Trabalho de Conclusão de Curso

Prevendo a flutuação de preço do Ether através da Análise de Sentimento do Twitter

Rodolfo Wagner Vieira Leite Moreira

Orientador:

Prof. Dr. Bruno Almeida Pimentel

Maceió, 2 de Dezembro de 2022

Rodolfo Wagner Vieira Leite Moreira

Prevendo a flutuação de preço do Ether através da Análise de Sentimento do Twitter

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Engenharia de Computação do Instituto de Computação da Universidade Federal de Alagoas.

Orientador:

Prof. Dr. Bruno Almeida Pimentel

Maceió, 2 de Dezembro de 2022

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Helena Cristina Pimentel do Vale – CRB4 –661

M838p Moreira, Rodolfo Wagner Vieira Leite.
Prevendo a flutuação de preço do Ether através da análise de sentimento do twitter / Rodolfo Wagner Vieira Leite Moreira. - 2022.
53 f : il.

Orientador: Bruno Almeida Pimentel
Monografia (Trabalho de Conclusão de Curso em Engenharia de Computação) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2022.

Bibliografia: f. 49-53.

1. Criptomoeda. 2. ETH (Ether). 3. Aprendizado de máquina. 4. Análise de sentimento. 5. Twitter. 6. Valence aware dictionary and entiment reasoner and humans apart, (VADER). I. Título.

CDU: 004

Agradecimentos

A Deus por ter me dado saúde e força para superar as dificuldades. Aos meus pais, pelo amor, incentivo e apoio incondicional e a minha esposa, Larissa, que sempre me apoiou e me manteve acreditando enquanto eu duvidava.

"A vida e o tempo são os dois maiores professores. A vida nos ensina a fazer bom uso do tempo enquanto o tempo nos ensina o valor da vida."

– Autor Desconhecido

Resumo

As criptomoedas, tecnologia inovadora que tem ganhado bastante adesão e visibilidade, ainda são cercadas de desconfianças, mas, à medida que o impacto social e econômico das criptomoedas continua a crescer rapidamente, o número de artigos em sites de notícias, reportagens televisivas e comentários nas redes sociais também crescem. Com a popularização e crescimento da rede social Twitter, as opniões e sentimentos expressos em forma de *tweets*, representam uma grande quantidade de dados que podem ser estudados. Neste trabalho, analisamos *tweets* com o objetivo de estabelecer se o sentimento destes é uma ferramenta útil para previsão de flutuações de preços numa criptomoeda chamada *ETH*. O conjunto de dados é constituído por 34.558 *tweets* recolhidos utilizando a biblioteca TWINT que consiste em *tweets* contendo o sinal "ETH". Os *tweets* recolhidos são convertidos em pontuações que representam sua positividade/negatividade, isto é feito por uma ferramenta de análise chamada VADER. A pontuação dos *tweets* é então somada para representar um sentimento coletivo por hora, funcionando como uma variável de predição para o preço da criptomoeda. Utilizamos em nossos experimentos técnicas de aprendizagem supervisionada treinando *Random Forest* e para efeitos de comparação, treinamos também modelo utilizando Regressão Linear Múltipla.

Palavras-chave: Criptomoedas, ETH, Aprendizado de Máquina, Análise de Sentimento, Twitter, VADER.

Abstract

Cryptocurrencies, an innovative technology that has gained a lot of traction and visibility, are still surrounded by mistrust, but as the social and economic impact of cryptocurrencies continues to grow rapidly, the number of articles on the web, television reports, and comments on social networks are also growing. With the popularization and growth of the social network Twitter, the opinions and sentiments expressed in the form of of tweets represent a large amount of data that can be studied. In this paper, we analyze tweets with the aim of establishing whether their sentiment is a useful tool for predicting price fluctuations in a cryptocurrency called ETH. The dataset consists of 34,558 tweets collected using the TWINT library, which consists of tweets containing the "ETH"sign. The collected tweets are converted into scores that represent positivity/negativity of the tweet, this is done by an analysis tool called VADER. The scores of the tweets are then summed up to represent a collective sentiment per hour, acting as a predictor variable for the price of cryptocurrency price. We use in our experiments supervised learning techniques training a Random Forest Classifier and for comparison purposes, we also trained a model using Multiple Linear Regression.

Keywords: Cryptocurrencies, ETH, Machine Learning, Sentiment Analysis, Twitter, VADER.

Lista de Abreviaturas e Siglas

ETH	*Ether*
VADER	*Valence Aware Dictionary and sEntiment Reasoner and Humans Apart*
TWINT	*Twitter Intelligence Tool*
RF	*Random Forest*
API	Application Programming Interface
CSV	*Comma Separated Values*
JSON	*JavaScript Object Notation*
MSE	*Mean Square Error*
URL	*Uniform Resource Locator*

Lista de Figuras

Sumário

1

Introdução

1.1 Motivação

No início da civilização humana, o comércio era baseado no escambo, que consistia basicamente na troca de mercadorias de valor abstrato e muito particular entre as partes, de tal forma que poderia resultar em impasses comerciais. Com o passar dos séculos a civilização evoluiu, as cidades cresceram e as interações humanas ficaram complexas, o escambo não era mais suficiente como sistema de troca então foram criadas as moedas. O comércio havia evoluído a fim de solucionar as necessidades humanas e continuou evoluindo constantemente até os tempos atuais, em que os bancos centralizam quase que a totalidade das operações financeiras e que além disso, são responsáveis pela segurança e confiabilidade de todo este sistema (CARVALHO FRANCISCO PRANCACIO ARAúJO, 2014).

Mundialmente, a dependência populacional nos bancos e em suas operações bancárias, assim como o grande avanço tecnológico, fomentou a consolidação das moedas digitais ou criptomoedas como forte alternativa ao sistema financeiro convencional (ORRELL; CHLUPATý, 2016). Criptomoedas são representações digitais de valor que existem puramente em formato eletrônico com base na tecnologia blockchain. É uma tecnologia revolucionária baseada em princípios descentralizados e criptográficos que funcionam como intermediários nas transações digitais. Isso torna os papéis de terceiros obsoletos. A tecnologia blockchain sustenta esta revolução através das redes de ponto-a-ponto (*peer-to-peer*) criando um registro de transações compartilhado e confiável, fornecendo segurança, privacidade e imutabilidade às operações financeiras realizadas. O *blockchain* pode ser definido como um livro distribuído ou um banco de dados descentralizado, acessível em uma rede e registra as transações como um bloco digital com carimbo de data/hora irrevogável. Por isso, tal rede ponto-a-ponto (*peer-to-peer*) dispensa a necessidade de um agente centralizador responsável para validar as transações, somente membros podem validar transações entre si. Como consequência, os custos operacionais demandados pelas instituições centralizadoras que

intermedeiam as operações financeiras são reduzidos, tornando as criptomoedas bastante atrativas em diferentes cenários.

Apresentada de uma forma totalmente diferente, o Bitcoin foi anunciado através de um artigo *white paper* de seu criador Satoshi Nakamoto, postado em um fórum aberto de discussões sobre criptografia. Sendo baseado em código fonte aberto, o sistema não pertence à ninguém, sua moeda não pode ser reproduzida fora das regras do sistema, resolvendo os problemas que eram encontrados nas tentativas prévias de criação de criptomoedas e também o problema das moedas em papel controladas por bancos e governos (NAKAMOTO, 2008). Nakamoto comenta como o mercado digital de transações estava muito dependente de empresas terceirizadas para avaliar as transferências e dar credibilidade às mesmas. Propôs então um sistema que não dependesse desse terceiro para confiar a transação, e sim um sistema de criptografia que gerasse a confiança necessária, como ele descreve quando diz "O que se precisa é um sistema de pagamento eletrônico baseado em prova criptográfica em vez de prova de confiança"(NAKAMOTO, 2008).

O contexto em que o Bitcoin foi introduzido não poderia ter sido melhor, em meio a crise de 2008 que atingia boa parte do mundo, mesmo que essa não tenha sido a única razão de sua criação. O mundo estava enfrentando a parte negativa dos chamados ciclos econômicos. O avanço considerável nas tecnologias digitais também foram um grande impulso para o desenvolvimento da moeda, visto que o poder computacional aumentava a cada ano que passava, juntamente com a rapidez e expansão da internet, que não só melhorava sua velocidade, como também o alcance em todo o mundo.

Após um tempo de amadurecimento no fórum em que foi postado, o Bitcoin foi oficialmente lançado no começo de 2009, sendo transmitido para a rede pelo próprio Satoshi, iniciando o chamado 'bloco gênese', primeiro bloco de dados do *blockchain* do Bitcoin. O sistema foi disponibilizado para *download* alguns dias depois, onde começou a mineração da moeda. Por um período de pelo menos um ano, a moeda foi apenas minerada, sem possuir um valor, pois ainda não havia sido comercializada por ninguém. Então em 2010 a primeira transação da moeda ocorreu, quando um usuário trocou seus 10.000 bitcoins por duas pizzas. A moeda passou a ter seu primeiro valor comercial e hoje em dia é a criptomoeda que mais ganhou notoriedade, apesar de não ser atrelada à nenhuma política econômica de um governo, tornou-se uma moeda confiável e consolidada no mundo todo (BöHME et al., 2015). O sucesso do Bitcoin se deve principalmente à tecnologia blockchain e sua ascensão fez com que alguns países a adotassem no comércio e em operações de câmbio (CHOKUN, 2013).

Após seu lançamento, o Bitcoin passou a ter várias outras criptomoedas concorrentes. Choran (2017) e Farell (2015) explicam: "desde o lançamento da anárquica pioneira criptomoeda, Bitcoin, para o público em janeiro de 2009, mais de 550 criptomoedas foram desenvolvidas, a maioria com quase nada de sucesso.". Hoje em dia já se estima que mais de 1000 existam, e novas criptomoedas surgem a todo o momento procurando um lugar ao sol,

sendo o Ether uma delas.

A proposta da plataforma *Ethereum* surgiu com Vitalik Buterin, um membro ativo da comunidade do próprio Bitcoin em 2013. Devido ao seu trabalho na comunidade ao longo dos anos, ele propôs uma nova plataforma baseada na tecnologia Blockchain, que pudesse fazer mais do que a plataforma do Bitcoin e que possuísse uma moeda própria (GERRING, 2016). A diferença está que a plataforma *Ethereum* foi projetada para ser mais acessível e flexível que o Bitcoin, podendo ser usada para criar contratos inteligentes (*Smart Contracts*) e também aplicativos descentralizados.

Foi formalmente anunciada em uma conferência sobre Bitcoin e teve seu sistema implementado com o apoio de Gavin Wood, co-fundador da plataforma. Todo o projeto foi financiado através da própria comunidade, utilizando Bitcoins, por entusiastas e programadores que acompanhavam a proposta do sistema. Estes Bitcoins eram trocados por sua própria moeda ou 'token', chamada *Ether*. A plataforma foi lançada 1 ano após o financiamento começar, em julho de 2015, atraindo os primeiros mineradores e também os primeiros utilizadores da plataforma. Hoje o Ether é a segunda maior criptomoeda em valor de mercado (REUTERS, 2021), consolidando-se no mercado internacional.

A popularização da internet resultou na produção em massa de conteúdo. Como consequência desse fenômeno, nascem as redes sociais que por sua capacidade de retroalimentação, isto é, produzem e consomem conteúdo, têm revolucionado as formas de relacionamentos, em todos os níveis, como por exemplo: relacionamentos interpessoais, amizade, namoro e casamento, cliente-empresa, captação de clientes, *feedback* dos serviços e publicidade (VRIES LISETTE, 2012). Formar opniões, dissipar informações, criar tendências, também são exemplos das inúmeras formas de atuação nesse ecossistema que trouxe o mundo a uma nova era.

Com tantas riquezas de informações disponíveis em domínio público, as opniões dos usuários, publicadas em seus perfis nas redes sociais, despertaram a atenção de empresas para análise de satisfação, em detrimento dos tradicionais formulários de questões, utilizados até então. Uma outra ocorrência derivada das redes é o engajamento em massa, popularmente donominado de "viralização", que detém o poder de alavancar rapidamente novas tecnologias, conteúdos, costumes, entre outros, podendo resultar em mudanças na dinâmica da sociedade (VRIES LISETTE, 2012).

As redes sociais são algumas das ferramentas mais atrativas na internet, uma vez que é possível compartilhar informações, promover a interação entre pessoas e disseminar ideias (TOMAéL MARIA INêS, 2005). Devido a esses motivos (WASSERMAN, 1994), é interessante permitir que essas informações sejam catalogadas e analisadas de forma a entender os diferentes grupos de usuários presentes na rede. Além disso, pode-se direcionar a coleta desses dados ao seguir um perfil específico de usuários ou participar de grupos temáticos, o que viabiliza o estudo de opniões que surgem nessas redes sociais.

