

UNIVERSIDADE FEDERAL FLUMINENSE
INSTITUTO DE CIÊNCIA E TECNOLOGIA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

JOÃO MACHADO DA SILVA JUNIOR

Criação de estimativas para o valor da criptomoeda Bitcoin utilizando
comentários extraídos de uma rede social

Rio das Ostras – RJ

2020

JOÃO MACHADO DA SILVA JUNIOR

Criação de estimativas para o valor da criptomoeda Bitcoin utilizando
comentários extraídos de uma rede social

Monografia apresentada ao Curso de
Bacharelado em Ciência da Computação
do Instituto de Ciência e Tecnologia,
como requisito parcial para obtenção do
Grau de Bacharel.

Orientador:

Prof. Dr. CARLOS BAZÍLIO MARTINS

Rio das Ostras – RJ

2020

JOÃO MACHADO DA SILVA JUNIOR

Criação de estimativas para o valor da criptomoeda Bitcoin utilizando
comentários extraídos de uma rede social

Monografia apresentada ao Curso de
Bacharelado em Ciência da Computação
do Instituto de Ciência e Tecnologia,
como requisito parcial para obtenção do
Grau de Bacharel.

Aprovado no mês de _____ do ano de _____

BANCA EXAMINADORA

Prof. Dr. CARLOS BAZÍLIO MARTINS - Orientador

UFF

Prof. NOME DO PROFESSOR

INSTITUIÇÃO

Prof. NOME DO PROFESSOR

INSTITUIÇÃO

Rio das Ostras – RJ

2020

Resumo

Bitcoin é uma criptomoeda descentralizada baseada em uma tecnologia ponto-a-ponto proposta como uma forma de dinheiro eletrônico global. Como a maioria das moedas contemporâneas, o Bitcoin é uma moeda fiduciária, isso significa que seu valor não está lastreado em nenhum metal ou outro recurso material, seu valor depende da confiança (fidúcia) das pessoas. Esse fato, adicionado ao fato de não haver nenhuma instituição reguladora que ofereça garantia de qualquer tipo para seus usuários, fizeram com que o Bitcoin se tornasse um ativo financeiro especulativo. Como todo ativo sujeito a especulação financeira, seu valor é influenciado pela opinião que seus investidores tem sobre sua tendência de preço. Este trabalho apresenta o desenvolvimento de uma aplicação capaz de extrair comentários sobre o Bitcoin de uma rede social, capaz também de capturar o histórico de transações da moeda, e propor múltiplos modelos de análise para estimar o valor futuro do Bitcoin a partir das informações capturadas. A aplicação utiliza técnicas de mineração de dados e de mineração de texto para abastecer um banco de dados. O desenvolvimento dos modelos de previsão propostos inclui processamento de linguagem natural, aprendizado de máquina, regressão logística e análise de sentimentos. O resultado das análises e uma descrição menos formal do projeto estão disponíveis em um website desenvolvido para a apresentação desses resultados. O website está disponível na rede mundial de computadores, no endereço <https://criptomante.online>.

Abstract

Bitcoin is a decentralized cryptocurrency based on a peer-to-peer technology proposed as an electronic cash system. Like most contemporary currencies, Bitcoin is a fiat currency, this means that Bitcoin has no intrinsic value, it has value because parties engaging in exchange agree to its value. This fact, added to the fact that there is no regulatory institution that offers guarantees of any kind to its users, has made Bitcoin a speculative asset. Like all speculative assets, its value is influenced by the opinion and thoughts of its investors about its price trends. This paper presents the development of an application capable of extracting content about Bitcoin from a social network, also capable of retrieve the currency's transaction history, and proposes four analysis models to estimate the future value of Bitcoin from this information. This project uses data mining and text mining techniques to supply a database. The development of the proposed forecasting models includes natural language processing, machine learning, logistic regression and sentiment analysis. The results of the analysis and a less formal description of the project are available on the project's website. The website is available on the internet at <https://criptomante.online>.

Lista de ilustrações

Figura 1: Etapas de aprovação de uma transação utilizando um bitcoin	3
Quadro 1: Criptomoedas com maior capitalização de mercado	4
Figura 2: Exemplo de aplicação de Bandas de Bollinger	6
Figura 3: Captura de tela da plataforma online desenvolvida pela empresa FinBrain.....	11
Figura 4: Diagrama apresentando os principais módulos do projeto.....	14
Figura 5: Fluxograma do processo de captura de transações com bitcoin.....	18
Quadro 2: Descrição das tuplas que definem um snapshot	20
Tabela 1: Tuplas de um snapshot com data de referência em 17 de outubro de 2020	21
Tabela 2: Snapshot com data de referência em 21 de junho de 2015	22
Figura 6: Gráfico exibindo a cotação do Bitcoin entre 18/10/2019 e 18/10/2020	22
Tabela 3: Snapshots com maior pontuação durante a quarta etapa	23
Figura 7: Resultado da estimativa obtida com a Análise por padrões Numéricos.....	25
Tabela 4: Frases do dia 19 de outubro de 2020 que demonstraram alguma tendência	26

Sumário

1	Introdução.....	1
2	Referencial Teórico	2
2.1	Financeiro.....	2
2.1.1	Criptomoedas	2
2.1.2	Bitcoin como Ativo Financeiro	5
2.1.3	Especulação Financeira	5
2.1.4	Princípio de Ondas de Elliott.....	7
2.2	Web Crawlers	7
2.3	Processamento de Linguagem Natural	7
2.4	Aprendizado de Máquina.....	8
2.4.1	Classificação Estatística	8
2.4.2	Regressão Logística.....	9
3	Trabalhos Relacionados.....	10
3.1	I Know First.....	10
3.2	FinBrain Technologies	11
4	Desenvolvimento.....	12
4.1	Tecnologias Utilizadas	14
4.1.1	Linguagens de programação.....	14
4.1.2	Bibliotecas e Frameworks	14
4.1.3	Banco de Dados.....	15
4.1.4	Spacy – Processador de Linguagem natural	15
4.1.5	Scikit-Learn – Motor de Aprendizado de Máquina	16
4.2	Mineração de Dados	16
4.2.1	Captura de Transações.....	16
4.2.2	Captura de comentários em rede social	18
4.3	Análise por padrões numéricos.....	19
4.3.1	Estrutura de dados Snapshot.....	19
4.3.2	Roteiro da análise	21
4.4	Análise por repetição de comentários.....	24
4.4.1	Roteiro.....	24

4.5 Análise de sentimentos	26
4.5.1 Tratamento de texto.....	27
4.5.2 Roteiro da análise	28
4.6 Análise consolidada.....	29
4.7 Aplicação Web e Deploy	29
5 Conclusão	31
5.1 Desafios.....	31
5.2 Resultados Obtidos.....	33
5.3 Sugestões de Trabalhos Futuros	33
Bibliografia.....	34

Capítulo 1

Introdução

Em 2008 uma pessoa chamada Satoshi Nakamoto publicou um artigo chamado *Bitcoin: Um Sistema de Dinheiro Eletrônico Ponto-a-Ponto* [2][44] em um grupo de e-mail do site metzdowd.com. Esse artigo propunha uma nova forma de dinheiro, eletrônico, descentralizado e independente de qualquer instituição ou governo, que poderia revolucionar o jeito como pagamos pelo o que compramos. Esse dinheiro eletrônico seria chamado de Bitcoin.

Doze anos já se passaram desde então. O Bitcoin já é razoavelmente bem conhecido em todo o mundo, embora poucas pessoas já tenham adquirido qualquer fração de um bitcoin. O Bitcoin nunca se tornou uma moeda popular para transações no dia a dia. Se tornou, contudo, um ativo para especulação financeira. É mais comum que seja adquirido como forma de investimento do que como moeda a ser utilizada em uma transação comercial.

Nesse trabalho desenvolvemos uma aplicação capaz de obter o histórico de preços do Bitcoin a partir de consultas a uma API da plataforma Bitcoincharts¹. Ao utilizar esse histórico, em conjunto com comentários de usuários na rede social Reddit², também extraídos pela aplicação desenvolvida nesse trabalho, foi possível elaborar quatro heurísticas para estimar o valor futuro do Bitcoin. As heurísticas apresentadas utilizam conceitos básicos e simples de especulação financeira, em especial o Princípio de Ondas de Elliott [22], a partir de seu histórico de valor, análise de sentimentos, aprendizado de máquina e processamento de linguagem natural.

O capítulo 2 dessa monografia irá apresentar o referencial teórico necessário para a compreensão do projeto. Na sessão 2.1 serão introduzidos conceitos de viés financeiro, como criptomoedas, especulação financeira e o Princípio de Ondas de Elliott. Na sessão 2.2 será apresentado o conceito de Web Crawlers. Na sessão 2.3 serão apresentados conceitos de processamento de linguagem natural e na sessão 2.4 serão apresentados conceitos relacionados a aprendizado de máquina. O capítulo 3, de título *Trabalhos Relacionados*, serão apresentadas duas iniciativas comerciais de prestação de serviço de consultoria de previsão de ações na bolsa de valores utilizando tecnologia.

No capítulo 4, de título *Desenvolvimento* serão apresentados os detalhes do desenvolvimento do projeto. A introdução do capítulo 4 irá apresentar a arquitetura geral do projeto. A sessão 4.1 irá apresentar as tecnologias de banco de dados, linguagens de programação, bibliotecas e frameworks utilizadas nesse projeto. A sessão 4.2 irá apresentar como esse projeto utilizou técnicas de mineração de dados para obter as informações necessárias para as análises desenvolvidas. As sessões 4.3 a 4.6 irão apresentar os quatro modelos de análise para previsão do valor futuro do Bitcoin propostos por esse projeto. A sessão 4.7 irá expor detalhes do desenvolvimento e deploy da página web criada para exposição dos dados produzidos.

O capítulo 5, de título *Conclusão*, irá expor os desafios encontrados durante o desenvolvimento do projeto na sessão 5.1. Na sessão 5.2 serão apresentados os resultados obtidos com o projeto, tanto no viés tecnológico quanto na precisão financeira das análises desenvolvidas. Por fim, na sessão 5.2 serão apresentadas algumas sugestões para trabalhos futuros. O código-fonte completo do projeto é aberto e está disponível em um repositório online, no endereço https://github.com/granpk/Criptomante_python

¹ Bitcoincharts é uma plataforma online que oferece informações financeiras sobre o Bitcoin. Está disponível no endereço <https://bitcoincharts.com>

² Reddit é uma rede social popular que possui um grande número de postagens sobre o Bitcoin. Está disponível no endereço <https://reddit.com>

Capítulo 2

Referencial Teórico

Esse capítulo irá introduzir os conceitos necessários para que se compreenda o trabalho desenvolvido. Esse projeto utiliza técnicas de mineração de dados, mineração de texto, aprendizado de máquina e processamento de linguagem natural para criar quatro modelos de análise para previsão do futuro valor do Bitcoin.

Nesse projeto, a criptomoeda Bitcoin será tratada como um ativo financeiro especulativo. Portanto, será necessário introduzir o que é uma criptomoeda e o que é o Bitcoin. Também será necessário introduzir conceitos financeiros básicos.

2.1 Financeiro

O referencial teórico financeiro apresentado nesse trabalho se concentra em explicar ao leitor os conceitos básicos do mercado especulativo e de criptomoedas necessários para a compreensão do projeto desenvolvido. Assim, será necessário introduzir o que são moedas fiduciárias; o que são criptomoedas; como o contexto tecnológico das criptomoedas afeta o seu valor; por que o bitcoin é usado como ativo de especulação financeira; o que é uma especulação financeira e o que é o Princípio de Ondas de Elliott que foi usado como base teórica para os modelos de previsão elaborados nesse projeto.

2.1.1 Criptomoedas

Em termos econômicos, moeda é tudo aquilo que é geralmente aceito para liquidar débitos e transações, isto é, para pagar pelos bens e serviços e para quitar obrigações. Ela é considerada o instrumento básico para que se possa operar no mercado, pois a moeda atua como meio de troca.

