UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO - CTC
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Rafael Calixto Ferreira de Araújo

CryptoBot: Processamento de Linguagem Natural para detecção de Sinais de *trading* automatizados

Florianópolis
2023

Rafael Calixto Ferreira de Araújo

CryptoBot: Processamento de Linguagem Natural para detecção de Sinais de *trading* automatizados

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Santa Catarina para a obtenção do título de mestre em Ciência da Computação.
Orientador: Prof. Alex Sandro Roschildt Pinto, Dr.
Coorientador: Prof. Mauri Ferrandin, Dr.

Florianópolis
2023

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

de Araújo, Rafael Calixto Ferreira
CryptoBot : Processamento de Linguagem Natural para detecção de Sinais de \textit{trading} automatizados / Rafael Calixto Ferreira de Araújo ; orientador, Alex Sandro Roschildt Pinto, coorientador, Mauri Ferrandin, 2023.
48 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2023.

Inclui referências.

1. Ciência da Computação. 2. Criptomoedas. 3. Redes Sociais. 4. Processamento de Linguagem Natural. 5. GoEmotions. I. Pinto, Alex Sandro Roschildt. II. Ferrandin, Mauri. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. IV. Título.

Rafael Calixto Ferreira de Araújo

CryptoBot: Processamento de Linguagem Natural para detecção de Sinais de *trading* automatizados

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof.(a) Jean Everson Martina, Dr(a).
Instituição PPGCC/UFSC

Prof.(a) Marcos Fagundes Caetano, Dr(a).
Instituição UNB

Prof.(a) Carlos Roberto Moratelli, Dr(a).
Instituição UFSC

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Ciência da Computação.

GOV.BR
Documento assinado digitalmente
PATRICIA DELLA MEA PLENTZ
Data: 26/04/2023 09:39:40-0300
CPF: ***.243.700-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de
Pós-Graduação

GOV.BR
Documento assinado digitalmente
ALEX SANDRO ROSCHILDT PINTO
Data: 26/04/2023 15:41:38-0300
CPF: ***.787.440-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Alex Sandro Roschildt Pinto, Dr.
Orientador

Florianópolis, 2023.

Este trabalho é dedicado à todos os meus professores,
ao meu filho Nolan, minha esposa Camila e aos meus
pais Alba e Roberto.

AGRADECIMENTOS

Agradeço em especial aos meus orientadores Alex e Mauri que me guiaram por esta jornada, agradeço também a secretaria do PPGCC por sempre me atender atenciosamente e me orientar. Agradeço aos envolvidos na criação do Dataset GoEmotions e a toda comunidade Open Source, em especial aos envolvidos nos projetos que utilizo neste projeto, como a linguagem Python e as bibliotecas Jupyter, Tweepy, Tensorflow, Pytorch, Pandas e outras. Agradeço também a plataforma Overleaf por fornecer gratuitamente uma aplicação que simplifica muito a criação de textos acadêmicos utilizando Latex. Acima de tudo, agradeço a minha esposa Camila por ter sido minha grande incentivadora durante todo o curso.

“It is wrong to blame
anyone for failing to forecast
accurately in an unpredictable world.
However, it seems fair to blame professionals
for believing they can succeed in an impossible task.”
(Kahneman, 2011)

RESUMO

A constante criação de novas tecnologias permite explorar conceitos antes ainda restritos ao campo teórico. As redes sociais são plataformas tecnológicas que permitem a geração de um volume de dados diário nunca antes visto, possibilitando captar a percepção do público geral sobre diversos temas, dentre eles ativos financeiros. Contudo, a captação da percepção coletiva não permite diretamente que seja realizada uma predição da volatilidade, havendo a necessidade de criar uma estrutura experimental que permita a transformação da percepção captada para uma métrica aplicável ao mercado financeiro. Neste trabalho foram utilizadas técnicas computacionais envolvendo Inteligência Artificial e Processamento de Linguagem Natural para criar um sistema capaz de gerar predições sobre o mercado de criptomoedas a partir da captação e processamento de dados da rede social *Twitter*. Para realizar este processo treinou-se um modelo de *Machine Learning* utilizando técnicas de *Deep Learning* e, como amostra para o treinamento do modelo, utilizou-se o *dataset* GoEmotions que possibilita o treinamento de modelos capazes de identificar 27 categorias de sentimentos, além do neutro. Contudo, a criação de um *pipeline* para o processo de extração e processamento dos dados também se fez necessário, sendo desenvolvido um algoritmo para execução desta atividade onde os dados foram transformados e estruturados para que as predições pudessem ser geradas. Com os dados captados da rede social e agora processados pelo *pipeline*, é realizada a identificação dos sentimentos presentes nos textos relacionados a cada criptomoeda, possibilitando que um score seja gerado para a predição do valor do cripto ativo no mercado a partir da percepção extraída das redes sociais. Para a geração deste *score* foi desenvolvido um cálculo experimental onde o sentimento extraído é correspondido a um valor tabelar e calculado juntamente com outros valores obtidos pelos metadados dos textos extraídos da rede social. A análise dos resultados obtidos foi realizada por meio da aplicação das principais métricas de avaliação de performances preditivas, como a acurácia, *recall*, precisão e *f-1 score*. Após comparar as predições geradas com as movimentações dos cripto ativos no mercado, foi constatado que o sistema apresentou valores para as métricas aplicadas acima de 55% para todas as criptomoedas analisadas, tendo a criptomoeda Ethereum apresentado a maior capacidade preditiva. Além disso, também foram analisadas diferentes janelas de tempo para investigar os intervalos que possuem maior capacidade preditiva, identificando os intervalos de 12 e 24 horas como os com as melhores performances.

Palavras-chave: Criptomoedas. Redes Sociais. Processamento de Linguagem Natural. GoEmotions.

ABSTRACT

The constant creation of new technologies allows us to explore concepts yet restricted to the theoric field. Social Medias are technological platforms that allow generating a volume of data never seen before, it turns possible to capture the general audience about several themes, including financial assets. However, the capture of the collective perception doesn't allow direct to forecast the volatility, being necessary to build an experimental structure to transform the perception captured into a metric that could be applied in the financial market. This work applied computational techniques with Artificial Intelligence and Natural Language Processing to create a system capable of generating forecasts over the crypto market from the capture and processing of data from the Social Media Twitter. To execute this process, a Machine Learning model was trained using techniques of Deep Learning and, as a sample to processing the training of the model, was used the dataset GoEmotions that executes the training process of models capable of identifying 27 categories of sentiment, besides neutral. However, building a pipeline to execute the extraction and processing of the data was necessary, thus was developed an algorithm to execute this activity where the data was transformed and structured to allow the forecasts to be generated. With the data captured from the social network and now processed by the pipeline, is executed the identification of the sentiments in the texts about each cryptocurrency, allowing generating a score to forecast the value of the cryptoasset in the market from the perception extracted from the Social Media. To generate this score has been developed an experimental formula where the extracted sentiment is corresponded to a table value and is computed in a formula with other values obtained from the metadata of the text extracted from the social media. The analysis of the results was done with the application of main metrics to evaluate forecast performances, such as accuracy, recall, precision, and f-1 score. After comparing the generated forecasts with the movements in the crypto market, it was found that the system presented values for the metrics applied above 55% for all the analyzed cryptocurrencies, being the cryptocurrency Ethereum presented the biggest forecast capacity. Besides that, were analyzed different time windows to investigate the intervals that have the bigger forecast capacity, identifying the intervals of 12 and 24 hours as the ones with better performances.