O *Twitter* [1], uma das redes sociais mais populares, permite a interação entre usuários, de maneira fácil, simples e objetiva, tendo se mostrado uma fonte rica em informações e opniões de usuários. O *Twitter* permite que usuários escrevam mensagens textuais de até 280 caracteres em seus perfis, geralmente expressando suas opiniões e sentimentos em relação a temas ou assuntos específicos. Particularmente, a grande disponibilidade de *tweets* sobre o mercado financeiro e o desenvolvimento de ferramentas computacionais para coleta possibilitou a análise desses textos para identificar tendências, expressas pelos seus usuários, e a partir disso, prever o comportamento do mercado de valores a partir das mensagens presentes nas redes sociais (KWAK et al., 2010).

Como a quantidade de mensagens textuais nas redes sociais cresce cada vez mais, tarefas de análise desses textos e as opiniões dos usuários tornam-se inviáveis se realizadas por um especialista humano. Nesse contexto, a área de análise de sentimentos, intersecção das área de mineração de textos e processamento de linguagem natural, pode contribuir com métodos e técnicas para extrair automaticamente conhecimento relevante e implícito de textos. Na literatura, a análise de sentimentos foi empregada em textos de diversos domínios do conhecimento, como na identificação de locais de crimes (CLARINDO JOãO PAULO,), no auxílio à decisão de vencedores de concursos de televisão (FIGUEREDO IGLESON F,), na classificação de filmes (TAVARES, 2017), entre outros. Por isso, essa pesquisa considera a oportunidade de aplicar técnicas de análise de sentimentos em conjuntos de tweets relacionados ao *Ether*.

1.2 Justificativa

Apesar de não haverem formas de usar criptomoedas de forma simples no dia-a-dia, as moedas passaram de ferramenta de marketing, para moedas capazes de servir de refúgio para países em situação de crise e hiperinflação, como na Venezuela (FRANCO. . . ,). Além de servir muito bem como proteção contra a inflação, as moedas permite facilidade em transações internacionais, antes impossíveis sem um ou mais bancos intermediários. A medida que a quantidade de *tokens* cresce e a tecnlogia fica mais conhecida, o valor das moedas mais conhecidas como Bitcoin e *Ether* sobem.

Apesar de historicamente o valor da moeda ter aumentado, dada sua natureza inovadora e seu real valor ainda por ser descoberto, as moedas são bastante volátéis. Ainda segundo o CoinMarketCap [2], a moeda já viu seu preço subir mais de 100% e depois voltar ao patamar original num prazo de apenas 1 mês. Tal volatilidade e possibilidade de ganhos expressivos em pouco tempo atrai vários *traders*, que são pessoas que compram ativos unicamente com a intenção de especular e vende-los por um preço maior do que o comprado. Normalmente

[1] Twitter <https://twitter.com/>
[2] CoinMarketCap <https://coinmarketcap.com/>

o prazo entre a compra e a venda é curto, variando de várias compras/vendas no mesmo dia (*day traders*) até um prazo de algumas semanas para os chamados *swing traders*.

Se tais operações podem ser realizadas por pessoas de forma lucrativa, deve haver também formas computacionais eficazes de prever posições lucrativas de operação. Com o Bitcoin, diversos estudos foram realizados com o objetivo de estabelecer uma correlação entre variáveis e o preço do Bitcoin (KRISTOUFEK, 2015) (KROLL; DAVEY; FELTEN, 2013) (SHAH; ZHANG, 2014) (KAMINSKI, 2014), nenhum dos estudos chegou em uma forma extremamente conclusiva. Entretanto, um pesquisador (KAMINSKI, 2014) tentou estabelecer uma correlação através de regressão entre emoções em *tweets* e os indicadores do mercado, concluíndo que o Twitter funcionava muito bem como um espelho do mercado.

Desde o início, a mídia social tem sido a plataforma para informações sobre criptomoedas. Pesquisadores usaram postagens de fóruns como Reddit e Coindesk para aplicar análise de sentimento e estimar flutuações de preço no Bitcoin (WOŁK, 2020). As evidências indicam que as variações de preço das criptomoedas são substancialmente influenciadas pelo sentimento das redes sociais (MAI et al., 2018).

O Twitter, é uma rede social bastante rica em informações, nele é permitido que usuários exprimam-se utilizando mensagens textuais de até 280 caracteres, quase sempre usados para expressar opniões sobre temas específicos (KWAK et al., 2010). O Twitter não apenas oferece informações sobre criptomoedas em tempo real, mas também fornece uma ampla fonte de perspectiva das pessoas em relação à uma criptomoeda. Uma vez que os investidores frequentemente compartilham seus sentimentos. Desta forma a plataforma é útil para gerar uma grande quantidade de sentimento na análise de dados (KWAK et al., 2010).

Apesar de na literatura existir diversas pesquisas que possuem o objetivo de prever posições lucrativas de operação de criptomoedas, e também o uso de análise de sentimentos como auxílio em tomadas de decisão, nenhuma teve um estudo satisfatório com objetivo do uso da análise de sentimentos como auxílio na predição da criptomoeda *Ether*, a segunda moeda com maior valor de mercado atualmente, somente atrás do Bitcoin.

1.3 Objetivos

Dado um grande número de dados de acesso livre do Twitter, contendo o sentimento de investidores e usuários de criptomoedas, o principal objetivo deste estudo é testar se a análise de sentimentos nos dados do Twitter é uma ferramenta útil para prever oscilações de preço de uma criptomoeda alternativa chamada Ether, respondendo as seguintes perguntas:

- **(i)** É possível realizar análise de sentimentos de *tweets* relacionados ao Ether?
- **(ii)** A análise de sentimentos é uma ferramenta útil para prever oscilações de preço do Ether?

Para respondermos a questão (i), os seguintes passos foram tomados: Primeiramente, a obtenção do conjunto de dados de *tweets* relacionados à criptomoeda ETH, em seguida foi aplicada técnicas de pré-processamento de texto e, subsequentemente, a classificação dos *tweets* com suas respectivas polaridades (positivo, negativo, neutro), por fim, os *tweets* foram agrupados em seções de uma hora, tendo suas polaridades somadas representando o sentimento sobre o ETH naquela hora.

Por outro lado, para respondermos a questão (ii) empregamos técnicas de Aprendizado de Máquina, criando e avaliando os modelos preditores de *Random Forests* e Regressão Linear Múltipla, utilizando nos dois casos datasets com e sem o *input* de sentimento para comparação.

1.4 Estrutura do trabalho

O restante dessa monografia está estruturada da seguinte maneira:

- **Capítulo 2 - Fundamentação teórica:** aspectos e referenciais teóricos que serviram de base para a pesquisa científica;
- **Capítulo 3 - Trabalhos Relacionados:** resumos de artigos relacionados à análise dos movimentos do mercado financeiro e análises de mensagens presentes no Twitter;
- **Capítulo 4 - Procedimentos Metodológicos:** expõe as principais informações relativas à implementação de técnicas de pré-processamento, caracterização, análise de sentimentos e modelos gerados;
- **Capítulo 5 - Resultados experimentais:** descreve os resultados obtidos a partir da realização de experimentos para validar a metodologia proposta em conjunto de tweets relacionados ao *ether*;
- **Capítulo 6 - Conclusões:** apresentação dos objetivos cumpridos, conclusão do trabalho, limitações encontradas e eventuais trabalhos futuros.

2

Fundamentação Teórica

2.1 Criptomoedas e a Tecnologia Blockchain

O começo de toda a vontade de criar um novo meio de pagamento que utilizasse a internet e fosse criptografado veio entre o final da década de 80 e começo da década de 90, onde surgiu os primeiros conceitos e algumas tentativas de criação de novas moedas que pudessem servir o propósito de substituir o dinheiro convencional (GRIFFITH, 2018). David Chaum, fundador da empresa *Digicash*, foi o primeiro a usar uma moeda digital que usava criptografia para a segurança das transações. A moeda *eCash* foi pioneira nesse aspecto e até utilizada por alguns bancos e também por smart cards, mas não se caracterizando exatamente como os as moedas digitais de hoje (CHUEN D. L. K.; GUO, 2017).

Desfrutando de certo sucesso, a empresa acabou tomando algumas decisões erradas e faliu, sendo vendida logo após, e a moeda sendo esquecida e descontinuada pelos seus novos donos. Apesar disso o sistema era promissor, pois além da criptografia, também utilizava o sistema de assinaturas cegas (*Blind Signatures*) para proteger a identidade dos utilizadores (CHUEN D. L. K.; GUO, 2017). Logo após a falência da *Digicash*, surgiram algumas outras empresas que tentaram suas próprias ideias para o dinheiro digital. Uma delas foi o *PayPal*, ainda que operasse de forma diferente do que é hoje.

Uma década depois, no começo do século 21, uma nova forma de dinheiro virtual veio à tona. Bancadas e garantidas por depósitos em barras de ouro, as moedas digitais de ouro (*Digital Gold Currency*) foram criadas em 1996, mas acabaram não sendo muito usadas até o final da década, quando começaram a ter sua vez no mercado (GRIFFITH, 2018). A principal contribuição dessa forma de moeda digital foi a tecnologia implantada nos métodos de pagamento. O sistema *e-Gold* foi o primeiro a ter alguma relevância no ramo de pagamentos eletrônicos, e contribuiu com tecnologias que são usadas até hoje por outros sistemas de e-commerce mais modernos.

Durando até o ano de 2008, e movimentando uma grande quantidade de dinheiro, a crise global atingiu também as moedas e plataformas de pagamento digitais, e esse sistema foi encerrado, e todos os métodos de pagamento que operavam via moedas digitais de ouro (DGC) foram liquidados. No meio de toda a fumaça e da batalha contra a crise, o interesse em cima das criptomoedas não recuou. A teoria que cerca as moedas digitais tinham o potencial de resolver e ajudar com alguns dos problemas que a crise tinha apresentado (CHUEN D. L. K.; GUO, 2017).

Alguns desses estudiosos se juntaram e começaram a elaborar a teoria que serviria de base para a tecnologia que se tornaria o blockchain e seria utilizada nas criptomoedas que temos hoje. Junto a eles, um membro de um grupo ativista que defendia o uso de criptografia pesada formulou um sistema novo reutilizável de checagem dos processos que os voluntários fariam. Esse novo sistema impediria os hoje conhecidos ataques DDoS (*Distributed Denial of Service* ou Recusa de Serviço Distribuída) e também ajudaria a impedir os *Spams*. A moeda digital Bitcoin, no seu blockchain usa um sistema que opera dessa forma chamado Hashcash (CHUEN D. L. K.; GUO, 2017).

Com os bancos centrais cada vez mais tentando suprimir os problemas gerando dinheiro da forma que queriam e causando mais problemas com uma distribuição de capital totalmente desigual, as criptomoedas tinham o potencial de ser uma alternativa, visto que sua geração seguia uma regra e sua quantidade podia ser limitada. Bem ao contrário do que estava acontecendo na economia real, os entusiastas começaram a acreditar no potencial desse sistema (CHUEN D. L. K.; GUO, 2017). Foi então que surgiu, nesse mesmo ano de 2008, a pioneira das moedas digitais que conhecemos hoje: o famoso Bitcoin.

O Bitcoin é que uma moeda criptografada e foi difundida por meio da internet. Essa moeda digital funciona de forma descentralizada através da tecnologia de rede *peer-to-peer* para realizar suas transações, sem interferência de bancos ou instituições financeiras. Todas as transações de bitcoins são registradas e validadas por meio da tecnologia blockchain, que pode ser definido como um banco de dados distribuído de registros onde cada movimentação é verificada por consenso da maioria das entidades participantes da rede (CROSBY MICHAEL, 2016).

A tecnologia Blockchain surgiu junto com a moeda Bitcoin em 2008. Mesmo que hoje em dia ainda algumas pessoas associem as duas coisas como sendo a mesma, o blockchain na verdade é o que garante o funcionamento do Bitcoin da forma que ele foi concebido. O Bitcoin acabou por ser a primeira moeda digital descentralizada que serviria para a utilização de todos sem restrições, mas a verdadeira ferramenta inovadora foi o sistema desenvolvido para garantir o registro e a segurança das transações (NAKAMOTO, 2008).

