Revista Brasileira de Computação Aplicada, Abril, 2020

DOI: 10.5335/rbca.v12i1.8831
Vol. 12, Nº 1, pp. 1–15
Homepage: seer.upf.br/index.php/rbca/index

ISSN 2176-6649

ARTIGO ORIGINAL

Predição de valores de moedas virtuais através da análise de sentimento de notícias e tweets

Prediction of cryptocurrency values using sentiment analysis of news and tweets

Wagner Resende Santos[1] and Hugo Bastos de Paula[1]

[1]PUC Minas

*wagnermecsantos@gmail.com; hugo@pucminas.br

Recebido: 05/11/2018. **Revisado:** 08/11/2019. **Aceito:** 23/03/2020.

Resumo

As moedas virtuais (ou criptomoedas) estão se tornando cada vez mais competitivas no mercado mundial, o que atrai investidores para obter lucros sobre as oscilações deste mercado. Esses investimentos são orientados por um princípio simples: comprar as moedas quando seu valor de mercado estiver prestes a subir, e vendê-las quando estiverem prestes a perder o valor. Várias são as informações que podem auxiliar o investidor na tentativa de prever esse movimento oscilatório – entre elas, notícias e comentários da própria comunidade sobre o desempenho da moeda. Entretanto, lidar com esse volume de informações e julgar qual informação é relevante pode ser um desafio. O objetivo deste trabalho é desenvolver um modelo de predição do movimento de preços de criptomoedas que utilize como base a percepção pública da população em relação à essas moedas. Foram realizadas predições tanto utilizando análise de sentimento de notícias quanto de *tweets*. Foram gerados modelos de previsão individuais para cada fonte de dados e um modelo que combina ambas as fontes. Os resultados obtidos alcançaram MDA de até 75% utilizando XGBoost a partir do modelo combinado de informações de notícias e *tweets*, sendo capaz de prever os resultados também em períodos de grandes oscilações.

Palavras-Chave: Bitcoin;Criptomoedas;Notícias;Predição;Twitter.

Abstract

Virtual currencies (or cryptocurrencies) are increasingly becoming more competitive in the global market, atracting investors seeking for profit on the oscillation of this market. These investiments are oriented by a simple principle: buy the currencies when their market value is about to rise, and sell them when their market value is about to drop. Several sources of information can be used to support the decision making process – such as news or comments in social networks on the topic itself. Nonetheless, to deal with the such a huge amount of information presents itself as a big challenge. The goal of this work is to develop a model that predicts the movement of cryptocurrencies' prices based on the public perception about the currencies. Prediction models were derived from each source of information and from the combination of both sources. Results obtained up to 75% MDA using the model induced with XGBoost, from the combination of the two sources, being able to predict the results even during periods of oscilation.

Keywords: Bitcoin;Cryptocurrencies;News;Prediction;Twitter.

1 Introdução

As moedas virtuais (ou criptomoedas) estão se tornando cada vez mais competitivas no mercado mundial. A comercialização dessas moedas está se tornando tão grande, que seu valor se equipara com a produção de bens e serviços de grandes países. O valor total de mercado das moedas virtuais registrado no final de 2017 foi de aproximadamente $600 bilhões, o que corresponde ao produto interno bruto da Argentina (*State of Blockchain 2018*, 2018).

Segundo *Q3 2017 Cryptocurrency Report* (2017), as 26 maiores moedas virtuais apresentaram lucro de 600% durante os três primeiros trimestres de 2017. Essa propensão a altos retornos traz investidores para o cenário dessas moedas, que procuram obter lucros sobre as oscilações do mercado. Esses investimentos são orientados por um princípio simples: vender as moedas quando estiverem prestes a perder valor, e comprá-las quando seu valor de mercado estiver prestes a subir, sendo que o maior desafio está em saber qual o melhor momento de efetuar essas transações de compra e venda.

Existem muitos fatores que influenciam no valor de uma moeda virtual, sendo que um dos principais fatores é a percepção pública em relação às moedas (Farell, 2015). Ataques a grandes corretoras de criptomoedas, por exemplo, podem contribuir para sua desvalorização, enquanto a adoção de moedas virtuais como forma de pagamento por grandes representantes do varejo podem ter um impacto positivo no seu valor.

Investidores podem procurar com frequência por notícias sobre determinada moeda virtual para saber o que está acontecendo no mercado da moeda, a fim de entender qual a percepção pública das pessoas e tomar decisões. Esse processo pode ser trabalhoso e necessitar de muito tempo da pessoa que investe, podendo ser até impraticável, se a quantidade de informação necessária para o correto entendimento do contexto for muito grande.

1.1 Objetivo

O objetivo deste trabalho é desenvolver um modelo de predição do movimento de preços de moedas virtuais que utilize como base a percepção pública da população em relação às moedas. Essa percepção pública será captada através da análise de sentimento de notícias e *tweets* que abordem a criptomoeda desejada. Os objetivos específicos do trabalho são:

- Definir qual moeda virtual será analisada;
- Extrair notícias e tweets para a moeda analisada;
- Realizar análise de sentimento nos textos extraídos, com o intuito de descobrir se o sentimento sobre o tópico é positivo, negativo ou neutro;
- Criar modelos de predição da tendência de valor da moeda virtual a partir do sentimento das notícias e *tweets*;
- Testar os modelos criados e *selecionar* o que possuir maior capacidade de predição.

O modelo criado poderá ser utilizado como um guia para pessoas que investem, que poderão saber previamente o provável valor da moeda em um futuro próximo, e tomar uma decisão de forma rápida. Este modelo também pode vir a ser utilizado por ferramentas de negociação automatizadas, que poderão decidir, por exemplo, entre investir mais ou vender a moeda, de acordo com a previsão obtida em comparação com o valor atual.

2 Fundamentos

2.1 Mercado de moedas virtuais

As moedas virtuais podem ser compradas e vendidas de forma direta entre duas pessoas. A princípio, não há necessidade de uma entidade intermediária para que essas transações ocorram. Porém, para que pessoas possam comprar e vender com mais facilidade e segurança, foram criadas as corretoras de criptomoedas, também chamadas de *exchanges*. Nas *exchanges* são criadas ordens de compra e de venda, através das quais pessoas podem comprar e vender moedas virtuais. Geralmente as corretoras de criptomoedas funcionam 24 horas por dia, sem pausas.

Cada *exchange* determina quais moedas virtuais são aceitas, além de determinar com quais moedas convencionais (como o Real por exemplo) irá trabalhar. Ao criar uma conta na corretora, as pessoas podem enviar dinheiro em uma moeda convencional para a corretora, e com este dinheiro, comprar moedas virtuais; ou realizar o oposto: enviar moedas virtuais para a corretora e vendê-las na plataforma.

O perfil de quem investe nesse tipo de bem é variado, porém existem algumas características que se destacam. Segundo relatório demográfico desenvolvido por *Bitcoin Demographics* (2018), das pessoas que investem no Bitcoin – a moeda virtual com maior valor de mercado no momento – identifica-se como maior interesse os usuários de Serviços Financeiros e Serviços de Investimento (8,14% da comunidade). Em segundo lugar na lista de interesse dessas pessoas está o desenvolvimento de software (3,72%). Além disso, ao se analisar as características dessas pessoas, tem-se como percentual predominante investidoras ávidas (7,29%) e tecnófilas (6,98%). A idade de quem investe é jovem, porém majoritariamente maior de 24 anos. Quase metade (45,71%) é composta por pessoas de 25 a 34 anos, e apenas 8,36% possuem entre 18 e 24 anos.

2.2 Seleção de parâmetros para predição

Nesta seção serão apresentados os dados que podem ser utilizados na previsão de ativos de acordo com o tipo de previsão que foi aplicado.

Em Zhu et al. (2017), foi realizada uma análise de fatores econômicos no preço do Bitcoin. Os fatores analisados foram preço do dólar, Dow Jones Industrial Average, Federal Funds Rate e preço do ouro. Foi concluído que todas as variáveis têm uma influência de longo termo no valor da moeda, sendo que o preço do

dólar possui maior influência e o preço do ouro possui menor influência. Além disso, descobriu-se causalidade entre o preço do dólar e o do Bitcoin no curto prazo.

No trabalho desenvolvido por Mern et al. (2017), foram utilizados indicadores econômicos como Down Jones Industrial Average, S&P 500, assim como dados analíticos do Google Trends para o termo "Bitcoin". Esses dados foram utilizados para se prever o preço da moeda do dia seguinte. O modelo de melhor resultados, que utilizou uma Rede Neural Convolucional, teve acurácia de 66,7%.

A partir do trabalho desenvolvido por Jang and Lee (2018), é possível ter um bom entendimento dos dados que podem ser utilizados na predição de valores de Bitcoin. A partir de 26 parâmetros diferentes utilizados, determinou-se que 16 deles apresentavam coeficientes de correlação satisfatórios para previsão dos preços de Bitcoin através de métodos de regressão. Em um dos modelos empregados, que utilizou redes neurais bayesianas, o Erro Absoluto Percentual Médio foi de 1,8%. Os 16 parametros utilizados na previsão foram os seguintes: volume de negociação do Bitcoin, tamanho médio do bloco na *blockchain*, média do tempo de conferência das transações, lucro de mineradores, preço do petróleo, índice VIX, índice FTSE100, preço do ouro, volume de negociação de dólares americanos, lucro percentual dos mineradores, transações de Bitcoin confirmadas por dia, índice SSE, cotação USD/CNY, cotação USD/JPY, cotação USD/CHF e *Hash Rate* do Bitcoin.

