

Isso fica evidente pela baixa taxa de erro observada nas tabelas onde mostram os resultados do RMSE e MAPE. E na alta acurácia, sensibilidade e especificidade que também são exibidas em tabelas com os resultados das métricas Acurracy, Sensitivity e Specificity. Logo abaixo duas tabelas com os resultados das métricas Acurracy e RMSE respectivamente.

Vale botar pelo menos duas imagens aqui, uma de cada, RSME e Acuraccy por exemplo

**{Trabalho 4 - Forecasting stock market movement direction with support vector machine}**

Nesta pesquisa (...) decidiram investigar a previsibilidade da direção do movimento semanal do índice NIKKEI 225, utilizando o algoritmo Máquina de Vetor de Suportes (SVM). O resultado dela foi comparado com outros 4 algoritmos, e a SVM se saiu melhor que todas elas. Por fim o autor desenvolveu um modelo final que combina os 5 algoritmos. Por fim propõe um modelo que integra todos os 4 algoritmos anterior, e esse se saiu melhor que todos os outros nos testes realizados.

Em relação ao dataset, os autores usaram o índice NIKKEI 225, que compõe 225 ações da bolsa de valores de Tóquio. O período da base de dados se estende de 1 de janeiro de 1990 até 31 de dezembro de 2002. Com um total de 676 pares de observações. Os autores recolheram esses dados da seção de finanças do Yahoo e do Pacific Exchange Rate Service. A base de dados foi dividida em duas partes, a primeira com 640 pares de observações destinada ao treino, para determinar as especificações e parâmetros do modelo, e a segunda com 36 partes para teste, para comparação das performances entre os modelos criados.

Dois dados macroeconômicos foram escolhidos como entrada para a SVM. São eles S&P 500 Index e o par de moedas JPY/USD. Os 4 algoritmos usados para comparar com ele foram: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Elman Backpropagation Neural Networks (EBNN) e Random Walk Model (RW). O modelo final, que combina todos esses anteriores se chama Modelo Combinatório.

O modelo combinatório usa um sistema de pesos que atribui pontos a cada modelo anterior baseado em sua performance na fase de treinamento. Aqueles que se saíram melhores receberam pesos maiores, logo eles obtiveram uma contribuição relativa maior para a pontuação final do deste modelo.

O desempenho dos modelos foi feito usando taxa de acerto (Hit Ratio). Dos modelos individuais, a SVM foi a que se saiu melhor, segundo o autor se deu ao fato de ela buscar minimizar o limite superior do erro de generalização ao invés de minimizar o erro de treinamento. O melhor modelo de todos, como é observado na tabela abaixo, foi o Modelo Combinatório. Segundo o autor diferentes métodos de classificação têm acesso a diferentes informações, por isso os resultados deles são diferentes, por isso que a combinação deles acaba por ser superior no resultado final.

3.5 Correlações entre os trabalhos e a pesquisa

Na tabela 1 apresentada logo abaixo, fizemos uma comparação dos 4 trabalhos e seus pontos mais relevantes para a nossa própria pesquisa. A coluna 1 se refere ao tipo de SVM, na literatura, observou-se que existem trabalhos que usam a SVM para problemas de classificação e outros para problemas de regressão relacionados ao mercado financeiro, porém nestas quatro pesquisas, todas a usam para problemas de classificação, até mesmo os trabalhos cujo o foco é Análise de Sentimentos.

O 3 primeiros trabalhos (\citeonline{trab1}; \citeonline{trab2}; \citeonline{trab3}) realizam uma junção de algortimos de aprendizagem de máquina (AM) aplicados em séries históricas, com análise de sentimentos (AS) em textos recolhidos de lugarem como blogs de notícias ou de redes sociais como o Twitter. Como é mostrado na coluna Áreas Abordadas da Tabela 1.

Além disso vale destacar que junto como o outro trabalho que analise apenas as séries históricas, as quatro pesquisas optaram por usar o algoritmo SVM para receber os dados trabalhados como entrada.

Com foco para o primeiro (\citeonline{trab1}) onde, além de realizar a junção de AM com AS, também implimentou uma estratégia de stop-loss para minimização dos riscos. O quarto trabalho (\citeonline{trab4}) nos auxiliará em relação ao treino e teste do dataset, as divisões dos dados e os parametros da SVM.

Na coluna do Dataset percebe-se que os ativos escolhidos pelos autores estão presentes nas bolsas de valores de seus respectivos países, isso já nos dá um direcionamento de onde vamos buscar o ativo para análise e estudo.