Keywords: Cryptocurrencies. Social Networking. Natural Language Processing. GoEmotions.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A internet é um ambiente onde muitas discussões e debates são promovidos, inicialmente concentradas em fóruns *on-line* e evoluindo para as recentes redes sociais onde as interações têm atingido velocidades e volumes nunca antes vistos. Dentre os muitos grandes debates presentes no mundo virtual desde o seu primórdio, a possibilidade de criação de uma moeda digital autônoma, nativa do ambiente virtual e que atendesse as demandas dos entusiastas do crescente ambiente virtual, atraiu o interesse não apenas de profissionais da computação, mas também de economistas renomados, como Milton Friedman (WALTON, 2014). Dentre muitas tentativas de criação deste ativo virtual, em um fórum de debates foi proposto em 2008 o sistema denominado Bitcoin (NAKAMOTO; BITCOIN, 2008), dando assim origem ao mundo das criptomoedas. Esta relação entre os debates virtuais e o próprio surgimento das criptomoedas é importante para entender como o mundo virtual possui uma forte relação com estes. No entanto, embora as criptomoedas talvez tenham sido precursoras nesta relação com o mercado financeiro, a relação das Redes Sociais com o mercado financeiro como um todo vem demonstrando cada vez mais uma forte relação, como no recente caso das ações da *GameStop* que foram fortemente afetadas por movimentos ocorridos na Rede Social *Reddit* (GIANSTEFANI; LONGO; RICCABONI, 2022). Seja no mercado tradicional ou no mercado *crypto* o poder de influência das Redes Sociais vem se mostrando cada vez mais evidente.

Em paralelo, com o crescimento da abrangência da comunicação no mundo virtual, diversos estudos surgiram buscando captar informações publicadas em diferentes portais, como sites de notícias (e comentários sobre as mesmas), fóruns e redes sociais, utilizando os dados coletados para alimentar algoritmos de predições de preços de ativos especulativos (NASSIRTOUSSI *et al.*, 2014), obtendo, em muitos casos, resultados significativos. A captação de dados públicos para predição do valor dos ativos financeiros pode seguir diferentes teorias da economia dependendo da abordagem e das fontes dos dados utilizados, sendo a captação das opiniões em redes sociais para predição do valor dos ativos uma aplicação do conceito de Sabedoria das Massas (GOLUB; JACKSON, 2010). Deste modo, a utilização de dados de interações públicas em ambientes virtuais, embora tenha surgido de forma independente ao desenvolvimento do mercado de criptoativos, parece mais promissora neste cenário do que no mercado tradicional.

Dentre as diversas técnicas aplicadas para obtenção de predições para a cotação dos ativos está a Análise de Sentimento que, em muitos casos, gera valores incorporados à formulas para criação de índices como o *fear and greed* (medo e ganância em português) (MOKNI; BOUTESKA; NAKHLI, 2022). A Análise de Sentimentos é parte do campo de estudos de Processamento de Linguagem Natural (PLN), uma

área antiga da computação que vem se desenvolvendo desde a época de Alan Turing (AGARWAL; SAXENA, 2019) e atualmente emprega recentes técnicas de Inteligência Artificial (IA) para obter resultados cada vez mais precisos. Com a aplicação da IA tornou-se treinar modelos que sejam capazes de identificar sentimentos em diferentes tipos de textos, como os textos informais publicados nas Redes Sociais. A própria classificação dos sentimentos pode variar de acordo com a literatura da psicologia, por isso este estudo se baseia na taxonomia proposta por Ekman onde os sentimentos são classificados em 27 diferentes categorias, além do neutro (EKMAN, 1992), apresentados na Tabela 3 presente no Capítulo 3. É importante ressaltar que, apesar desta categorização deixar claro que o neutro não é um sentimento em si, logo não poderia ser considerado uma categoria de sentimento, neste trabalho, para simplificar o entendimento, esta categorização será referida simplismente como 28 categorias de sentimentos. A aplicação desta categorização de sentimentos visa contribuir com a Análise de Sentimentos aplicada ao mercado de criptomoedas, pois não foi identificada na literatura nenhum outro estudo que aplique a mesma taxonomia para este propósito. Outra contribuição proposta é a aplicação de uma fórmula apresentada no Capítulo 3 baseada na fórmula para cálculo da relevância de páginas da internet, o Pagerank (BRIN; PAGE, 1998). Esta fórmula foi desenvolvida para aplicação neste trabalho e pode ser aprimorada em trabalhos futuros. Além disso, há ainda uma contribuição geral da utilização de dados obtidos com a Análise de Sentimentos para a predição de valores para criptoativos por meio da geração de *scores* de predição da volatilidade para as criptomoedas, desenvolvendo um sistema capaz de processar os dados das redes sociais e estruturá-los para sua utilização e um sistema de cálculos preditivos.

Por fim, a avaliação dos resultados é realizada utilizando métricas consolidadas na literatura, sendo estas a correlação, precisão, acurácia, *recall* e *f1-score*. Estas métricas foram aplicadas em diferentes intervalos de tempo para investigar qual janela de tempo entre as publicações nas Redes Sociais e o impacto no Mercado Financeiro poderia ter a maior capacidade preditiva. Os resultados obtidos por estas métricas foram promissores tendo todas as criptomoedas obtido uma performance preditiva acima de 55% na janela de tempo de melhor performance.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Este trabalho tem como objetivo avaliar a contribuição da identificação de sentimentos segundo a taxonomia de Ekman (1992) nas redes sociais para a geração de índices para a predição do comportamento de criptomoedas no mercado financeiro.

1.1.2 Objetivos Específicos

1. Realizar um estudo sobre a aplicação de uma classificação mais detalhada dos sentimentos encontrados nos textos de redes sociais do que os trabalhos encontrados atualmente na literatura.
2. Desenvolver um algoritmo para extração de dados relacionados à criptoativos em redes sociais.
3. Implementar um modelo de IA para a identificação de sentimentos complexos em textos de redes sociais.
4. Desenvolver um pipeline capaz de relacionar métricas fornecidas pelas redes sociais com a identificação do sentimento presente no texto.
5. Avaliar a performance do índice gerado para a predição da volatilidade das criptomoedas.

1.2 ESTRUTURA DO TRABALHO

Este trabalho foi estruturado em diferentes fases de desenvolvimento. Primeiramente foi realizada uma Revisão da Literatura, apresentada no Capítulo 2. Em seguida foi dado início ao Desenvolvimento da Aplicação, começando pela Delimitação do Escopo, definição da Metodologia e desenho da Arquitetura da Solução para enfim iniciar o desenvolvimento da aplicação, criando os módulos que serão utilizados para a análise dos dados do *Twitter*. Após este estágio foi realizado o Treinamento do Modelo de IA utilizado para a identificação dos sentimentos nos textos coletados e aplicada a fórmula para cálculo da predição do valor das criptomoedas. Por fim, foi realizada a análise dos resultados comparando os índices gerados com os dados de mercado coletados da *exchange*. Todo este processo foi registrado no Capítulo 3.

A Conclusão final do trabalho e os Trabalhos Futuros mapeados são apresentados no Capítulo 4. Os capítulos apresentados neste trabalho estão listados abaixo.

1. Introdução.
2. Revisão Bibliográfica.
3. Desenvolvimento.
4. Conclusão.

2 REVISÃO BIBLIOGRÁFICA

2.1 TECNOLOGIAS EMPREGADAS

2.1.1 Redes Sociais

As Redes Sociais são plataformas com sistemas que permitem aos seus usuários compartilharem informações diversas entre si. Cada Rede Social possui um perfil específico direcionando os seus participantes a tratar de temas específicos ou a postar conteúdos em um formato específico. Uma forte característica das Redes Sociais é a sua fácil acessibilidade, permitindo que qualquer pessoa possa fazer parte e contribuir com conteúdos diversos. Em 2004 a Rede Social *MySpace* já possuia um milhão de Usuários Ativos Mensalmente (UAM), já em 2019 o *Facebook* possuiam mais de 2,4 bilhão de UAM, demonstrando o rápido crescimento e adoção das Redes Sociais (ORTIZ-OSPINA; ROSER, 2023). As empresas criadoras das plataformas das Redes Sociais buscam gerar lucros por meio do processamento dos dados postados pelos usuários, utilizando-os para entregar materiais publicitários ou oferecer serviços diversos (SUSANTO *et al.*, 2023). No entanto, estas também disponibilizam APIs que permitem a qualquer pessoa explorar o conteúdo postado nas Redes Sociais, gerando uma grande gama de estudos e experimentos a partir do processamento do seu conteúdo. Neste estudo será utilizado a API do *Twitter* para coletar os dados desta Rede Social e, utilizando técnicas de PLN, realizar Análise de Sentimentos relacionadas as criptomoedas para predição do seu valor no mercado *crypto*.