As moedas contemporâneas são chamadas de moedas fiduciárias porque seu funcionamento baseia-se na confiança (fidúcia) das pessoas nas instituições que emitem e regulam a moeda. Ao contrário de outros tipos de moedas mais antigos, moedas fiduciárias não tem seu valor lastreado em nenhum metal e não possui nenhum valor intrínseco. [1]

Em 2008 um artigo chamado *Bitcoin: Um Sistema de Dinheiro Eletrônico Ponto-a-Ponto* foi publicado em uma lista de e-mails aberta relacionada com o site metzdowd.com. O artigo era assinado por alguém chamado Satoshi Nakamoto (presumidamente um pseudônimo) e descrevia um sistema de dinheiro eletrônico baseado em transações ponto-a-ponto. [2]

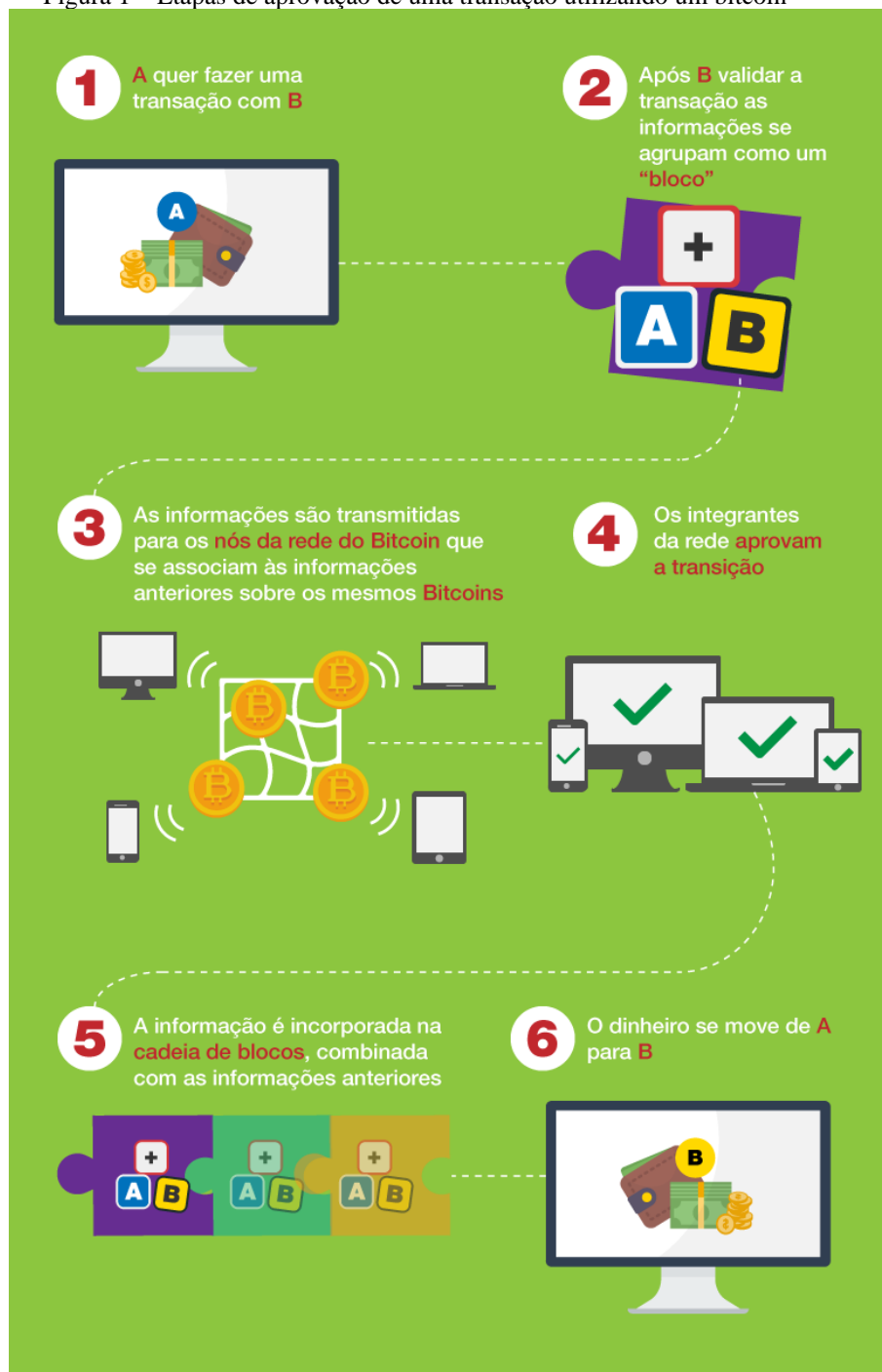
No mesmo artigo, Nakamoto propõe uma tecnologia contendo um modelo banco de dados distribuído e seguro que posteriormente veio a ser chamado de Blockchain. Segundo NAKAMOTO (2008b), o Blockchain funcionaria como um livro-razão público onde se registra todas as transações com a criptomoeda. Sua segurança está baseada no fato de que as transações armazenadas na rede distribuída podem ser validadas por qualquer nó da rede utilizando certos algoritmos. E, portanto, para que houvesse qualquer fraude seria necessário obter o consenso de outros nós da rede.

O escopo do Blockchain proposto por Nakamoto possui as seguintes características:

- Transações podem ser agrupadas. Esse agrupamento se chama **bloco**.
- Cada bloco possui referência para um bloco mais antigo. Dessa forma, é possível percorrer uma cadeia de blocos desde o bloco atual até o primeiro bloco da rede (chamado de bloco gênese).
- Todos os nós da rede são capazes de validar um bloco a partir de sua cadeia, utilizando certos algoritmos.
- Após um nó validar um novo bloco, ele propaga o bloco validado pela rede. Isso é conhecido como adicionar o bloco à Blockchain

- A propagação do bloco evita que ocorra um problema econômico chamado de gasto-duplo, onde o mesmo ativo financeiro é usado como troca em duas transações distintas.
- Existem sites que fornecem o serviço de consulta de quantas ‘validações’ um bloco recebeu. Não há regra sobre a quantidade necessária de validações para que um bloco seja considerado válido. Mas a maioria das plataformas consideram que um bloco com seis validações é um bloco que foi suficientemente validado para que o beneficiário reconheça que recebeu o pagamento.
- Os usuários recebem bitcoins gratuitos como incentivo por fornecerem seus computadores para serem nós da rede, validando novos blocos e realizando outras ações de rotina. Essa atividade é chamada de Mineração de Bitcoin, e não terá seu processo aprofundado nesse projeto.

Figura 1 – Etapas de aprovação de uma transação utilizando um bitcoin













Fonte: WIRECARD (2016) [3]

O código-fonte da primeira versão do banco de dados Blockchain foi tornado aberto por Satoshi Nakamoto. Desde então, o Blockchain ganhou popularidade e contribuidores ao redor do mundo. Eventualmente, bancos de dados derivados do Blockchain de Nakamoto foram usados como base para a criação de outras criptomoedas, popularmente chamadas de altcoin [4].

As motivações para o desenvolvimento e uso de novas criptomoedas são as mais diversas. Por exemplo, a criptomoeda *Litecoin* utiliza uma versão mais leve do Blockchain e isso permite maior velocidade para as transações [5]. Já a *Gridcoin* utiliza uma versão da Blockchain onde parte de sua rede de nós é usada para executar tarefas para projetos científicos de pesquisadores que precisam de poder computacional, mas não possuem os recursos financeiros necessários [6].

Segundo o site coinmarketcap.com, a tabela abaixo expressa as 10 criptomoedas de maior **capitalização de mercado** (valor atual da moeda multiplicado pela quantidade de moedas em circulação).

Quadro 1 – Criptomoedas com maior capitalização de mercado

#▲	Nome	Preço	Cap. de Mercado ⓘ
1	 Bitcoin BTC	\$16,257.89	\$301,394,837,735
2	 Ethereum ETH	\$472.67	\$53,608,486,554
3	 Tether USDT	\$1.00	\$17,840,013,540
4	 XRP XRP	\$0.264599	\$11,993,761,976
5	 Chainlink LINK	\$12.81	\$5,017,088,985
6	 Bitcoin Cash BCH	\$257.94	\$4,790,303,660
7	 Litecoin LTC	\$65.25	\$4,298,338,785
8	 Binance Coin BNB	\$28.10	\$4,057,675,924
9	 Polkadot DOT	\$4.53	\$3,978,105,489
10	 Cardano ADA	\$0.106073	\$3,300,179,948

Fonte: COINMARKETCAP (2020) [7]

2.1.2 Bitcoin como Ativo Financeiro

A primeira transação comercial conhecida utilizando Bitcoin ocorreu em 2010 quando o programador Laszlo Hanyecz comprou duas pizzas pelo valor de 10.000 bitcoins [8].

Por sua natureza descentralizada e de difícil rastreabilidade, o Bitcoin ganhou popularidade como moeda para tráfico de produtos ilegais e outras atividades criminosas. A plataforma de comércio ilegal de drogas Silk Road utilizou exclusivamente Bitcoins entre fevereiro de 2011 até outubro de 2013, quando seus responsáveis foram identificados pelo FBI. Durante esse período, estima-se um movimento de 9,9 milhões de bitcoins. Essas atividades ilegais com bitcoin contribuíram para a valorização da moeda. Que avançou de \$0,30 em 2011 para \$ 770 em 2013 [9] [10].

O valor do Bitcoin teve ciclos de quedas e aumentos acentuados desde então. Esses movimentos foram causados principalmente por fenômenos que afetavam a confiabilidade que o público tinha sobre a moeda. Um bug no Blockchain causou uma queda no valor do Bitcoin de 23% em um único dia em 2013 [11]. Posteriormente, outras quedas foram causadas por consequência da proibição do governo chinês de transações com Bitcoin em seu território [12].

De acordo com o economista Mark T. Williams, a volatilidade do Bitcoin é sete vezes maior do que a do ouro e dezoito vezes maior do que a do Dólar Americano [13]. A Fundação Bitcoin, formada por defensores e entusiastas da criptomoeda, afirma que sua alta volatilidade se deve a sua baixa liquidez.

Atualmente há inúmeros sites que oferecem o serviço de corretagem de bitcoin. Esses sites oferecem uma interface gráfica para a venda e compra da criptomoeda. No Brasil, alguns sites oferecem inclusive informes de rendimento para auxiliar na declaração de imposto de renda. Nesse projeto, estudaremos principalmente transações no site estado-unidense <https://www.bitstamp.net/>, em operação desde 2011.

2.1.3 Especulação Financeira

Especulação Financeira é a compra de ativos com objetivos que não sejam o uso do ativo, mas sim a expectativa de obtenção de lucros em uma venda futura de maior valor [14]. Muitos especuladores prestam pouca atenção ao valor intrínseco do ativo que estão comprando. Esses se interessam principalmente pelo movimento nos preços dos ativos para obter lucros.

O uso do Bitcoin como moeda de troca, com o objetivo de realizar transações comerciais, não se enquadra como especulação financeira. Porém, a aquisição de bitcoins com o objetivo de obter lucros em uma venda futura de maior valor se enquadra como uma operação de especulação financeira.

Como o Bitcoin é uma moeda fiduciária sem nenhum governo que possa intervir em sua precificação, o Bitcoin passou a ser considerado um ativo sem nenhum valor intrínseco, e, portanto, ideal para a especulação financeira. O movimento em seus valores se deve em grande parte às expectativas de seus especuladores [15].

Existem diversas técnicas de análise para realizar o estudo de viabilidade de lucro com a compra e venda de determinado ativo. Em geral essas técnicas podem ser classificadas em três tipos: Análise Técnica, Análise Fundamentalista e Análise de Sentimentos [16].

A Análise Fundamentalista é o estudo da situação financeira, econômica e mercadológica de uma empresa, setor econômico, moeda ou qualquer outro ativo. Para o estudo de uma empresa, por exemplo, seria necessário estudar a área de atuação da empresa, seus concorrentes, suas demonstrações contábeis e as possíveis novidades tecnológicas que poderiam afetar sua área de negócio [17].

Quando aplicada ao Bitcoin, a Análise Fundamentalista estudaria o caráter tecnológico da criptomoeda. Como suas características de escalabilidade e de segurança. Outro fator que pode ser estudado é a popularização do uso do Bitcoin como moeda para transações comerciais, que aumentaria a sua demanda e poderia diminuir a volatilidade de seu preço.

A Análise Técnica é o estudo dos movimentos de preço de determinado ativo, baseado nos ciclos de oferta e procura. Nessa análise há interesse apenas nos padrões de movimento do preço, e não há interesse no motivo que levou a esses movimentos. Pode ser vista tanto como uma forma de psicologia social aplicada como pesquisa de opinião, onde padrões gráficos visuais ou estatísticos percebidos, seriam como fotografias do comportamento dos participantes do mercado em determinado momento [17].

Como exemplo de aplicação de análise técnica, segue abaixo a aplicação da técnica Bandas de Bollinger em um gráfico. Essa é uma das mais tradicionais análises técnicas utilizadas para que investidores saibam quando é a hora de comprar ou vender um ativo. Nesse gráfico, o ativo em estudo são ações da empresa Vale do Rio Doce. A variação das ações ao longo de um dia é exibida no formato de um retângulo verde ou vermelho. As bandas são dadas pelas seguintes fórmulas:

- Banda superior: Média Móvel Simples (20 dias) + (2xDesvio Padrão (20 dias))
- Banda inferior: Média Móvel Simples (20 dias) – (2xDesvio Padrão (20 dias))
- Linha central: Média Móvel Simples (20 dias)

Figura 2 – Exemplo de aplicação de Bandas de Bollinger



Fonte: MAISRETORNO (2019) [19]

Uma interpretação comum para o gráfico acima é comprar o ativo quando o valor do ativo atingir a banda inferior, e vender o ativo quando o valor do ativo atingir a banda superior. Note que não há interesse em entender o motivo de variação no preço do ativo.

A análise proposta nesse trabalho nomeada de *Análise por Padrões Numéricos* se classifica como uma técnica de Análise Técnica. Existem muitas técnicas do tipo “análise técnica” disponíveis na literatura financeira.

A Análise de Sentimentos se aproveita do fato de que o mercado não é movido apenas por mudanças fundamentais dos ativos, mas também pelos sentimentos e opiniões das pessoas, frequentemente de curto-prazo. O uso de inteligência artificial, aprendizado de máquina e processamento de linguagem natural pode fornecer técnicas sistematizadas de análises de sentimentos.

O termo “*análise de sentimentos*” escrito nessa sessão se refere a uma técnica financeira, e sua aplicabilidade não é dependente do uso da computação. O mesmo termo é usado como nome de uma junção de técnicas de processamento de linguagem natural e inteligência artificial que tem como objetivo classificar publicações de usuários sobre um serviço, produto, objeto ou tema usando algum algoritmo classificador [18].

2.1.4 Princípio de Ondas de Elliott

O Princípio de Ondas de Elliott é um conceito de Análise Técnica proposto por Ralph Nelson Elliott na década de 1930 que propõe que há ciclicidade não apenas no valor de ativos do mercado financeiro, mas também na psicologia dos investidores, nas opiniões populares e em outros fatores coletivos e sociais. Atualmente esse princípio é bem conhecido na literatura financeira e é um conceito importante para a análise técnica e para a análise de sentimentos [21].

Elliott propôs que a psicologia coletiva se move entre otimismo e pessimismo em ciclos perceptíveis. E que esse movimento de emoções gera padrões cíclicos nos valores de ações na bolsa de valores. No mesmo livro em que Elliott propôs esse princípio, ele também propõe um modelo de análise técnica para identificar esses ciclos e como usá-los para prever o valor futuro de um ativo qualquer.