Outro parâmetro frequentemente utilizado na previsão de ativos é a opinião pública em relação a determinado ativo. Em Mittal and Goel (2011) foi realizada previsão do valor do *Down Jones Industrial Average* (DJIA), sendo que os parâmetros utilizados foram o valor do DJIA nos últimos 3 dias, assim como o sentimento de *tweets* no mesmo período. Também são encontrados projetos em que o sentimento do título da notícia é utilizado, como visto em Coinanalysis (2018), enquanto outros utilizam o conteúdo da notícia, como visto em Daultani (2017).

2.3 Modelos para regressão e previsão de séries temporais

Diversos algoritmos podem ser utilizados na previsão de ativos, sendo que o modelo mais adequado varia de acordo com o problema a ser resolvido. A seguir são apresentados alguns desses modelos de acordo com a sua utilidade para o trabalho atual.

2.3.1 *ARIMA*

ARIMA é um tipo de modelo estatístico para previsão de séries temporais que é um acrônimo para *AutoRegressive Integrated Moving Average*, ou modelo auto-regressivo integrado de médias móveis . A parte auto-regressiva do modelo ARIMA consiste na relação entre o valor de uma variável e os seus valores passados. A média móvel no modelo diz respeito à dependência entre o valor da variável e o erro residual de uma média móvel usada nas observações passadas. A parte integrada consiste na diferenciação dos valores da variável com seus valores anteriores (Karakoyun and Osman, 2018).

Os parâmetros do ARIMA são os seguintes:

- p: O valor do período de defasagem
- d: o número de vezes que os valores são diferenciados
- q: o tamanho da janela de média móvel

O modelo matemático pode ser definido da seguinte forma:

$$y_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \phi_1 e_{t-1} - \ldots - \phi_1 e_{t-1} \quad (1)$$

2.3.2 *LSTM*

As redes LSTM (*Long Short Term Memory* - memória de longo e curto prazo) são um tipo de rede neural recorrente que permite a preservação de pesos para serem realizadas operações de *forward-propagation* e *back-propagation* através das camadas, e pode ser utilizada para descobrir dependências de longo termo (Karakoyun and Osman, 2018).

Após a análise realizada em Karakoyun and Osman (2018), descobriu-se que o algoritmo LSTM possui melhores resultados que o ARIMA na previsão de valores da moeda Bitcoin.

2.3.3 *SVM*

As Máquinas de Vetores de Suporte, ou *Support Vector Machine* (SVM) foram desenvolvidas para modelar relações lineares entre dados. Existem implementações de SVM para classificação em problemas de duas classes e várias classes, além de possuir implementações para regressão, chamadas de *SVR* (Support Vector Regression - Máquinas de Vetores de Regressão) (Ghosh and Purkayastha, 2017).

2.3.4 *Random Forest*

O algoritmo de *Random Forst* (Floresta Aleatória) é uma implementação focada na melhoria de métodos de árvore de decisão. Quando árvores de decisão são treinadas até uma grande profundidade, elas tendem a ficar com sobreajuste, funcionando bem com o conjunto de dados de treinamento, porém não mostra capacidade de prever novos resultados. A Floresta Aleatória procura diminuir o efeito do sobreajuste treinando diversas árvores de decisão diferentes e criando uma média das múltiplas árvores de decisão treinadas. O algoritmo constantemente seleciona partes aleatórias do conjunto de dados para treinar as árvores, por isso o nome (Ghosh and Purkayastha, 2017).

2.3.5 *Extreme Gradient Boosting (XGboost)*

O XGboost (*Extreme Gradient Boosting* - Impulsionador de Gradiente Extremo) segue o mesmo princípio de *Gradient Boosting*, porém inclui penalidades de regressão na equação de impulsionamento de gradiente. Assim como o impulsionamento de gradiente, o XGboost atua na melhoria de modelos de árvore de decisão, porém utiliza a estrutura do hardware do computador para melhorar o tempo de computação e melhorar utilização de memória, o que é muito importante ao trabalhar com

impulsionamento de árvores (Ghosh and Purkayastha, 2017).

Em Ghosh and Purkayastha (2017), foi realizada uma comparação dos algoritmos de XGboost, *Random Forest* e SVM no processo de regressão. Concluiu-se que o XGboost obteve melhores resultados do que os outros dois algoritmos para a previsão de valores de ativos, utilizando como parâmetros média móvel exponencial, índice de força relativa e Média de Amplitude de Variação.

2.4 Predição de ativos baseada em análise de sentimento

No âmbito de pesquisas recentes, registraram-se grandes esforços no desenvolvimento de modelos que sejam capazes de prever a tendência de ativos, principalmente ações. Muitos deles utilizam indicadores técnicos, enquanto outros mostraram forte relação entre notícias e *tweets* de uma determinada empresa e o valor das suas ações. A seguir serão apresentados trabalhos anteriores que utilizam a análise de sentimento de textos para determinar a tendência de ativos.

Daultani (2017) coletou artigos do jornal New York Times de um período de 10 anos, e classificou o sentimento das notícias em positivo, negativo e neutro. Também foram coletados os valores do índice de ações *Dow Jones Industrial Average* (DJIA) para este período. Então, foram utilizados os sentimentos das notícias para prever os valores das ações de acordo com o DJIA, ou seja, para prever as mudanças no mercado de ações dos Estados Unidos como um todo. O autor comparou os resultados de predição utilizando os modelos *Random Forest*, *Logistic Regression* e *Multi-Layer Perceptron* (MLP), e obteve melhores resultados com o MLP.

No trabalho de Pagolu et al. (2016), foram coletados textos de *tweets* que expressavam a opinião de pessoas sobre a empresa Microsoft e/ou seus produtos e serviços. Os *tweets* foram classificados em positivos, negativos e neutros. No trabalho foram coletadas as tendências das ações da Microsoft, atribuindo um valor numérico de 0 quando o valor da ação caiu em relação ao seu valor no dia anterior, e no caso contrário, atribuindo um valor de 1. Utilizando um modelo de regressão logística, obtiveram-se resultados com uma precisão de 69.01%.

Em Mittal and Goel (2011), foram coletados *tweets* de junho a dezembro de 2009 para prever os valores do índice *Dow Jones Industrial Average* (DJIA). A análise de sentimentos realizada separou o sentimento dos *tweets* em 4 classes: calmo (*calm*), feliz (*happy*), alerta (*alert*) e amável (*kind*). Foram utilizados como parâmetro o resultado das análises de sentimento dos últimos 3 dias, além dos valores do DJIA nos últimos 3 dias. Descobriu-se que os sentimentos calmo e feliz obtiveram maior correlação com o valor do DJIA. Utilizando um modelo baseado em *Self Organizing Fuzzy Neural Networks* (SOFNN), foi obtido um percentual de sucesso da predição da direção das ações de 75,56%.

Em Coinanalysis (2018), foi feita a utilização de LSTM na análise do sentimento do título de notícias relacionadas a Bitcoin. Também foi utilizada uma segunda LSTM para realizar a predição do valor do Bitcoin no dia seguinte. O autor utilizou dados de notícias dos 10 últimos dias, juntamente com o preço atual, para realizar essa previsão. Obteve-se uma taxa de erro número percentual de 1.22%.

2.5 Métricas de avaliação para predição de séries temporais

2.5.1 *RMSE*

A métrica *Root Mean Squared Error*, ou Raiz do Erro Quadrático Médio é geralmente a forma de medição de erro médio mais utilizada para se avaliar a precisão de um modelo (Willmott and Matsuura, 2005). A métrica é definida pela seguinte fórmula:

$$\sqrt{\frac{\sum_{t=1}^{N}(P_t - R_t)^2}{N}} \tag{2}$$

Sendo que N é o número de itens na série temporal, R_t é o valor real na posição t e P_t é o valor da predição na posição t. A ideia de elevar a diferença a 2 é para remover o sinal e fazer com que a magnitude dos erros influenciem a medida da média dos erros (Willmott and Matsuura, 2005).

2.5.2 *MAE*

O Erro Médio Absoluto, ou *Mean Absolute Error*, é a média da diferença entre o valor da predição (P_t) e o valor real (R_t), sendo definida pela seguinte fórmula:

$$MAE = \frac{\sum_{t=1}^{N}|R_t - P_t|}{N} \tag{3}$$

Segundo Willmott and Matsuura (2005), MAE é a métrica mais natural para se medir a magnitude do erro médio, não apresenta ambiguidade, e deve ser escolhida em detrimento de RMSE.

2.5.3 *MAPE*

A métrica Mean Absolute Percentage Error, ou Erro Médio Percentual Absoluto, possui o mesmo princípio da MAE, porém é apresentada em valores percentuais. A fórmula para a métrica pode ser definida da seguinte forma (Jang and Lee, 2018):

$$MAPE = \frac{\sum_{t=1}^{N}\left|\frac{R_t - P_t}{R_t}\right|}{N} \tag{4}$$

2.5.4 *MDA*

A métrica MDA (*Mean Directional Accuracy* - Precisão Direcional Média) se difere de outras métricas de avaliação de modelos de regressão no sentido de mensurar a capacidade do modelo de prever o aumento ou queda do valor, ou seja, capacidade de prever a direção futura do valor (Blaskowitz and Herwartz, 2009). A equação será

construída a seguir, utilizando como base (Bergmeir et al., 2014).