2.1.2 Criptomoedas

As criptomoedas são algoritmos que funcionam em redes descentralizadas de usuários provendo uma plataforma que funciona como moeda digital, permitindo o armazenamento e transmissão de valores, assim como o registro histórico das transações com protocolos de segurança que garantem a integridade dos registros. O Bitcoin foi a primeira criptomoeda criada e foi publicado como um projeto *open source*, possibilitando a criação de diversas outras criptomoedas a partir de *forks* realizados no repositório do projeto e aprimoramentos adicionados, gerando um novo projeto (ANTONOPOULOS, 2014). Com a difusão das criptomoedas, diversas empresas surgiram provendo ambientes onde seria possível realizar a compra e venda de diversas criptomoedas como em uma casa de câmbio. Estas empresas, conhecidas como *exchanges*, providenciam dados sobre as negociações realizadas onde é possível se verificar a cotação de cada criptomoeda em tempo real. Com o lançamento de novas criptomoedas, foram apresentadas novas tecnologias e novos conceitos para estas, como o caso da criptomoeda Ethereum que implementou os *smart contracts* possibilitando transações automatizadas por meio de algoritmos persistidos na *blockchain*. Outras criptomoe-

das implementaram uma *blockchain* criptografada, permitindo transações anônimas, como é o caso de criptomoedas como a DASH e a Monero. Diversos outros projetos trazem diferentes tipos de inovação às criptomoedas, no entando, a criptomoeda com maior volume de *marketcap* e popularidade permanece sendo, atualmente, o Bitcoin. No presente trabalho serão analisadas apenas as criptomoedas Bitcoin, Ethereum e Binance Coin, não sendo exploradas as tecnologias relacionadas a estas, mas apenas o comportamento do seu valor no mercado financeiro.

2.1.3 Mercado Crypto

Após o surgimento das criptomoedas, estas começaram a ser comercializadas gerando uma demanda crescente no mercado. Inicialmente as primeiras transações de compra e venda foram feitas de maneira informal e não regulada por nenhum Órgão financeiro. Contudo este mercado começou a crescer rapidamente e logo começaram a surgir empresas que proveram plataformas especializadas em realizar operações de compra e venda de criptomoedas, transformando-as em ordens de compra e venda que se comunicam dentro da plataforma e são registrados formando o livro de ordens, gerando assim mais um segmento do mercado financeiro, o segmento de cripto-ativos (WĄTOREK *et al.*, 2021). Atualmente grandes empresas mantém plataformas para realizar transações de compra e venda de criptomoedas, fornecendo dados atualizados constantemente sobre os valores das cotações de cada criptomoeda que podem ser processados por meio de APIs disponibilizadas por cada empresa. Estes dados geralmente são disponibilizados em um formato chamado de OHLCV que é um acrônimo formado pelas palavras *Open*, *High*, *Low*, *Close* e *Volume*. Cada palavra destas representa uma informação de sobre o valor do ativo financeiro dentro de uma janela de tempo, sendo estes, seguindo a ordem descrita, o valor no início, o valor mais alto, o valor mais baixo, valor no final e o volume total transacionado dentro de cada janela de tempo. Este formato de dados permite a geração de gráficos de *candle*, muito utilizados no mercado financeiro em geral. Neste trabalho será realizado o processamento de dados sobre os valores das criptomoedas extraídos da API da empresa Binance para avaliação da performance preditiva do *score* gerado pelo algoritmo desenvolvido.

2.1.4 Análise de Sentimentos

A análise de sentimentos consiste na aplicação de diferentes técnicas que permitem a extração do sentimento presente em um texto. Esta técnica é amplamente empregada em redes sociais e comentários em diversos tipos de plataformas para detectar o sentimento dominante em cada comentário. Além disso, a análise de sentimentos pode ser aplicada em diferentes segmentos e em diversos negócios, como sistemas de recomendação e índices de *customer experience* (LIN; LUO, 2020). Diferentes técnicas podem ser aplicadas para gerar a análise de sentimentos, podendo

ser feita apenas com algoritmos que geram uma *bag of words* de palavras avaliadas como positivas ou negativas em um texto, gerando classificações limitadas as classes positivo, negativo e neutro por meio de um cálculo simples entre a quantidade e peso de termos positivos subtraída pela quantidade e peso de termos negativos. Outras abordagens mais sofisticadas utilizam *words embeddings* associadas a algoritmos de *machine learning* para identificação dos sentimentos (RUDKOWSKY *et al.*, 2018), podendo gerar classificações com maior número de classes de sentimento, porém ainda um número reduzindo, geralmente de três a cinco casos. Outras abordagens aplicam técnicas de PLN associadas com algoritmos de *deep learning*, conforme descrito na próxima subseção. Com o desenvolvimento de técnicas mais avançadas de IA, algoritmos vem sendo desenvolvidos para permitir a identificação de sentimentos de forma mais detalhada. Este trabalho utiliza 28 categorias de sentimento usando com base o estudo de Paul Ekman (EKMAN, 1992).

2.1.5 Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural (PLN) consiste em um conjunto de técnicas e ferramentas computacionais que possuem o objetivo de extrair informações úteis de textos. Atualmente o PLN é aplicado em diversas áreas distintas como negócios, educação, comércio, saúde, política e interações sociais (BAHJA, 2020). Este é um campo de estudos antigo na computação, tendo iniciado ainda com Alan Turing com a publicação do seu artigo "*Machine and Intelligence*"onde descreve o Teste de Turing, conhecido para avaliação do nível da IA (CHOPRA; PRASHAR; SAIN, 2013). Para realizar a extração das informações pode-se utilizar algoritmos proprietários que realizam o tratamento de *strings* e geram *outputs* por meio de árvores de decisões ou técnicas de IA, como o *machine learning* que pode funcionar com modelos de treinamento supervisionado como o *Suport Vector Machine* (SVM) e o Naïve Bayes, ou modelos de treinamento não supervisionados, como o KNN ou métodos de clusterização. Técnicas de *deep learning* também são aplicadas, sendo estas as mais eficientes atualmente para a aplicação em algoritmos de PLN. Dentre as técnicas mais aplicadas estão as redes recorrentes como *Long Short Term Memory (LSTM) network* e *Recurrent Neural Network (RNN)*, no entanto o estado-da-arte para o desenvolvimento de algoritmos de PLN são os novos modelos de representação de linguagem como o *Bidirectional Encoder Representations from Transformers (BERT)*. Há também diversas ferramentas dedicadas a esta área, como as bibliotecas *Spacy* e *NLTK* para a linguagem Python (BENITEZ-ANDRADES *et al.*, 2022). Os textos atuais que circulam pelas Redes Sociais trazem muitos desafios para a aplicação de técnicas de PLN, como a utilização de *emoticons*, imagens que complementam a mensagem do texto, diferentes expressões e gírias de acordo com cada grupo ou contexto e erros de sintaxe, propositais ou não (FELDMAN, 2013). Contudo o crescimento de empresas que têm como negócio a

extração de dados da internet, como no caso das empresas provedoras das Redes Sociais, vêm fomentando o desenvolvimento de novas técnicas e estudos sobre PLN, buscando assim ser mais eficiente na geração de valor em diferentes tipos de negócios (SRINIVASA-DESIKAN, 2018).