Sua base de raciocínio é a de que a emoção surge primeiro que a ação. Por isso, a representação gráfica de uma série histórica de cotações de um ativo nada mais é do que a oscilação de humor do coletivo numa tentativa desesperada de encontrar sua precificação.

Segundo ELLIOTT (2020), o público age de forma emocional, subjetiva e impulsiva, tomando decisões em condições de ignorância e incerteza, e, na maioria das vezes, assumindo a chamada *atitude de manada*.

Existem diversas técnicas de análise fundamentadas nesse princípio em maior ou em menor grau, como por exemplo o Modelo de Sentimentos de Thovallo [22]. De forma semelhante, nesse trabalho o Princípio de Ondas de Elliott será um conceito importante na construção das análises que o trabalho propõe. Contudo, o método específico de interpretação de gráficos que Elliott propôs em seu livro não será o método aplicado por esse trabalho.

2.2 Web Crawlers

No contexto de software, um *bot* é uma aplicação de software concebida para simular ações humanas repetidas vezes de maneira padrão, da mesma forma como faria um robô. Um uso comum de bots é para a execução de testes automatizados, onde a interação humano-computador é definida em scripts e simulada pelo bot. Nesse caso, é possível verificar se a aplicação está fornecendo o resultado esperado para um conjunto de ações de usuário pré-definido.

Um Web Crawler é um bot que navega pela rede mundial de computadores, tipicamente para propósitos de indexação de conteúdo. Acredita-se que metade do tráfego de internet atualmente seja realizado por bots [23].

Além do uso de Web Crawler para indexação de conteúdo, normalmente utilizado por buscadores, outro uso comum é para mineração de dados. Uma vez que é possível simular o comportamento de um usuário acessando determinado website, é possível criar um Web Crawler para ler, categorizar e armazenar em um banco de dados todas as informações deste website que estariam disponíveis para esse usuário.

Este trabalho utiliza um Web Crawler para realizar a captura de comentários de usuários da plataforma Reddit.

2.3 Processamento de Linguagem Natural

Processamento de linguagem natural é uma subárea da ciência da computação que estuda as interações entre computadores e linguagens humanas naturais. Em particular, como criar programas capazes de gerar sentenças em linguagem humana natural ou capazes de extrair informação a partir de sentenças em linguagem humana natural [20]

O desenvolvimento de aplicações capazes de extrair informação a partir de sentenças em linguagem humana natural frequentemente aplica técnicas de tratamento de texto que diminuem ou agrupam partes das sentenças com o objetivo de facilitar a compreensão das sentenças. Essas são chamadas de técnicas de normalização. Segue abaixo algumas técnicas comuns:

- **Stemização:** É o processo de reduzir palavras flexionadas (ou às vezes derivadas) ao seu tronco (stem), base ou raiz, geralmente uma forma da palavra escrita. Exemplo: A palavra “meninas” se reduziria a “menin”, assim como “meninos” e “menininhos” [24].
- **Lematização:** Técnica de converter uma palavra em seu Lema. Embora seja semelhante ao processo de stemização, o processo de lematização é mais avançado pois compreende o sentido e o contexto de cada palavra. Por exemplo: A palavra “meninas” ainda possui o lema “menin”, porém a palavra “melhor” possui o lema “bom”. Além disso, é uma prática comum que todos os pronomes, pessoal ou de tratamento, sejam reduzidos ao lema “pronome”. [24]
- **Remoção de palavras-vazias:** Técnica de remoção de palavras que possuem pouco ou nenhum significado relevante para um software específico. Não existe uma lista universal para palavras-vazias, pois sua definição depende do objetivo do software. Embora seja comum que artigos ou palavras muito comuns sejam incluídos nessa lista. Exemplo: A frase “Eu tenho certeza absoluta que chegarei na hora marcada” pode ser reduzida a “Eu tenho certeza que chegarei na hora marcada” ou até mesmo a “Eu chegarei na hora marcada”.

Embora existam outras técnicas, as técnicas acima são as técnicas mais comuns e são as técnicas empregadas por este trabalho para o tratamento das mensagens capturadas da rede social, e posteriormente, seu uso no modelo de aprendizado de máquina.

2.4 Aprendizado de Máquina

Aprendizado de máquina é um subcampo da ciência da computação que evoluiu do estudo de reconhecimento de padrões. O aprendizado de máquina explora o estudo e construção de algoritmos que podem aprender de seus erros e fazer previsões sobre dados. Tais algoritmos operam construindo um modelo a partir de uma entrada amostral [25].

Existem diversas técnicas de aprendizado de máquina. Nesse projeto, serão empregadas técnicas de *classificação estatística* e *regressão logística* que se enquadram em um grupo de técnicas chamado de *aprendizado supervisionado*.

Em técnicas do tipo *Aprendizado supervisionado* são apresentadas ao computador exemplos de entradas e saídas desejadas, fornecidas por um "professor". O objetivo é que o algoritmo analise as entradas e as saídas fornecidas e procure aprender uma regra geral que mapeie as entradas para as saídas. Como o mapeamento é aprendido a partir de uma *amostra de treinamento*, a aplicabilidade dessa técnica fornece melhores resultados de precisão conforme aumentamos o tamanho da amostra de treinamento.

2.4.1 Classificação Estatística

Classificação estatística é o problema de identificar a qual de um conjunto de categorias uma nova observação pertence, com base em um conjunto de dados de treinamento contendo observações cuja categoria é conhecida. Alguns exemplos de classificação estatística podem ser atribuir um determinado e-mail à classe “spam” ou “não spam”, dado que há um conjunto de e-mails cuja classificação é conhecida e que serviram de treinamento para o software.

Classificação Estatística é considerada uma instância de aprendizado supervisionado, e é um exemplo de reconhecimento de padrões.

Um algoritmo que implementa a classificação, especialmente em uma implementação concreta, é conhecido como classificador. O termo “classificador” às vezes também se refere à função matemática, implementada por um algoritmo de classificação, que mapeia os dados de entrada para uma categoria [26].

Existem diversos algoritmos classificadores. Este projeto apresenta uma solução de uso de um classificador oferecido pela biblioteca *scikit-learn* baseado em *regressão logística*. Este é um dos algoritmos mais usados em problemas de análise de sentimentos.

2.4.2 Regressão Logística

A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas [28].

É usada para modelar a probabilidade de uma determinada classe ou evento existir, como aprovação/reprovação, vitória/perda, vivo/morto ou saudável/doente. Como exemplo nesse trabalho, regressão logística é usada para determinar se o valor do Bitcoin irá subir/diminuir.

A regressão logística é um modelo estatístico que, em sua forma básica, usa uma função logística para modelar uma variável dependente.

O próprio modelo de regressão logística simplesmente modela a probabilidade de saída em termos de entrada e não executa classificação estatística, portanto não é um classificador, pois é incapaz de tomar decisão [28]. Embora possa ser usado para criar um classificador, por exemplo, escolhendo um valor de corte (Exemplo: 0.5) e classificando entradas com probabilidade maior que o ponto de corte de uma classe, abaixo do ponto de corte de outra. Essa é uma maneira comum de criar um classificador binário.

O algoritmo completo de regressão logística e suas equações matemáticas não serão apresentados nesse documento pois sua explicação seria muito extensa e fugiria ao propósito do trabalho. Seu desenvolvimento está encapsulado pela biblioteca *scikit-learn* e seu entendimento completo dos passos intermediários não é necessário para o entendimento ou replicação do projeto [30].

Capítulo 3

Trabalhos Relacionados

3.1 I Know First

A I Know First é uma empresa fintech que fornece soluções de previsão algorítmica para o mercado de ações. A empresa fornece previsões diárias de investimento com base em um algoritmo de aprendizado de máquina. O algoritmo foi desenvolvido pelo Dr. Lipa Roitman, um cientista com mais de 20 anos de pesquisa e experiência em inteligência artificial e campos de aprendizado de máquina. O Dr. Lipa Roitman possui longo histórico em modelagem computacional de processos, desenvolvimento de produtos e desenvolvimento de processos.

A tecnologia subjacente do algoritmo é baseada em inteligência artificial, aprendizado de máquina e incorpora elementos de redes neurais artificiais e algoritmos genéticos por meio dos quais é possível obter previsões para o mercado de ações.

O algoritmo gera previsões diárias de mercado para ações, commodities, ETFs, taxas de juros, moedas e índices mundiais para investimentos de tempo de curto, médio e longo prazo.

O sistema produz a tendência prevista como um número, positivo ou negativo, junto com um gráfico de ondas que prevê como as ondas se sobreporão à tendência. Isso ajuda o negociante a decidir em qual direção negociar, em que ponto entrar na operação e quando sair.

Como o modelo é 100% empírico, os resultados são baseados apenas em dados factuais, evitando, assim, quaisquer preconceitos ou emoções que possam acompanhar as suposições derivadas de humanos. O fator humano está envolvido apenas na construção da estrutura matemática e no fornecimento do conjunto inicial de entradas e saídas para o sistema.

O índice Bovespa é um indicador da média de desempenho nos das principais ações da bolsa de valores brasileira. É normal que empresas que forneçam serviços de consultoria financeira forneçam a seus clientes potenciais dados comparativos entre o desempenho de sua consultoria e o índice Bovespa, que representa o desempenho médio da bolsa.

A empresa I Know First realizou previsões e análises do mercado de ações brasileiro entre 19 de junho de 2019 e 26 de dezembro de 2019. Ao longo desse tempo, o índice Bovespa teve desempenho positivo de 2,03%.

No mesmo período, a empresa I Know First afirma ter obtido desempenho de 5,15%, mais do que duas vezes melhor do que o desempenho médio da bolsa. Afirma ainda que se forem considerados apenas os cinco ativos que foram indicados com maior destaque pela empresa, o desempenho foi de lucro superior a 12% [31].

A empresa I Know First possui uma página na rede mundial de computadores onde expõe o seu portfólio. A página está disponível no seguinte endereço: <https://iknowfirst.com/>.

3.2 FinBrain Technologies

FinBrain Technologies é uma empresa estadunidense, sediada em Wall Street, Nova York. A empresa oferece serviços de previsão de valores de ativos financeiros, como ações em bolsas de valores de vários países, commodities e até criptomoedas.

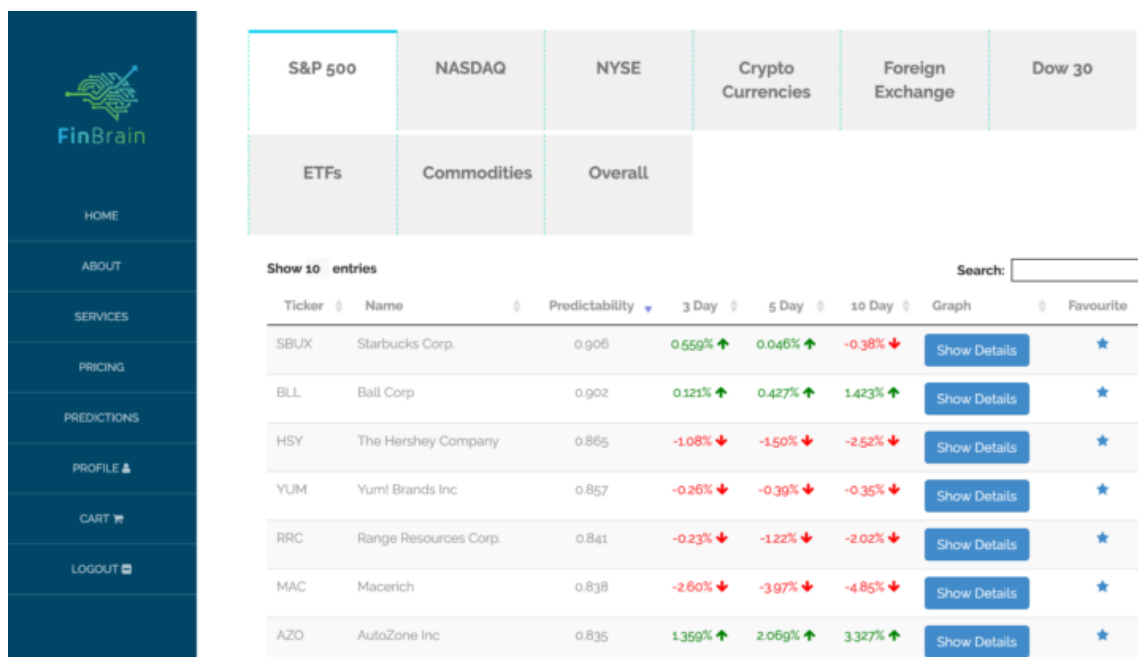
A FinBrain Technologies desenvolveu um algoritmo utilizando aprendizado profundo e redes neurais para realizar as previsões. Segundo o blog da empresa, os algoritmos coletam, organizam e alimentam os conjuntos de dados financeiros para as redes neurais profundas a fim de analisar e prever os preços das ações. Adicionalmente, foi elaborado um meio de cálculo para um valor para medir o quão previsíveis os estoques são para um determinado período de tempo. Isso é importante para realizar as negociações mais seguras que maximizem os retornos e minimizem os riscos.

De acordo com o blog oficial da empresa FinBrain:

FinBrain construiu um algoritmo de análise de previsibilidade especial que calcula o quão previsível um ativo é, incorporando o Expoente de Hurst, a Volatilidade e o Coeficiente de Determinação (R ao quadrado). Ambos os conceitos são combinados em um único algoritmo para fornecer um valor de indicador de previsibilidade robusto que ajuda nossos clientes a realizar negociações seguras. FINBRAIN (2019) [32]

Ao contratar os serviços da empresa, um cliente obtém acesso a uma plataforma com previsões para os valores de ativos financeiros em diferentes datar, e com um indicador de previsibilidade. Segue abaixo um *printscreen* da plataforma disponível no site da empresa.