Utilizando a função indicadora $I[\ldots]$ a direção real (DR_t) e a direção da predição (DP_t) são dadas por:

$$DR_t = I[(R_{t+h} - R_t) > 0] \tag{5}$$

$$DP_t = I[(P_{t+h} - R_t) > 0] \tag{6}$$

Em que t é a posição no conjunto de dados e h é o número de posições no futuro a se prever (por exemplo, 3 dias). R_t é o valor atual na série, P_{t+h} é o valor da previsão no tempo $t + h$, e R_{t+h} é o valor real no tempo $t + h$. Assim, o erro direcional (DE) pode ser definido da seguinte forma:

$$DE_t = I[DR_t = DP_t] \tag{7}$$

Utilizando o DE, a podemos encontrar a precisão direcional (DA) da seguinte forma:

$$DA_t = \begin{cases} a & para \quad DE_t = 1 \\ b & para \quad DE_t = 0 \end{cases} \tag{8}$$

Ou seja, uma predição correta da direção resulta no valor a, que denota um acerto, e uma predição incorreta resulta no valor b, que denota um erro. Segundo Blaskowitz and Herwartz (2009), valores comuns para a e b são (a, b) = (1, 0) ou (a, b) = (1, −1). No primeiro caso, DA_t é idêntico a DE_t. No segundo caso, b é uma penalidade. A partir do erro direcional podemos então definir a Precisão Direcional Média, ou MDA, a partir da seguinte equação:

$$MDA = \frac{1}{N} \sum_{t=1}^{N} DA_t \tag{9}$$

2.5.5 *MDV*

A métrica MDA é útil para definir se o algoritmo é capaz de prever a direção do ativo, mas não leva em consideração o grau de alteração, ou seja, não mensura o valor econômico da predição. O modelo pode ser capaz de prever a direção em casos de baixa volatilidade, mas pode falhar quando a volatilidade está alta, o que diminui seu valor econômico (Bergmeir et al., 2014). Para obter essa informação de valor econômico, é utilizado o valor da predição direcional (DV), que é uma multiplicação do DA (precisão direcional) pelo valor absoluto da alteração de valor. A DV média (MDV) é definida da seguinte forma:

$$MDV = \frac{1}{N} \sum_{t=1}^{N} (|R_{t+h} - R_t|) DA_t \tag{10}$$

Como já dito anteriormente, ao se escolher a opção (a, b) = (1, −1) na obtenção do DA_t, penaliza-se os erros. Nesse caso, a métrica MDV é então utilizada para se identificar o valor médio dos ganhos monetários implicados pelos acertos predição do modelo. Quando o valor do MDV é negativo, significa que o valor econômico dos erros se sobrepôs em relação ao valor econômico dos acertos, gerando prejuízo.

3 Metodologia

Este trabalho utiliza três fontes de dados para realizar a predição: *tweets* escritos na rede social *Twitter*, notícias sobre moedas virtuais e dados de valores das moedas virtuais. A coleta e pré-processamento de dados é realizada de forma diferente para cada fonte. Após o pré-processamento, é realizado o processo de transformação, que é seguido da mineração de dados (predição) e análise dos resultados. A Fig. 1 apresenta o fluxo do processo de aprendizagem de máquina.

3.1 Coleta e processamento de notícias e tweets

A coleta de notícias foi realizada utilizando a *Application Programming Interface* (API) chamada News API[1]. A News API possui artigos de mais de 30000 fontes, e possui um plano grátis para utilização sem fins comerciais (NewsAPI, 2018). Os *endpoints* da API proveem diversos filtros, entre eles palavras-chave e data mínima e máxima da notícia. Os principais *endpoints* da API são os de *top headlines*, ou notícias mais famosas, e *everything*, que inclui qualquer notícia disponível na API, independente de ser famosa ou não. Os *endpoints* citados retornam uma lista de notícias, contendo título, *Uniform Resource Locator* (URL) de acesso ao artigo, data de publicação, entre outros dados.

A coleta de notícias foi realizada utilizando a *Application Programming Interface* (API) chamada News API[2]. A News API possui artigos de mais de 30000 fontes, e possui um plano grátis para utilização sem fins comerciais (NewsAPI, 2018). Os *endpoints* da API proveem diversos filtros, entre eles palavras-chave e data mínima e máxima da notícia. Os principais *endpoints* da API são os de *top headlines*, ou notícias mais famosas, e *everything*, que inclui qualquer notícia disponível na API, independente de ser famosa ou não. Os *endpoints* citados retornam uma lista de notícias, contendo título, *Uniform Resource Locator* (URL) de acesso ao artigo, data de publicação, entre outros dados. De todas as notícias de Bitcoin coletadas, os veículos com maiores percentuais de notícias são focados em assuntos relacionados a finanças e negócios. Dentre eles podemos separar alguns com foco em criptomoedas, como o CCN (Cryptocurrency News and Breaking Updates), NewsBTC e Bitcoinist. Aproximadamente 50% das notícias foram compostas por veículos com menos de 1% do total de notícias, sendo estes classificados como "Outros". A Fig. 2 apresenta os veículos com maior número de notícias, assim como o percentual em relação ao total de notícias coletadas.

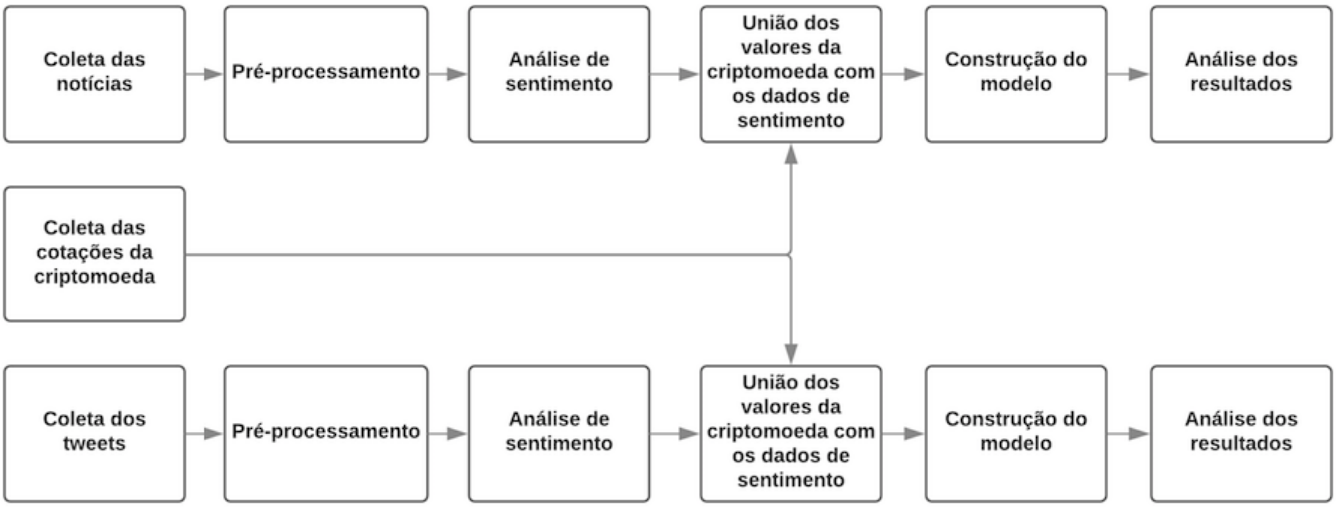

Figura 1: Processo de aprendizado de máquina

Para se obter o conteúdo das páginas das notícias, foi utilizada uma ferramenta de automatização de acesso a páginas *web* chamada Puppeteer[3]. O Puppeteer é uma biblioteca escrita em Javascript (Node.js) que permite acessar páginas *web* de forma automatizada através dos navegadores Chrome ou Chromium (Puppeteer, 2018). Utilizando a biblioteca, pode-se realizar ações como interagir com a página, gerar capturas de tela da página e obter o conteúdo de *Hypertext Markup Language* (HTML) da página. Como a ferramenta abre o nevagador Chrome ou Chromium, códigos Javascript são executados ao acessar uma página, permitindo a obtenção de conteúdos que são dinamicamente adicionados à página através de Javascript. Outra utilidade de se abrir o navegador é que *websites* que possuem barreiras contra robôs automatizados (ou *bots*) têm menos chance de bloquear o acesso à página, pois a interação é quase idêntia à interação que um usuário real teria.

Após extrair o HTML da notícia, é necessário extrair o conteúdo da notícia e ignorar os metadados presentes na linguagem HTML. Para esse processo foi utilizada a biblioteca unfluff[4], escrita em Javascript.

Já a coleta de *tweets* foi realizada utilizando a API do Twitter. Foi utilizada a biblioteca Twit[5], escrita em javascript, para auxiliar na realização das requisições à API. Foram realizadas requisições à API do Twitter a cada 30 minutos para encontrar os *tweets* que eram relacionados ao Bitcoin.

Como existem robôs no Twitter, foi feito um processo de filtragem dos *tweets* que pertenciam a robôs. A heurística de identificação de bots utilizada foi a vista em Stenqvist and Lönnö (2017), que consiste em remover *tweets* que contenham determinadas *hashtags*, palavras, bigramas e trigramas de palavras. A Tabela 1 apresenta os símbolos que foram utilizados nessa identificação de robôs.

3.2 Coleta e processamento de valores de moedas virtuais

A coleta dos valores de moedas virtuais foi feita utilizando a API CryptoCompare[6]. O principal motivo para a escolha desta fonte de dados em detrimento de outras foi a existência de um índice agregado. Caso fosse utilizado o valor da moeda de uma corretora específica, os dados estariam sujeitos a eventuais problemas que a corretora venha a enfrentar, como interrupção das transações devido a ataques de negação de serviço. A CryptoCompare possui um índice que utiliza várias corretoras para calcular o valor de mercado de uma moeda virtual. O índice é chamado CCCAGG, e utiliza uma média ponderada dos valores das corretoras, de acordo com o volume que a corretora possui. O índice aplica uma série de táticas para garantir que mudanças fora da curva não afetem negativamente o valor da moeda, como detecção de *outliers* e penalidade a corretoras que suspendem suas transações (CryptoCompare, 2018).