O blockchain opera de forma par-a-par ou *peer-to-peer* (P2P) em uma rede compartilhada, onde todos os usuários podem acessar os registros para conferência e podem adicionar novos blocos de informações. Ele não possui um servidor principal e toda a informação que roda na rede é criptografada, com os próprios usuários garantindo sua veracidade, o

que dificulta para qualquer hacker atacar a rede. O próprio nome do sistema serve como esclarecimento de como ele funciona. "*Block*" traduzindo para Blocos e "*Chain*" traduzindo para Corrente, nos entrega a ideia. Tudo é uma grande corrente de blocos de dados que necessitam de seus anteriores para serem válidos.

Cada bloco deve possuir as informações sobre o que esse bloco está operando, deve conter a referência ao seu bloco anterior, o Hash (de forma simples, seria a impressão digital do bloco anterior), deve conter seu próprio *Hash*, a data em que está sendo gerado, e dependendo de como funcionar especificamente o blockchain, mais informações são necessárias para tornar o bloco válido. Com todas essas informações juntas, o bloco entrará na corrente do blockchain não só validando a transação contida nele, mas como todas as outras transações posteriores, mantendo a integridade da corrente.

O blockchain segue alguns conceitos e propriedades que o tornam essa tecnologia revolucionária. Seus conceitos partem do "*Shared Ledger*", que é um sistema grava todas as informações que passam pela rede. Ele também é compartilhado entre todos os usuários, todos possuem uma cópia do registro completo. Outro conceito é o das permissões que em uma rede de blockchain podemos operá-la de forma livre, sem necessidade de permissão (*Permissionless*) para acessar as operações, ou podemos utilizar algumas restrições de acesso, tornando-a mais restrita a alguns usuários apenas. Esse método é utilizado por empresas de forma privada, quando necessitam colocar mais informações dentro dos blocos e não querem isso acessível a todos. Dessa forma o blockchain pode ser configurado na forma de um "smart contract", e as informações mais sensíveis só ativam para os usuários prédeterminados, ou seja, todos podem ver que A e B estão fazendo negócios, mas não podem ver os dados mais sensíveis da transação.

O terceiro conceito é o do consenso. O consenso diz que em uma rede compartilhada, se a maioria dos usuários se mantiver honesta, a corrente vai ser confiável. Para garantir que os usuários continuem corretos, são adotados alguns sistemas na criação dos blocos, como as Provas de Trabalho (*Proof of Work*) ou as Provas de Aposta (*Proof of Stake*) que garantem a honestidade da rede. A prova de trabalho requer que os usuários gastem seus recursos, eles sendo energia e hardware, para resolver vários problemas matemáticos, para assim poder gerar seu bloco. Já a prova de aposta requer que os usuários possuam certa quantidade da moeda consigo para poder gerar blocos.

O ultimo conceito é o do próprio *Smart Contract*. O smart contract é uma forma de garantir certas condições para que ocorram as transações de forma segura, pois só será executado o contrato conforme as condições forem atendidas. Um exemplo dessas regras foi descrito acima, dando permissão somente para alguns usuários terem acesso a certas informações. Dessa forma o usuário deve provar que possui a autorização necessária para acessar as informações, assim ativando a condição do *smart contract*.

No que cerca as propriedades do blockchain, temos 4 exemplos do que é necessário para a rede funcionar corretamente. O primeiro é o tamanho da rede (*Network Size*), isso é im-

portante para que a rede sobreviva por muito tempo, e é garantido normalmente pelo meio da mineração, no caso do Bitcoin, ou pelas taxas cobradas, no caso das outras moedas e plataformas, funcionando como um atrativo (WITTE, 2016). O segundo é a profundidade do Blockchain (*Blockchain Depth*). Quanto mais profunda for a corrente, mais difícil será de alguém conseguir interferir nela, pois terá que reconstruir todos os blocos existentes para que possa realmente ter algum efeito. Uma forma de entender essa segunda propriedade está na terceira propriedade, chamada de Ataque de 51% (51% *Attack*). A teoria diz que se um grupo malicioso conseguir ter o poder de processamento computacional que mais da metade dos usuários honestos, esclarecendo que não necessariamente 51%, eles poderiam interferir diretamente na corrente, alterando tanto os blocos novos quanto os antigos. A probabilidade de isso acontecer em qualquer dos grandes Blockchains em circulação hoje é muito baixa, visto que o investimento para ter esse tipo de poder de processamento é muito alto (WITTE, 2016).

A última propriedade é a do roubo (*Theft*). Enquanto é muito difícil para alguém conseguir interferir na corrente do Blockchain, nada impede que se você possuir algum valor de criptomoeda em uma carteira digital, essa carteira digital não possa ser hackeada, pois ela não está vinculada com a corrente blockchain em si. As criptomoedas roubadas podem ser gastas diretamente da carteira digital, sem violar nenhuma regra do Blockchain (WITTE, 2016).

A pioneira das moedas digitais que conhecemos hoje é o famoso Bitcoin. A criação do Bitcoin é misteriosa, até hoje não se sabe se foi criada por uma pessoa ou por um grupo de pessoas usando o nome "Satoshi Nakamoto"em 2009 (NAKAMOTO, 2008) [1]. O Bitcoin é uma coleção de conceitos e tecnologias que formam a base de dinheiro digital. As unidades de moeda chamadas bitcoins são usadas para armazenar e transmitir valor entre os participantes da rede bitcoin. Os usuários de Bitcoin se comunicam usando o protocolo bitcoin principalmente via Internet. A pilha de protocolos bitcoin, pode ser executada em uma ampla gama de dispositivos de computação, incluindo laptops e smartphones, tornando a tecnologia facilmente acessível (ANTONOPOULOS, 2014).

As moedas são geradas em blocos, através de pessoas que usam softwares de mineração digital. A quantidade de fundos disponibilizada é ajustada em uma crescente previsível e controlada – apenas 21 milhões de Bitcoins serão criadas até o ano 2140. Os cálculos feitos pelos “mineradores” ajudam a verificar as transações de toda a rede (CABRAL, 2013). Todo usuário controla seus próprios fundos, por meio de uma chave privada criptográfica (FRANCO, 2014). Dessa forma, quando um usuário deseja gastar alguns fundos, ele deve usar essa chave privada para assinar uma mensagem informando para quem deseja enviar os fundos, bem como o valor a ser enviado. O usuário transmite essa mensagem assinada para a rede e todos os participantes da rede recebem uma cópia dela. Cada nó pode veri-

[1] Bitcoin <https://en.wikipedia.org/wiki/Bitcoin>

ficar independentemente a validade da mensagem e atualizar seu banco de dados interno adequadamente.

Após o lançamento do Bitcoin e ao passar dos anos, muito mais *blockchains* foram criadas com o objetivo de criação de criptomoedas, hoje existem mais de 12.000 criptomoedas em existência [2]. Entretanto, apesar da tecnologia da *blockchain* ter sido criada para o lançamento de criptomoedas, mais especificamente o Bitcoin, o uso da mesma é alvo de intensas discussões e incentivado por governos como o da Alemanha, motivado pelas diversas oportunidades de aplicação.

As aplicações da tecnoloia têm alterado de forma disruptiva a forma como funcionam os mercados financeiros, cadeias de fornecedores e as relações com os consumidores. (WALPORT, 2015) Em Tapscott (TAPSCOTT D.; TAPSCOTT, 2016), aplicações diversas são apresentadas e é feito até a comparação do surgimento da internet ao surgimento da *Blockchain*, evidenciando o poder da ferramenta em seus campos de aplicação. A *Blockchain* otimiza a forma como as negociações são feitas redefinindo a confiança no mundo digital e eliminando a necessidade de intermediários entre transações. De acordo com (FRONI A. A.; MEULEN,) a fase inicial onde a tecnologia passa por fortes especulações já foi superada, sendo esperado que se torne uma tecnologia padrão em cerca de dois a cinco anos.

[2] https://www.fool.com/investing/stock-market/market-sectors/financials/cryptocurrency-stocks/how-many-cryptocurrencies-are-there/

2.2 O Ethereum

Desde seu surgimento em 2009, o Bitcoin e seu núcleo operacional *blockchain*, fundaram uma nova era para as transações digitais ponto a ponto, de acordo com (WOOD, 2014) o bitcoin não só mostrou o poder dos mecanismos de consenso mas também o respeito aos contratos que tornam possível a criação de um sistema de transferência de valor descentralizado, compartilhado ao redor do mundo e virtualmente livre para uso. Mas foi só em 2013 que o verdadeiro poder da *blockchain* iria ser reconhecido.

A proposta da plataforma *Ethereum* surgiu com Vitalik Buterin, um membro ativo da comunidade do próprio Bitcoin em 2013. Devido ao seu trabalho na comunidade ao longo dos anos, ele propôs uma nova plataforma baseada na tecnologia Blockchain, que pudesse fazer mais do que a plataforma do Bitcoin e que possuísse uma moeda própria (GERRING, 2016). A diferença está que a plataforma *Ethereum* foi projetada para ser mais acessível e flexível que o Bitcoin, podendo ser usada para criar contratos inteligentes (*Smart Contracts*) e também aplicativos descentralizados.

Foi formalmente anunciada em uma conferência sobre Bitcoin e teve seu sistema implementado com o apoio de Gavin Wood, co-fundador da plataforma. Todo o projeto foi financiado através da própria comunidade, utilizando Bitcoins, por entusiastas e programadores que acompanhavam a proposta do sistema. Estes Bitcoins eram trocados por sua própria moeda ou 'token', chamada *Ether*. O montante foi utilizado para o pagamento de todos os débitos legais e também para pagamento dos programadores originais envolvidos, pois muitos largaram empregos para trabalhar na plataforma, assegurando o andamento do projeto, mas sem ver qualquer garantia de retorno em troca.

A plataforma foi lançada 1 ano após o financiamento começar, em julho de 2015, atraindo os primeiros mineradores e também os primeiros utilizadores da plataforma. No final de 2015, em novembro, aconteceu a primeira conferência de desenvolvedores relacionada a plataforma. Não só estavam presentes desenvolvedores, curiosos, e empresários, como também compareceram representantes de grandes empresas, como Microsoft e IBM, provando o crescente interesse na plataforma (GERRING, 2016).

A plataforma Ethereum basicamente é um sistema de programação. É chamado de *Ethereum Virtual Machine* (Máquina Virtual Ethereum), e pode ser entendido como um sistema operacional semelhante ao Windows. Foi construído através de algumas linguagens de programação existentes, e é capaz de rodar algoritmos de várias complexidades. Isso permite a criação de aplicativos e de contratos inteligentes, que não dependem de alguém para administrá-los. Podem até mesmo ser criadas novas criptomoedas que usem o sistema ou plataformas inteiras que sigam algum outro propósito específico.

A plataforma foi responsável pelo lançamento de centenas de outras criptomoedas e projetos descentralizados nos últimos anos por meio de um novo mecanismo de captação de recursos chamado de 'Oferta Inicial de Moedas' ou (ICO) e o Ether, desta forma, é uti-

lizado para realizar operações dentro da *Ethereum Virtual Machine* (EVM), possibilitando uma compra de poder computacional dos usuários ao redor do mundo para que as movimentações possam ser realizadas (ANDONI M.; ROBU, 2017).

O artigo de (WOOD, 2014) foi publicado depois do artigo de (VITALIK, 2017) que apresentou a proposta inicial do Ethereum, e acrescentou conceitos importantes como cofundador, criador da linguagem Solidity e coordenador de tecnologia da fundação. Além disso, apresentou os conceitos de mudanças de estado e algoritmos relacionados à prova de trabalho e consenso, que são a base de todo o sistema da *Blockchain* do Ethereum.

O Ether é hoje a segunda criptomoeda de maior capitalização do mundo, ficando atrás apenas do Bitcoin [3].

[3] https://coinmarketcap.com/

2.3 Twitter

O Twitter foi lançado em 2006 como uma aplicação que mistura conceitos de rede social e *blogging*. REF(RUSSEL, Mathew A. Mining the social web: Data Mining Facebook, Twitter, LinkedIn,Google+, GitHub and More. 2 ed. Sebastopol: O'reilly Media, Inc., 2013.) cita a curiosidade humana e a necessidade de compartilhar idéias e experiências, fazer perguntas, interagir de maneira rápida. O Twitter propicia de maneira dinâmica que todos estes aspectos sejam possíveis.

Com um limite de 280 caracteres por postagem ou *tweets*, os usuários podem incorporar os símbolos " @ " e " # " em suas mensagenspara garantir que membros específicos recebam as informações que estão divulgando. As pessoas usam o símbolo de *hashtag* () antes de uma palavra-chave ou frase relevante nos *tweets* que publicam para classificá-los e facilitar a exibição deles na busca do Twitter, enquanto que o símbolo de arroba(@) é usado antes para se referir a uma conta específica no Twitter.