Figura 3 – Captura de tela da plataforma online desenvolvida pela empresa FinBrain



Fonte: FINBRAIN (2019) [32]

O site oficial da empresa na rede mundial de computadores está disponível nesse endereço: <https://finbrain.tech/>. O blog oficial da empresa, com maiores informações sobre como utilizar as previsões fornecidas, está disponível nesse endereço: <https://blog.finbrain.tech/>.

Capítulo 4

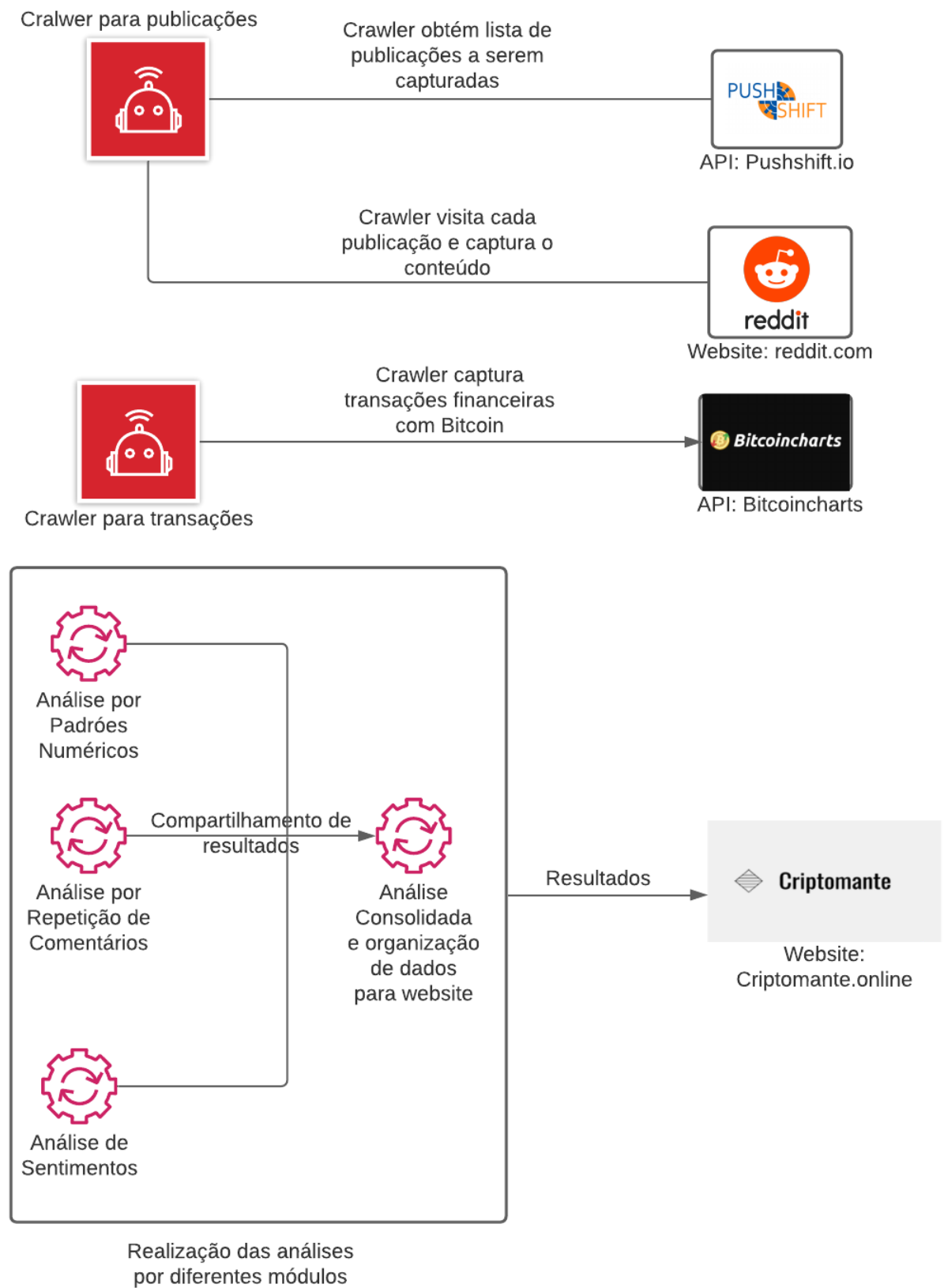
Desenvolvimento

O desenvolvimento desse projeto inclui a produção de vários softwares. De forma geral, os componentes que fazem parte desse projeto são:

- Uma aplicação responsável por minerar dados da internet e realizar as previsões e projeções. A aplicação pode ser dividida nos seguintes módulos:
 - Um módulo responsável por obter as cotações do Bitcoin.
 - Um módulo responsável por indexar os tópicos de todas as publicações da categoria *bitcoin* da rede social Reddit.
 - Um módulo responsável por visitar cada tópico indexado e capturar as mensagens publicadas pelos usuários, utilizando técnicas tradicionais de webcrawling.
 - Um módulo responsável pela interpretação das cotações capturadas e geração da previsão mais fundamental, chamada de *Análise por padrões numéricos*.
 - Um módulo responsável pelo tratamento do texto capturado, utilizando processamento de linguagem natural, para facilitar a execução das heurísticas de previsão baseadas em textos de usuários.
 - Um módulo responsável pela execução da previsão chamada de *Análise por Repetição de Comentários*.
 - Um módulo responsável pela execução de técnicas de aprendizado de máquina.
 - Um módulo responsável pela execução da previsão chamada de *Análise de Sentimentos*.
 - Um módulo responsável por interpretar os resultados das análises e organizar estes resultados para sua exibição no website
- Um website, disponível no endereço Criptomante.online, que contém os seguintes componentes:
 - Uma aplicação Django para o tratamento de requisições web.
 - Uma coleção de arquivos .html, contendo modelos (templates) para as páginas do site, que serão processados pelo motor de processamento de templates do framework Django.
 - Uma coleção de arquivos .js, contendo modelos (templates) para o código-fonte javascript que serão usados para a geração de gráficos. Estes também serão processados pelo motor de processamento de templates do framework Django.
- Um servidor particular, contratado da empresa BH Servers, responsável por hospedar o site, hospedar o banco de dados e executar os crawlers e análises 24 horas por dia.

O diagrama abaixo apresenta um esboço geral dos principais módulos desenvolvidos neste projeto. Cada módulo será apresentado individualmente nesse capítulo.

Figura 4 – Diagrama apresentando os principais módulos do projeto



4.1 Tecnologias Utilizadas

4.1.1 Linguagens de programação

A principal linguagem de programação utilizada no trabalho, responsável por toda a lógica backend da aplicação, é a linguagem Python.

Essa linguagem foi escolhida pelos seguintes motivos:

- Tem baixa verbosidade e alta produtividade.
- É orientada a objetos. Fato que facilita a modelagem.
- É de fácil integração com bancos de dados.
- Oferece bibliotecas nativas para desenvolvimento multithreading.
- O motor de processamento de linguagem natural Spacy é oferecido como um módulo Python.
- O motor de aprendizado de máquina Scikit-Learn é oferecido como um módulo Python.
- Possui ampla gama ferramentas úteis em webcrawling, como Selenium e BeautifulSoup.
- Possui o framework Django, que oferece um modo simples de desenvolver aplicações web.
- É amplamente recomendada como principal linguagem de programação para o desenvolvimento de aplicações com inteligência artificial. [\[33\]](#) [\[34\]](#) [\[35\]](#)
- O autor desse projeto já possuía alguns anos de experiência profissional com essa linguagem e com a maioria das bibliotecas utilizadas no desenvolvimento do projeto.

Adicionalmente, para a camada front-end, foi necessário utilizar CSS3 e HTML para a produção e design das páginas web. Assim como javascript para a exibição dos gráficos disponíveis no site.

4.1.2 Bibliotecas e Frameworks

As principais bibliotecas e frameworks utilizados no projeto são:

- Django: É um framework para desenvolvimento rápido para web, escrito em Python, que utiliza o padrão model-template-view (MTV). Nesse projeto, Django é utilizado para o tratamento das requisições web e construção dos arquivos html e javascript. É o principal componente que tornou possível a construção do site [\[36\]](#).
- Statistics: Biblioteca Python que oferece funções para técnicas estatísticas comuns. Como cálculo de média, de mediana, e etc [\[37\]](#).
- Numpy: Biblioteca com recursos para cálculos com arrays multidimensionais [\[38\]](#).
- Pandas: Biblioteca com recursos para análise e manipulação de conjuntos de dados (datasets). Nesse projeto é utilizada nas rotinas de aprendizado de máquina [\[39\]](#).
- Threading e Multiprocessing: Bibliotecas usadas para a execução de tarefas de forma paralela.
- Scikit-learn: Conjunto de bibliotecas com foco em Aprendizado de Máquina. Utilizado no desenvolvimento da Análise de Sentimentos [\[40\]](#).
- Spacy: Processador de Linguagem Natural [\[41\]](#).

- Requests e Wget: Usados para realizar requisições a APIs e para a leitura dos tópicos. Ao longo do desenvolvimento, a biblioteca Requests se mostrou mais performática do que o uso da ferramenta Selenium, que é mais tradicional no desenvolvimento de webcrawlers.
- Gzip: usada para a descompactação do arquivo .zip com as cotações do Bitcoin.
- BeautifulSoup: Biblioteca com recursos para a leitura e interpretação de arquivos .html. Nesse projeto é usado no processo de mineração de texto [\[42\]](#).

4.1.3 Banco de Dados

A escolha do banco de dados adequado é um fator decisivo para a viabilidade do projeto.

Logo no início do planejamento do projeto foi percebida a necessidade de dividir as mensagens dos usuários nas redes sociais em trechos textuais menores, e mais padronizados, para melhorar a qualidade das previsões realizadas. Foi percebido também que esse tratamento não poderia ser feito a cada execução dos modelos heurísticos, pois demanda muito poder computacional tratar todas as mensagens capturadas. Por exemplo, em novembro de 2020 as publicações capturadas da rede social Reddit já se aproximam de somar 2 bilhões de caracteres e banco de dados já soma 21GB de informação. Portanto, seria mais inteligente tratar cada mensagem somente uma vez, e armazenar o texto tratado no banco de dados. Isso resulta em ter o texto armazenado de forma não-tratada, e de forma tratada. Essas duas formas de texto possuem relação uma com a outra. Essa foi a primeira motivação de usar um banco de dados relacional. E, portanto, usar um banco de dados SQL.

Uma vez decidido usar um banco de dados SQL, o PostgreSQL foi escolhido pelos seguintes motivos:

- É gratuito
- É considerado um banco de dados de excelente performance [\[43\]](#).
- Possui excelente ganho de performance em operações de leituras e em relacionamentos (joins) com a implantação de índices. Embora precisem ser criados manualmente pelo programador, são especialmente úteis quando o programador tem bom conhecimento de como se distribui o volume de dados das tabelas do banco.
- Permite o uso de índices baseados em retornos de funções. Permite, por exemplo, indexar uma tabela pelo resultado do hash md5 de um campo textual da tabela. Esse recurso foi usado nesse projeto para realizar buscas em tabelas.
- O autor desse projeto já possuía vários anos de experiência profissional com PostgreSQL, tendo executado inclusive função de DBA com esse sistema.
- Durante a planejamento do projeto, houve o interesse de usar recursos de Full Text Search em PostgreSQL para encontrar semelhança entre frases. Contudo essa ideia foi abandonada com a possibilidade de uso de um motor de processamento de linguagem natural.

4.1.4 Spacy – Processador de Linguagem natural

Spacy é uma biblioteca open source para processamento de linguagem natural avançado, escritos com as linguagens de programação Python e Cython.

Spacy foi projetado especificamente para uso em produção e ajuda a criar aplicativos que processam e “entendem” grandes volumes de texto. Ele pode ser usado para construir sistemas de extração de informações ou de compreensão de linguagem natural, ou para pré-processar texto para aprendizado profundo.

Spacy é frequentemente comparado com a biblioteca Natural Language Toolkit (NLTK), que também oferece recursos para processamento de linguagem natural. Uma diferenciação comum para as bibliotecas é que Spacy tem características mais prontas para o uso imediato em produção, enquanto a biblioteca NLTK é mais útil para experimentação e para fins pedagógicos.

Nesse projeto, a biblioteca Spacy é usada para realizar as técnicas apresentadas no capítulo 2.3 desse documento.

4.1.5 Scikit-Learn – Motor de Aprendizado de Máquina

A scikit-learn é uma biblioteca de aprendizado de máquina de código aberto para a linguagem de programação Python. Ela oferece suporte ao aprendizado supervisionado e não supervisionado. Ele também fornece várias ferramentas para ajuste de modelo, pré-processamento de dados, seleção e avaliação de modelo e muitos outros utilitários.

A scikit-learn é utilizada nesse projeto para a execução de tarefas de aprendizado de máquina, classificação estatística e regressão logística (Veja o capítulo 2.4) necessárias para a conclusão de uma das heurísticas apresentadas, chamada de *análise de sentimentos*.

A “entrada” para as funções de análise é um conjunto de dados (dataset) definido pela biblioteca Pandas, contendo todo o texto resultante do tratamento por processamento de linguagem natural realizado pela biblioteca Spacy.

A biblioteca scikit-learn oferece recursos para a simples execução de treinamento de um modelo de aprendizado de máquina utilizando regressão logística, e também oferece métricas de precisão para o modelo, como verdadeiros/falsos positivos/negativos.