Através da API CryptoCompare, foram salvos os valores de fechamento de cada hora das moedas virtuais analisadas.

3.3 Análise de sentimento dos textos coletados

Para determinar o posicionamento das pessoas em relação às moedas analisadas, foi utilizado o recurso de análise de sentimento. Os sentimentos em relação às moedas foram classificados em positivo, negativo e neutro. Para a realização dessa análise, foi utilizada a API Rosette[7]. a API é utilizada por grandes empresas como Airbnb, Amazon e Microsoft, e possui capacidade de analisar o sentimento tanto de *tweets* quanto de notícias (Corp., 2018). A API possui planos especiais para

[2]https://newsapi.org/
[3]https://github.com/GoogleChrome/puppeteer
[4]https://www.npmjs.com/package/unfluff
[5]https://www.npmjs.com/package/twit
[6]https://www.cryptocompare.com/api/

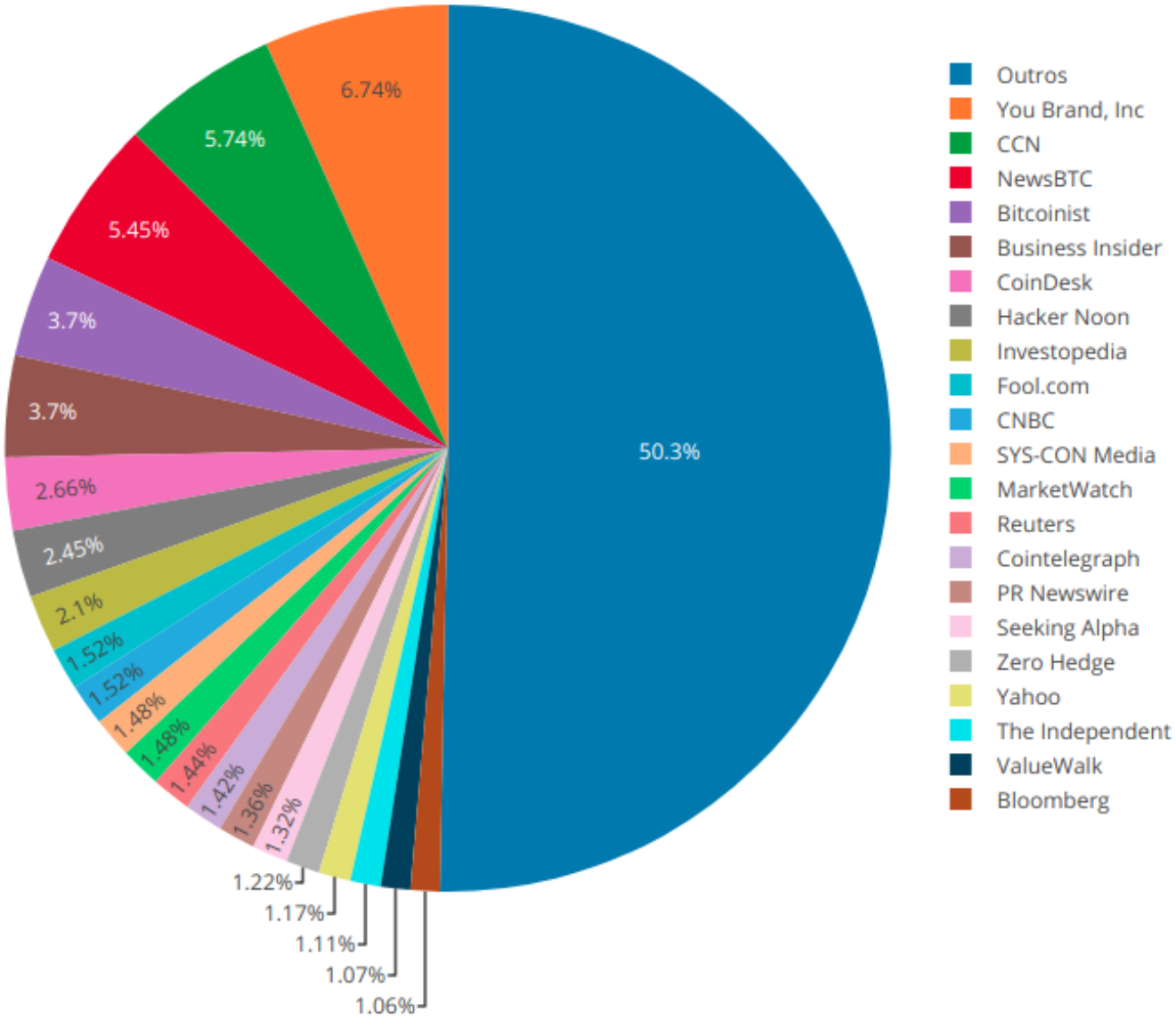

Figura 2: Percentuais de cada veículo de notícias

pesquisas acadêmicas, com limite de 100000 requisições por mês, sem a necessidade de pagar pelo seu uso.

A API provê tanto o sentimento do documento (texto inteiro) quanto o sentimento em relação a uma entidade. Utilizou-se o sentimento da entidade para obter-se especificamente o posicionamento das pessoas em relação à moeda retratada no texto.

Foi realizada análise de sentimento apenas de textos escritos em inglês, devido ao fato de a API não ter suporte para análise de sentimento de textos em português e também devido ao fato de existir mais conteúdo (notícias e tweets) em inglês do que em português em relação a moedas virtuais.

3.4 Apresentação dos dados coletados e transformados

A Tabela 2 apresenta os dados iniciais utilizados na extração, que são URLs de notícias relacionadas à moeda Bitcoin.

A seguir encontra-se o texto final da notícia extraído, após se obter o HTML da página e então extrair o texto do HTML. O texto de exemplo abaixo foi extraído a partir do seguinte URL de notícia: http://news.abs-cbn.com/business/01/01/18/bangko-sentral-on-bitcoins-study-it-very-closely.

[7]https://www.rosette.com/capability/sentiment-analysis

Tabela 1: Símbolos identificados como suspeitos

Tipo de símbolo	Lista de símbolos utilizados
hashtags	#mpgvip, #freebitcoin, #livescore, #makeyourownlane, #footballcoin
palavras	{entertaining, subscribe}
bigramas	{free, bitcoin}, {current, price}, {bitcoin, price}, {earn, bitcoin}
trigramas	{start, trading, bitcoin}

Fonte: criado pelo autor

Tabela 2: URLs de notícias de Bitcoin

URL da notícia
https://www.youbrandinc.com/blockchain/blockchain-predictions-for-2018/
https://www.stocktrader.com/2017/12/31/weekly-market-recap-dec-31-2017/
https://www.inc.com/melanie-curtin/science-says-happy-couples-have-these-13-characteristics.html
https://www.malaysiakini.com/news/407175
http://www.thestar.com.my/tech/tech-news/2018/01/01/ten-top-tech-trends-in-2018/
http://www.scmp.com/week-asia/business/article/2126189/why-south-korea-suddenly-terrified-bitcoin
http://news.abs-cbn.com/business/01/01/18/bangko-sentral-on-bitcoins-study-it-very-closely
http://bitcoinist.com/2017-is-over-what-does-2018-hold-for-bitcoin/
http://www.newsbtc.com/2018/01/01/bitcoin-cash-price-technical-analysis-bch-usd-preparing-break/

Fonte: criado pelo autor

> MANILA - Bangko Sentral ng Pilipinas Deputy Governor Diwa Guinigundo has advised the public to be aware of the risks of investing in bitcoin after the cryptocurrency soared to record highs last year.
>
> Asked if bitcoin had a future in the Philippines in 2018, Guinigundo told ANC: "I don't think so. It is something that people will need to understand very closely."
>
> The BSP is "accepting"cryptocurrencies "but at the same time, providing advisories to the general public that they have to be conscious of the opaqueness of transactions involving Bitcoins."
>
> "This is something we should think about very closely,"he said.
>
> The BSP released guidelines on cryptocurrencies including bitcoin earlier this year. Digital currency transactions have grown to $6 million daily from $2 million to $3 million a few years ago, according to central bank data.

A Tabela 3 apresenta os dados da moeda Bitcoin após as transformações, ou seja, após utilizar os textos para se extrair os sentimentos e extrair os preços da moeda. A primeira coluna apresenta a data das notícias, no formado Ano-Mês-Dia. A segunda coluna a contagem de sentimentos positivos em relação à moeda nesse dia, a terceira a contagem de sentimentos neutros e a quarta a contagem de sentimentos negativos. A penúltima coluna apresenta o valor de fechamento do Bitcoin em dólares na date de publicação da notícia, e a última mostra o valor do Bitcoin no dia seguinte.

A coleta de tweets começou no dia 16/03/2018, portanto, foram utilizados tweets e notícias a partir do dia 16/03/2018, ignorando as notícias de datas anteriores que haviam sido coletadas. A data final da coleta foi 12/08/2018, totalizando então 131 dias de coleta. Nesse período foi realizada análise do sentimento de 224798 tweets e 28505 notícias.

3.5 Construção do modelo

Para realizar o processo de mineração dos dados, mais especificamente predição de valores e predição da direção dos valores, foi escolhida a linguagem Python, devido ao grande volume de material e bibliotecas que possuem o objetivo de auxiliar a prática da mineração de dados. A principal biblioteca utilizada foi a scikit-learn, que possui um conjunto de ferramentas de mineração e análise de dados.