Na tabela 1, a coluna do Tam. Dataset mostra qual o número de dias de negociação escolhidos pelos autores, o tamanho da base de dados escolhido por eles. Isso nos dá uma ideia de quantos dados precisaremos recolher para a nossa própria pesquisa.

As técnicas escolhidas para análise de performance dos modelos criados são exibidas na coluna Métricas. Com esses dados temos uma noção de quais as principais métricas usadas na literatura.

A coluna Acurácia mostra o resultado dos modelos criados pelos autores no processo de classificação do preço dos ativos. Todas as pesquisas obtiveram resultados consistentes, mostrando que há viabilidade de investir tempo e recurso nesta área de pesquisa.

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

\section{Trabalho 1 -Forecasting Stock Market Movement Direction Using

Sentiment Analysis and Support Vector Machine }

No trabalho proposto por \citeonline{trab1} é desenvolvido um método que integra a análise de sentimentos em um sistema que usa aprendizagem de máqina baseado em Máquina de Vetor de Suportes (SVM). O intuito é prever a direção do movimento do SSE 50 Index, o índice da Bolsa de Valores de Xangai, e auxiliar o investidor na hora da tomada de decisão.

Com a ideia antecipar a tendência do mercado de ações, os autores implementaram uma SVM com fivefold cross validation. Para elimitar o viés de antecipação causada por esse método, uma abordagem com rolling window é utilizada. Por último é implementado uma estratégia de stop-loss, para reduzir e limitar as perdas e por consequência potencializar os ganhos dos investidores.

Para o DataSet o autor escolheu explorar o mercado da China, mais especificamente o índice da Bolsa de Valores de Xangai, o SSE 50 Index. Os dados recolhidos em relação aos preços foram: máxima e mínima do dia, preço de abertura e de fechamento, volume negociado, volume em RMB mudança em RMB e mudança em porcentagem. Eles foram retirados do Wind Economic Database. Já os dados textuais foram retirados dos fóruns Sina stock e Eastmoney stock. O período do DataSet vai de 17 de junho de 2014, até 7 de junho de 2016, somando um total de 485 dias de negociação.

A pesquisa realiza 3 abordagens diferentes, são elas: SVM com Fivefold Cross Validation, SVM com Rolling Windows e Logistic Regression (LR) com Fivefold Cross Validation. Para cada uma foram feitos 2 experimentos, o primeiro apenas com os dados históricos do ativo em questão. O segundo é uma combinação do primeiro com as características dos sentimentos. Tudo isso é observado na imagem abaixo. Uma fórmula de Acurácia é usada para o cálculo da performace dos algoritimos.

\begin{center}

Resultado das predições dos modelos propostos

\includegraphics[width=15cm]{images/imagem 1 do artigo 1.PNG}

Fonte: \citeonline{trab1}

\end{center}

A SVM com fivefold cross validation serviu para determinar a melhor função de kernel e os melhores parâmetros. Para o experimento 1 são: função de kernel RBF, C=256, Y=0,9942. Para o experimento 2 são: função de kernel RBF, C=181.0193, Y=0,005524.

O modelo com o melhor resultado foi a SVM com Rolling Windows. Nele a ideia é selecionar a melhor janela, usa-se N dias anteriores pra prever a direção do movimento dos próximos dias. Os parâmetros para os experimentos 1 e 2 são os mesmos do anterior. Repita em loop o procedimento alterando o valor de N até que a SVM atinja a maior precisão com os parâmetros a função de kernel que já temos selecionados. No experimento 1 a janela continua ideal é 68 com precisão de 71,33 porcento, para o experimento 2 é 76, com precisão de 89,93 porcento. É evidente que a adição dos sentimentos na SVM elevou bastante os resultados no segundo experimento.

AL é uma medida usada pra calculo dos pontos de estoque, por exemplo se um ativo é comprado pelo preço de 100 e vendido por 150, temos 50 pontos de lucro, logo o valor de Al é 50 pontos de estoque. Um calculo de drawdrown (MDD) também é feito, MDD é o declínio, o rebaixamento máximo de uma série de um pico até um vale durante um período de tempo determinado. Expected maximum drawdown (EDD) é uma expectativa da média de perdas máximas.

Os dados acima são importantes pois este trabalho propôs uma estratégia de stop-loss, a ideia dela é minimizar os prejuízos em períodos de perdas. Na tabela abaixo o último item representa os resultados da adição do stop-loss.