4.2 Mineração de Dados

A mineração de dados é formada por um conjunto de ferramentas e técnicas que são capazes de explorar um conjunto de dados, extraindo ou ajudando a evidenciar padrões nestes dados e auxiliando na descoberta de conhecimento.

Um processo fundamental para a viabilidade de mineração de dados é o processo de *Recuperação de Informação (Information Retrieval)*. Neste processo, são aplicadas técnicas para consultar, catalogar e associar informações que estejam em algum conjunto de dados.

Este projeto de monografia contém o desenvolvimento de uma análise do valor histórico da Bitcoin e de conteúdo publicado em uma rede social, e cria heurísticas e métricas baseadas nesses dados. Portanto a recuperação de informação é uma parte vital para o trabalho.

Durante o desenvolvimento da aplicação, a recuperação de informação foi dividida em duas rotinas. A primeira rotina é responsável por obter e catalogar o valor histórico da Bitcoin, enquanto a segunda rotina é responsável por obter e catalogar publicações de usuários na rede social Reddit.

4.2.1 Captura de Transações

A captura e armazenamento de transações (compras e vendas) de bitcoin é uma parte fundamental desse projeto, necessária para que as análises possam ser realizadas.

O maior fornecedor mundial de informações gratuitas sobre transações de bitcoins é a plataforma bitcoincharts.com. Essa plataforma oferece listagem de todas as compras e vendas de bitcoin dos principais fazedores de mercado que operam com o ativo. Nesse projeto, a captura é focada em transações registradas pelo fazedor de mercado Bitstamp, um dos maiores e mais antigos do mundo.

Transações realizadas na última semana são disponibilizadas por um API JSON, enquanto transações mais antigas estão disponíveis em formato CSV no endereço <https://api.bitcoincharts.com/v1/csv/>.

Neste projeto, foi implementado em python um programa capaz de recuperar essas informações e Snapshot é uma estrutura de dados proposta para a recuperação de transação. Assim, decide se deverá usar a API que fornece informações apenas dos últimos 7 dias, ou se deve usar os artefatos em CSV oferecidos pela plataforma Bitcoincharts.

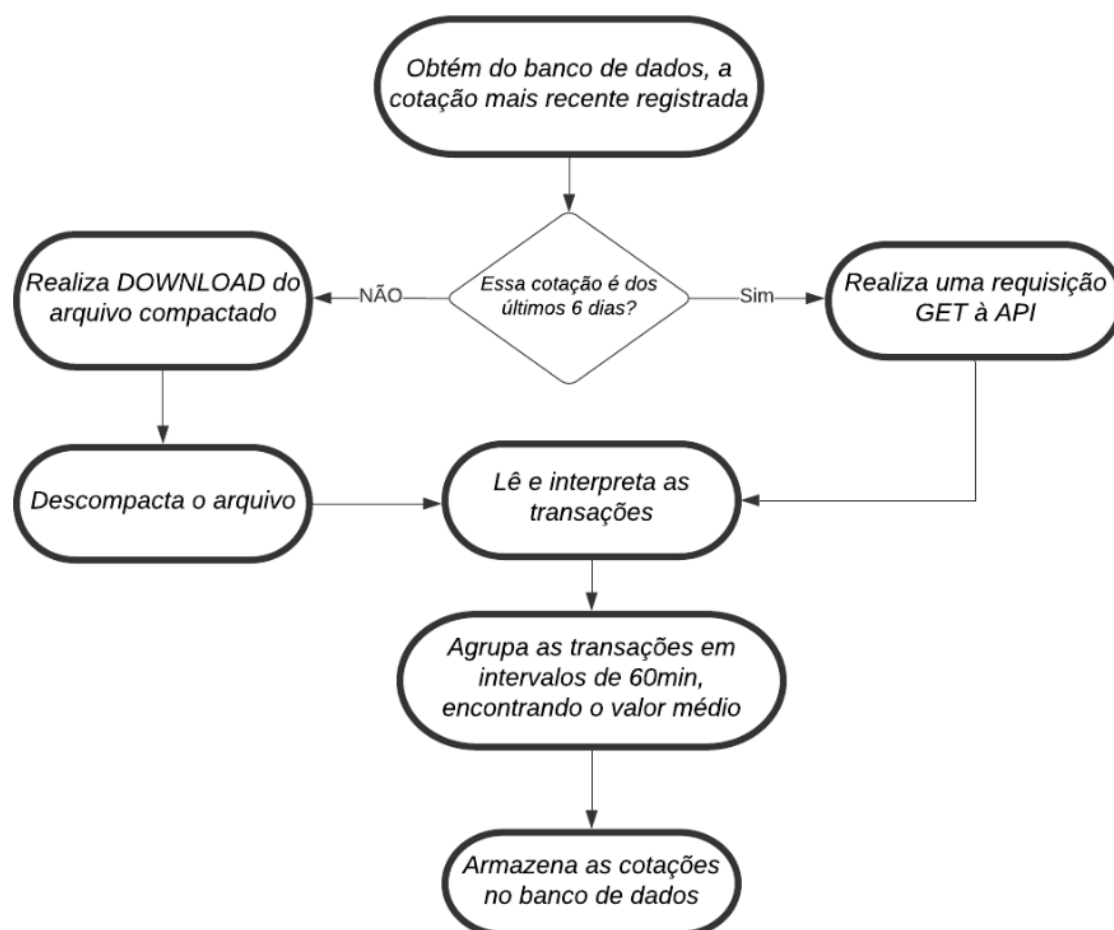
Caso seja necessário obter apenas os últimos 7 dias de transações, o programa utiliza a biblioteca *requests* para realizar uma requisição GET à API e obter as transações. E então armazena em memória local.

Caso seja necessário obter transações mais antigas, então o programa utiliza a biblioteca *wget* para fazer download do arquivo .CSV compactado em formato ZIP. Então, o programa utiliza a biblioteca *gzip* para descompactar o arquivo, e então lê e armazena em memória local as transações listadas no arquivo .CSV.

Por fim, o programa agrupa transações em intervalos de 60 minutos, obtendo o valor médio de compra (cotação) da Bitcoin naquele intervalo, e armazena a cotação no banco de dados.

O fluxograma abaixo oferece um esboço do processo.

Figura 5 – Fluxograma do processo de captura de transações com bitcoin.



4.2.2 Captura de comentários em rede social

A captura de comentários e publicações sobre a criptomoeda Bitcoin em uma rede social é uma parte fundamental desse projeto.

Nos próximos capítulos serão apresentadas heurísticas de previsão para a Bitcoin que são parcialmente baseadas na ciclicidade de comentários publicados em uma rede social.

Reddit é uma rede social gratuita que se destaca em ter todo seu conteúdo rigidamente dividido em sessões do site especializadas em algum tema. Cada sessão, popularmente chamada de subreddit, possui conteúdo especializado no tema da sessão, fazendo com que a rede social funcione como um agregador de fóruns de discussão independentes.

Especialmente para discussões sobre a Bitcoin, existem as seguintes sessões:

- reddit.com/r/Bitcoin/ - Principal subreddit para a criptomoeda. 1.6 milhões de usuários
- reddit.com/r/btc/ - Subreddit secundária para a criptomoeda. 323mil usuários
- reddit.com/r/BitcoinMarkets - Subreddit de viés financeiro-especulativo. 160 mil usuários

Nesse projeto, são capturados comentários publicados na sessão [reddit.com/r/Bitcoin.](https://reddit.com/r/Bitcoin/), pois possui o maior volume de usuários e de publicações. Nesse projeto foi implementado em python um programa capaz de capturar publicações dessa sessão e armazenar o conteúdo em um banco de dados Postgresql.

1. O programa consulta o banco de dados e obtém a data da última mensagem capturada.
2. O programa então utiliza a biblioteca requests para enviar uma requisição para a API pushshift.io, solicitando a lista de publicações posteriores a data obtida no passo 1.
3. Cada requisição fornece 25 publicações. Para que todas sejam lidas, é necessário realizar novas requisições solicitando uma nova data como data mínima de interesse.
4. O programa registra o endereço (URL) de cada publicação obtida em uma tabela do banco de dados.
5. O programa, utilizando multithreading, busca do banco de dados todas as URLs que ainda não foram visitadas. Cada URL será tratada por uma thread independente.
6. O programa utiliza novamente a biblioteca requests para obter o código fonte HTML da página.
7. O programa utiliza a biblioteca BeautifulSoup para interpretar o HTML capturado, e extrair o texto da publicação principal, assim como todos os comentários da publicação.
8. O programa registra no banco de dados os textos capturados no passo 7, mantendo a informação de vínculo com a URL da publicação principal. Enfim, o programa registra no banco de dados que a URL foi visitada, e encerra a thread dessa URL específica.
9. Por fim, quando não houver mais em banco de dados URLs que não foram visitadas, e todas as threads auxiliares estiverem encerradas, o programa é encerrado.

4.3 Análise por padrões numéricos

Os principais componentes desse projeto são os métodos propostos para criar estimativas para o valor futuro da bitcoin. Nesse projeto, todas as análises foram baseadas no valor do Bitcoin em dólares americanos.

Entre os métodos contidos nesse trabalho, o mais fundamental é chamado de *Análise por padrões numéricos*. Esse método pode ser considerado como o mais fundamental porque age como valor base para a *análise consolidada*.

A análise por padrões numéricos, como todas as análises propostas nesse projeto, se baseia na ciclicidade dos valores de ativos especulativos proposta pelo princípio de ondas de Elliott, apresentado na sessão 2.1.4.

Antes da apresentação do roteiro da metodologia desenvolvida será necessário apresentar uma estrutura de dados desenvolvida com o objetivo de descrever a variação do valor da bitcoin em um intervalo de tempo. Essa estrutura de dados será chamada de *snapshot*.

4.3.1 Estrutura de dados Snapshot

Snapshot é uma estrutura de dados proposta nesse projeto como meio para armazenar e descrever a variação do valor do Bitcoin em um intervalo de tempo vizinho a uma data que está sendo analisada.

A princípio, é necessário definir as seguintes nomenclaturas:

- **Data referência** do snapshot: A data base de um snapshot, é a data descrita pelo snapshot.
- **Snapshot**: Estrutura de dados que descreve a variação do valor da bitcoin em torno da vizinhança da data referência.

Um Snapshot é uma estrutura contendo 13 tuplas ordenadas. Cada tupla é formada por uma sigla, uma data vizinha à data referência e um valor numérico que representa a variação no valor da bitcoin entre a data dessa tupla e a data da tupla seguinte.

As siglas das tuplas que compõem um snapshot são iguais em todos os snapshots. Enquanto cada uma das 13 datas que compõem um snapshot pode ser descrita pela fórmula abaixo.

Quadro 2 – Descrição das tuplas que definem um snapshot

Sigla	Fórmula para o datetime	Fórmula para Variação
DR	Data Referência do snapshot	$\frac{\text{Valor}(DR) - \text{Valor}(DR - 1D)}{\text{Valor}(DR - 1D)}$
DR -1D	Data Referência – 1 Dia	$\frac{\text{Valor}(DR - 1D) - \text{Valor}(DR - 5D)}{\text{Valor}(DR - 5D)}$
DR -5D	Data Referência – 5 Dias	$\frac{\text{Valor}(DR - 5D) - \text{Valor}(DR - 10D)}{\text{Valor}(DR - 10D)}$
DR -10D	Data Referência – 10 Dias	$\frac{\text{Valor}(DR - 10D) - \text{Valor}(DR - 15D)}{\text{Valor}(DR - 15D)}$
DR -15D	Data Referência – 15 Dias	$\frac{\text{Valor}(DR - 15D) - \text{Valor}(DR - 20D)}{\text{Valor}(DR - 20D)}$
DR -20D	Data Referência – 20 Dias	$\frac{\text{Valor}(DR - 20D) - \text{Valor}(DR - 25D)}{\text{Valor}(DR - 25D)}$
DR -25D	Data Referência – 25 Dias	$\frac{\text{Valor}(DR - 25D) - \text{Valor}(DR - 1M)}{\text{Valor}(DR - 1M)}$
DR -1M	Data Referência – 1 Mês	$\frac{\text{Valor}(DR - 1M) - \text{Valor}(DR - 2M)}{\text{Valor}(DR - 2M)}$
DR -2M	Data Referência – 2 Meses	$\frac{\text{Valor}(DR - 2M) - \text{Valor}(DR - 3M)}{\text{Valor}(DR - 3M)}$
DR -3M	Data Referência – 3 Meses	$\frac{\text{Valor}(DR - 3M) - \text{Valor}(DR - 4M)}{\text{Valor}(DR - 4M)}$
DR -4M	Data Referência – 4 Meses	$\frac{\text{Valor}(DR - 4M) - \text{Valor}(DR - 5M)}{\text{Valor}(DR - 5M)}$
DR -5M	Data Referência – 5 Meses	$\frac{\text{Valor}(DR - 5M) - \text{Valor}(DR - 6M)}{\text{Valor}(DR - 6M)}$
DR -6M	Data Referência – 6 meses	$\frac{\text{Valor}(DR - 6M) - \text{Valor}(DR - 12M)}{\text{Valor}(DR - 12M)}$

Segue abaixo um exemplo de um snapshot com data de referência em 18 de outubro de 2020, contendo o valor da cotação na data.

Tabela 1 – Tuplas de um snapshot com data de referência em 17 de outubro de 2020

Sigla	Data	Variação	Cotação na Data
DR	18/10/2020	0.81%	\$ 11441,54
DR-1D	17/10/2020	-0.71%	\$ 11349,42
DR-5D	13/10/2020	6.37%	\$ 11430,77
DR-10D	08/10/2020	1.83%	\$ 10745,97
DR-15D	03/10/2020	-2.70%	\$ 10553,04
DR-20D	28/09/2020	4.62%	\$ 10845,46
DR-25D	23/09/2020	-5.09%	\$ 10366,17
DR-1M	18/09/2020	-9.40%	\$ 10922,35
DR-2M	18/08/2020	31.49%	\$ 12055,45
DR-3M	18/07/2020	-2.30%	\$ 9168,17
DR-4M	18/06/2020	-2.97%	\$ 9384,47
DR-5M	18/05/2020	34.61%	\$ 9671,36
DR-6M	18/04/2020	-9.32%	\$ 7184,73

4.3.2 Roteiro da análise

A análise por padrões numéricos pode ser dividida em cinco etapas principais.