Para fins de comparação, foi utilizada mais de uma técnica de aprendizado de máquina. Para a prática da regressão, foram utilizadas as estratégias LSTM e Árvores de Decisão (utilizando a técnica de melhoria Gradient Boosting). O modelo baseado em Gradient Boosting foi criado a partir da biblioteca xgboost, enquanto o modelo de LSTM foi criado a partir da biblioteca keras. Foram escolhidos estes modelos devido aos resultados satisfatórios obtidos em outros trabalhos de proprósito parecido.

3.6 Análise dos resultados

A etapa de análise dos resultados será realizada a partir de métricas de avaliação de predição estudadas no trabalho. Foram escolhidas as métricas MDA e MDV nessa avaliação, pois elas são capazes de identificar qual é a capacidade do modelo de prever a direção do valor (e o valor econômico dessa previsão), o que vai de acordo com o objetivo do trabalho.

A métrica MDA será utilizada com os valores de a e b definidos como (a,b) = (1,0), com o objetivo de saber em média quantas vezes o modelo acertou a direção do valor da moeda.

Já a métrica MDV será utilizada com os valores de a e b definidos como (a, b) = (1, −1), com o intuito de penalizar os erros, fazendo com que a métrica nos mostre qual o valor econômico da previsão. Dessa forma podemos, por exemplo, detectar falhas de um modelo

Tabela 3: Dados de moeda transformados

dia	positivos	neutros	negativos	preço atual (US$)	próximo preço (US$)
2018-01-01	49	78	40	14754.13	15156.62
2018-01-02	138	187	120	15156.62	15180.08
2018-01-03	164	191	105	15180.08	16954.78
2018-01-04	113	195	97	16954.78	17172.3
2018-01-05	125	142	100	17172.3	16228.16
2018-01-06	37	63	29	16228.16	14976.17
2018-01-07	66	59	47	14976.17	14468.5
2018-01-08	133	184	134	14468.5	14919.49
2018-01-09	171	184	125	14919.49	13308.06

Fonte: criado pelo autor

que funciona bem na maioria dos casos, mas que erra em situações de alta volatilidade.

A fim de se entender qual a melhoria do modelo criado em relação a um modelo trivial, será criado um modelo base que, ao realizar a previsão, repete a direção do valor anterior na série temporal. As mesmas métricas serão utilizadas neste modelo base e comparadas com o modelo criado.

4 Desenvolvimento

A moeda escolhida para a realização do trabalho foi o Bitcoin, por ser a moeda que possui maior volume de notícias e por ser mais tratada em outros trabalhos de predição de moedas virtuais.

Para entender a correlação entre os sentimentos dos *tweets* e o preço do Bitcoin, foi utilizado o coeficiente de correlação Pearson. Foi feita uma análise do resultado da correlação para diferentes cenários de atraso do sentimento em relação ao preço da moeda, a fim de determinar para quantas horas ou dias no futuro o sentimento da moeda possui maior correlação com o preço. O valor de sentimento utilizado é uma agregação, representada pela diferença entre o número de sentimentos positivos e negativos, ou seja, essa hipótese considera que o impacto do sentimento positivo é oposto ao impacto do sentimento negativo, e o sentimento neutro não traz impacto no valor. A Fig. 3 apresenta uma comparação de um atraso em dias e um atraso em horas, a fim de determinar qual a melhor unidade de tempo a ser utilizada na predição.

Segundo Zhang (2013), sentimentos de tweets geralmente necessitam ser analisados em períodos mais longos, como um dia ou uma semana, e pode-se notar este comportamento com os dados utilizados. A correlação foi maior ao utilizar um agrupamento dos sentimentos e valores de fechamento de um dia, não hora a hora.

A maior correlação encontrada foi quando o atraso é de 0 dias (valores de sentimento e de preço pertencem ao mesmo dia), com resultado de −0.194. Como o interesse é realizar previsões futuras do valor do Bitcoin, será utilizado o segundo melhor valor de correlação (−0.193), que pertence ao atraso de 1 dia, ou seja, serão utilizados os sentimentos do dia atual para prever o preço do dia seguinte.

Outra hipótese considerada é a de que os sentimentos positivos, negativos e neutros possuem diferentes impactos no valor de uma criptomoeda, portanto, foi realizada uma segunda análise de correlação de acordo com o atraso em dias, dessa vez para cada tipo de sentimento de forma separada. A Fig. 4 apresenta a comparação das correlações de cada um dos tipos de sentimento.

Nota-se que individualmente cada tipo de sentimento pode apresentar maior correlação em relação à agregação dos valores realizada anteriormente. O sentimento positivo chega a −0.32 com atraso de 0 dias, e o sentimento negativo chega a −0.26. O sentimento neutro possui a menor correlação, chegando ao máximo de 0.12. Um resultado interessante é que ambos os sentimentos positivos e negativos apresentaram uma relação inversamente proporcional em relação ao preço, o que é contrário ao que era esperado.

Como o melhor valor de correlação nesse caso pertence ao sentimento positivo com atraso de 0 dias, será utilizado o mesmo atraso de 1 dia ao realizar predições com os três tipos de sentimento como entrada.

Foi também realizado o mesmo processo de checagem da correlação dos diferentes tipos de sentimentos em relação ao preço do Bitcoin, porém desta vez a entrada foram os sentimentos das notícias relacionadas à criptomoeda. A Fig. 5 apresenta a comparação das correlações de cada um dos tipos de sentimento.

A partir da análise do Gráfico acima nota-se que os sentimentos positivos e negativos de notícias possuem maior correlação com o valor do Bitcoin, sendo que o positivo chega a um ponto alto de 0.46 para 2 dias de atraso, o neutro 0.49 também para 2 dias de atraso, e o negativo 0.38 para 10 dias de atraso. Como os sentimentos positivo e negativo tiveram maior correlação que o negativo e ambos têm como melhor valor 2 dias de atraso, este será o atraso utilizado nas predições baseadas em sentimentos de notícias.

A partir da pesquisa realizada, os modelos que foram considerados melhores para o objetivo do trabalho foram o SVM, *Random Forest* e XGBoost. Para validação dos modelos, foi utilizado o recurso de divisão entre base de treinamento e base de teste, com 70% da base para treinamento e 30% para teste. O Algoritmo da Fig. 6 apresenta a função de separação entre as bases de treinamento e de teste. As bases são separadas na linha 14 através da função train_test_split, passando o parâmetro random_state constante para que o resultado possa ser repetido. Na linha 22 são retornados tanto os dados sem a coluna de data para a predição, e os dados com a coluna de data para se exibir os gráficos