O limite de perdas foi definido em 95 pontos de estoque. Ou seja, se as perdas, no dia de negociação em questão, atingirem 95 pontos, a negociação aberta será fechada e finalizada. Na tabela é mostrado que essa estratégia melhorou os resultados, tanto dos pontos em Al quanto da redução dos valores de MDD e EMD. Pois em sequencias grande de perdas o stop-loss as limita, e menos perdas significa um lucro líquido maior.

\begin{center}

Performace dos modelos propostos

\includegraphics[width=14cm]{images/imagem 2 do artigo 1.PNG}

Fonte: \citeonline{trab1}

\end{center}

\section{Trabalho 2 - Sentiment analysis on social media for stock movement prediction }

No trabalho proposto por \citeonline{trab2}, o objetivo foi criar um modelo para previsão do preço de ações baseado nos sentimentos coletados dos textos, e como a adição de sentimentos ao modelo melhora os resultados, em comparação com o modelo que usa apenas os preços históricos. O diferencial deste trabalho é que em vez de analisar apenas -os sentimentos gerais, aqui é coletado também os sentimentos dos temas específicos de cada uma das 18 empresas analisadas neste trabalho, mostrando resultados superiores do modelo contendo os sentimentos, em relação ao modelo que analise apenas os preços históricos.

Duas bases de dados foram utilizadas, do período de 23 de julho de 2012 até 19 de julho de 2013. Elas foram divididas da seguinte forma: dos 249 dias de transações recolhidas, 171 foram destinadas para treino e 71 para testes. A métrica usada para avaliação do método proposto foi a Accuracy.

A primeira é o histórico de preços de 18 ativos, onde foram coletados do Yahoo Finance, e as informações de cada um são: preço de abertura, máxima, mínima, fechamento e fechamento ajustado. Dentre elas algumas empresas como: DELL, Amazon, Apple, Coca Cola, Ebay, Nvidia, Microsoft, Google, dentre outras.

A segunda base é o conjunto de dados de informações de humor. Onde foram coletados 18 quadros de mensagens das 18 ações, a fonte foi o Yahoo Finance Message Board. A vantagem destes quadros é que os usuários normalmente discutem notícias e previsões sobre as ações em questão, além disso um ponto importantíssimo obervado pelo autor é que em 15,6 porcento das mensagens coletadas possuem marcações de sentimentos nelas, são 5 tipos de marcações existentes: Compra forte, Compra, Espera,Venda e Venda Forte. Abaixo uma imagem com exemplo de texto coletado.

\begin{center}

Uma mensagem do quadro de mensagens da Apple (AAPL)

\includegraphics[width=16cm]{images/imagem 1 do artigo 2.PNG}

Fonte: \citeonline{trab2}

\end{center}

Para o desenvolvimento do algoritmo os autores escolheram a Máquina de Vetor de Suportes (SVM) com kernel linear como modelo para a previsão. Em cima disso foram desenvolvidos seis conjuntos de modelos distintos. Como é mostrado na imagem abaixo.

\begin{center}

Figura 3. Os conjuntos de modelos criados para predição

\includegraphics[width=15cm]{images/imagem 2 do artigo 2.PNG}

Fonte: \citeonline{trab2}

\end{center}

O sexto modelo Aspect-based sentiment obteve o melhor resultado entre todos eles. Aqui foi adicionado à SVM que utiliza os preços históricos, os sentimentos extraídos de cada mensagem do DataSet formado. Cada mensagem é representada como uma lista de tópicos e seus valores de sentimento correspondentes. No Aspect-based sentiment o tópico é substantivo da frase. A Imagem logo a seguir, de um exemplo de texto da base de dados, exemplifica isso, 'profit' (cuja a tradução é lucro) é o tópico, e 'up' (nesta frase tem sentido de 'irá aumentar ') é o sentimento positivo.

\begin{center}

Figura 4. Um exemplo de texto com o tópico e seu sentimento correspondente

\includegraphics[width=14cm]{images/imagem 3 do artigo 2.PNG}

Fonte: \citeonline{trab2}

\end{center}

Depois da divisão das mensagens em frases, a ferramenta Stanford CoreNLP é usada para fazer a marcação de partes do discurso em corpus e para lemantização de cada palavra em cada frase. Os substantivos consecutivos foram extraídos como tópicos das frases (os apareceram menos de 10 vezes, formam descartados).