Desde seu lançamento, o Twitter cresceu rapidamente em popularidade. Um dos primeiros exemplos desta popularidade é o de Janeiro de 2009, onde um avião de uma companhia americana fez um pouco de emergência no rio Hudson. Uma imagem postada no Twitter saiu primeiramente nas notícias que as nas mídias tradicionais [4]. Twitter possui 330 milhões de usuários ativos por mês e 1.3 bilhões de contas já foram criadas, 83% dos líderes mundiais possuem uma conta no Twitter e 500 milhões de *tweets* são enviados por dia [5].

(BARBOSA, 2012) pontuam que o modelo de interação do Twitter induz os usuários a compartilhar e expressar continuamente suas opniões e sentimentos, que são propagados para seus respectivos seguidores, fazendo do Twitter uma excelente fonte para coletar informações de texto em tópicos. As mensagens também são mais fáceis de serem analisadas devido so limite de tamanho porque os autores são diretos ao ponto (FRANCO, 2014). Assim é mais fácil obter alta precisão na análise de sentimentos (LIU, 2012). O Twitter tornou uma mina de ouro para dados opinativos, pois podem ser usados em experimentos de análise de opinião e análise de sentimentos (PAK A.; PAROUBEK, 2010). Portanto o Twitter será importante para extração de *tweets* que envolvem o assunto Ethereum para análise de influência do valor da moeda.

[4] <https://www.brandwatch.com/blog/twitter-stats-and-statistics/>
[5] <http://www.iflscience.com/technology/how-much-data-does-the-world-generateevery-minute/>

2.4 Análise de sentimentos

Com o advento da internet e os avanços tecnológicos subsequentes, foram criados incontáveis meios para entreter a humanidade. A tentativa que se mostrou mais acertada foi, até o momento, a criação das redes sociais. A adesão em massa alcançada nas redes tem gerado, como consequência, grandes massas de dados (GANTZ, 2007).

Partindo do pressuposto que a maioria desses dados está disponível publicamente nas redes. Com consequência, grandes empresas, pesquisadores, entre outros, iniciaram a empreitada de buscar conhecimento, informação, *feedback*, por meio dessas interações publicadas por usuários de todo o mundo. A partir da necessidade de transformar esses dados em resultados surge a área de pesquisa, nomeado análise de sentimento.

Indurkhya e Demerau (INDURKHYA NITIN; DAMERAU, 2010) citam que as opniões são tão importantes que, onde quer que se queira tomar decisões, as pessoas desejam ouvir a opnião de outros. Por conta disso, existem pesquisadores que já se utilizaram da análise de sentimentos para mensurar sentimentos acerca de uma determinada ação do mercado financeiro, para assim prever sua valorização ou desvalorização e também do mercado de criptomoedas, como Abraham (ABRAHAM DANIEL HIGDON; IBARRA., 2018) e Stenqvist e Lönnö (STENQVIST; LöNNö.,).

O nome “análise de sentimentos” é de alguma forma, autoexplicativo, porém não obstante a isso é possível definir que: são técnicas cujo objetivo é extrair automaticamente informações subjetivas de textos escritos em linguagem natural (BENEVENUTO FABRíCIO,). É um campo de estudo onde ocorre a intersecção entre diversos outros campos de pesquisa, tais como processamento de computação linguística, aprendizado de máquina e mineração de dados (YUE, 2019).

A aplicação de técnicas de análise de sentimentos viabiliza a transformação de textos em efetivo conhecimento, e consequentemente, geram inteligência. Esse conceito é amplamente estudado em disciplinas como sistemas de informação, onde é desenvolvida a capacidade de diferenciar dados, informação e conhecimento. Tais definições podem ser descritas de N formas, porém neste trabalho foram usadas quatro: positivo, negativo, neutro e compound.

Inicialmente, o conceito de dados é definido da seguinte forma: São códigos que constituem a matéria prima da informação, ou seja, é o conteúdo que ainda não apresenta relevância. Os dados representam um ou mais significados de um sistema que isoladamente não podem transmitir uma mensagem ou representar algum conhecimento (VALENTIM, 2002). O resultado do processamento de dados são as informações (VALENTIM, 2002). As informações têm significado e podem contribuir no processo de tomada de decisões.

Por outro lado o conhecimento é o ato ou efeito de abstrair ideia ou noção de alguma coisa, como por exemplo: conhecimento das leis; conhecimento de um fato (obter informação); conhecimento de um documento; conhecimento da estrutura e função de determina-

dos sistemas. O saber, a instrução ou domínio científico estão relacionados com o conhecimento (SILVA, 2022).

Esta etapa, de análise de sentimentos, desperta o questionamento: “Qual é a importância de saber o que pensam as pessoas?” (“*What other people think*”) (PANG BO, 2008). Essa questão é bastante significativa e vai de encontro com as possibilidades geradas a partir da aplicação de técnicas de análise de sentimentos. Sendo assim, se faz necessário apontar exemplos, de possíveis aplicações da técnica, como: avaliar a receptividade do consumidor, perante produtos e serviços oferecidos (*feedback*), mapear áreas de risco, observar novas necessidades de mercado e finalmente, prever eventos significativos.

Assim como Lamon (LAMON; REDONDO.,) que extraiu sentimentos de notícias e redes sociais para prever preços de criptomoedas. Matta, Lunesu e Marchesi (MATTA M.,) extraíram sentimentos de pesquisas na internet e tweets para analisar se o aumento do preço do Bitcoin estava ligado a quantidade de menções positivas.

Por fim, Gomes (GOMES, 2013) pontua que apesar da análise de sentimentos ser apresentada por grande parte da literatura como estudo computacional de sentimentos, sua aplicação é muito ampla, pois se tratando de um problema de classificação, ela pode ser utilizada para classificar dados textuais mesmo se o texto não denotar algum sentimento.

2.5 VADER

VADER (Dicionário de Conhecimento de Valência e Raciocinador de Sentimentos) é uma ferramenta de análise de sentimentos feita especificamente para sentimentos expressos nas redes sociais, o mesmo funciona baseando-se em um dicionário em inglês que mapeia características léxicas que são associados às suas medidas de intensidade de emoção. As medidas de sentimento possuem uma escala de -4 (polaridade extremamente negativa)"a "4 (polaridade extremamente positiva)"sendo o "0"sua polaridade neutra.

Desenvolvido por Hutto (HUTTO C.; GILBERT, 2014) o mesmo é utilizado por alguns dos pesquisadores que já trabalharam em temas parecidos com o deste trabalho, como Stenqvist e Lönnö (STENQVIST; LöNNö.,). O dicionário léxico utilizado pelo VADER é um conjunto de palavras usadas para comunicação textual que foi construído a partir de dicionários bem formados, como o LIWC (TAUSCZIK Y. R.; PENNEBAKER, 2010), ANEW (BRADLEY, 1999) e GI (STONE P. J.; DUNPHY,). Ainda foram adicionadas siglas, gírias e emoticons que também expressam sentimentos.

Em sua análise de desempenho, o VADER foi comparado com onze outras ferramentas de análise semântica e obteve um desempenho consistente, ficando entre os melhores em todos os casos de teste e superou todas as outras técnicas no domínio de texto de redes sociais.

2.6 Aprendizado de Máquina para Problemas de Classificação

Existem diversas subáreas do campo conhecido como aprendizado de máquina. Esse campo é, de forma geral, dividido em três grandes áreas.

A primeira área é Aprendizado supervisionado corresponde ao problema de aprender a prever à qual classe uma determinada instância (representada por um conjunto de features) pertence. Esse tipo de método é conhecido como um método de classificação. Outro tipo de aprendizado supervisionado diz respeito a algoritmos de regressão. Nesses, o objetivo não é predizer uma classe, e sim um valor numérico contínuo. Em aprendizado supervisionado, existe um conjunto de treinamento com dados na forma: (x_i, y_i). Onde, $x_i \, \epsilon \, R^d$ é um vetor com *d features* descrevendo uma instância, e y_i é a sua classe correspondente. O objetivo é aprender um modelo de predição f : X $\rightarrow$ Y capaz de fazer predições corretas.

A segunda área da aprendizado de máquina é conhecida como aprendizado não supervisionado. Aqui, o objetivo é encontrar padrões nos dados, tais como grupos de instâncias similares de acordo com seus atributos. Por fim, existe também uma área conhecida como aprendizado por reforço, na qual o objetivo é construir um sistema de recompensas para um agente com punições e recompensas, de forma a permitir com que o agente aprenda automaticamente ações que sejam mais adequadas em cada contexto/estado. Neste trabalho, iremos focar em algoritmos de aprendizado supervisionado para classificação, visto que nosso objetivo é utilizar tais métodos para classificar tweets de acordo com o tipo de expressão política ou opinião expressa pelo seu criador.

Para resolução de problemas de aprendizado de máquina do tipo supervisionado ou não-supervisionado, precisamos ter à disposição um conjunto de dados de treinamento, os quais correspondem a um conjunto de instâncias descritas por um conjunto de features. Essas features podem ser contínuas, discretas ou binárias (KOTSIANTIS, 2017). Quando falamos de aprendizado por reforço (SUTTON R. S.; BARTO, 2018), temos um agente que tem como objetivo aprender por repetição qual sequência de ações gera a maior soma total de recompensas. Já quando falamos de aprendizado não-supervisionado (JAIN A. K.; MURTY, 1999), nos referimos a problemas em que as instâncias não estão sendo mapeadas para classes. O objetivo do algoritmo, ao invés disso, é identificar grupos de similaridade entre as instâncias. Neste trabalho, como previamente mencionado, iremos focar em técnicas de aprendizado supervisionado, nas quais temos um conjunto de dados com instâncias e suas respostas/classes esperadas, e onde o objetivo do algoritmo é generalizar os dados de treinamento para descobrir qual o padrão que produz respostas corretas para novas instâncias.

No caso específico de algoritmos de classificação, conforme mencionado anteriormente, queremos identificar à qual classe uma determinada instância pertence (KOTSIANTIS, 2017). Diversos algoritmos se propõem a resolver esse tipo de problema. Neste capítulo, ire-

mos focar nas técnicas de regressão logística e florestas aleatórias, visto que elas frequentemente consistem no estado-da-arte no que diz respeito a aplicações práticas/reais de aprendizado de máquina. Após discutir esses algoritmos (nas seções seguintes), iremos, nos capítulos que seguem, avaliar diferentes variantes desses algoritmos no contexto específico de nossa aplicação de interesse, a fim de identificar a maneira mais eficaz de construir um modelo de análise de sentimentos que nos permita efetuar as análises de interesse.

2.6.1 Regressão Logística

Regressão logística é um modelo estatístico que calcula a probabilidade de um evento ocorrer dadas informações de entrada descrevendo uma determinada situação ou contexto; por exemplo, prever a classe de um tweet dadas as palavras que o compõem (CRAMER, 2002).

Cada instância de treinamento é um par (x_i, y_i), onde x_i é um vetor representando as features da instância i de treinamento: $x_i = [x_1, x_2, ..., x_d]$, onde $d \epsilon R_d$ é o número de atributos da instância, e onde y_i é a saída desejada. O algoritmo aprende pesos w através de técnicas de gradiente descendente, a qual ajusta os pesos iterativamente de forma a otimizar uma função objetivo (tipicamente entropia cruzada) que, dados pesos w, retorna a performance da regressão logística—ou seja, a acurácia do algoritmo resultante do uso dos pesos w. A probabilidade estimada pelo modelo é calculada da seguinte forma. Primeiro, calculamos $t = w_0 + w_1 x_1 + w_2 x_2 + ... + w_d x_d$. O modelo de predição, então, se baseia no uso de uma função logística/sigmoide, a qual recebe como entrada t e produz como saída um valor entre 0 e 1, expressando a probabilidade da instancia $x = [x_1, ..., x_d]$ pertencer à classe positiva. A equação logística, responsável por prever a probabilidade da classe, é dada por:

$$\sigma(t) = 1/1 + e^t$$

No contexto de aprendizado de máquina, quando aplicado a um conjunto de dados, esse tipo modelo é capaz de prever a probabilidade de uma dada instância pertencer a uma determinada classe—ou até mesmo a várias delas.

2.6.2 Florestas Aleatórias

Um árvore de decisão é um algoritmo de aprendizado de máquina que recebe como entrada um conjunto de dados e realiza divisões nesses dados, com base em testes e comparações nos valores de suas features. As divisões são feitas de acordo com critérios que tentam identificar o menor número de testes até que a classe de uma dada instância possa ser determinada. O critério mais comum para divisão dos dados é o ganho de informação. O algoritmo de treinamento, a cada momento, determina qual o atributo que auxilia a forma máxima a reduzir a incerteza do modelo acerca da classe da instância.