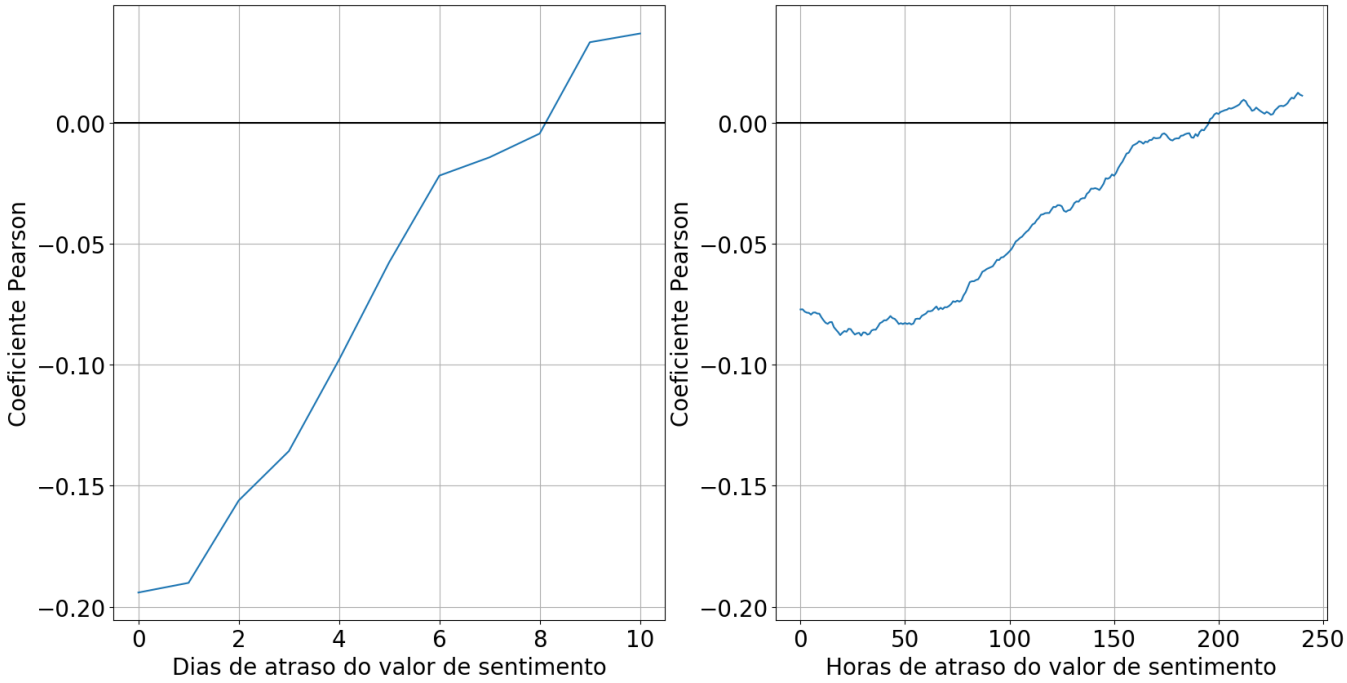

Figura 3: Correlação do sentimento agregado de *tweets* em dias e horas

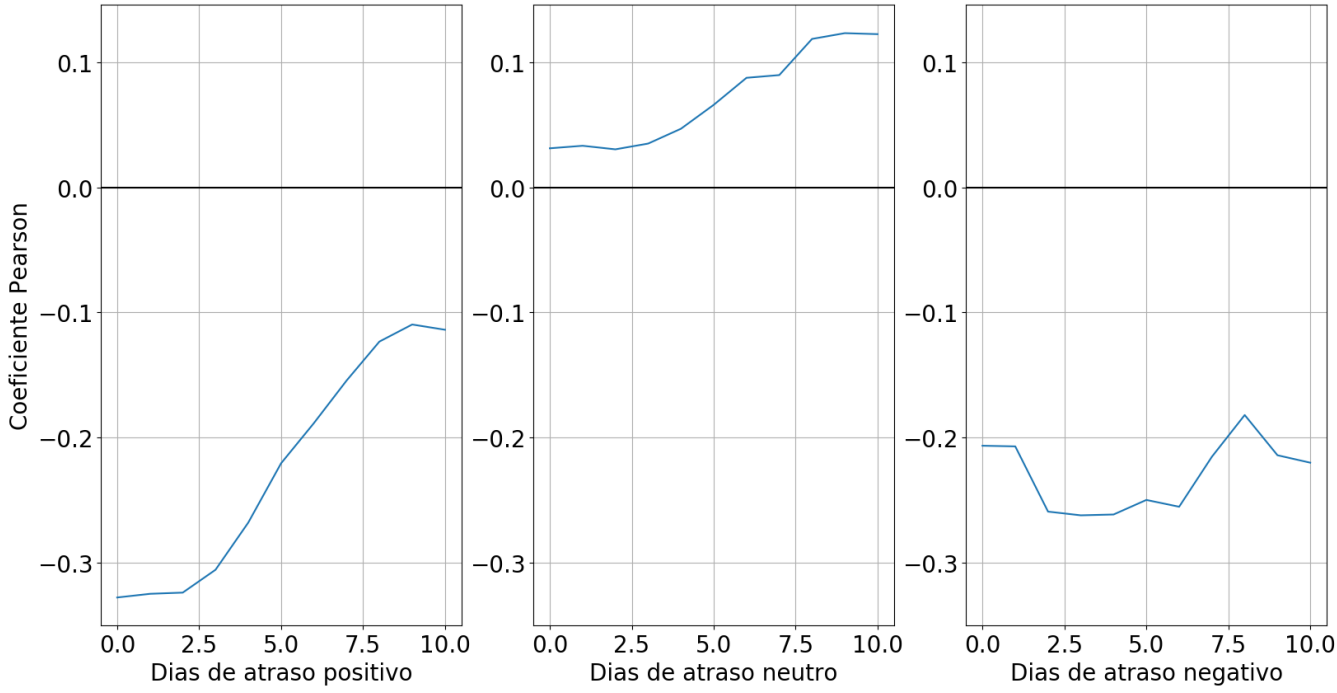

Figura 4: Correlação de cada sentimento de *tweets* em dias

comparando valor previsto e valor real. Este procedimento de separação de bases de treinamento e de teste foi utilizado por todos os modelos testados.

Em todos os modelos, foi utilizado o método *grid search* para otimização dos parâmetros. O Algoritmo da Fig. 7 apresenta o processo de seleção dos melhores parâmetros no modelo XGBoost. Na função param_selection foram definidas faixas de valores para os parâmetros gamma, max_depth, min_child_weight, subsample, colsample_bytree, reg_alpha e learning_rate. O resultado é extraído na linha 25, e após isso é feito o processo de treinamento e teste do modelo.

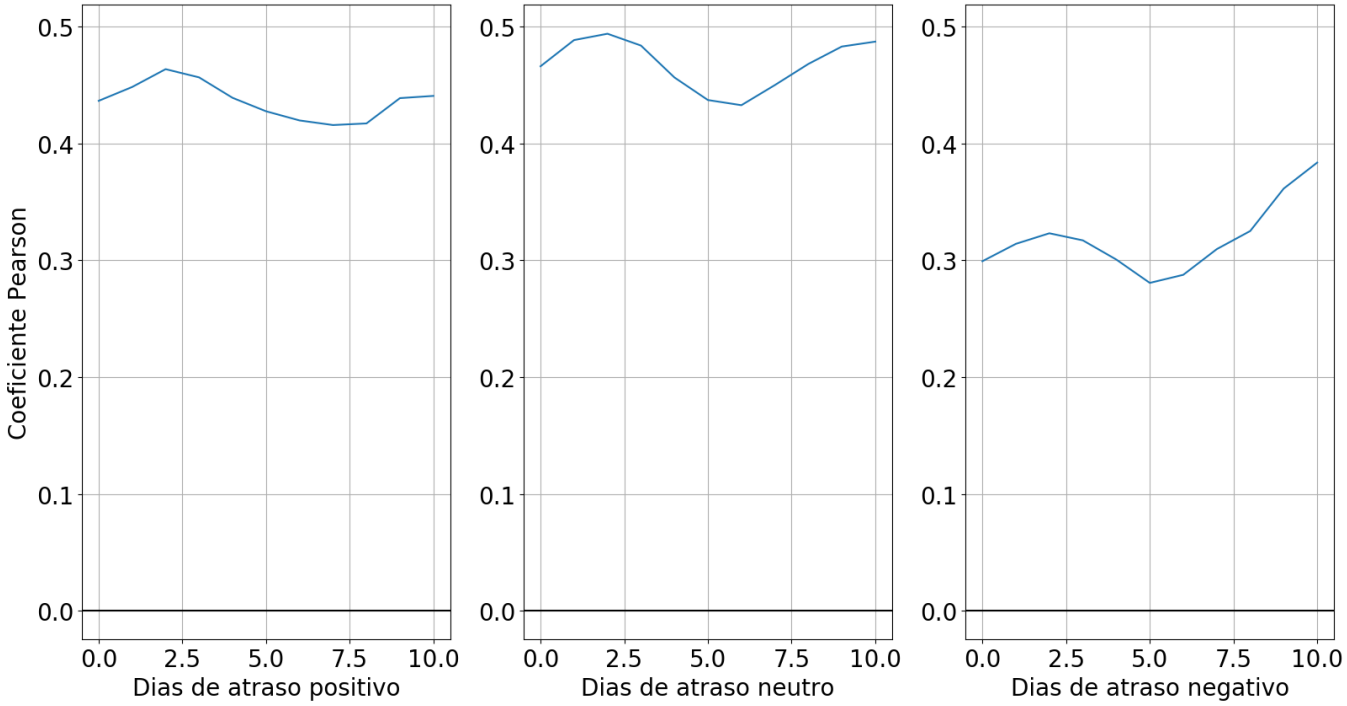

Figura 5: Correlação de cada sentimento de notícias em dias

```python
import numpy as np
from sklearn.model_selection import train_test_split

def load_daily_train_test_split():
    dataset = np.genfromtxt(
        'data/daily-bitcoin-data-news.csv',
        delimiter=",", skip_header=1, dtype=object
    )

    X = dataset[:,0:5]
    Y = dataset[:,5]
    Y = Y.astype(np.float)

    X_train, X_test, y_train, y_test = train_test_split(
        X, Y, test_size=0.3,
        random_state=7, shuffle=False
    )

    X_train_without_date = X_train[:,1:].astype(np.float)
    X_test_without_date = X_test[:,1:].astype(np.float)

    return X_train, X_test, y_train, y_test,
        X_train_without_date, X_test_without_date
```

Figura 6: Separação de bases de treinamento e de teste

5 Resultados

Na Tabela 4 estão apresentados os resultados utilizando os seguintes modelos: XGBoost, SVM, *Random Forest* e o modelo base. Como já dito anteriormente, o modelo base realiza previsões que mantêm a última direção do valor da moeda, ou seja, possui uma estratégia trivial para ser utilizada como base de comparação.

A partir da tabela, podemos notar que o modelo utilizando XGBoost obteve o melhor resultado entre os modelos analisados de acordo com o objetivo principal. O MAPE foi um pouco superior ao do modelo base,

```python
from xgboost import XGBRegressor
from sklearn.model_selection import train_test_split,
    GridSearchCV
from utils import load_daily_train_test_split

X_train, X_test, y_train, y_test,
X_train_without_date,
X_test_without_date = load_daily_train_test_split()

def param_selection(X, y):
    gamma = [i/10.0 for i in range(0,5)]
    max_depth = [3, 4, 5, 6, 7, 8, 9, 10]
    min_child_weight = [1, 2, 3, 4, 5, 6]
    subsample = [i/10.0 for i in range(6,10)]
    colsample_bytree = [i/10.0 for i in range(6,10)]
    reg_alpha = [1e-5, 1e-2, 0.1, 1, 100]
    learning_rate = [0.1, 0.2, 0.3]
    param_grid = {
        'max_depth': max_depth,
        'min_child_weight' : min_child_weight,
        'gamma': gamma, 'subsample': subsample,
        'colsample_bytree': colsample_bytree,
        'reg_alpha': reg_alpha,
        'learning_rate': learning_rate
    }
    grid_search = GridSearchCV(
        XGBRegressor(booster='gbtree', n_estimators=140,
            objective='reg:linear', learning_rate=0.056
        ),
        param_grid,
        cv=None
    )
    grid_search.fit(X, y)
    return grid_search.best_params_

best_params = param_selection(X_train_without_date, y_train)
```

Figura 7: Seleção de parâmetros com GridSearch

porém tanto o MDA quando o MDV apresentaram melhores resultados, demonstrando melhor capacidade do modelo de identificar a direção futura do Bitcoin. A **??**