Como foi apresentado na sessão 4.3.1, esse projeto inclui um módulo de captura de transações da bitcoin realizadas no fazedor de mercado Bitstamp.

A **primeira etapa** da análise é consultar nosso banco de dados de cotações de Bitcoin e obter as cotações dos últimos 12 meses.

Segue abaixo um gráfico exibindo a cotação da Bitcoin nos 12 meses anteriores à data de 18 de outubro de 2020.

Figura 6 – Gráfico exibindo a cotação do Bitcoin entre 18/10/2019 e 18/10/2020



A **segunda etapa** da análise é construir um snapshot tendo como data de referência a data atual.

Um exemplo de um snapshot para a data de 18 de outubro de 2020 foi demonstrado na sessão anterior.

O objetivo da análise é encontrar snapshots com características semelhantes ao snapshot da data atual, e, portanto, usar essa semelhança para estimar um snapshot futuro.

A **terceira etapa** da análise é construir snapshots para todos os dias desde 13 de setembro de 2011 (Data da primeira transação realizada no site Bitstamp fornecida pela plataforma Bitcoincharts).

A seguir, a quarta etapa da análise consiste em comparar os snapshots obtidos na terceira etapa com o snapshot da data atual. Cada snapshot receberá uma pontuação.

A pontuação de um snapshot será igual a quantidade de tuplas em que a variação desse snapshot for igual a variação do snapshot atual, com uma margem de tolerância de 10%, para a mesma sigla.

O seguinte código-fonte Python expressa o método de pontuação:

```
def calcular_semelhanca_entre_snapshots(snapshot_atual:Snapshot, outro_snapshot:Snapshot)
    pontuacao=0
    for sigla in snapshot_atual.siglas:
        diferenca = abs(outro_snapshot.variacao[sigla]/snapshot_atual.variacao[sigla])
        if (diferenca>=0.9 and diferenca<=1.1):
            pontuacao=pontuacao+1
    return pontuacao
```

A partir da pontuação de um snapshot, é definido um peso para o mesmo, a ser usado na quinta etapa.

O peso de um snapshot é definido pela seguinte fórmula exponencial:

$$Peso = Pontuação^{Pontuação}$$

A seguir está no snapshot da data 21 de junho de 2015, que durante a análise para a data de 18 de outubro de 2020 obteve pontuação igual a 4 e peso igual a 256. Em negrito estão as quatro tuplas que forneceram pontos a este snapshot, por possuírem variação igual à variação do snapshot da data de 18 de outubro de 2020, dentro de uma margem de erro de 10%.

Tabela 2 – Snapshot com data de referência em 21 de junho de 2015

Sigla	Data	Variação	Cotação na Data
DR	21/06/2015	0.34%	\$ 243,27
DR-1D	20/06/2015	-0.44%	\$ 242,45
DR-5D	16/06/2015	6.38%	\$ 243,52
DR-10D	11/06/2015	1.95%	\$ 228,91
DR-15D	06/06/2015	-0.30%	\$ 224,53
DR-20D	01/06/2015	-4.42%	\$ 225,21
DR-25D	27/05/2015	-0.56%	\$ 235,63
DR-1M	21/05/2015	2.38%	\$ 235,40
DR-2M	21/04/2015	-11.33%	\$ 229,94
DR-3M	21/03/2015	5.55%	\$ 259,32
DR-4M	21/02/2015	12.46%	\$ 245,69
DR-5M	21/01/2015	-32.55%	\$ 218,48
DR-6M	21/12/2014	-45.19%	\$ 323,92

A **quinta etapa** da análise consiste em realizar a previsão para o futuro valor da bitcoin baseada nos dados obtidos na etapa anterior.

Para que possamos prever o valor da bitcoin X dias futuros à data atual, primeiro criamos uma **lista de variações** vazia.

A seguir, verificamos todos os snapshots que obtiveram peso maior do que zero no passo 4. E para a data referência de cada snapshot, verificamos a variação do valor da bitcoin entre a data de referência e X dias depois.

Segue abaixo os snapshots de maior pontuação para a análise da data de 18 de outubro de 2020, assim como a variação entre a data de referência do snapshot e 30 dias depois.

Tabela 3 – Snapshots com maior pontuação durante a quarta etapa.

Data Referência	Cotação na data	Cotação depois de 30 dias	Variação em 30 dias	Peso
21/06/2015	\$ 243,27	\$ 278,50	14.48%	256
20/02/2016	\$ 433,48	\$ 409,85	-5.45%	256
29/04/2017	\$ 1330,56	\$ 2259,97	69.85%	256
04/05/2017	\$ 1542,73	\$ 2528,04	63.87%	256
17/06/2020	\$ 9420,71	\$ 9133,46	-3.05%	256
12/10/2020	\$ 11469,66	\$ 0,00	0.00%	256
15/10/2020	\$ 11417,04	\$ 0,00	0.00%	256
16/10/2020	\$ 11329,40	\$ 0,00	0.00%	256
04/06/2014	\$ 640,79	\$ 631,92	-1.38%	27
05/09/2014	\$ 481,13	\$ 304,97	-36.61%	27

A seguir para cada snapshot que conseguiu pontuação na quarta etapa, obtemos a sua variação entre a data de referência e X dias futuros a essa data (Apresentados na quarta coluna da tabela acima). E inserimos essa variação em nossa **lista de variações**.

Cada variação será inserida repetidamente múltiplas vezes na lista. A quantidade de repetições será igual ao peso do snapshot. Considerando apenas os 10 snapshots exibidos acima, a lista de variações da análise do dia 18 de outubro de 2020 teria pelo menos 1.644 elementos.

A variação prevista para X dias futuros será igual à mediana da lista de variações.

Nesse trabalho, foi decidido utilizar a mediana, ao invés da média, para diminuir a relevância de variações bruscas (outliers) no valor da bitcoin, que ocorreram algumas vezes durante sua história.

Uma forma resumida de descrever essa previsão é afirmar que a previsão de variação entre a data atual e uma data futura que esteja a X dias no futuro, é feita pela mediana ponderada de um vetor composto com a variação do valor da bitcoin entre a data referência de cada snapshot que pontuou no passo quatro e X dias futuros à data de referência. O fator de ponderamento da mediana será o peso de cada snapshot.

A mediana obtida para a análise em 18 de outubro de 2020 foi de 4,4. Portanto, estimamos que haverá um aumento de 4,4% entre essa data e 18 de novembro de 2020.

Calculando previsões em diversas datas futuras, podemos conseguir o seguinte gráfico para os três meses seguintes.

Figura 7 – Resultado da estimativa obtida com a análise por padrões numéricos



4.4 Análise por repetição de comentários

O segundo método de análise e previsão apresentado por esse trabalho é chamado de **Análise por repetição de comentários**. Esse também será o primeiro método a fazer uso de comentários e publicações extraídos de redes sociais conforme descrito na sessão 4.3.2.

Nesse método, estamos interessados em explorar a ciclicidade em pensamentos e opiniões de entusiastas da criptomoeda assim como sua relação com a ciclicidade natural do valor da moeda.

A forma mais simples de buscar ciclicidade de comentários é encontrar repetições de comentários ao longo do tempo e buscar se há comentários específicos que mostram tendência de coincidência com datas em que o valor da criptomoeda subiu ou caiu.

4.4.1 Roteiro

O **primeiro passo** da análise é obter de nosso banco de dados todas as mensagens trocadas nas últimas 24 horas.

Segue abaixo um exemplo de uma publicação de um usuário no dia 19 de outubro de 200.

“Very interesting. I don’t agree with the storage aspect though. Bullion dealers do offer a storage service, sure, but if you’re going to compare a bitcoiner that has set up their own wallet to a goldbug with the same level of personal involvement, I’ll tell you that goldbug has physical possession of that gold and isn’t paying \$451/100oz a year. I love bitcoin but it’s still clearly in infancy, it’ll be decades before it has a chance of being superior to gold irl, not just theory. The big reason is right now, there are about 7 billion people on the planet that know gold has value and a majority feel its probably the most valuable thing there is even if they have never owned, used, or even held any gold in their hand in their lives. Bitcoin is only a few years out of being some weird, internet thing people used to buy drugs. Everyone knows gold, and in the large scheme of things practically no one knows about bitcoin... yet. ”

O **segundo passo** é dividir cada mensagem em frases menores. Essa divisão é realizada utilizando o motor de processamento de linguagem natural Spacy.

A mensagem acima, ao ser dividida, dá origem às seguintes frases:

- *Very interesting.*
- *I don't agree with the storage aspect though.*
- *Bullion dealers do offer a storage service, sure, but if you're going to compare a bitcoiner that has set up their own wallet to a goldbug with the same level of personal involvement, I'll tell you that goldbug has physical possession of that gold and isn't paying \$451/100oz a year.*
- *I love bitcoin*
- *but it's still clearly in infancy, it'll be decades before it has a chance of being superior to gold irl, not just theory.*
- *The big reason is right now, there are about 7 billion people on the planet that know gold has value and a majority feel its probably the most valuable thing there is even if they have never owned, used, or even held any gold in their hand in their lives.*
- *Bitcoin is only a few years out of being some weird, internet thing people used to buy drugs.*
- *Everyone knows gold, and in the large scheme of things practically no one knows about bitcoin... yet.*

O **terceiro passo** é fazer um pequeno tratamento nas frases obtidas. Nesse tratamento, o texto será inteiramente convertido para caracteres minúsculos. Também serão removidos das extremidades do texto (início ou final) caracteres que não sejam uma letra ou um sinal de pontuação.










As frases acima, depois de tratadas, se tornam:

- *very interesting.*
- *i don't agree with the storage aspect though.*
- *bullion dealers do offer a storage service, sure, but if you're going to compare a bitcoiner that has set up their own wallet to a goldbug with the same level of personal involvement, i'll tell you that goldbug has physical possession of that gold and isn't paying \$451/100oz a year.*
- *i love bitcoin*
- *but it's still clearly in infancy, it'll be decades before it has a chance of being superior to gold irl, not just theory.*
- *the big reason is right now, there are about 7 billion people on the planet that know gold has value and a majority feel its probably the most valuable thing there is even if they have never owned, used, or even held any gold in their hand in their lives.*
- *bitcoin is only a few years out of being some weird, internet thing people used to buy drugs.*
- *everyone knows gold, and in the large scheme of things practically no one knows about bitcoin... yet.*

O **quarto passo** é buscar em nosso banco de dados ocorrências anteriores de cada frase obtidas no terceiro passo. E para cada ocorrência anterior, verificar se foi em um dia em que o valor do Bitcoin aumentou ou diminuiu. São considerados nessa etapa apenas dias em que a variação foi de pelo menos 5% desde o horário de publicação da mensagem até as 24 horas seguintes.

Segue abaixo algumas das frases publicadas em 19 de outubro de 2020 que demonstraram maior expressividade de tendência:

Tabela 4 – Frases do dia 19 de outubro de 2020 que demonstraram alguma tendência

Frase	Vezes em que precedeu uma queda	Vezes em que precedeu um aumento	Tendência
understandable.	3	8	 72.73%
time is money.	4	2	 66.67%
thanks for the info	36	60	 62.50%
are you 12?	6	4	 60.00%
not my fault	3	2	 60.00%
you should be fine.	4	6	 60.00%
i don't know.	250	352	 58.47%
not your keys, not your bitcoin.	64	86	 57.33%
good point.	342	456	 57.14%

O **quinto passo** é contar quantas das frases extraídas de mensagens publicadas nas últimas 24 horas demonstraram tendência positiva, e quantas demonstraram tendência negativa.

O resultado da análise será positivo, caso o número de frases com tendência positiva seja maior do que o número de frases com tendência negativa. E vice-versa. A análise será considerada inconclusiva se as contagens forem iguais.

No caso do dia 19 de outubro de 2020, houve mais frases com tendência positiva.

4.5 Análise de sentimentos

A próxima análise a ser apresentada é um exemplo de aplicação de análise de sentimentos utilizando um motor de linguagem natural e um classificador.

Nessa análise utilizamos um motor de aprendizado de máquina para treinar um classificador capaz de avaliar se as mensagens publicadas em uma rede social implicam uma tendência de aumento ou de queda no preço da Bitcoin.

O modelo foi projetado de forma que responda a seguinte pergunta: “Caso haja uma variação de pelo menos 5% no valor do Bitcoin nas próximas 24 horas, essa variação será um aumento?”

Como um modelo de aprendizado de máquina, que é treinado pelo histórico publicações de usuários, esse método depende fortemente da ciclicidade de opiniões, e portanto, está ancorado no **princípio de ondas de Elliott**.

4.5.1 Tratamento de texto

Em implementações de modelos de aprendizado de máquina sobre dados que representam texto em linguagem natural, é comum que seja realizado algum procedimento de normalização e tratamento dos dados que serão usados para o treinamento.

Por isso, foi necessário realizar um tratamento de texto mais avançado do que o tratamento apresentado na sessão 4.5. Nesse projeto, os seguintes tratamentos foram aplicados:

- Separação de mensagens longas em frases menores.
- Remoção de caracteres especiais e símbolos das extremidades das frases.
- Lematização
- Remoção de Stop Words
- Conversão de todo o texto para caracteres minúsculos

Veja, por exemplo, a seguinte mensagem, publicada na rede social em 10 de janeiro de 2014.