Na Figura 2.1, apresentamos um exemplo de conjunto de dados no qual a classe (correspondendo à predição sobre se uma pessoa vai jogar tênis ou não) é feita com base em quatro atributos descrevendo o dia em questão. Esse conjunto de dados poderia ser fornecido como entrada para o treinamento de uma árvore de decisão.

Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

Figura 2.1: Exemplo de um conjunto de dados de entrada para uma árvore de decisão

Já na Figura 2.2 apresentamos uma possível árvore de decisão treinada sobre os dados da Figura 2.1.

Embora seja possível treinar uma única árvore de decisão, esse tipo de modelo, frequentemente, não generaliza bem quando apresentado a dados novos de teste. Nesse caso, podemos construir uma floresta de árvores. Uma floresta aleatória é um algoritmo que constrói suas predições através da combinação das predições de diferentes árvores de decisão (BREIMAN, 2001). A intuição que motiva esse modelo consiste na ideia de que é possível obter melhores predições se combinarmos árvores de decisão treinadas com base em diferentes subconjuntos dos dados de treinamento.

Florestas aleatórias têm uma vantagem importante em relação a outros modelos estatísticos, tais como redes neurais, visto que seus modelos são interpretáveis e, portanto, se torna mais fácil para usuárias do algoritmo entenderem por quê determinadas classificações estão sendo feitas. Além disso, florestas aleatórias são frequentemente o estado-da-arte em vários problemas reais desafiadores (GOLDSTEIN, 2010).

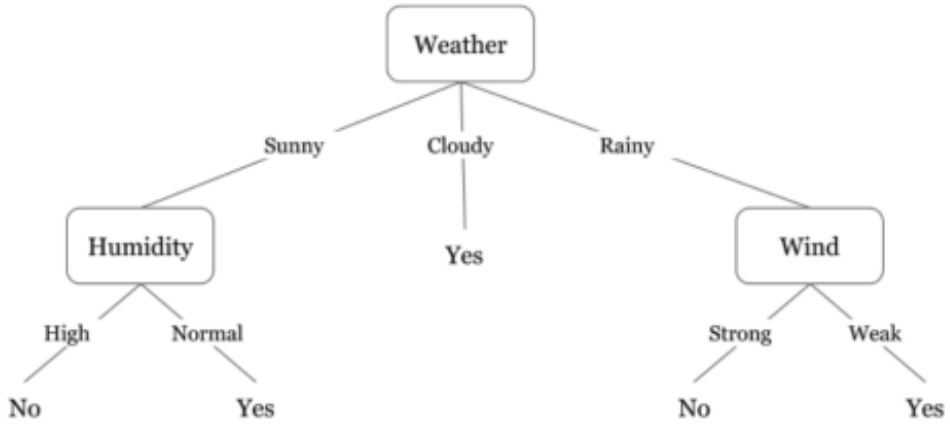

Figura 2.2: Exemplo de uma árvore de decisão.

3

Trabalhos Relacionados

Diversos pesquisadores estudaram a aplicação de análise de sentimentos com o intuito de prever opnião, o sentimento, ou a avaliação de um grupo de pessoas sobre um determinado tópico, assim como outros pesquisadores fizeram uso de aprendizagem de máquina para construir um preditor de eventos. Neste capítulo, iremos discutir alguns trabalhos relacionados ao nosso, isto é, que também fizeram uso das ferramentas e técnicas utilizadas neste trabalho.

3.0.1 Mineração de Textos

Tan (1999) e Gomes (2013) pontuam a relevância da aplicação de métodos de Mineração de Textos para a extração de conhecimento em base de dados textuais, uma vez que a Mineração de Dados é comumente aplicada em dados que possuem certo nível de estruturação e contemplam apenas uma parte limitada de dados que as organizações possuem, ou seja, dados estruturados (GOMES, 2013).

Gomes (2013) aplica a Mineração de Textos em seu trabalho na busca de extrair conhecimento acerca de notícias de economia de Portugal. Assim, o autor visita sítios de notícias sobre economia de seu país para representar o sentimento expresso nas notícias, analisando os textos dos títulos das notícias.

Rodrigues Babosa et al (2012) utilizam os processos da Mineração de Textos para explorar *tweets* que falam sobre as eleições presidenciais do Brasil do ano de 2010, afim de traçar o sentimento online da população expresso nos *tweets*, classificando-os em positivos, negativos, neutros ou ambíguos, e correlacionar a classificação dos tweets aos fatos que ocorriam no Brasil relacionados às eleições, como debates políticos, por exemplo (BARBOSA, 2012).

Neste trabalho, será aplicado o processo de Mineração de Textos semelhante ao trabalho de Gomes (2013), no entanto a aplicação será na rede social Twitter. A escolha de usar essa rede social é pelo alcance global da mesma, que possui milhões de usuários cadastrados. Ao

contrário de Gomes (2013), que restringe o alcance do estudo a apenas a um lugar. Os textos a serem minerados compõem um *tweet*, que é uma sequência de caracteres publicada pelos usuários, podendo conter outros tipos de dados anexados.

Este trabalho assemelha-se ao de Barbosa et al (2012), por utilizar *hashtags* para determinar o sentimento expresso em um tweet, correlacionando os resultados aos de oscilação do *Ether*. Entretanto, difere-se por considerar outras palavras do *tweet* que possam expressar um sentimento, mesmo que estas não estejam marcadas como uma *hashtag*.

Jain (JAIN, 2018) também extraiu e analisou *tweets* usando um modelo de regressão multi-linear para prever o preço futuro de duas criptomoedas. Sua metodologia extrai e analisa tweets que são filtradas com o nome das criptomoedas, assim como dados de preços simultâneos extraídos do *Coindesk* [1] e explora recursos significativos mapeados com os preços simultâneos das criptomoedas para criar curvas de previsão que podem prever os preços das criptomoedas em um futuro próximo.

Para a análise de sentimento, os autores utilizaram o TextBlob [2] que é uma biblioteca Python. Para a análise dos dados, utilizaram o modelo de regressão linear múltiplo. Os autores propuseram um *framework* que funciona em duas fases: fase de treinamento e fase de detecção. Basicamente, a frase de treinamento consiste em coletar os dados do Twitter, com relação às criptomoedas, convertê-los para o formato CSV e então, analisados pela polaridade. A quantidade de tweets positivos, negativos e neutros são contados e armazenados. Esses números contados são mapeados com a média do preços que ocorrem na duração correspondente de duas horas. O modelo então só é validado se o resutado é aceitável, pronto para ser usado para prever o futuro preço das criptomoedas. Caso contrário, um novo modelo é formado e o processo de treinamento e teste é repetido até que se obtenha um modelo aceitável. Na fase de detecção, *tweets* em tempo real são usados como entrada para o modelo e o modelo faz previsão do preço médio pela duração de duas horas.

Os resultados mostraram que o preço do Bitcoin não é muito afetado pelos sentimentos dos tweets em comparação ao preço do Litecoin. Os autores acreditam que a flutuação no preço do Bitcoin depende de outros fatores como custo de mineração e fator econômico.

Fung (FUNG, 2019) descreve em seu artigo estratégias de *trading* utilizando notícias e *tweets* relacionados a Bitcoin. Os autores usaram análise de sentimento para cada notícia e tweets utilizando VADER e alguns modelos de ajuste para prever o mercado Bitcoin.

Para a estratégia de *trading*, os autores escolheram três modelos a partir da observação e testes dos dados empíricos. O primeiro modelo (Modelo 1) foi regressão logística com base na análise de sentimentos das notícias, o segundo modelo (Modelo 2) foi regressão logística com base na análise de sentimentos dos *tweets* e o terceiro modelo (Modelo 3) foi uma regressão linear de série temporal sobre preço e volume de *tweets*.

A partir dos dados empíricos, os autores observaram que existe um viés positivo com

[1] Coindesk <https://www.coindesk.com/>

[2] TextBlob <https://textblob.readthedocs.io/en/dev/>

relação ao sentimento de notícias e *tweets*. Principalmente, devido à prevalência de criptomoedas nos últimos anos, a maioria das notícias relacionadas exerce uma visão otimista.

Assemelhando-se ao presente trabalho no modo de utilização da ferramenta de análise de dados e modelos de referência.

3.0.2 Análise de Sentimentos

Gomes (2013) ressalta que, com a popularização da Internet, as pessoas geram um imenso volume de dados a cada segundo. O desafio é saber como manipular essa grande quantidade de informação gerada e investigar como as organizações podem se beneficiar dessas informações, considerando que 80% das informações das organizações estão contidas em documentos de texto (TAN, 1999). Em seu trabalho, Gomes (2013) analisa títulos de notícias sobre economia extraídas de endereços feeds RSS. O autor propõe um modelo de Análise de Sentimentos que polariza as notícias em positivas, negativas ou neutras, além de fornecer um documento que guie as organizações a como procederem para extrair conhecimento de dados textuais.

No artigo proposto por Melo and Figueiredo (MELO T. DE; FIGUEIREDO, 2021), os autores têm como objetivo analisar a correlação entre os tópicos descritos, no que diz respeito à pandemia de COVID-19 no Brasil em publicações de usuários do Twitter e em publicações de mídias tradicionais de notícias. Nesse trabalho, os autores utilizaram técnicas de *Topic Modeling*, em particular, *Namely Entity Recognition* (MOHIT, 2014) e análise de sentimentos, através da ferramenta VADER (HUTTO C.; GILBERT, 2014).

Já no trabalho proposto por Yaqub (YAQUB, 2020), o autor analisa o sentimento do tweets publicados pelo ex-presidente dos Estados Unidos da América, Donald Trump, durante o início da pandemia COVID-19. Para realizar a análise de sentimentos desses tweets, o autor utilizou assim como no trabalho citado anteriormente—a ferramenta proposta por Hutto and Gilbert (2014) para desenvolver seu analisador automático de sentimentos.

No artigo proposto por Prastyo (PRASTYO, 2020), os autores analisam o sentimento de usuários do Twitter, através de tweets, em relação à forma como a pandemia COVID19 foi gerenciada pelo governo indonésio. Esse trabalho está relacionado ao presente trabalho no que se refere à análise de sentimentos de tweets relacionados à pandemia COVID-19 e à atuação de governos no gerenciamento da pandemia em seus respectivos países. Em contrapartida, difere-se em seus objetivos: enquanto nós queremos correlacionar tweets sobre o *ETH* à oscilação de valores , os autores daquele trabalho se propõem a analisar a opinião da população em relação à atuação do governo indonésio no contexto da pandemia. Além disso, os trabalhos se diferem nos métodos utilizados, os autores Prastyo et al. (2020) desenvolveram um analisador de sentimentos com as classes negativo e positivo e utilizando uma técnica diferente, baseada em *Support Vector Machine*.

Com a popularização da Internet, as pessoas geram um imenso volume de dados a cada

segundo, ressalta Gomes (GOMES, 2013). Saber como manipular esta quantidade de informação gerada e investigar como as organizações podem se beneficiar dessas informações é um desafio. Considerando que 80% das informações das organizações estão contidas em documentos de texto (TAN, 1999), Gomes (GOMES, 2013) analisa títulos de notícias sobre economia e propõe um modelo de Análise de Sentimentos que polariza as noticias em positivas, negativas e neutras.

Rodrigues (BARBOSA, 2012) pontua que o modelo de interação do Twitter induz os usuários a compartilharem e expressarem continuamente suas opniões e sentimentos, que são propagados para seus seguidores. Em seu trabalho, fez uso da classificação das *hashtags*, que são bastante utilizadas na plataforma do Twitter para associação de discussões na qual deseja indexar sua postagem. No presente trabalho também fizemos o uso das *hashtags*, utilizando-as para análise de sentimentos.

Abraham, Hidgon e Nelson (ABRAHAM J.; HIGDON, 2018), utilizando dados do Twitter e do Google Trends, apresentam um método de prever mudanças em preços do Bitcoin e Ethereum. A ferramenta VADER (*Valence Aware Dictionary and sEntiment Reasoner*) foi utilizada pelos autores, um método muito utilizado para o contexto de mídias sociais. Nesse caso, o método determina se os tweets geralmente são positivos ou negativos em suas opiniões sobre criptomoedas determinando se os tweets são positivos ou negativos em suas opniões sobre criptomoedas.

Para a análise dos dados, os autores fazem um detalhamento sobre quais dados são adequados como entradas do modelo de regressão linear múltipla. Isso inclui determinar quantos dos tweets possuem de fato algum sentimento e estabelecer uma relação entre o sentimento dos tweets sobre as criptomoedas e as mudanças nos preços das criptomoedas. Pois, se não contiver uma relação entre estas métricas e a mudança de preços, então não serão consideradas como entradas do modelo.