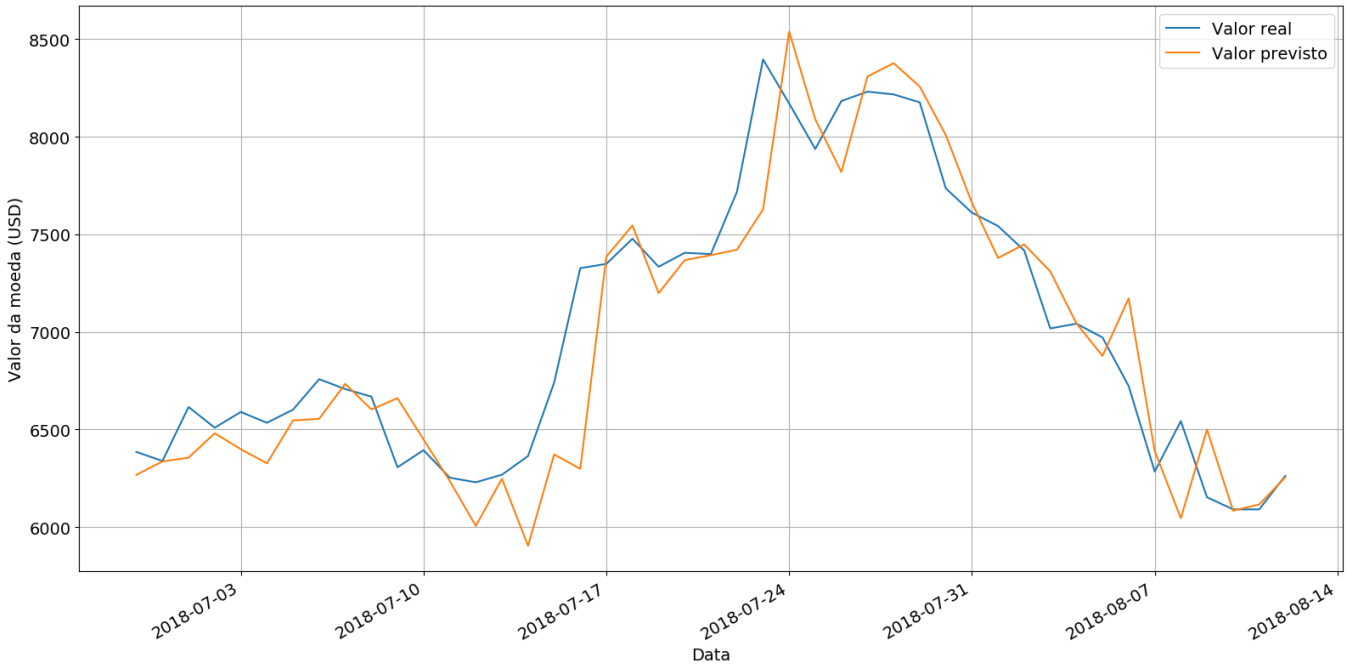

Figura 8: Predição de valores do Bitcoin com base em tweets utilizando XGBoost

Tabela 4: Comparação dos modelos de predição utilizando *tweets*

Modelo	MAPE	MDA	MDV
Base	2.46%	55%	6.4
SVM	3.68%	52%	−11.32
Random Forest	4.09%	48%	−41.32
XGBoost	2.69%	73%	59.86

Fonte: criado pelo autor

mostra os resultados da predição do XGBoost. Nota-se pelo gráfico que o modelo não alcança valores absolutos muito próximos ao real, mas a tendência de subida ou queda é identificada com precisão na maioria dos casos, tanto em situações de alta volatilidade quanto baixa. A correta identificação da tendência também em situações de baixa volatilidade é demonstrada pelo MDV do XGBoost, que alcançou um valor positivo e maior do que todos os outros modelos.

Na Tabela 5 também são apresentados os resultados dos modelos Base, SVM, *Random Forest* e XGBoost, mas dessa vez utilizando como parâmetro os sentimentos de notícias.

Tabela 5: Comparação dos modelos de predição utilizando notícias

Modelo	MAPE	MDA	MDV
Base	3.57%	45%	5.72
SVM	3.65%	52%	13.18
Random Forest	3.94%	52%	14.57
XGBoost	4.63%	66%	44.66

Fonte: criado pelo autor

A partir da tabela, podemos notar que o modelo utilizando XGBoost também obteve o melhor resultado entre os modelos analisados de acordo com o objetivo principal (MDA). O comportamento é parecido com a predição utilizando tweets, a diferença é que a a taxa de sucesso do MDA foi inferior na predição baseada em notícias. A Fig. 9 apresenta os resultados da predição utilizando o XGBoost. Nota-se pelo gráfico e pelos resultados das métricas que a predição da tendência do valor foi correta em mais da metade dos casos, porém com maiores erros em situações de alta volatilidade, assim como maiores erros nos valores absolutos, quando comparado com o modelo anterior baseado no sentimento de *tweets*.

Supõe-se que os melhores resultados foram obtidos a partir de sentimentos de *tweets* devido ao maior volume de dados gerado em um único dia, e devido à velocidade com a qual a manifestação pública é evidenciada na rede social.

Procurou-se também realizar um terceiro cenário, que realiza uma fusão do cenário baseado em tweets com o cenário baseado em notícias. Neste cenário foi utilizado apenas o XGBoost, que foi o melhor modelo obtido até então. Primeiro foi criado um modelo que utiliza tanto os sentimentos dos *tweets* quanto os sentimentos das notícias, depois foi criado um modelo que utilizou como entradas o resultado da predição baseada em *tweets* e o resultado da predição baseada em notícias. O primeiro não apresentou melhores resultados, já o segundo apresentou MDA de 75%, MDV de 94.31 e 3.38% de MAPE, o que é superior aos resultados obtidos anteriormente. A Fig. 10 apresenta a predição utilizando esse modelo de fusão.