2.1.6 PageRank

A equação denominada *PageRank* foi criada para gerar um *ranking* das páginas mapeadas pelo algoritmo da Google na *web*. O intuito da equação é gerar um *score* para cada página de acordo com o número de menções encontradas em outras páginas e ao *PageRank* de cada página onde a mensão foi encontrada. Sendo assim, quando maior o número de *links* encontrados na *web* que levem a uma determinada página, maior será considerado o grau de relevância da página na *web*. As variáveis que compõem a equação são as seguintes, *PR* que é o valor do *PageRank*, *C* que é o número de *links* encontrados na página, *A* é a página que está sendo *ranqueada*, *T* é a página verificada em busca de *links* e *d* que é uma valor de amortização, sendo $0 > d < 1$. A equação então se dá da seguinte forma: $PR(A) = (1 - d) + d(\Sigma \frac{PR(Tn)}{C(Tn)})$ (BRIN; PAGE, 1998). Com base nessa lógica de considerar o número de menções a uma página e o peso de cada menção, neste trabalho foi desenvolvida uma equação que gera um *score* de predição para o valor das criptomoedas com base no número de menções e ao peso de cada menção, calculado com base nos metadados do *tweet*, que será detalhada na seção 3.6

2.1.7 Métricas

As métricas utilizadas para avaliar a performance das predições geradas pelo algoritmo são métricas comumente aplicadas para avaliar a performance de algoritmos de *machine learning* e outras formas de geração de predições. Estas métricas são geradas com base na matriz de confusão onde são contabilizados os casos de verdadeiro positivo (VP), falso positivo (FP), verdadeiro negativo (VN) e falso negativo (FN), conforme apresentado na Tabela 1. (YACOUBY; AXMAN, 2020). A acurácia, gerada pela fórmula ($A = \frac{\Sigma(VP+VN)}{\Sigma(VP+FP+FN+VN)}$) demonstra uma performance mais geral para a predição, considerando todos os resultados obtidos. Já a precisão, gerada pela fórmula ($P = \frac{\Sigma VP}{\Sigma(VP+FP)}$), se concentra na assertividade considerando apenas os casos positivos. Esta métrica é relevante quando se deseja analisar predições em que os casos negativos não causam um grande impacto, não havendo grandes prejuízos ao deixar de identificar casos positivos. Porém, quando um caso positivo é identificado, é necessário que ocorra uma baixa taxa de erro, pois uma identificação incorreta pode causar grandes danos. O *recall*, gerado pela fórmula ($R = \frac{\Sigma VP}{\Sigma(VP+FN)}$), considera os casos de identificações realizadas corretamente e de identificações que não foram realizadas. Neste caso, deixar de identificar um caso positivo causa um dano maior do

que identificar de forma incorreta. Este caso pode ser ilustrado pela situação quando um médico prescreve um remédio para um paciente mesmo sem estar completamente seguro sobre a doença que o atinge, pois sabe que o remédio não causará danos a sua saúde. Por fim, o *f1-score*, dado pela fórmula ($F = \frac{2xPxR}{P+R}$), busca um equilíbrio entre a precisão e o *recall*, mas elevando mais o peso de casos onde o caso positivo não foi identificado.

Tabela 1 – Matriz de Confusão

	Cima	**Baixo**
Cima	Verdadeiro Positivo (VP)	Falso Positivo (FP)
Baixo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

2.2 FUNDAMENTAÇÃO TEÓRICA

Buscando por revisões sistemáticas realizadas anteriormente sobre o tema em portais voltados para publicações acadêmicas como o *Journal of Computer Science*, *IEEE* e o *Google Scholar*, foi encontrado no artigo (NASSIRTOUSSI *et al.*, 2014) uma revisão focada na aplicação de técnicas de mineração de textos para predição de preços de ativos em diferentes mercados financeiros. Esta revisão é importante para contextualização das principais abordagens, técnicas, tecnologias e fontes de dados utilizadas. Porém, como a publicação do trabalho ocorreu em 2014, é necessário realizar atualizações para incorporação de elementos recentes que surgiram nesta área.

Buscando por artigos mais recentes foi identificado em (GROSS-KLUSSMANN; KÖNIG; EBNER, 2019) uma abordagem utilizando a análise de sentimentos para geração de um índice para ser utilizado como massa de dados para o treinamento de modelos de *machine learning* utilizando dados do *Twitter* e aplicando diferentes técnicas. O artigo ressalta que este tipo de abordagem ainda possui poucos estudos ao relatar que não foram encontrados trabalhos semelhantes utilizando a mesma fonte de dados e técnicas aplicadas para a predição de valores de ativos no mercado financeiro tradicional. Embora o artigo traga muitas semelhanças com a proposta deste trabalho, algumas diferenças fundamentais podem ser observadas, como o fato de não ter sido utilizado uma massa de dados classificados para o treinamento do modelo de *machine learning*, aplicando assim técnicas de aprendizagem não supervisionadas como *K-Means* para gerar dados para treinar um modelo *Support Vector Machine* (SVM). No trabalho apresentado por Gross-Klussmann et al. (2019) foi realizada a identificação que quais usuários do *Twitter* possuíam maiores taxas de assertividade para que estes tivessem maior peso no cálculo da predição do valor dos ativos, no entanto, no presente trabalho este desenvolvimento ficou mapeado apenas como um possível trabalho futuro.

No trabalho (RAHEMAN *et al.*, 2022) é realizada uma abordagem semelhante a este trabalho, utilizando como fonte de dados o Twitter e o Reddit e aplicando diferentes abordagens de PLN com técnicas baseadas em *lexicons* e textitn-grams para gerar uma análise de sentimentos que é aplicada como um componente no cálculo da predição do valor das criptomoedas. Outro ponto em comum é a utilização de um modelo treinado a partir de uma massa de dados com avaliações de textos de redes sociais para servir como classificador. O algoritmo BERT também foi utilizado no trabalho apresentado por Raheman et al. (2022), contudo foram utilizados datasets com dados extraídos de redes sociais, porém não classificados em uma taxonomia específica, como no presente trabalho. Por fim, este artigo utiliza a classificação de sentimentos apenas como positivo, negativo, neutro e contraditório que é limitada por não explorar nuances de cada sentimento, enquanto no presente trabalho é utilizada a classificação em 28 categorias de sentimento permitindo criar um sistema de pesos que permitirá um cálculo mais detalhado e que poderá receber ajustes finos posteriormente.

Com o foco apenas no processo de análise de sentimentos, em (KANE *et al.*, 2022) é realizado o treinamento de um modelo para reconhecimento das emoções utilizando como massa de dados para o treinamento o *dataset* GoEmotions (DEMSZKY *et al.*, 2020) e a arquitetura BERT (DEVLIN *et al.*, 2018). Este estudo traz semelhanças com o processo de treinamento do modelo proposto, porém se limita apenas a geração do modelo e ao teste de identificação dos sentimentos segundo a taxonomia de Ekman (1992) enquanto o presente trabalho utiliza os sentimentos identificados para aplicação em uma fórmula de precisão do comportamento das criptomoedas no mercado financeiro e avalia os resultados obtidos.

No trabalho de Aslam (2022) é realizada a abordagem mais próxima ao trabalho proposto que foi identificada na revisão de literatura. A utilização do Twitter como fonte de dados, o foco nas criptomoedas, a abordagem de detecção dos sentimentos relacionados às criptomoedas e a aplicação de técnicas de *deep learning* para geração de um modelo treinado para reconhecimento dos sentimentos são pontos em comum em ambos os trabalhos. No entanto, os sentimentos foram classificados apenas como positivos e negativos, gerando um *score* que varia de -1 à 1 para a classificação de cada sentimento presente no texto, sendo o valor do *score*, quanto mais próximo à -1, maior a prevalência do sentimento negativo e quanto mais próximo ao valor 1, maior a prevalência do sentimento positivo. Este sistema resulta em uma classificação em apenas três categorias, sendo estas positivo, negativo e neutro, ocorrendo de forma gradativa por utilizar o intervalo que varia de -1 à 1. Em comparação ao presente trabalho, aqui será utilizada uma classificação mais específica aplicando a categorização de 28 sentimentos permitindo que sejam testados diferentes pesos para cada categoria diferente. Com isso, busca-se deixar em aberto a possibilidade de aplicar diferentes combinações de valores para os pesos das categorias, buscando encontrar

a combinação que ofereça a melhor performance preditiva para o *score*.