Com a lista de tópicos formadas, os valores dos sentimentos de cada frase são extraídos. Neste processo as palavras que contém opiniões de cada frase analisada são mineradas usando um recurso léxico chamado SentiWordNet, ele serve para classificar as palavras atribuindo-lhes uma nota para o grau de positividade, negatividade e objetividade. Por fim essas pontuações obtidas de cada tópico naquela data de transação ‘t’ foram usadas como dados de entrada para o modelo de SVM.

O modelo Aspect-based sentiment obteve o melhor resultado médio para as 18 ações, com 54,41\%, como é exibido na figura de número 5 logo abaixo. Para algumas ações em especifico ele foi ainda maior, como por exemplo 71,05\% na Amazon (AMZN) e 64,47\% para a Dell (DELL). Os resultados também mostraram a evolução dos modelos quando adicionados nele a análise de sentimentos, o Aspect-based sentiment obteve 2,07\% médio a mais que o Price Only (algoritmo mostrado da figura 3 que usa apenas os preços históricos para predição do preço dos ativos).

\begin{center}

Figura 5. Tabela com os resultados de todos os modelos criados

\includegraphics[width=15cm]{images/imagem 33 do artigo 2.PNG}

Fonte: \citeonline{trab2}

\end{center}

\section{Trabalho 3 - Social Media Aided Sentiment Analysis for

Stock Prediction}

\citeonline{trab3} propuseram um método que emprega medidas combinadas de Indicadores da Análise Técnica (TI), Indicadores Fundamentais (FI) e Indicadores de Sentimentos (SI). São usados para previsão do preço das ações de 5 empresas distintas.

Com o intuito de auxiliar o investidor na compra e na venda de determinado ativo da bolsa de valores. O algoritmo escolhido foi a Máquina de Vetor de Suportes (SVM). No geral foram implementados vários algoritmos de SVM, e para os vários modelos desses algoritmos, foram também usados vários tipos de indicadores, que serviram como entrada para cada SVM.

Os indicadores de TI foram RSI, Money Flow Index (MFI), Exponential Moving Average (EMA), Stochastic Oscillator (SO) e Moving Average Convergence and Divergence (MACD), neste caso esses dados foram obtidos dos preços históricos dos ativos. Os de FI foram: Preço do ouro (GP), Taxa de Câmbio (ER) e o Preço do pretóleo bruto (COP). Por fim para o SI foi criado uma tabela com várias palavras chaves que possuem relação com o mercado financeiro, e foram classificadas como positiva ou negativas, alguns exemplos de palavras chave são: touro, urso, perigo, baixa, dentre outras. A seguir a imagem ilustrando o diagrama do algoritmo.

\begin{center}

Figura 6. Diagrama do modelo de predição

\includegraphics[width=15cm]{images/imagem 1 do artigo 3.PNG}

Fonte: \citeonline{trab3}

\end{center}

O Dataset utilizado referente ao histórico do preço das ações foi retirado das bolsas de valores National Stock Exchange (NSE) e Bombay Stock Exchange (BSE), e do Yahoo Finance. O Dataset referente aos textos que contém os sentimentos das pessoas foi retirado do Money Control e do Twitter. As empresas escolhidas foram: Infosys, Wipro, Hindustan Computer Limited (HCL), Tata Consultancy Services (TCS) e a Tech Mahindra, junto com os índices de ações SENSEX e Nifty da BSE e o índice BSE. Por último o preço do ouro (GP), taxa de câmbio (ER) e o preço do petróleo bruto (COP) também foram coletados para o dataset.

Logo depois, na parte de pré-processamento, os dados recolhidos são tratados, como remoção de stop-words, informações sem relevância para a opinião expressa no texto, e lematização. Para o histórico de preços, foi usado as informações de: Máxima do preço, Mínima, Abertura, Fechamento, e o Volume. O preço de fechamento foi escolhido para ser o valor previsto pelos algoritmos.

Sobre as regras, se o modelo prever que a tendência do dia (t+1) for de alta, então invista no ativo, se a tendência for de baixa, então venda, e por último se não houver tendência prevista, não venda.

Os resultados dos modelos de SVM foram calculados usando as métricas: Acurracy, Sensitivity, Specificity, RMSE e MAPE. No geral o modelo que combina os 3 tipos de indicadores (CI) se saiu melhor do que todos os outros que usavam indicadores individualmente (TI, SI e FI) ou uma combinação de dois deles (TI\&SI e FI\&SI).