A descoberta foi que em um ambiente em que os preços estão caindo, a análise de sentimentos é menos eficaz para alterações de preços de criptomoedas, pois mesmo quando os preços caem, as pessoas que twittam sobre criptomoedas publicam tweets tendenciosos a positivos, pois possuem interesse nelas além da oportunidade de investimento. A correlação com o presente trabalho se dá através do uso de análise de sentimentos para prever mudanças nos preços de criptomoedas.

Valencia, Gómez e Valdés (VALENCIA F.; GÓMEZ-ESPINOSA, 2019) estudam o comportamento dos mercados por meio da aplicação de técnicas de análise de sentimentos e aprendizado de máquina para a tarefa de previsão do mercado de ações. Para isso, propõe usar ferramentas de aprendizado de máquina e dados disponíveis de mídia social para prever o preço das criptomoedas Bitcoin, Ethereum, Ripple e Litecoin.

Os autores ainda comparam a utilização de RN, SVM e RF ao utilizar elementos do Twitter e dados do mercado como características (*features*) de entradas. Para a análise de sentimentos, os autores utilizaram o VADER. Para o aprendizado de máquina, O perceptron multica-

madas (PMCs) foi utilizado, que é um tipo de RN. Para avaliar cada modelo foram utilizados as métricas de acurácia, precisão, *recall* e *f1-score*.

Os resultados mostraram que é possível prever o mercado de criptomoedas usando aprendizado de máquina e análise de sentimentos, onde os dados do Twitter por si só poderiam ser usados para prever certas criptomoedas e que a RN supera os outros modelos.

Neste trabalho, as opiniões dos usuários do Twitter sobre o *ETH* serão polarizadas utilizando a mesma categorização: positivas, negativas, neutras ou ambíguas através do VADER. Entretanto, a classificação será feita por meio de uma implementação do algoritmo de classificação *Random Forests* e Regressão Logística.

Os trabalhos comentados acima possuem em comum com o presente trabalho o uso da análise de sentimentos para construção de uma base de predição, utilizando e ressaltando a importância de dados de redes socias. Sendo de grande importância como referências para o trabalho proposto.

4

Procedimentos Metodológicos

O objetivo deste capítulo será detalhar toda a metodologia executada, iremos detalhar como foi feita a coleta, o pré-processamento, a rotulação, o uso das ferramentas, o treinamento dos modelos e o cálculo das métricas no contexto desde projeto. Nas seções a seguir estão descritos em detalhes a execução, desafios e aprendizados de cada etapa.

Nas seções a seguir estão descritos em detalhes a execução, desafios e aprendizados de cada etapa.

4.1 Coleta dos dados

Conseguir um amplo e qualificado conjunto de textos é uma tarefa difícil para aplicação em projetos de mineração de textos. É uma prática comum de grandes plataformas oferecerem API's de desenvolvimento, com o objetivo de facilitar a interação de desenvolvedores aos serviços ofertados e fomentar a criação de um ecossistema de serviços da própria plataforma. Como exemplos da utilização dessas API's temos a possibilidade de autenticação em um site utilizando a sua conta do Google ou Apple, criação de aplicações que utilizam dados financeiros através da API da Yahoo Finance ou busca de passagens aéreas através da API do SkyScanner. Como o projeto é baseado na mineração de *tweets*, buscamos utilizar a API do Twitter [1].

Apesar da API do Twitter oferecer planos gratuitos, o limite de *tweets* a serem coletados diariamente são de 3.200, volume no qual seria pouco representativo para os objetivos deste trabalho, entretanto, o volume pode ser aumentado através de uma assinatura mensal mas foi decidido encontrar outro método para obtenção deste maior volume de dados.

Durante a pesquisa de um novo método para obtenção dos dados foi encontrada a ferramenta chamada*Twint* [2]. O *Twint* é uma ferramenta avançada para obtenção de dados do

[1] Twitter Developer <https://developer.twitter.com/en>
[2] Twint Library <https://pypi.org/project/twint/>

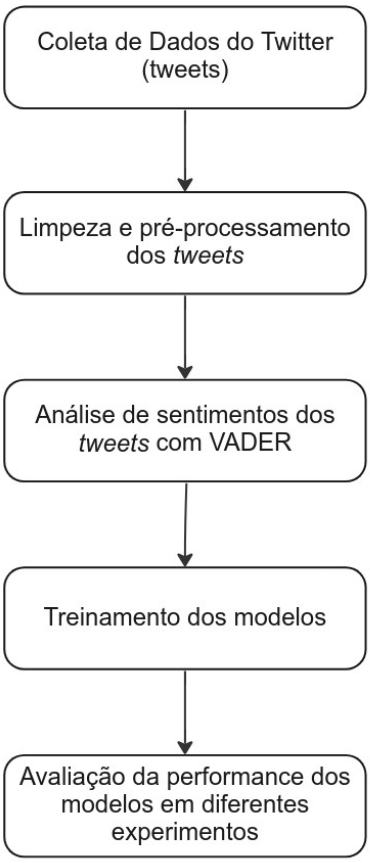

Figura 4.1: Fluxograma da metodologia aplicada no projeto

Twitter, escrita em Python, utiliza os operadores de busca do Twitter para buscar dados das mais variadas formas e dependendo do modo na qual será configurada pode buscar *tweets* de usuários e tópicos específicos, *hashtags*, *trends* ou separar informações sensíveis como números de telefone e e-mails.

O *Twint* possui os seguintes benefícios em relação ao Twitter:

i Sem restrições à quantidade de *tweets* a serem buscados.

ii Configuração rápida, sem necessidade de criação de conta e implantação de projeto na plataforma de desenvolvedor do Twitter.

iii Utilização de forma gratuita.

iv Oferece opções fáceis de transformação dos *tweets* em diferentes formatos - CSV, JSON, SQLite e Elasticsearch.

Sendo o Twint a ferramenta escolhida, o mesmo foi configurado para pesquisar *tweets* em inglês, entre as datas 06/07/2022 e 09/07/2022 e que contivessem 'ETH' no corpo da

mensagem, sendo 'ETH' o acrônimo da criptomoeda *Ethereum*. No total foram encontrados 34.558 *tweets* que foram armazenados em um JSON, contendo o *tweet*, a data e horário de postagem e o número identificador do mesmo.

Ao final desta etapa, após transformações, obtivemos um *dataframe* com 34.558 entradas e duas colunas, uma contendo a data de criação do *tweet* e a outra com o conteúdo com os tweets extraídos. A Figura 4.2 ilustra o *dataframe* criado.

	created_at	tweet
0	2022-07-09 11:59:59+00:00	💰 Sold 💰 🎶 Dream by @sofractures 🤝 Bought for ...
1	2022-07-09 11:59:59+00:00	@bertm You don't need to hurt me like this 😭 1...
2	2022-07-09 11:59:55+00:00	#XVSUSDT Bear Alert! 5X Volume Price: 5....
3	2022-07-09 11:59:54+00:00	@wasabi_eth Let's go
4	2022-07-09 11:59:53+00:00	👀✨✨✨ https://t.co/7POFYC5ygH
...	...	...
34553	2022-07-09 01:35:32+00:00	@NFTonyP @dgtlemissions @NateDigital @timbo_et...
34554	2022-07-09 01:35:32+00:00	@melfarra123 Yeahhh $ETH Amazing
34555	2022-07-09 01:35:31+00:00	Merch? Utility? What what??? https://t.co/l...
34556	2022-07-09 01:35:29+00:00	#PublicMint #NFTs https://t.co/MYeBvCW3nn
34557	2022-07-09 01:35:28+00:00	Come chillll #GodHatesNFTees https://t.co/LX8...

34558 rows × 2 columns

Figura 4.2: Exemplo dos *tweets* extraídos, sem pré-processamento.

4.2 Pré-processamento e Limpeza

Nesta etapa, os *tweets* passaram por um pré-processamento, onde foram descartados conteúdos considerados irrelevantes para o processo de classificação de texto. Dentre as diversas técnicas de pré-processamento de texto existentes na literatura, foram aplicadas no trabalho:

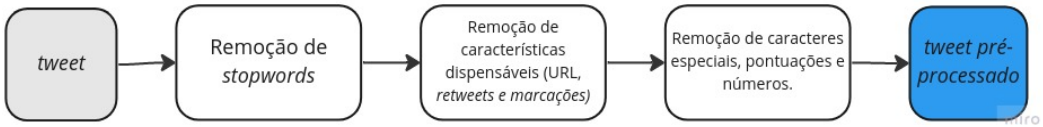

Figura 4.3: Etapas do pré-processamento do *tweet*

Remoção de Stopwords: *Stopwords* (palavras vazias) são importantes para o entendimento de frases, dão sentido às mesmas e fazem ligações léxicas essenciais, mas computacionalmente as mesmas ocupam muito espaço em memória, aumentando o tempo de processamento. Ao remover estas palavras contribuímos para a construção do *corpus*, conjunto estruturado de termos. O processo de remoção foi realizado na língua inglesa e em cada *tweet*.

Remoção de URL's: As URL's não possuem um papel importante na identificação do sentimento do *tweet*, não contribuindo na execução do algoritmo de polaridade as mesmas tornam-se desnecessárias para os próximos passos. É importante citar que durante a remoção é necessário atentar no formato das URL's, sendo necessário remover as que iniciam com 'http' e 'https'.

Remoção de *Retweets*: *Retweets* são marcações aplicadas nos *tweets* para identificar que aquele *tweet* originalmente foi postado por outro usuário, a marcação possui o formato 'RT @', seguido do identificador do usuário que postou o *tweet*. Como esta marcação não representa uma informação importante para execução do algoritmo de polaridade, todos os *retweets*

Remoção de Marcações: Marcações no *tweet* possuem o formato '@ + identificador do usuário' e assim como URL's e *retweets* não adiciona informações para o processamento das polaridades, sendo interessante sua remoção, que foi realizada em nossa base.

Remoção de Caracteres Especiais, Números e Pontuações: Caracteres especiais, pontuações e números, assim como as outras características acima removidas, não são informações necessárias para que o algoritmo de polaridade consiga obter uma melhor classificação, portanto estas características também foram removidas de todos os *tweets* da base.

Ao fim do pré-processamento e limpeza, obtivemos um *Dataframe* com todos os *tweets* já pré-processados. A Figura 4.4 ilustra o *dataframe* obtido após a limpeza.

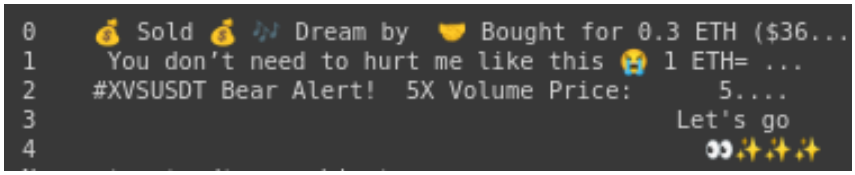

Figura 4.4: Exemplo dos *tweets* extraídos, após pré-processamento.

4.3 Análise de sentimentos

A escolha do classificador mais adequado para a tarefa de análise de sentimentos nos *tweets* é de suma importância para este projeto, pois é a partir da análise dos sentimentos que iremos obter o insumo para testar nossas hipóteses. A ferramenta VADER [3],de código aberto, foi utilizada em linguagem Python e com uma fácil instalação e documentação bem detalhada, foi simples seguir o processo de configuração.

O *score* calculado retorna uma pontuação no intervalo de -1 a 1, do mais negativo ao mais positivo. O *score* de uma frase é calculado somando-se as pontuações de cada palavra listada no dicionário VADER, este sendo especialmente útil para análise de textos provenientes da internet, como por exemplo contendo expressões utilizadas online e também avaliando emojis.

Palavras individuais têm um *score* entre -4 e 4, entretanto os criadores do VADER aplicaram uma normalização ao total para mapeá-lo para um valor entre -1 e 1. Os valores que os autores utilizaram foram:

Sentimento positivo: $score \geq 0.05$;

Sentimento neutro: $score > -0.05$ e $score < 0.05$;

Sentimento negativo: $score \leq -0.05$;

Com a aplicação do algoritmo em nossa base, adicionamos mais 4 colunas ao *dataframe* dos *tweets*, sendo elas: positivo(pos), negativo(neg), neutro(neu) e *score* (compound). A Figura 4.6 ilustra os valores da análise de sentimentos consolidados por hora.