Outros dois artigos que realizam levantamentos relacionados ao tema do presente trabalho foram identificados, como o trabalho apresentado por Özkaynar (2022) que explora diversas tecnologias recentes aplicadas no mercado financeiro, como as criptomoedas e suas tecnologias associadas, como a Blockchain, os *Smart Contracts* e os *Non-Fungible Tokens* (NFTs). Também são abordadas tecnologias como o Metaverso e os mercados que se utilizam das Redes Sociais. A abordagem utilizada no artigo difere do presente trabalho por focar em analisar o comportamento dos consumidores frente aos produtos do mercado financeiro que se utilizam das tecnologias citadas, não investigando possíveis ferramentas de predição do mercado financeiro como no presente trabalho. Em Tran (2022) é realizada uma investigação das técnicas de PLN e abordagens empregadas em trabalhos de análise de sentimentos para a predição de valores de criptomoedas.

Todos os trabalhos mencionados, incluindo os mais relevantes presentes na revisão bibliográfica, foram organizados na Tabela 2. A revisão da literatura demonstra que existem diversos artigos analisando distintas formas de captação e avaliação de textos publicados na internet para geração de predição de valores de ativos no mercado financeiro. Contudo, também é demonstrado que diversas técnicas e abordagens são possíveis, podendo variar a fonte de dados, as tecnologias utilizadas para processamento dos dados, os ativos avaliados e o método de predição do seu valor. Por isso, o resultado da busca por artigos semelhantes, não surpreende ao retornar trabalhos com diversos pontos em comum, porém sem nenhum deles cobrir a mesma abordagem adotada neste trabalho.

Neste capítulo foram apresentadas as principais tecnologias que serão exploradas neste trabalho assim como os principais conceitos teóricos que foram identificados por meio da revisão bibliográfica. Este conteúdo fundamentará todo o experimento realizado e sua avaliação, sendo fundamental para gerar uma análise crítica sobre todo o trabalho desenvolvido.

Tabela 2 – Revisão bibliográfica

Referência	**Tipo de texto**	**Fonte de dados**	**Nº de textos**	**Sentimentos**
(TRAN, 2022)	Tweets e posts	Notícias e Redes Sociais	NM	Diversas Abordagens
(ÖZKAYNAR, 2022)	Notícias	Notícias	NM	Não utiliza Análise de Sentimentos
(RAHEMAN *et al.*, 2022)	Tweets e posts	Twitter e Reddit	100,000	Positivo, Negativo, Neutro e Contraditório
(KANE *et al.*, 2022)	posts	Reddit	58,000	Raiva, Aversão, Medo, Alegria, Neutro, Tristeza e Surpresa
(ASLAM *et al.*, 2022)	Tweets	Twitter	40,000	Felicidade, Tristeza, Surpresa, Raiva e Medo
(SEONG; NAM, 2021)	Notícias	Notícias	1,397,800	Positivo e Negativo
(GROSS-KLUSSMANN; KÖNIG; EBNER, 2019)	Tweets	Twitter	6,800	Positivo e Negativo
(BURNIE; YILMAZ, 2019)	Submissions	Reddit	338,415	NA
(ZHANG *et al.*, 2016)	Post	Weibo	139,855	Positivo, Negativo e Neutro
(LIU *et al.*, 2015)	Tweets	Twitter	NM	Positivo e Negativo
(CHATRATH *et al.*, 2014)	Macroeconomic news	Bloomberg	NM	NA
(JIN *et al.*, 2013)	Notícias Gerais	Bloomberg	361,782	Positivo e Negativo
(YU; DUAN; CAO, 2013)	Daily conventional e Social Media	Blogs, forums, notícias e Twitter	52,746	NA
(VU *et al.*, 2012)	Tweets	Twitter	5,001,460	Positivo e Negativo

A sigla NM é uma abreviação para Não Mencionado
A sigla NA é uma abreviatura de Não se Aplica. Nestes casos os artigos não aplicavam Análise de Sentimentos nas análises, eles realizaram um mapeamento de termos específicos relacionados as notícias para realizar as predições

3 DESENVOLVIMENTO

Neste capítulo é apresentada a delimitação do escopo, a metodologia e os resultados obtidos no desenvolvimento do projeto de pesquisa. A delimitação do escopo apresenta os elementos que serão trabalhados neste estudo e a motivação que levou a sua seleção. Em um universo tão vasto como o deste campo de estudo, explorar todos os elementos possível demandaria um esforço muito maior do que o suportado para a produção deste estudo, então esta reflexão se faz necessária para que sejam englobados os elementos que possam trazer um cenário representativo para o desenvolvimento deste estudo e que permitam futuras expansões para o mesmo. Então, será detalhada a arquitetura da solução desenvolvida para suportar todos os elementos que compõem a aplicação. A metodologia, apresentada na sequência, nos permite traçar um planejamento para os experimentos que serão realizados, permitindo a aplicação da metodologia científica para o desenvolvimento do estudo. As seções seguintes, Captação dos dados e Processamento dos dados, detalham o *pipeline* desenvolvido do momento da captação dos dados, passando pelos tratamentos recebidos e chegando até a estrutura que será utilizada no cálculo da predição do valor das criptomoedas. Desta forma, a próxima seção apresentada é Cálculo do Índice de Predição detalhando a fórmula proposta por este trabalho que possui o intúito de gerar predições para o mercado das criptomoedas. São apresentadas ainda as seções Disponibilização dos Dados, que detalha como os dados gerados foram armazenados para realizar análises futuras, e Treinamento do Modelo, que descreve todo o processo de treinamento do modelo de *deep learning* utilizando o *dataset* GoEmotions. Por fim, na seção Resultados é realizada a análise dos resultados obtidos nos permitindo avaliar o impacto do estudo no mercado financeiro das criptomoedas, a análise é realizada aplicando métricas que nos permitam comparar os resultados com outros estudos semelhantes ou com desdobramentos futuros deste mesmo trabalho. O capítulo se encerra com a seção Discussão que avalia todo o experimento e as possíveis contribuições deste trabalho para as áreas de conhecimento envolvidas.

3.1 DELIMITAÇÃO DE ESCOPO

Atualmente existem diversas plataformas de mídias sociais, cada um delas possui suas características de formato, perfil de usuários e temas comuns para as publicações. A análise desenvolvida neste trabalho selecionou a rede social que será utilizada como fonte de dados com base nos seguintes critérios:

1. Número de usuários;
2. Volume de debates relacionados às criptomoedas (sendo considerado relevante para este estudo apenas debates que citem diretamente as criptomoe-

das);

3. Acessibilidade aos dados (textos) e metadados do *post* das Redes Sociais.

Para selecionar as plataformas candidatas segundo o critétio (1), foi utilizado como referência o *ranking* das plataformas com maior volume de UAM (STATISTA, 2022). Com isto, foi formada uma lista de Redes Sociais composta por Facebook, Sina Weibo, Twitter, Reddit e Quora. Considerando os critérios (1) e (3), foram descartados os fóruns sobre criptomoedas pela quantidade inferior de usuários e pela falta de metadados quando comparados às redes sociais.

Utilizando o critério (2) foi realizada a exclusão do Facebook devido a falta de debates relevantes focados no mercado de criptomoedas. Por fim, utilizando o critério (3) foi excluída a plataforma Sina Weibo por ser focada em textos utilizando idiomas provenientes da China, fazendo com que o conteúdo esteja restrito a região, por serem idiomas ainda pouco difundidos globalmente. Com a lista resultante contendo as plataformas Twitter, Reddit e Quora, foi selecionada apenas a primeira plataforma para aplicação deste estudo, podendo ser as demais exploradas em estudos futuros.