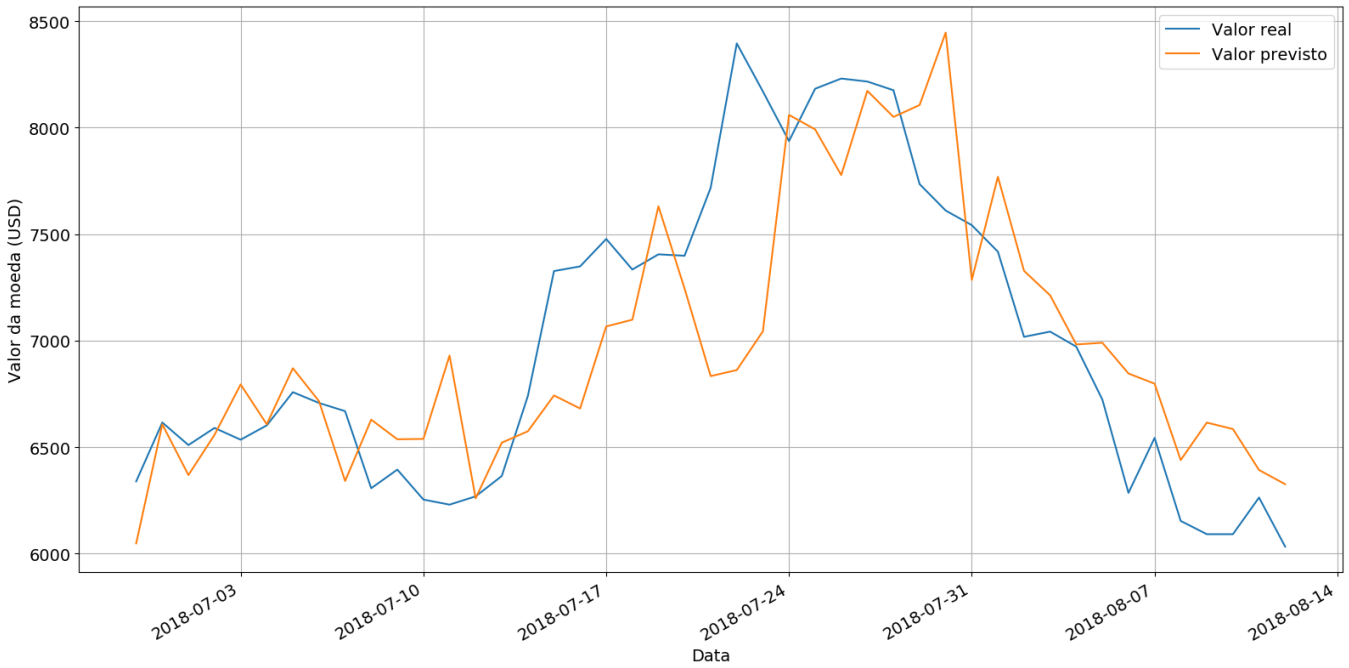

Figura 9: Predição de valores do Bitcoin com base em notícias utilizando XGBoost

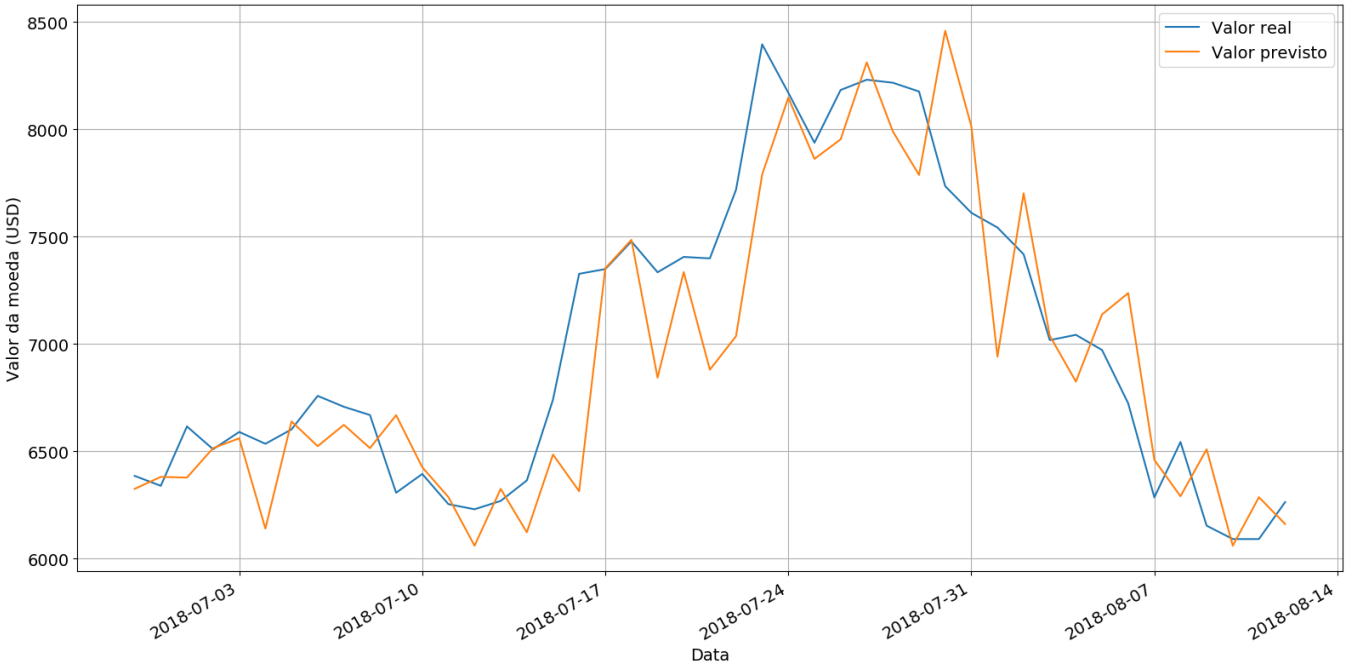

Figura 10: Predição de valores do Bitcoin com base na fusão dos dois modelos XGBoost

5.1 Conclusão

De forma similar ao trabalho realizado em Ghosh and Purkayastha (2017), este trabalho trouxe uma análise comparativa dos modelos baseados em XGBoost, SVM e Random Forest, com o diferencial de utilizar estes modelos para a predição do movimento de preços no mercado do Bitcoin a partir da análise de sentimento de tweets e notícias.

Em Valencia et al. (2019) também foram utilizados modelos baseados em SVM e Random Forest para a predição do movimento de criptomoedas com base em sentimentos de tweets e preços do Bitcoin, porém tratando o problema como classificação. Foram obtidos para SVM e Random Forest resultados de acurácia de 55% e 44%, respectivamente. Neste trabalho, obtiveram-se

resultados parecidos na acurácia do movimento dos preços (através do MDA), com resultados de 52% (SVM) e 48% (Random Forest), mesmo que aplicando estes modelos na forma de técnicas de regressão.

Assim como em Ghosh and Purkayastha (2017), o modelo do XGBoost demonstrou melhor capacidade de prever os movimentos de subida e descida em comparação com o SVM e Random Forest. Tanto utilizando como entrada os sentimentos de notícias quanto utilizando sentimentos de tweets, o XGBoost teve melhores resultados na predição do valor do Bitcoin.

A predição baseada em tweets obteve melhores resultados na métrica MDA em relação à predição baseada em notícias. Supõe-se que os melhores resultados foram obtidos a partir de sentimentos de tweets devido ao maior volume de dados gerado em um único dia, gerando mais textos a em analisados, e devido à velocidade com a qual a manifestação pública é evidenciada na rede social.

O trabalho de Li et al. (2019) apresentou também uma estratégia baseada em XGBoost para prever valores de uma criptomoeda (ZClassic) com bons resultados (coeficience de correlação de Pearson de 0.806 entre valores previstos e dados de teste). Nota-se que no trabalho atual os valores previstos ficaram percentualmente mais próximos dos valores reais. Além disso, pode-se destacar um diferencial do trabalho atual de ter aplicado filtragem de tweets de *bots*, o que auxilia a obter melhores resultados agregados da análise de sentimentos. Um ponto negativo é que, apesar de ter sido realizada uma coleta de mais dias, foram analisados menos pontos de dados já que cada dia representava um ponto de dado, diferentemente de Li et al. (2019) que utilizou uma hora como ponto de dado.

Ao analisar os resultados das métricas e o gráfico da predição do modelo de fusão, nota-se que apesar de o modelo não ter realizado uma predição muito próxima em termos de valor absoluto (MAPE de 3.38%), o modelo foi capaz de prever na maioria dos casos a movimentação de subida ou queda do Bitcoin, com resultado de 75% na métrica MDA.

5.2 Trabalhos futuros

Como possíveis trabalhos futuros, pode-se apontar:

- Realizar predições no mercado de criptomoedas alternativas (*alt-coins*), como foi feito em Li et al. (2019).
- Analisar transações de compra e venda em corretoras para identificar tendências no mercado e utilizá-las na predição.
- Implementar um robô de negociação automatizado, que poderá utilizar o resultado da predição do modelo diariamente para decidir, por exemplo, entre comprar Bitcoins ou realizar uma venda.

Referências

Bergmeir, C., Costantini, M. and Benítez, J. M. (2014). On the usefulness of cross-validation for directional forecast evaluation, *Computational Statistics & Data Analysis* **76**: 132 – 143. https://doi.org/10.1016/j.csda.2014.02.001.

Bitcoin Demographics (2018). Disponível em https://coin.dance/stats#demographics.

Blaskowitz, O. and Herwartz, H. (2009). On economic evaluation of directional forecasts, *Technical report*, Berlim. Disponível em https://ideas.repec.org/p/hum/wpaper/sfb649dp2009-052.html.

Coinanalysis (2018). Predict cryptocurrency prices based on news and historical price data. Disponível em http://tiny.cc/urivmz.

Corp., B. T. (2018). Rosette: an adaptable platform for text analytics and discovery. Disponível em https://www.rosette.com.

CryptoCompare (2018). Cccagg index methodology. Disponível em https://www.cryptocompare.com/media/12318004/cccagg.pdf.

Daultani, D. (2017). Stock predictions through news sentiment analysis. Disponível em https://software.intel.com/en-us/blogs/2017/07/14/stock-predictions-through-news-sentiment-analysis.

Farell, R. (2015). An analysis of the cryptocurrency industry. Disponível em https://repository.upenn.edu/wharton_research_scholars/130.

Ghosh, R. and Purkayastha, P. (2017). Forecasting profitability in equity trades using random forest, support vector machine and xgboost, *10th International Conference on Recent Trends in Engineering Science and Management*, Conference World, Ontário, pp. 476–486. Disponível em http://data.conferenceworld.in/Newton/Index.pdf.

Jang, H. and Lee, J. (2018). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information, *IEEE Access* **6**: 5427–5437. https://doi.org/10.1109/ACCESS.2017.2779181.

Karakoyun, E. S. and Osman, A. (2018). Comparison of arima time series model and lstm deep learning algorithm for bitcoin price forecasting, *Proceedings of MAC 2018 in Prague*, MAC Prague consulting s.r.o., Praga, pp. 171–179.

Li, T. R., Chamrajnagar, A. S., Fong, X. R., Rizik, N. R. and Fu, F. (2019). Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model, *Frontiers in Physics* **7**: 98. https://doi.org/10.3389/fphy.2019.00098.

Mern, J., Anderson, S. and Poothokaran, J. (2017). Using bitcoin ledger network data to predict the price of bitcoin. Disponível em http://cs229.stanford.edu/proj2017/final-reports/5228421.pdf.

Mittal, A. and Goel, A. (2011). Stock prediction using twitter sentiment analysis. Disponível em http://tiny.cc/rkivmz.

NewsAPI (2018). News api. Disponível em https://newsapi.org.

Pagolu, V. S., Challa, K. N. R., Panda, G. and Majhi, B. (2016). Sentiment analysis of twitter data for predicting stock market movements, *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, IEEE, Parlakhemundi, pp. 1345–1350. https://doi.org/10.1109/SCOPES.2016.7955659.

Puppeteer (2018). Puppeteer. Disponível em https://github.com/GoogleChrome/puppeteer.

Q3 2017 Cryptocurrency Report (2017). *Technical report*. Disponível em https://www.coingecko.com/buzz/q3-2017-cryptocurrency-report.

State of Blockchain 2018 (2018). *Technical report*, Nova York. Disponível em https://www.coindesk.com/research/state-blockchain-2018.

Stenqvist, E. and Lönnö, J. (2017). *Predicting Bitcoin price fluctuation with Twitter sentiment analysis*, PhD thesis. Disponível em http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-209191.

Valencia, F., Gómez-Espinosa, A. and Valdés-Aguirre, B. (2019). Price movement prediction of cryptocurrencies using sentiment analysis and machine learning, *Entropy* **21**. https://doi.org/10.3390/e21060589.

Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, *Climate Research* **30**: 79–82. Disponível em https://www.jstor.org/stable/24869236.

Zhang, L. (2013). Sentiment analysis on twitter with stock price and significant keyword correlation, *Technical report*, Austin. Disponível em https://repositories.lib.utexas.edu/handle/2152/20057.

Zhu, Y., Dickinson, D. and Li, J. (2017). Analysis on the influence factors of bitcoin's price based on vec model, *Financial Innovation* **3**(1): 3. https://doi.org/10.1186/s40854-017-0054-0.