(BOLLEN J., 2011) enfatizam a importância das mídias sociais e do sentimento de notícias na previsão dos preços de negociações de criptoativos em curtos períodos, sejam hora em hora até diariamente. Assim como (WIJHE, 2016), iremos realizar transformações nos dados para obtermos um sentimento geral de hora em hora dos *tweets*, a média dos *scores* dos *tweets* dentro de cada hora servirá para demonstrar a o sentimento de hora em hora.

[3]VADER - Sentiment Analysis <https://github.com/cjhutto/vaderSentiment>

	created_at	Compound	Positive	Negative	Neutral
0	2022-07-09 01:00:00+00:00	0.123474	0.105749	0.029243	0.678053
1	2022-07-09 02:00:00+00:00	0.123346	0.103559	0.029620	0.662350
2	2022-07-09 03:00:00+00:00	0.127591	0.099261	0.027894	0.669419
3	2022-07-09 04:00:00+00:00	0.133839	0.103883	0.028456	0.675964
4	2022-07-09 05:00:00+00:00	0.133420	0.100493	0.036541	0.677944
5	2022-07-09 06:00:00+00:00	0.140770	0.104860	0.027626	0.690620
6	2022-07-09 07:00:00+00:00	0.175076	0.129879	0.031087	0.699205
7	2022-07-09 08:00:00+00:00	0.151581	0.109698	0.029139	0.689929
8	2022-07-09 09:00:00+00:00	0.143436	0.106776	0.028270	0.685252
9	2022-07-09 10:00:00+00:00	0.161756	0.116108	0.028007	0.672831
10	2022-07-09 11:00:00+00:00	0.153436	0.111679	0.030873	0.704346

Figura 4.5: Exemplo dos sentimentos calculados consolidados por hora.

4.4 Dados históricos do ETH

As informações históricas dos valores diários de abertura, fechamento, volumee qualquer outra informação diária do ETH são de suma importância para o projeto, pois são com estas informações que iremos alimentar nosso preditor e um dos desafios foi encontrá-las de forma consolidada. Iniciamos uma busca de Datasets confiáveis na plataforma Kaggle [4], um dos maiores portais de sobre *datascience* do mundo e encontramos *datasets* com informações antigas de cerca de cinco anos, muito desatualizadas para o objetivo deste projeto. Outros *datasets* mais novos possuíam lacunas de informação no histórico ou estavam consolidados mensalmente, como o objetivo era a busca de dados consolidados de hora em hora, não encontramos *datasets* que satisfizessem nossa necessidade.

A estratégia a seguir foi buscar as informações na fonte, *exchanges* de criptoativos, mas logo dado o início às buscas vimos que poucas possuíam API para consulta e as que continham, uma taxa de utilização era cobrada para obtenção dos dados.

Por fim, encontramos o site *Criptodata Dowload*, especializado em reunir informações históricas dos mercados de criptoativos consolidadas em diferentes espaços de tempo e de diferentes exchanges, disponibilizando-as gratuitamente através de vários formatos para *download*. Fizemos o *download* em csv dos dados provenientes de uma das mais conhecidas *exchanges* de criptoativos do mundo, a Binance [5], uma gigante asiática. As informações do ETH foram obtidas de hora em hora e importadas em um *dataframe* para melhor manipulação.

Na Figura 4.6 podemos conferir um exemplo das informações obtidas do ETH.

[4] Kaggle.com <https://www.kaggle.com/>
[5] Binance <https://www.binance.com/en>

	date	symbol	open	high	low	close	Volume ETH	Volume USDT	tradecount
0	2022-07-09 11:00:00	ETH/USDT	1224.75	1226.43	1205.27	1210.28	31844.3187	3.871337e+07	29255
1	2022-07-09 10:00:00	ETH/USDT	1219.93	1225.25	1216.65	1224.76	15617.8301	1.908562e+07	17174
2	2022-07-09 09:00:00	ETH/USDT	1227.20	1227.20	1216.38	1219.92	16882.4966	2.062530e+07	18590
3	2022-07-09 08:00:00	ETH/USDT	1225.01	1228.25	1218.47	1227.20	16628.2669	2.035268e+07	19237
4	2022-07-09 07:00:00	ETH/USDT	1217.78	1227.40	1217.48	1225.02	24371.0713	2.980203e+07	23825

Figura 4.6: Dataframe com informações estatísticas diárias do ETH.

4.5 Treinamento do modelo de previsão

Para treinamento dos modelos, iremos utilizar as informações consolidadas de hora em hora obtidas da análise de sentimentos junto aos dados históricos do ETH, entretanto, foram utilizadas as seguintes colunas:

- Valor de abertura
- Valor da maior alta
- Valor da maior baixa
- Valor de fechamento

Iniciamos por consolidar em um só *dataframe* os dados de *score* da análise de sentimentos e o histórico de desempenho do ETH de hora em hora, assim como resultante um *dataframe* pronto para ser utilizado em nosso modelo escolhido, um exemplo pode ser visto na figura 4.7.

	compound	open	high	low	close
compound	1.000000	0.531806	0.811483	0.575893	0.624932
open	0.531806	1.000000	0.834355	0.336096	0.332523
high	0.811483	0.834355	1.000000	0.590405	0.511478
low	0.575893	0.336096	0.590405	1.000000	0.845480
close	0.624932	0.332523	0.511478	0.845480	1.000000

Figura 4.7: Dataframe com as estatísticas e sentimento consolidado por hora

O procedimento para criação do modelo de previsão seguiu o protocolo fundamental de aprendizado de máquina, o conjunto de dados foi dividido em duas seções, 70% para treinamento e 30% para teste. A validação cruzada não foi incluída no modelo, no entanto, explorar mais seu uso em trabalhos futuros pode determinar se um modelo mais generalizado é produzido.

Os modelos de regressão linear múltipla e de *random forests* foram selecionados para modelagem do algoritmo, escolha devido a fortes métricas de correlação. Cada modelo será treinado de duas formas, uma com o input do *score* da análise dos sentimentos dos *tweets* e outro sem esta informação.

Os modelos utilizaram as informações do *dataframe* construído para predição do valor de fechamento de cada hora. Após o treinamento a parcela separada para teste foi utilizada para avaliação e o cálculo da MSE foi realizado para comparação entre os resultados obtidos.

5

Resultados experimentais

5.1 Resultados

No decorrer deste capítulo são apresentados os resultados por intermédio de experimentos. Note que, aqui, utilizamos o termo hipótese no sentido de que são possibilidades que gostaríamos de verificar empiricamente se podem ser confirmadas ou não de acordo com as predições feitas pelos nossos modelos. Por exemplo, é possível confirmar empiricamente que a análise de sentimentos é uma ferramenta útil para prever oscilações de preço do Ether? Essa é uma hipótese que será avaliada. Note que estamos utilizando o termo hipótese de forma diferente do uso do termo hipótese no contexto estatístico.

5.1.1 É possível realizar análise de sentimentos de *tweets* relacionados ao Ether?

Nesta seção, iremos apresentar dados relacionados à análise de sentimentos, de hora em hora, relacionados ao Ether.

Durante o procedimento de obtenção dos dados, consguimos reunir 34.558 *tweets* para análise. Na figura 5.1, podemos conferir um exemplo dos *tweets* antes de passarem por qualquer tipo de transformação. É perceptível que os *tweets* possuem características distintas de uma comunicação escrita realizada através da internet, com símbolos, emojis e acrônimos que constantemente referenciam o mercado cripto. Então, foi necessário realizar um pré-processamento cauteloso para não remover quaisquer referências ao Ether feita pelos usuários.

Para execução do algoritmo de análise de sentimentos possuir maior fidelidade ao sentimento expressado, foi necessário remover traços nos quais não seriam computados pelo mesmo. Executamos a limpeza removendo quaisquer marcações a outros perfis, espaços em branco, caracteres especiais e *retweets*, como pode ser observado na figura 5.2.

	created_at	tweet
0	2022-07-09 11:59:59+00:00	💰 Sold 💰 🎶 Dream by @sofractures 🤝 Bought for ...
1	2022-07-09 11:59:59+00:00	@bertm You don't need to hurt me like this 😭 1...
2	2022-07-09 11:59:55+00:00	#XVSUSDT Bear Alert! 5X Volume Price: 5....
3	2022-07-09 11:59:54+00:00	@wasabi_eth Let's go
4	2022-07-09 11:59:53+00:00	👀✨✨✨ https://t.co/7POFYC5ygH
...	...	...
34553	2022-07-09 01:35:32+00:00	@NFTonyP @dgtlemissions @NateDigital @timbo_et...
34554	2022-07-09 01:35:32+00:00	@melfarra123 Yeahhh $ETH Amazing
34555	2022-07-09 01:35:31+00:00	Merch? Utility? What what??? https://t.co/I...
34556	2022-07-09 01:35:29+00:00	#PublicMint #NFTs https://t.co/MYeBvCW3nn
34557	2022-07-09 01:35:28+00:00	Come chillll #GodHatesNFTees https://t.co/LX8...

34558 rows × 2 columns

Figura 5.1: Exemplo dos *tweets* extraídos, sem pré-processamento.

	tweet
0	💰 Sold 💰 🎶 Dream by 🤝 Bought for 0.3 ETH ($36...
1	You don't need to hurt me like this 😭 1 ETH= ...
2	#XVSUSDT Bear Alert! 5X Volume Price: 5....
3	Let's go
4	👀✨✨✨

Figura 5.2: Exemplo dos *tweets* extraídos, após pré-processamento.

O VADER, algoritmo utilizado para análise dos tweets, mostrou-se adequado para aplicação neste projeto. O mesmo possui como especialidade a análise de textos de redes sociais, conseguindo extrair informações de gírias e emojis utilizados. Após sua execução, conseguimos as informações de sentimento de cada tweet quantificadas, como mostra a figura 5.3. Os resultados da análise de sentimentos indicaram que, no geral, a opnião pública sobre o Ether, em sua maior parte, permanece neutra no espaço de tempo analisado. Os valores variam entre -0.61 e 0.70, significando que a cada hora o sentimento médio oscila entre negativo e positivo, a questão central, porém é se as pontuações de sentimento têm uma qualidade preditiva na previsão do flutuações de preços.

tweet	Compound	Positive	Negative	Neutral
💰 Sold 💰 🎶 Dream by 🤝 Bought for 0.3 ETH ($36...	0.2500	0.125	0.00	0.875
You don't need to hurt me like this 😭 1 ETH= ...	-0.6124	0.119	0.31	0.571
#XVSUSDT Bear Alert! 5X Volume Price: 5....	0.3595	0.128	0.00	0.872
Let's go	0.0000	0.000	0.00	1.000
👀✨✨✨	0.7096	0.873	0.00	0.127

Figura 5.3: Exemplo dos *tweets* extraídos, após análise de sentimentos.

Portanto é possível obter evidências de que é possível aplicar a análise de sentimentos

de *tweets* relacionados à criptomoeda Ether.

5.1.2 A análise de sentimentos é uma ferramenta útil para prever oscilações de preço do Ether?

No total dois modelos foram utilizados para prever oscilações de preço da criptomoeda Ether. Para metrificar fizemos o uso do cálculo da *Mean Square Error* ou MSE.

MSE é o valor médio da diferença quadrada entre a previsão e o valor real como pode ser visto na figura 5.4, seu resultado pode ser interpretado como quão bom o modelo atua na previsão dos valores do conjunto de teste.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$$

Figura 5.4: Formulação da MSE

Os dois modelos seguiram seguiram o método de avalição Hold out, o conjunto de dados foi dividido em duas partes, 70% para treinamento e 30% para teste.

O modelo treinado sem a análise de sentimentos não possui a coluna com os *scores* de sentimento de cada hora, portanto a predição ocorreu somente sobre os valores históricos do Ether descritos no capítulo anterior. Já os modelos treinados com a análise de sentimento, possuíam a coluna com os *scores* de sentimento.

O resultado do cálculo da MSE para os dois modelos e os quatro experimentos pode ser visto abaixo:

Modelo	MSE sem A.Sentimento	com A. Sentimento
Multiple Linear Regression	113.83	114.83
Random Forest	2.5	3

Pelo resultado da MSE dos modelos, é possível perceber que o modelo de regressão linear múltipla é significantemente pior em predizer os valores que o modelo de *randon forest*, dado o alto valor da MSE.

Comparando os modelos com o uso ou não da análise de sentimentos, podemos perceber que quando se dá o uso do *input* de sentimentos, o resultado da MSE tende a aumentar, sendo possível obter evidências que a análise de sentimentos, empiricamente, não foi útil para prever as oscilações de preço nos experimentos conduzidos neste projeto.