Dentre os possíveis idiomas a serem explorados neste estudo foi selecionado apenas o inglês pela maior disponibilidade de modelos de IA neste idioma e a maior multiculturalidade presente nos textos publicados nas redes sociais já que este idioma é amplamente difundido e dominado por pessoas de diversos países, mesmo aqueles onde não é o idioma oficial do país. Outros idiomas podem ser incorporados aos estudos em desdobramentos futuros.

Para selecionar as criptomoedas que serão monitoradas na análise nas redes sociais foi realizado o levantamento com base no *ranking* das criptomoedas com maior *market cap* (COINMARKETCAP, 2022) e utilizado como único critério de exclusão as criptomoedas que possuem lastro em outro ativo, como é o caso da categoria de criptomoedas lastreadas em moedas fiduciárias conhecidas como *stablecoins*. Este foi o único critério selecionado para as criptomoedas por ser o único caso onde o comportamento do seu valor é, na verdade, reflexo do comportamento de outro ativo. Deste modo, a lista inicial de criptomoedas é formada por Bitcoin, Ethereum, Tether, Binance Coin (BNB) e USD Coin. Duas destas criptomoedas são classificadas como *stablecoins*, Tether e USD Coin, sendo excluídas da análise. A lista final de criptomoedas a serem monitoradas neste estudo consiste em Bitcoin, Ethereum e Binance Coin (BNB).

A performance dos índices gerados, isto é, a assertividade da predição gerada pelo algoritmo, foi comparada aos valores do mercado cripto obtidos da *exchange* Binance. Esta *exchange* foi selecionada para comparação dos valores, isto é, o valor de predição gerado pelo algoritmo desenvolvido neste trabalho e os valores de mercado registrados pela *exchange* por ser a única a conter todas as três criptomoedas. Uma particularidade desta *exchange* é que ela não possui o valor da cotação das moedas

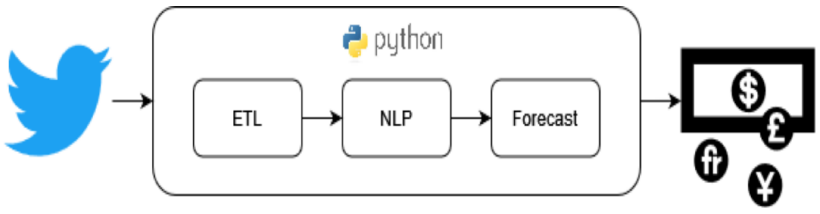
diretamente para o dólar, contudo é possível fazer a cotação para a criptomoeda Tether que é lastreada em dólar, sendo assim fornecida uma cotação em dólar de forma indireta. Como o valor da cotação das criptomoedas possui pouca variação entre as *exchanges*, optou-se por utilizar apenas os valores obtidos desta *exchange*.

Para a análise dos resultados foram utilizadas métricas populares para processo de aferição da performance de *scores* preditivos (YACOUBY; AXMAN, 2020), como a correlação, acurácia, precisão, *f1-score* e *recall*. Este trabalho não envolve a investigação de graus de causalidade entre as redes sociais e o mercado de criptomoedas, podendo este estudo, assim como a aplicação de outras métricas para aferição da performance, serem realizados em trabalhos futuros.

3.2 ARQUITETURA DA SOLUÇÃO

A solução desenvolvida neste trabalho consiste em um *pipeline* de dados com uma estrutura para aquisição dos dados, um algoritmo de processamento dos dados e uma estrutura de entrega das predições geradas. As próximas seções detalham as etapas realizadas para o desenvolvimento e avaliação da aplicação. A Figura 1 ilustra a arquitetura da aplicação proposta, assim como a maneira que os diferentes componentes se comunicam. Nela estão representadas diferentes etapas no processo de geração do índice para predição. A primeira etapa, nomeada como ETL (termo apropriado do conceito de *Extract, Transform and Load*), responsável pela aquisição, formatação e estruturação dos dados no Twitter e posterior envio dos mesmos para o processo de PLN. A segunda etapa, nomeada de NLP (Natural Language Processing, termo inglês para o Processamento de Linguagem Natural), consiste nos processos relativos ao tratamento e identificação das emoções presentes no texto. Por fim, a etapa nomeada de *Forecast* consiste na realização do cálculo para predição da volatilidade das criptomoedas.

Figura 1 – Arquitetura da Aplicação



Fonte: Do Autor

3.3 METODOLOGIA

Este trabalho se iniciou com uma revisão da literatura envolvendo a utilização de textos coletados da internet para predição de preços de ativos e investimentos. Como

foi detalhado no capítulo de Revisão da Literatura, foi encontrada uma revisão sistemática sobre o tema (NASSIRTOUSSI *et al.*, 2014) com um levantamento consistente sobre o tema, porém com alguns anos desde a sua realização. Desta forma, optou-se por utilizar a revisão sistemática como base e atualizar a mesma com novas referências encontradas. Buscando uma maior proximidade com o tema desenvolvido neste trabalho, foram identificados artigos que aplicaram técnicas de análise de sentimentos para a predição de preços de criptoativos. É importante ressaltar que os artigos mais recentes também sofrem o impacto causado pelo cenário da pandemia de COVID-19, tendo os resultados relacionados ao comportamento do mercado referente à este período específico (MOKNI; BOUTESKA; NAKHLI, 2022).

Após a realização do levantamento bibliográfico foi iniciado o desenvolvimento da aplicação, selecionando as ferramentas à serem aplicadas. Primeiramente será apresentada uma visão geral sobre a aplicação e as subseções a seguir servirão para o detalhamento de cada componente desenvolvido. O primeiro estágio para o desenvolvimento do *pipeline* de dados da aplicação se deu com a definição das ferramentas e tecnologias para a construção da estrutura responsável pela captação de dados do Twitter, desta forma foi selecionada a biblioteca *open source* da linguagem Python, *tweepy*. Após a conclusão do desenvolvimento desta estrutura, a aplicação já possuía uma função capaz de fornecer continuamente (em *loop*, não em *streaming*) dados capturados do Twitter em um formato JSON contendo a mensagem extraída e os metadados referentes a mesma. Contudo, antes de realizar a análise dos dados era necessário realizar alguns tratamentos nos mesmos, por isso foi desenvolvido uma função para recebimento e estruturação dos dados, selecionando apenas aqueles a serem utilizados para o cálculo da predição da volatilidade, realizando também o pré-processamento de alguns dados para ajustes relacionados a sua formatação. O objeto devolvido pela função *search_tweets* da biblioteca *tweepy* é do tipo dicionário, contendo como chaves os metadados e o texto de cada *tweet*. A quantidade de metadados fornecida pela API do Twitter é bastante volumosa (ROESSLEIN, 2009), porém neste trabalho foram utilizados apenas os seguintes metadados: *user.id_str*, *user.screen_name*, *user.followers_count*, *retweet_count*, *favorited*, *created_at*, *id_str*, *entities.hashtags*, *entities.user_mentions*, *entities.urls*, *metadata.result_type* e *retweeted_status*. Estes metadados foram selecionados por permitirem a identificação do usuário autor do *tweet* e por trazer dados que serão usados na fórmula desenvolvida e apresentada na seção 3.6. Neste estágio foram aplicadas as técnicas de remoção de *stop words* (palavras descartadas por serem consideradas sem valor em processos de PLN) e conversão das strings com valores de datas para o formato *timestamp*, utilizando funções *core* do Python, que permitirá a sua utilização na equação apresentada mais adiante.