Isso fica evidente pela baixa taxa de erro observada nas tabelas onde mostram os resultados do RMSE e MAPE. E na alta acurácia, sensibilidade e especificidade que também são exibidas em tabelas com os resultados das métricas Acurracy, Sensitivity e Specificity. Logo abaixo duas tabelas com os resultados das métricas Acurracy e RMSE respectivamente.

\begin{center}

Figura 7. Acurácia dos modelos de previsão criados

\includegraphics[width=15cm]{images/imagem 2 do artigo 3.PNG}

Fonte: \citeonline{trab3}

\end{center}

\begin{center}

Figura 8. RMSE dos modelos de previsão criados

\includegraphics[width=15cm]{images/imagem 3 do artigo 3.PNG}

Fonte: \citeonline{trab3}

\end{center}

\section{Trabalho 4 - Forecasting stock market movement direction with

support vector machine}

Nesta pesquisa \citeonline{trab4} decidiram investigar a previsibilidade da direção do movimento semanal do índice NIKKEI 225, utilizando o algoritmo Máquina de Vetor de Suportes (SVM). O resultado dela foi comparado com outros 4 algoritmos, e a SVM se saiu melhor que todas elas. Por fim o autor desenvolveu um modelo final que combina os 5 algoritmos. Por fim propõe um modelo que integra todos os 4 algoritmos anterior, e esse se saiu melhor que todos os outros nos testes realizados.

Em relação ao dataset, os autores usaram o índice NIKKEI 225, que compõe 225 ações da bolsa de valores de Tóquio. O período da base de dados se estende de 1 de janeiro de 1990 até 31 de dezembro de 2002. Com um total de 676 pares de observações. Eles recolheram esses dados da seção de finanças do Yahoo e do Pacific Exchange Rate Service. A base de dados foi dividida em duas partes, a primeira com 640 pares de observações destinada ao treino, para determinar as especificações e parâmetros do modelo, e a segunda com 36 partes para teste, para comparação das performances entre os modelos criados.

Dois dados macroeconômicos foram escolhidos como entrada para a SVM. São eles S\&P 500 Index e o par de moedas JPY/USD. Os 4 algoritmos usados para comparar com ele são: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Elman Backpropagation Neural Networks (EBNN) e Random Walk Model (RW). O modelo final, que combina todos esses anteriores se chama Modelo Combinatório.

O modelo combinatório usa um sistema de pesos que atribui pontos a cada modelo anterior baseado em sua performance na fase de treinamento. Aqueles que se saíram melhores receberam pesos maiores, logo eles obtiveram uma contribuição relativa maior para a pontuação final do deste modelo.

O desempenho dos modelos foi feito usando taxa de acerto (Hit Ratio). Dos modelos individuais, a SVM foi a que se saiu melhor, segundo o autor se deu ao fato de ela buscar minimizar o limite superior do erro de generalização ao invés de minimizar o erro de treinamento. O melhor modelo de todos, como é observado na tabela abaixo, foi o Modelo Combinatório. Segundo o autor diferentes métodos de classificação têm acesso a diferentes informações, por isso os resultados deles são diferentes, por isso que a combinação deles acaba por ser superior no resultado final.

\begin{center}

Figura 9. Performace da previsão de todos os modelos criados

\includegraphics[width=16cm]{images/imagem 1 do artigo 4.PNG}

Fonte: \citeonline{trab4}

\end{center}

3.3Social Media Aided Sentiment Analysis for Stock Prediction

Nirmaladevi e Krishnamoorthy (2019) propuseram um método que emprega medidas combinadas de Indicadores da Análise Técnica (TI), Indicadores Fundamentais (FI) e Indicadores de Sentimentos (SI), que são descritos nos próximo parágrafo. São usados para previsão do preço das ações de 5 empresas: Infosys, Wipro, Hindustan Computer Limited (HCL), Tata Consultancy Services (TCS) and Tech Mahindra. Elas estão localizadas nas Bolsas de Valores da Índia.

Com o intuito de auxiliar o investidor na compra e na venda ativos da bolsa de valores. O algoritmo escolhido foi a Máquina de Vetor de Suportes (SVM do inglês Support VectorMachine). No geral o autor implementou modelos de SVM que usaram os indicadores descritos anteriormente como entrada para o algoritmo. Ele mesclou entre, modelos que usaram individualmente cada indicador como entrada, e modelos que usaram combinação de dois ou mais indicadores para saber qual teria o melhor resultado possível.