*But the people's concerns regarding control are legitimate.
We know from history that any kind of centralized power and control of anything, much less money, will be abused.
Absolute power corrupts absolutely.
Sure, the mining pool operators have no incentive to torpedo faith in bitcoin now.
But what if those centralized mining pools are compromised by other interests.
Perhaps a government or competing corporation or organized crime organization might want to take down bitcoin.
Where would ill wishers of bitcoin attack? The area of most centralization.
That centralized point could be the exchanges, it could be the ASIC mining manufacturers, or in our example, it could be the owners of the mining pools.
How would they attack the miners? Many ways. They could hack the pool owners' systems.
An outside group could bribe the pool owners by offering them more wealth than they're making mining.
An entity could threaten, harass, or blackmail mining pool owners.
This threat could take the form of legal action, or in some countries, physical violence to the mining pool owners and their families.
TL;DR The 51% issue needs to be closely monitored by bitcoin stakeholders.*

A mensagem acima, após realizar o tratamento mencionado, se divide nas seguintes frases:

- *people concern control legitimate*
- *know history kind centralized power control money abuse*
- *absolute power corrupt absolutely*
- *sure mining pool operator incentive torpedo faith bitcoin*
- *centralized mining pool compromise interest*
- *government compete corporation organize crime organization want bitcoin*
- *ill wisher bitcoin attack*
- *area centralization*
- *centralized point exchange asic mining manufacturer example owner mining pool*
- *attack miner*

- *way*
- *hack pool owner system*
- *outside group bribe pool owner offer wealth make mine*
- *entity threaten harass blackmail mining pool owner*
- *threat form legal action country physical violence mining pool owner family*
- *issue need closely monitor bitcoin stakeholder*

As frases tratadas, apesar de difícil leitura para um ser humano, oferecem melhores resultados como conteúdo de treinamento para um modelo de aprendizado de máquina.

4.5.2 Roteiro da análise

O **primeiro passo** da análise é obter todas as mensagens armazenadas no banco de dados, e agrupar por data.

O **segundo passo** é classificar cada data de acordo com a variação da bitcoin. Datas que antecederam um aumento de pelo menos 5% no valor da são classificadas como positivas. Datas que antecederam uma queda são classificadas como negativas. Descarta-se datas que não antecederam queda ou aumento, por ter havido uma estabilidade do valor da Bitcoin no período.

O **terceiro passo** é aplicar o tratamento de texto explicado na sessão 4.6.1 para dividir as mensagens em frases menores e lematizadas. Mantendo ainda as frases agrupadas por data.

O **quarto passo** é utilizar a biblioteca Scikit-learn para criar um modelo de classificação por regressão logística.

O **quinto passo** é dividir nossos dados de treinamento em dois sub-conjuntos. O primeiro possui 90% dos dados, e o segundo possui apenas 10% dos dados.

O **sexto passo** é utilizar o conjunto maior, com 90% dos dados, para treinar nosso modelo.

O **sétimo passo** é utilizar o modelo, depois de treinado, para avaliar os textos das frases do sub-conjunto menor. Como a classificação real (se precedeu uma queda ou aumento no valor da bitcoin) já era conhecida, é possível calcular a precisão de nosso modelo.

O trecho de código-fonte Python abaixo representa do quarto ao sétimo passo.

```

34 def construir_modelo(dataframe):
35     #Classificador de Regressão Logística. Com método SAGA para solucionar o problema de minimização
36     classificador = LogisticRegression(max_iter=999999999, n_jobs=6, solver='saga')
37
38     #Tokenizador simples. Pois o texto já estava pré-tratado
39     vetorizador = TfidfVectorizer(tokenizer = tokenizer, ngram_range=(1,1))
40
41     classifier = classificador
42     tfidf_vector = vetorizador
43
44     X = dataframe['texto'] # Eixo X, indica o texto de entrada. Sendo as frases lematizadas
45     ylabels = dataframe['tendencia'] # Eixo Y, indica o a classificacao positiva ou negativa de cada frase.
46
47     #Dividimos os dados de entrada. Sendo 90% para treinamento. E 10% para testar a qualidade do treinamento
48     X_train, X_test, y_train, y_test = train_test_split(X, ylabels, test_size=0.1)
49
50     # Criamos um Pipeline, que indica a sequência em que as coisas são feitas
51     pipe = Pipeline([('vectorizer', tfidf_vector),
52                     ('classifier', classifier)],
53                    verbose=True)
54
55     #Realizando a previsão.
56     pipe.fit(X_train, y_train)
57     predicted = pipe.predict(X_test)
58
59     #Preparando saída
60     saida = dict()
61     matriz = confusion_matrix(y_test, predicted)
62     saida["TP"] = matriz[1][1] #Verdadeiros Positivos
63     saida["FP"] = matriz[0][1] #Falsos Positivos
64     saida["TN"] = matriz[0][0] #Verdadeiros Negativos
65     saida["FN"] = matriz[1][0] #Falsos negativos

```

O **oitavo passo** é utilizar nosso modelo para testar a data de hoje, fornecendo as frases tratadas publicadas nas últimas 24 horas. Nosso classificador então classificará como um aumento ou uma queda.

4.6 Análise consolidada

A última metodologia de previsão para o preço do valor da bitcoin proposta nesse projeto é chamada de *análise consolidada*.

Essa metodologia combina os resultados obtidos nas três metodologias apresentadas anteriormente.

Seu valor é calculado da seguinte forma:

- Caso a Análise por repetição de comentários e a Análise de sentimentos tenham simultaneamente resultados positivos, então o valor previsto da Análise Consolidada será igual ao valor previsto pela Análise Numérica com um acréscimo de 5% em todas as datas.
- Caso a Análise por repetição de comentários e a Análise de sentimentos tenham simultaneamente resultados negativos, então o valor previsto da Análise Consolidada será igual ao valor previsto pela Análise Numérica com um decréscimo de 5% em todas as datas.
- Caso a Análise por repetição de comentários e a Análise de sentimentos tenham resultados opostos, então o valor previsto pela Análise Consolidada será igual ao valor previsto pela Análise Numérica.

O gráfico em destaque na página principal do website <https://criptomante.online/> é o resultado da Análise Consolidada.

4.7 Aplicação Web e Deploy

Para a execução das rotinas apresentadas nesse trabalho foi contratado um servidor do tipo Virtual Private Server (VPS) da empresa brasileira BhServers. Nesse servidor, diversas rotinas são executadas 24 horas por dia. As rotinas executadas pelo servidor são as seguintes:

1. Execução de extração de publicações da rede social Reddit.
2. Execução de extração de transações de compras e vendas de bitcoin.
3. Execução da rotina de Análise por Padrões Numéricos.
4. Execução da rotina de Análise por Repetição de Comentários
5. Execução da rotina de Análise de Sentimentos
6. Execução de uma rotina que organiza os resultados das análises anteriores para a apresentação dos gráficos e tabelas presentes no website.

No mesmo servidor foi hospedado o website criptomante.online.

O website foi desenvolvido em Python utilizando o framework Django. Os gráficos disponíveis no website estão programados em javascript utilizando as bibliotecas eCharts e D3.js.

O site foi publicado no servidor utilizando o serviço IIS da Microsoft. A comunicação do serviço IIS e o código-fonte python foi feita usando FastCgi e wFastCgi.

O termo “criptomante” vem das palavras “cripto” e “mante”. A palavra “cripto” vem da palavra grega *kryptos*, que significa “escondido”, e faz referência a relação entre a Bitcoin com o estudo de criptografias. A palavra “mante” vem da palavra grega “*mántis*”, que exprime a noção de pessoa que pratica adivinhação (ex.: cartomante), e faz referência ao objetivo do trabalho de prever o valor futuro da Bitcoin.

O código fonte do website e de todas as rotinas está disponível no repositório https://github.com/granpk/Criptomante_python.

Capítulo 5

Conclusão

5.1 Desafios

Segue abaixo alguns dos desafios encontrados ao longo do desenvolvimento desse trabalho, assim como as medidas tomadas.

- Foram encontrados vários desafios relacionados a captura de todas as publicações de usuários na categoria ‘Bitcoin’ do site Reddit. Mais especificamente, desafios em obter publicações muito antigas:
 - O site oferece duas formas de navegação. A primeira utiliza uma técnica chamada infinite scroll. É uma forma de ‘paginação infinita’, onde a próxima ‘página’ de conteúdo é anexada ao final da página quando o usuário realiza alguma ação (Como descer o cursor até o final da página). Infelizmente esse método faz com que a memória necessária para a visualização de uma página com quase um milhão de publicações seja maior do que o servidor conseguiria suportar.
 - Outra forma de navegação que o site permite, utilizando uma configuração específica, é pela navegação tradicional de ‘páginas’ de conteúdo, onde é possível carregar uma página de cada vez, com 27 publicações. Porém, só é possível navegar por uma quantidade limitada de páginas. Não é possível, por exemplo, utilizar o botão de “próximo” para navegar até publicações de 2016.
 - O site Reddit possui uma API que permite a extração de conteúdo. Infelizmente, em 2018 houve uma atualização na API que fez com que o método utilizado para a leitura de publicações antigas parasse de funcionar.
 - Como solução para os desafios acima, foi necessário utilizar a API do site pushshift.io, que permite navegação por todas as publicações do site Reddit.
- Depois de resolvido o desafio em obter as publicações, houve um novo desafio na leitura de comentários da publicação. O site Reddit utiliza eventos em javascript para o carregamento de comentários e respostas de comentários. Dessa forma, o seria necessário utilizar Selenium ou outra ferramenta semelhante para automatizar o uso de um navegador para carregar o site e extrair as informações desejadas. Porém, essa abordagem tornaria a extração mais pesada e prejudicaria a velocidade de extração, mesmo com multithreading. Como solução, ao utilizar o domínio old.reddit.com é possível navegar por uma versão do site que utiliza uma navegação menos dinâmica. E dessa forma, foi possível utilizar apenas a biblioteca *requests* para obter uma extração mais rápida do conteúdo
- Na data de desenvolvimento desse projeto, não havia API gratuita na internet para a extração de transações de compra e venda de Bitcoin para datas muito anteriores. Assim, foi necessário adicionar à rotina de extração de dados a capacidade de realizar download de um arquivo .zip contendo dados históricos sobre a moeda.
- A bitcoin, assim como a maioria dos commodities, tem seu valor lastreado no dólar. De forma que, caso o dólar aumente, todos os operadores de Bitcoin brasileiros aumentam o valor de suas transações na mesma medida. Dessa forma, a cotação de bitcoin em real brasileiro sofre variações pelo desempenho do real, pelo desempenho do dólar e pelo desempenho da própria Bitcoin. Como nesse trabalho o objeto de estudo é a Bitcoin, e não o dólar, foi tomada a decisão de que todos os valores de todas as cotações fossem em dólar. Assim, todos os gráficos presentes no site estão em dólares americanos.

- Complementando o item acima, a extração de publicações foi realizada sobre publicações feitas no idioma inglês. Isso foi feito por dois fatores: O primeiro fator é que as cotações capturadas são em dólar americano, e o segundo fator é que há muito mais conteúdo em inglês sobre a criptomoeda em redes sociais do que há em português.
- Durante o desenvolvimento do projeto foi percebido que os dados capturados somavam um grande volume. Com dezenas de milhões de transações e de frases capturadas. E essas informações deveriam ser cruzadas e relacionadas entre si a cada execução das análises. Por isso, os seguintes procedimentos foram realizados para a melhoria de performance das análises:
 - Ao separar uma mensagem de um usuário em frases, e posteriormente ao realizar o tratamento da frase, os resultados de cada etapa foram armazenados no banco de dados. Para que nas análises seguintes não seja necessário realizar novamente o tratamento textual.
 - Para cada frase capturada, a quantidade de vezes em que a frase precede um aumento, queda ou estabilidade no preço do Bitcoin é armazenada em uma tabela. Isso foi feito para que não seja necessário realizar essa contagem mais de uma vez para a mesma incidência de uma frase.
 - A análise por repetição de comentários compara incidências repetidas de frases publicadas recentemente com frases publicadas anteriormente. É comum que centenas de frases sejam publicadas todos os dias, e buscar as incidências dessas centenas de frases na tabela `frases_ocorrencias` se mostrou algo que levava quase uma hora. A abordagem natural para esse problema é criar um índice de árvore binária contendo o texto da frase, pois buscas em índices desse tipo são muito mais rápidas, porém quase todos os SGBDs, incluindo o Postgres, são limitados em relação a índices em campos textuais de tamanhos muito grande. Por isso, foi usado um recurso do SGBD Postgres que é criação de índice com resultado de funções. Assim, utilizando a função `md5(texto)`. E ao buscar uma frase nessa tabela, é utilizada uma comparação do resultado da função `md5()` da frase. Isso deixou as queries centenas de vezes mais rápidas.
 - O agrupamento de transações de compra e venda de bitcoin em intervalos de 60 minutos foi feito para ganho de performance nas operações. Assim, foi criada a tabela “cotações”, contendo o valor médio do Bitcoin nesse intervalo (Média do valor de cada transação ponderada pelo volume da transação), o volume total no intervalo e a quantidade de transações contempladas.
 - Várias rotinas, desde a extração de dados, o tratamento de texto e a execução de análises foram programadas utilizando técnicas de paralelismo e são executadas em multithreading.
 - Para acelerar o carregamento das páginas do site, foram criadas tabelas de poucas linhas contendo os dados a serem exibidos nos gráficos presentes no site. Assim, suas consultas são mais rápidas. O tratamento dessas tabelas é feito como última rotina de um ciclo de rotinas agendadas no servidor, listadas na sessão 4.7.
 - A construção do modelo de análise de sentimentos é realizada a cada 7 dias, pois é um processo demorado e poderia prejudicar o tempo necessário para realizar o ciclo de rotinas presente na sessão 4.7. Assim, a análise de sentimentos está sempre avaliando comentários realizados nas últimas 24 horas baseado em um modelo de dados de até no máximo 7 dias atrás.
- Como desafio pessoal do aluno, o trabalho foi feito nos finais de semana pois o aluno atuou durante seu desenvolvimento como programador sênior e gerente de projetos em uma fábrica de softwares tradicional do Rio de Janeiro. Com expedientes que ultrapassavam facilmente as 8 horas diárias previstas. Outro desafio é que anteriormente ao trabalho o aluno jamais havia trabalhado diretamente com o desenvolvimento da camada frontend de um site.