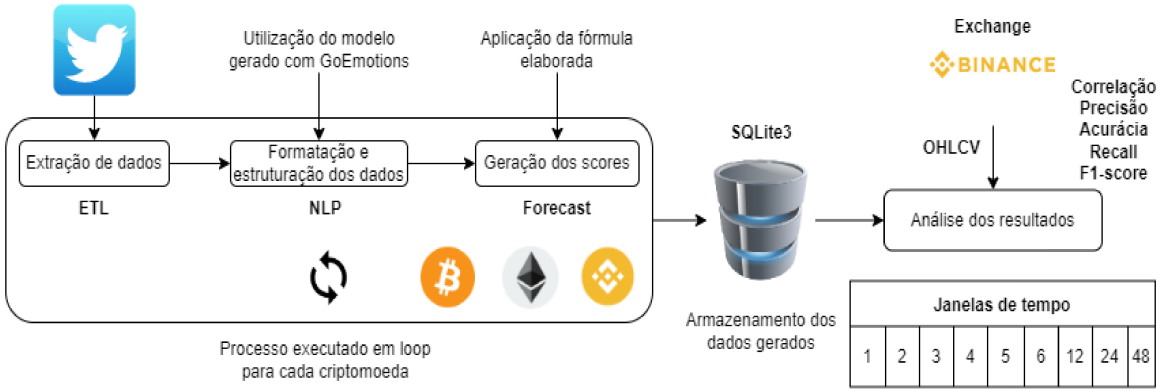
A estrutura desenvolvida para o recebimento dos valores consiste em uma

classe com atributos onde são armazenados os valores a serem utilizados no cálculo de volatilidade, dentre estes valores está o valor do peso relacionado ao sentimento encontrado no texto, sendo gerado a partir da identificação dos sentimentos sentimentos presentes no texto utilizando um modelo de IA e a correspondência do sentimento identificado ao seu valor atribuído, detalhado na Tabela 3. Para gerar este modelo de IA, foi realizado o desenvolvimento de um algoritmo para treinamento de um modelo voltado para PLN utilizando a arquitetura BERT (DEVLIN *et al.*, 2018) e o dataset GoEmotions (DEMSZKY *et al.*, 2020), este processo será detalhado na subseção 3.8. Com o modelo gerado, foram obtidos os indicadores e gráficos relacionados à performance durante processo de treinamento, sendo utilizados mais tarde no capítulo de conclusão deste trabalho. Com o modelo treinado, este foi adicionado ao *pipeline*, representado na Figura 2, realizando o reconhecimento dos sentimentos presentes nos *tweets* e adicionando ao atributo da classe desenvolvida.

Tabela 3 – Sentimentos

índice	Sentimento	Peso
0	Admiração	8
1	Diversão	4
4	Aprovação	7
7	Curiosidade	5
8	Desejo	9
13	Excitação	10
15	Gratidão	6
17	Alegria	1
18	Amor	2
20	Otimismo	10
21	Orgulho	8
22	Realização	3
23	Alívio	4
24	Remorso	-10
25	Tristeza	-3
26	Surpresa	1
2	Raiva	-7
3	Aborrecimento	-9
5	Cautela	-8
6	Confusão	-5
9	Decepção	-10
10	Desaprovação	-9
11	Nojo	-5
12	Constrangimento	-6
14	Medo	-8
16	Pesar	-2
19	Nervosismo	-4
27	Neutro	0

Figura 2 – Fluxo geral de atividades



Fonte: Do Autor

Após o objeto ser instanciado com os valores carregados nos atributos uma função para realização do cálculo da volatilidade (para cada criptomoeda) recebe os valores necessários para a predição e retorna um *score* que será utilizado para tentar gerar predições sobre o valor da criptomoeda no mercado. Com o *score* gerado será buscada uma relação entre o valor do *score* e o valor de mercado das criptomoedas, podendo o *score* representar a porcentagem de valorização ou desvalorização do criptoativo em relação ao seu valor atual, o valor bruto em dólares para valorização ou desvalorização, podendo estar em escalas diferentes dependendo do valor atual da criptomoeda ou apenas uma indicação da direção que o valor irá seguir (valorização x desvalorização). Para investigar as duas primeiras possibilidades será verificada a correlação entre os *scores* gerados, pois para que este corresponda ao percentual ou o valor bruto de valorização ou desvalorização, é necessário que exista uma alta correlação entre o *score* gerado e o valor histórico de mercado das criptomoedas. Então, este *score* será armazenado em um banco de dados para que as predições possam ser resgatadas e comparadas com a dinâmica apresenta pelo mercado de criptoativos para as criptomoedas monitoradas (Bitcoin, Ethereum e Binance Coin).

A aferição da performance dos índices gerados com o histórico do mercado foi realizada por meio da comparação com oito janelas de tempo diferentes, sendo elas 0, 1, 2, 3, 6, 12, 24 e 48 horas de diferença entre a predição e os dados do mercado. Estas janelas de tempo foram estabelecidas por entender que a influência das redes sociais no mercado de criptomoedas pode demorar a ter efeito e esse intervalo de tempo precisa ser investigado. Para cada intervalo de tempo foram aplicadas as seguintes métricas: Correlação, acurácia, *recall*, precisão e *f1-score*. Estas métricas foram aplicadas por serem populares e amplamente utilizadas em diferentes aferições de performance (YACOUBY; AXMAN, 2020).

Por fim, após a análise dos resultados, foi realizada a documentação descrevendo todo o processo e os resultados obtidos.

3.4 CAPTAÇÃO DOS DADOS

O Twitter oferece uma API que permite realizar a captura de *tweets* na plataforma segundo os critérios de busca definidos pelo usuário. Para facilitar a interação com a API do Twitter foi utilizada a biblioteca tweepy (ROESSLEIN, 2009) para a linguagem Python que oferece diversas funções para aquisição e tratamento de dados. A busca inicial por textos relacionados às criptomoedas na plataforma é realizada utilizando a função *search_tweets*, retornando os *tweets* identificados pela busca no formato descrito na seção 3.3. Os parâmetros inseridos na função para realização da busca por *tweets* relacionados são os seguintes:

- Termo buscado, onde será passado como parâmetro o nome de cada criptomoeda (Bitcoin, Ethereum e Binance Coin) a cada iteração do algoritmo;
- Idioma dos tweets, que foi definido como inglês pelas delimitações deste trabalho;
- Número de resultados máximos retornados, sendo definido o valor de 200 que é o valor máximo permitido pela API;
- Modo de apresentação dos *tweets*, sendo este parâmetro definido como *extended*, indicando para a função que deverá ser retornado todo o conteúdo do *tweet* e não apenas uma amostra, que é o parâmetro *default*.

3.5 PROCESSAMENTO DOS DADOS

O processamento dos dados é onde de fato ocorre a parte mais complexa da aplicação. Após os dados dos *tweets* terem sido coletados, estes serão processados por uma função que realizará o tratamento e armazenamento dos dados que serão utilizados para o cálculo da predição. Como resultado é retornado uma estrutura em forma de dicionário conforme descrito na seção 3.3. Este dicionário retornado contém os seguintes dados:

- Identificador do usuário;
- Nome de exibição do usuário;
- Número de seguidores que o usuário possui;
- Número de usuários que favoritaram a conta do criador do *tweet*;
- Identificador do *tweet*;
- Número de *retweets* que a mensagem possui;
- Data de criação em timestamp;
- Texto do *tweet* tratado;
- Hashtags incluídas;

- Usuários mencionados;
- URLs incluídas;
- Tipo de *tweet*;
- Marcação para saber se a mensagem é original ou um *retweet*;
- Sentimentos presentes.