6

Conclusão

Devido à evolução da internet e à popularização das redes sociais, que disponibilizaram de maneira pública, constante e imediata a opnião de milhões de usuários, entende-se que é possível coletar opniões de usuários em tempo real sobre praticamente quaisquer tópicos. Nesse sentido, diversos autores relatam os benefícios potenciais do uso de análise automática de sentimento (FELDMAN, 2013).

No entanto, há desafios relacionados à coleta e à limpeza de tais dados e análises. Existem, por exemplo, limitações relacionadas à filtragem de informações que representam opniões dos usuários de propagandas que na maioria das vezes são publicadas com uma alta frequência por *bots* nas redes sociais.

Este projeto estudou a aplicação de técnicas de análise de sentimentos para verificar se é possível levantar dados significantes dos *tweets* relacionados ao Ether, e também se existe relação entre os sentimentos dos *tweets* e a oscilação de preço. A principal motivação se deve ao fato do Ether ter se popularizado nos últimos anos, como também pela grande quantidade de informações e opniões disponibilizadas pela rede social *Twitter*.

Nesse sentido, pudemos observar as capacidade e as limitações que técnicas de análise automática de sentimentos possuem e também sua aplicação como *input* em treinamento de modelos de predição. Em particular, observamos que não é possível encontrar tais correlações com as oscilações de valores do Ether quando analisamos a ocorrência de variações temporais de *tweets* negativos e positivos. Por outro lado, pudemos também observar que é possível utilizar ferramentas de análise de sentimentos para gerar informações importantes sobre o Ether.

Como evolução do trabalho, entendemos que seria interessante a análise de sentimentos aplicada em espaços de tempo maiores, assim como o uso de modelos de predição diferentes. Para obter resultados mais conclusivos, é importante usar métricas de avaliação mais robustas acompanhadas do uso da técnica de *K-Fold Cross Validation* para fornecimento

de uma média e desvio padrão. Em decorrência do desenvolvimento deste trabalho, durante a elaboração do projeto, foram observadas diversas possibilidades de aplicação das técnicas estudadas na obtenção de resultados em aplicações distintas das trabalhadas anteriormente. Análise de qualidade de serviços prestados, correlação da opnião pública com o aumento do faturamento da venda de um produto, *feedback* de consumidores nas mais diversas áreas, são trabalhos promissores e não possuem muitas divergências em relação ao projeto atual.

Referências Bibliográficas

ABRAHAM DANIEL HIGDON, J. N. J.; IBARRA., J. Sentiment analysis and opinion mining. *Cryptocurrency price prediction using tweet volumes and sentiment analysis.*, 2018.

ABRAHAM J.; HIGDON, D. N. J. Cryptocurrency price prediction using tweet volumes and sentiment analysis. *Data Science Review*, 2018.

ANDONI M.; ROBU, V. F. D. Crypto-control your own energy supply. *Nature*, 2017.

ANTONOPOULOS, A. M. Mastering bitcoin: Unlocking digital crypto-currencies. *O'Reilly Media*, 2014.

BARBOSA, G. A. e. a. R. Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment. *ACM ANNUAL CONFERENCE EXTENDED ABSTRACTS ON HUMAN FACTORS IN COMPUTING SYSTEMS EXTENDED ABSTRACTS*, 2012.

BENEVENUTO FABRíCIO, F. R. e. M. A. Métodos para análise de sentimentos em mídias sociais. *Brazilian Symposium on Multimedia and the Web (Webmedia)*.

BOLLEN J., M. H. . Z. X. Twitter mood predicts the stock market. *Journal of computational science*, v. 2, 2011.

BRADLEY, M. M. e. a. Affective norms for english words (anew): Instruction manual and affective ratings. 1999.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 10, 2001.

BöHME, R. et al. Bitcoin: Economics, technology, and governance. *Journal of Economic Perspectives*, v. 29, n. 2, p. 213–38, May 2015. Disponível em: <https://www.aeaweb.org/articles?id=10.1257/jep.29.2.213>.

CABRAL, R. Tudo sobre o bitcoin: a história, os usos e a política por trás da moeda forte digital. 2013.

CARVALHO FRANCISCO PRANCACIO ARAúJO, J. B. L. J. M. V. Reflexões econômicas: dinheiro, economia e sociedade. *Informe Econômico*, v. 31, n. 1, p. 45, 2014.

CHOKUN, J. *Who accepts bitcoins as payment. List of Companies, Stores, Shops," https://99bitcoins.com/who-accepts-bitcoins-payment-companiesstores-take-bitcoins/*. 2013.

CHUEN D. L. K.; GUO, L. W. Y. Cryptocurrency: A new investment opportunity? 2017.

CLARINDO JOãO PAULO, F. C. e. A. L. F. Detecção de casos de violência patrimonial a partir do twitter.

CRAMER, J. The origins of logistic regression. *Tinbergen Institute Discussion Papers*, 2002.

CROSBY MICHAEL, P. P. S. V. V. K. e. a. Blockchain technology: Beyond bitcoin. *Applied Innovation*, p. 6–10, 2016.

FELDMAN, R. Techniques and applications for sentiment analysis. *Commun. ACM, Association for Computing Machinery*, v. 56, n. 4, p. 82–89, April 2013.

FIGUEREDO IGLESON F, L. B. M. e. L. A. S. Investigando a influência de tweets em programas de votação popular no brasil.

FRANCO, 2017. <https://g1.globo.com/mundo/noticia/venezuelanos-investem-em-bitcoin-para-encarar-desemprego-hiperinflacao-e-falta-de-notas.ghtml>. Accessed: 2022-11-30.

FRANCO, P. Understanding bitcoin: Cryptography, engineering and economics. *Wiley*, 2014.

FRONI A. A.; MEULEN, R. V. D. G. Identifies three megatrends that will drive digital business into the next decade.

FUNG, P. M. Y. e. a. Bitcoin trading strategies using news and tweets with sentiment analysis. 2019.

GANTZ, J. e. D. R. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView*, 2007.

GOLDSTEIN, B. A. e. a. An application of random forests to a genome-wide association dataset: methodological considerations new findings. *BMC genetics*, v. 11, n. 1, p. 1–13, 2010.

GOMES, H. J. C. Text mining: análise de sentimentos na classificação de notícias. *Information Systems and Technologies (CISTI)*, p. 82–89, April 2013.

GRIFFITH, K. A quick history of cryptocurrencies bbtc – before bitcoin. 2018. Disponível em: <https://bitcoinmagazine.com/articles/quick-history-cryptocurrenciesbbtc-bitcoin-1397682630/>>.

HUTTO C.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 2014.

INDURKHYA NITIN; DAMERAU, F. J. Handbook of natural language processing. *CRC Press*, n. 2, 2010.

JAIN, A. e. a. Forecasting price of cryptocurrencies using tweets sentiment analysis. *International Conference on Contemporary Computing*, 2018.

JAIN A. K.; MURTY, M. N. F. P. J. Data clustering: A review. *ACM Comput. Surv*, 1999.

KAMINSKI, J. *Nowcasting the Bitcoin Market with Twitter Signals*. 2014.

KOTSIANTIS, S. B. Supervised machine learning: A review of classification techniques. *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies., 2017.

KRISTOUFEK, L. What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis. *PLoS ONE*, v. 10, n. 4, p. 1–15, 2015.

KROLL, J. A.; DAVEY, I. C.; FELTEN, E. W. The economics of bitcoin mining, or bitcoin in the presence of adversaries. In: . [S.l.: s.n.], 2013.

KWAK, H. et al. What is twitter, a social network or a news media? In: *Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA: Association for Computing Machinery, 2010. (WWW '10), p. 591–600. ISBN 9781605587998. Disponível em: <https://doi.org/10.1145/1772690.1772751>.

LAMON, E. N. C.; REDONDO., E. Cryptocurrency price prediction using news and social media sentiment. *Morgan Claypool Publishers*.

LIU, B. Sentiment analysis and opinion mining. *Morgan Claypool Publishers*, 2012.

MAI, F. et al. How does social media impact bitcoin value? a test of the silent majority hypothesis. *Journal of Management Information Systems*, Routledge, v. 35, n. 1, p. 19–52, 2018. Disponível em: <https://doi.org/10.1080/07421222.2018.1440774>.

MATTA M., L. I. . M. M. Bitcoin spread prediction using social and web search media. *Umap workshops*.

MELO T. DE; FIGUEIREDO, C. M. S. Comparing news articles and tweets about covid19 in brazil: Sentiment analysis and topic modeling approach. *JMIR Public Health Surveill*, 2021.

MOHIT, B. Named entity recognition. in: . natural language processing of semitic languages. *Heidelberg: Springer Berlin Heidelberg*, 2014.

NAKAMOTO, S. *Bitcoin: A peer-to-peer electronic cash system," http://bitcoin.org/bitcoin.pdf*. 2008.

ORRELL, D.; CHLUPATý, R. *The Evolution of Money*. Columbia University Press, 2016. Disponível em: <http://www.jstor.org/stable/10.7312/orre17372>.

PAK A.; PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. in: Proceedings. *LREC*, 2010.

PANG BO, L. L. e. a. Opinion mining and sentiment analysis. *Foundations and Trends*, p. 1–135, 2008.

PRASTYO, P. H. Tweets responding to the indonesian government's handling of covid-19: Sentiment analysis using svm with normalized poly kernel. *Journal of Information Systems Engineering and Business Intelligence*, 2020.

REUTERS. *Ethereum, the second-biggest cryptocurrency," https://www.euronews.com/next/2021/05/10/ethereum-the-second-biggest-cryptocurrency-soars-above-4-000-to-hit-a-new-record-high*. 2021.

SHAH, D.; ZHANG, K. Bayesian regression and bitcoin. *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, Sep 2014. Disponível em: <http://dx.doi.org/10.1109/ALLERTON.2014.7028484>.

SILVA, H. M. d. Sociedade da informação. 2022. Disponível em: <http://www.profcordella.com.br/unisanta/textos/tgs21_dados_info_conhec.htm>.

STENQVIST, E.; LöNNö., J. Predicting bitcoin price fluctuation with twitter sentiment analysis.

STONE P. J.; DUNPHY, D. C. S. M. S. The general inquirer: A computer approach to content analysis. *MIT press*.

SUTTON R. S.; BARTO, A. G. Reinforcement learning: An introduction. *Bradford Book*, 2018.

TAN, A.-H. Text mining: The state of the art and the challenges. *WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES*, 1999.

TAPSCOTT D.; TAPSCOTT, A. Blockchain revolution: How the technology behind bitcoin is changing money. *Sage Publications*, 2016.

TAUSCZIK Y. R.; PENNEBAKER, J. W. The psychological meaning of words: Liwc andcomputerized text analysis methods. *Journal of Language and Social Psychology*, 2010.

TAVARES, R. e. G. P. G. Classificação de filmes: uma abordagem utilizando o liwc. *6o Brazilian Workshop on Social Network Analysis and Mining(BraSNAM 2017)*, 2017.

TOMAéL MARIA INêS, A. R. A. e. I. G. D. C. Das redes sociais à inovação. *Ciência da informação*, v. 34, n. 2, 2005.

VALENCIA F.; GÓMEZ-ESPINOSA, A. V.-A. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy*, 2019.

VALENTIM, M. L. P. e. a. Inteligência competitiva em organizações: dado, informação e conhecimento. *DataGramaZero*, p. 1–13, 2002.

VITALIK, B. A next generation smart contract decentralized application platform. *Ethereum White Papper*, 2017.

VRIES LISETTE, S. G. e. P. S. L. D. Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of interactive marketing*, v. 26, n. 2, p. 83–91, 2012.

WALPORT, M. Distributed ledger technology: Beyond block chain. *Government Office for Science*, 2015.

WASSERMAN, S. Advances in social network analysis: Research in the social and behavioral sciences. *Sage*, v. 2, 1994.

WIJHE, S. V. Using sentiment analysis on twitter to predict the price fluctuations of solana. *Sage*, 2016.

WITTE, J. The blockchain: a gentle four page introduction. *Record Currency Management*, 2016.

WOOD, G. Ethereum: a secure decentralised generalised transaction ledger. *Ethereum Project Yellow Paper*, 2014.

WOŁK, K. Advanced social media sentiment analysis for short-term cryptocurrency price prediction. *Expert Systems*, v. 37, n. 2, p. e12493, 2020. E12493 EXSY-Apr-19-215.R1. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12493>.

YAQUB, U. Tweeting during the covid-19 pandemic: Sentiment analysis of twitter messages by president trump. *Digit. Gov.: Res. Pract.*, 2020.

YUE, L. e. a. A survey of sentiment analysis in social media. *Knowl. Inf. Syst., SpringerVerlag*, 2019.