5.2 Resultados Obtidos

Entre os resultados obtidos com o desenvolvimento desse trabalho, destacam-se:

- A publicação do site `criptomante.online`, contendo hospedagem própria, domínio próprio, certificado SSL e possibilidade de visualização das análises publicadas nesse documento.
- A cada 7 dias o modelo de análise de sentimentos é reconstruído e é realizado novamente o processo de separação de uma amostra aleatória de dados para verificação da precisão do novo modelo. E então o resultado é exibido no site. Os resultados obtidos até agora sempre estiveram acima de 60%, e já chegaram até a 63%. Assim sendo, a análise proposta assim como os métodos de tratamento de texto e de transações se demonstraram coerentes com o real movimento no valor do ativo, sendo um resultado positivo, principalmente considerando que o aluno não possui experiência com o mercado financeiro e realizou as análises baseado em ideias apresentadas no referencial teórico desse documento.
- Durante o desenvolvimento do trabalho, o aluno aprofundou produziu pela primeira vez um website sozinho. Realizou o desenvolvimento da camada frontend de um site pela primeira vez (Embora já tenha trabalhado com a camada backend anteriormente).
- O aluno também teve sua primeira experiência fazendo deploy de um site. Desde o aluguel de um servidor, configuração do mesmo, compra de um domínio, compra de um certificado SSL, configuração de DNS, do serviço IIS, etc.
- O aluno aprofundou seus conhecimentos sobre crawlers, web scraping e sobre técnicas de mineração de dados.
- O aluno teve seu primeiro contato prático com aprendizado de máquina, classificadores e outros conceitos de inteligência artificial.

5.3 Sugestões de Trabalhos Futuros

Durante o desenvolvimento desse trabalho, a única análise que teve sua precisão sistematicamente analisada foi a análise de sentimentos. Assim sendo, seria interessante analisar a precisão das demais análises propostas pelo trabalho. Inclusive a análise consolidada.

Como a maioria dos trabalhos relacionados com aprendizado de máquina, a qualidade das análises está ligada com a quantidade de dados analisadas. Por isso pode haver ganho na precisão do resultado ao expandir as rotinas de extração de dados para extrair dados de outras redes sociais e fóruns que sejam sobre a criptomoeda Bitcoin.

Adicionalmente, embora esse trabalho utilize fortemente os argumentos do Princípio de Ondas de Elliott, esse trabalho não utilizou o exato modelo de identificação de ondas proposto por Elliott, pois o trabalho decidiu seguir com técnicas de aprendizado de máquina. Contudo, o modelo de identificação de ondas proposto por Elliott poderia se tornar uma nova análise a ser de alguma forma considerada durante a análise consolidadas. Porém, o modelo é majoritariamente “visual” sobre o gráfico do ativo, e criar um algoritmo sistematizado pode ser um desafio considerável.

Uma aplicação útil para o trabalho é no uso de robôs de investimentos. São ferramentas que realizam de forma automática a negociação de ativos financeiros utilizando análises técnicas sobre o gráfico do ativo e outras técnicas que possam ser expressas por algoritmos. Em 2009 estimava-se que até 73% do volume de negociações de ações nos estados unidos fossem feitas por esse tipo de robôs. Assim sendo, os resultados das análises propostas por esse trabalho poderiam ser usados como um fator adicional a ser considerado por robôs de investimento.

Bibliografia

- [1] FONTENELE, Mariana. **Bitcoin, Ouro e o Papel-moeda**: Uma análise jurídica comparativa da criptomoeda. Orientador: Dra. Ana Frazão. 2017. Monografia (Bacharelado em Direito) - Faculdade de Direito, Universidade de Brasília, Brasília, 2017.
- [2] NAKAMOTO, Satoshi. **Bitcoin: A Peer-to-Peer Electronic Cash System**, 2008a. Disponível em: <https://bitcoin.org/bitcoin.pdf>. Acesso em: 17 nov. 2020.
- [3] WIRECARD. **Pagamentos digitais: como a tecnologia os torna mais seguros?**, 2016. Disponível em: <https://wirecard.com.br/blog/pagamentos-digitais-como-a-tecnologia-os-torna-mais-seguros/>. Acesso em: 17 nov. 2020.
- [4] COMETTI, Natalia. **Um Estudo Sobre A Tecnologia Blockchain E Sua Aplicação Em Sistemas De Votação**. Orientador: Ruy José Guerra Barretto de Queiroz. 2016. TRABALHO DE GRADUAÇÃO (Bacharelado em Ciência da Computação) - Universidade Federal de Pernambuco, Recife, 2016.
- [5] **O QUE É LITECOIN?** [S. l.], 2020. Disponível em: <https://litecoin.org/pt/>. Acesso em: 17 nov. 2020.
- [6] **GRIDCOIN WHITE PAPER: The computation power of a blockchain driving science & data analysis**. 2020. Disponível em: <https://gridcoin.us/assets/img/whitepaper.pdf>. Acesso em: 17 nov. 2020.
- [7] COINMARKETCAP. **Cryptocurrency Prices, Charts and Market Capitalizations**. [S. l.], 2020. Disponível em: <https://coinmarketcap.com/pt-br/>. Acesso em: 17 nov. 2020.
- [8] ÉPOCA (ed.). **Há oito anos, duas pizzas foram vendidas por bitcoins. Hoje, valem R\$ 300 milhões**. [S. l.], 2018. Disponível em: <https://epocanegocios.globo.com/Dinheiro/noticia/2018/05/ha-oito-anos-duas-pizzas-foram-vendidas-com-bitcoin-hoje-elas-valem-r-300-milhoes.html>. Acesso em: 17 nov. 2020.
- [9] BILTON, Nick. **American Kingpin**: The Epic Hunt for the Criminal Mastermind Behind the Silk Road. Portfolio, 2017.
- [10] **US Seizes \$1bn in bitcoin linked to Silk Road site**. The Guardian, 2020. Disponível em: <https://www.theguardian.com/technology/2020/nov/06/us-seizes-1bn-in-bitcoin-linked-to-silk-road-site>. Acesso em: 17 nov. 2020.
- [11] TIMOTHY B., Lee. **Major glitch in Bitcoin network sparks sell-off; price temporarily falls 23%**. [S. l.], 2013. Disponível em: <https://arstechnica.com/information-technology/2013/03/major-glitch-in-bitcoin-network-sparks-sell-off-price-temporarily-falls-23/>. Acesso em: 17 nov. 2020.
- [12] KELION, Leo. **Bitcoin sinks after China restricts yuan exchanges**. BBC News, 2013. Disponível em: <https://www.bbc.com/news/technology-25428866>. Acesso em: 17 nov. 2020.
- [13] WILLIAMS, Mark T. **Virtual Currencies: Bitcoin Risk**. Washington, D.C.: World Bank Conference, 2014.
- [14] PATU, Gustavo. **A Especulação Financeira**. Publifolha, 2001. ISBN 857402287X.
- [15] DALEY, Beth. **Bitcoin is a highly speculative investment. Why caution is required**. The Conversation, 2017. Disponível em: <https://theconversation.com/bitcoin-is-a-highly-speculative-investment-why-caution-is-required-88440>. Acesso em: 17 nov. 2020.
- [16] FARHADLOO, Mohsen. **Fundamentals of Sentiment Analysis and Its Applications**. Springer International Publishing Switzerland, 2016.

- [17] MEIRELLES, Rodrigo. **ESPECULANDO NA BOLSA DE VALORES**: estudos de caso utilizando a análise técnica e a fundamentalista. Orientador: Ângelo Cardoso Pereira. 2010. Trabalho de Conclusão de Curso (Graduação em Engenharia de Produção) - Universidade Federal de Juiz de Fora, Juiz de Fora, 2010.
- [18] NASUKAWA, T. **Sentiment analysis**: capturing favorability using natural language processing. In the Proceedings of the 2nd International Conference on Knowledge 2003. pp. 70-77.
- [19] MAISRETORNO. **Bandas de Bollinger**. Mais Retorno, 2019. Disponível em: <https://maisretorno.com/blog/termos/b/bandas-de-bollinger>. Acesso em: 17 nov. 2020.
- [20] BARROS, Flávia; ROBIN, Jacques. **Processamento de Linguagem Natural**. Universidade Federal de Pernambuco. 1997.
- [21] ELLIOTT, Ralph Nelson. **The Wave Principle**. 1ª. Ed.: BN Publishing, 2020. ISBN 9787364654667.
- [22] **SENTIMENT Model**. Elliott Wave Analytics, 2020. Disponível em: <http://www.Elliottwaveanalytics.com/services/the-thovallo-sentiment-model/>. Acesso em: 17 nov. 2020.
- [23] ZEIFMAN, Igal. **Bot Traffic Report 2016**. Imperva, 2017. Disponível em: <https://www.imperva.com/blog/bot-traffic-report-2016/>. Acesso em: 17 nov. 2020.
- [24] LUCCA, J. L. **Lematização versus Stemming**. Núcleo Interinstitucional de Linguística Computacional NILC – ICMC – USP. 2002.
- [25] CARVALHO, André. **Inteligência Artificial**: Uma Abordagem de Aprendizado de Máquina. 1ª. Ed. Novatec Editora, 2011. ISBN 8521618808.
- [26] KIM, Minsoo; COLLEGE, Pomona. **Statistical Classification**. 2010.
- [28] PAULA, Gilberto. **Modelos de Regressão com Apoio Computacional**. Instituto de Matemática e Estatística, Universidade de São Paulo, 2011.
- [30] **LOGISTIC Regression (aka logit, MaxEnt) classifier**. Scikit-learn.org, 2020. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. Acesso em: 17 nov. 2020.
- [31] **BEST Brazilian Stocks: Daily Forecast and Global Model Performance Evaluation Report**. I Know First, 2020. Disponível em: <https://iknowfirst.com/best-brazilian-stocks-daily-forecast-and-global-model-performance-evaluation-report>. Acesso em: 17 nov. 2020.
- [32] FINBRAIN. **FINBRAIN'S Stock Predictability Indicator**. FinBrain, 2019. Disponível em: <https://blog.finbrain.tech/2019/07/29/finbrains-stock-predictability-indicator/>. Acesso em: 17 nov. 2020.
- [33] ZOLA, Andrew. **The 5 Best Programming Languages for AI**. Springboard Blog, 2018. Disponível em: <https://www.springboard.com/blog/best-programming-language-for-ai/>. Acesso em: 17 nov. 2020.
- [34] COSTA, Claire D. **Top Programming Languages for AI Engineers in 2020**. Towards Data Science, 2020. Disponível em: <https://towardsdatascience.com/top-programming-languages-for-ai-engineers-in-2020-33a9f16a80b0>. Acesso em: 17 nov. 2020.
- [35] RAHMAN, Jahidur. **Top 5 Best Programming Languages for Artificial Intelligence field**. Geeks for geeks, 2020. Disponível em: <https://www.geeksforgeeks.org/top-5-best-programming-languages-for-artificial-intelligence-field/>. Acesso em: 17 nov. 2020.
- [36] **WHY Django?** Django Software Foundation, 2020. Disponível em: <https://www.djangoproject.com/start/overview/>. Acesso em: 17 nov. 2020.

- [37] **STATISTICS - Mathematical statistics functions**. Python Software Foundation, 2020. Disponível em: <https://docs.python.org/3/library/statistics.html>. Acesso em: 17 nov. 2020.
- [38] **WHAT is NumPy?** The SciPy community, 2020. Disponível em: <https://numpy.org/doc/stable/user/whatisnumpy.html>. Acesso em: 17 nov. 2020.
- [39] **PANDAS**. [S. l.], 2020. Disponível em: <https://pandas.pydata.org/>. Acesso em: 17 nov. 2020.
- [40] **GETTING Started**. Scikit-learn.org, 2020. Disponível em: https://scikit-learn.org/stable/getting_started.html. Acesso em: 17 nov. 2020.
- [41] **SPACY 101: Everything you need to know**. ExplosionAI GmbH, 2020. Disponível em: <https://spacy.io/usage/spacy-101>. Acesso em: 17 nov. 2020.
- [42] RICHARDSON, Leonard. **Beautiful Soup Documentation**. [S. l.], 2020. Disponível em: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Acesso em: 17 nov. 2020.
- [43] ABUBAKAR, Yusuf. **BENCHMARKING POPULAR OPEN SOURCE RDBMS: A PERFORMANCE EVALUATION FOR IT PROFESSIONALS**. [S. l.], 2014. Disponível em: https://www.researchgate.net/publication/316132763_BENCHMARKING_POPULAR_OPEN_SOURCE_RDBMS_A_PERFORMANCE_EVALUATION_FOR_IT_PROFESSIONALS_BENCHMARKING_POPULAR_OPEN_SOURCE_RDBMS_A_PERFORMANCE_EVALUATION_FOR_IT_PROFESSIONALS. Acesso em: 17 nov. 2020.
- [44] NAKAMOTO, Satoshi. **Bitcoin P2P e-cash paper**, 2008. Disponível em: <https://www.metzdowd.com/pipermail/cryptography/2008-October/014810.html>. Acesso em: 22 nov. 2020b.