Para realizar a análise dos textos é necessário executar o tratamento dos mesmos retirando *stop words*, pontuações, aspas, marcações de *retweet*, quebras de linha, URLs, e-mails, menções à outros usuários e outros componentes que possam gerar ruídos na execução do PLN. Para isso é utilizada a biblioteca Spacy (SPACY, 2023) que é um *toolkit* para Processamento de Linguagem Natural, fornecendo ferramentas que facilitam o tratamento do texto. Contudo o Spacy não possui nenhuma ferramenta para tratamento de *emoticons* no texto, por isso foi utilizada a biblioteca Emoji que permite a fácil identificação e remoção dos *emojis* existentes no texto. Desta forma, após o texto ser formatado pelo Spacy, são também removidos possíveis *emojis* e o texto tratado é finalmente adicionado a estrutura final. Por fim, o processamento dos dados também converte a data de criação do *tweet* de um valor *datetime* em *string* para uma valor inteiro em *timestamp* (em segundos) utilizando a biblioteca nativa do Python *datetime*.

A análise dos sentimentos presentes no texto é realizada por um modelo de IA, indicando quais emoções dentre as 28 categorizadas estão presentes no texto. O processo de treinamento do modelo e demais especificidades serão detalhados na seção Treinamento do Modelo. Com todos os valores gerados e convertidos para o formato adequado, o cálculo de predição é realizado por uma função que recebe estes valores como *input* e retorna o *score* de predição.

3.6 CÁLCULO DO ÍNDICE DE PREDIÇÃO

O cálculo do índice para a predição da volatilidade das criptomoedas se dá pela média dos valores extraídos de cada *tweet* para cada uma das criptomoedas que é monitorada. Cada *tweet* tem o valor inicial igual a 1 e é multiplicado segundo a fórmula Equação (1).

$$P = \frac{(1 + (Nf * 0{,}001) + (Nrt * 0{,}01)) * (S * 0{,}1)}{1 + ((Dtn - Dtc) * 0{,}001)} \tag{1}$$

Onde P é a predição, Nf é o número de seguidores, Nrt é o número de *retweets*, S é o sentimento presente, Dtc é a data de criação do *tweet* e Dtn é a data atual, ambos em *timestamp*. Todas as variáveis sofrem valores de amortização de acordo com o impacto que estas devem ter na fórmula, seguindo sempre a escala de divisão por 10^n. Isto é, Nf tem um peso de 10^{-3} somando-se a Nrt que tem um peso de 10^{-2}. Estes

pesos foram obtidos levando em consideração a contribuição destas variáveis para a difusão do *tweet* e a escala de cada valor na plataforma Twitter. Isto é, enquanto cada *tweet* costuma ter um número muito maior de seguidores ligados ao seu autor do que um número de *retweets*, o número de *retweets* é uma métrica muito mais significativa para avaliar o quanto a mensagem foi difundida do que o número de seguidores do autor do *tweet*. Por fim, a variável S recebe uma amortização de 10^{-1} apenas para diminuir a escala do valor geral do índice gerado e o valor do divisor recebe uma amortização para transformar o valor em *timestamp* para milissegundos.

A fórmula desenvolvida é uma adaptação do algoritmo PageRank (BRIN; PAGE, 1998) utilizado para gerar *scores* indicadores da relevância de uma página *web* pela Google no início de suas atividades. Da mesma forma que a fórmula original tem o intuito de identificar a relevância de uma página *web* para realizar o seu ranqueamento, a adaptação da fórmula neste trabalho visa identificar a percepção da relevância das criptomoedas entendendo que esta relevância percebida também se reflete em ações no mercado crypto. O ponto diferencial entre o universo do ranqueamento das páginas *web* e a indentificação da volatilidade das criptomoedas é que o primeiro possui apenas valores positivos, sendo $P \in \mathbb{Q}_{>0}$, enquanto o segundo pode possuir tanto valores positivos quanto negativos, sendo $P \in \mathbb{Q}$. Isto é, as páginas *web* possuem o seu valor de relevância recalculado constantemente tomando como base o valor zero, enquanto as criptomoedas tem o seu valor atual tomado como base e então o cálculo realizado indica a direção do movimento (valorização ou desvalorização) da criptomoeda.

A Equação (1) considera o número de seguidores que o usuário, autor do *tweet*, possui. Este valor é um indicador do potencial de alcance que a sua publicação possui, sendo amortizado pela constante 0,001 para que sejam adicionados valores significativos a cada mil seguidores que o perfil possua. Após a amortização o valor é somado ao número de *retweet* da mensagem. Este valor demonstra o quanto a mensagem está sendo compartilhada na rede, sendo também amortizado ao ser multiplicado pela constante 0,01, sendo considerado um valor expressivo a cada centena de *retweets*. O valor obtido a partir da soma entre estas duas variáveis é multiplicado pelo peso do sentimento que sofre uma amortização de 0,1 para aproximar o valor da escala da primeira casa decimal. Por fim, o resultado obtido pelo cálculo é dividido pelo tempo em segundos que o *tweet* foi realizado amortizado pela constante 0,001 e somado à constante 1, fazendo com que o valor gerado pelo dividendo seja gradualmente decrescido dependendo de quanto tempo se passou desde a publicação do *tweet*. Desta forma, o cálculo privilegia a observação dos usuários sobre o momento atual, dando um menor peso para publicações anteriores.

A Tabela 3 apresenta o valor atribuído como peso de cada sentimento. Resultados positivos trazem uma previsão de alta para o valor da criptomoeda enquanto resultados negativos trazem uma previsão de queda. O algoritmo pode apontar mais

de um sentimento presente no texto, neste caso é feita uma média no valor dos sentimentos presentes.

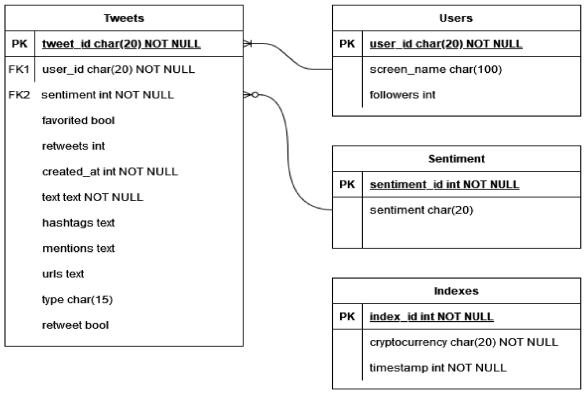
É importante ressaltar que para o algoritmo gerar uma divisão por zero seria necessário receber um *tweet* com o mesmo *timestamp* atual, o que apontaria um dado incorreto enviado pelo Twitter, já que seria impossível a operação ocorrer ao mesmo tempo considerando a latência entre os sistemas. No entanto, esta possibilidade ainda permanece, embora extremamente improvável. Considerando que nesta fase de desenvolvimento da aplicação, é preferível que o erro ocorra (caso exista) para que possam ser analisados os motivos dos dados inconsistentes estarem sendo recebidos, ficando a possibilidade de desenvolvimento de um tratamento para este tipo de erro para um futuro aprimoramento da aplicação. Quanto ao dividendo, este pode se tornar zero quando o sentimento dominante do texto for neutro ou a média dos sentimentos for igual a zero, gerando uma previsão final de valor 1 que é o *score* recebido pela mera menção à criptomoeda.

Para realizar a análise da performance preditiva do índice gerado, foram aplicadas as métricas de precisão, acurácia, *recall* e *f-1 score*, além de uma análise da correlação entre o índice gerado para cada criptomoeda e os valores de volatilidade do mercado. Para facilitar o cálculo e a geração de gráficos para exibir a performance do algoritmo, foi desenvolvido um algoritmo utilizando o *Jupyter Notebook* para realizar automaticamente o cálculo das métricas e a geração dos gráficos, facilitando o processo de análise da performance preditiva comparada com diferentes intervalos de tempo, sendo estes o mesmo momento ou uma, duas, três, seis, doze, vinte e quatro e quarenta e oito horas após a predição. Estas janelas de tempo foram definidas pois, assumindo que ocorra uma relação de causalidade entre o conteúdos das Redes Sociais e o Mercado de Criptomoedas, este impacto pode ocorrer após um determinado período após o conteúdo ser postado na Rede Social. Também foram aplicadas as mesmas métricas ao processo de treinamento do modelo de reconhecimento de sentimentos complexos, porém, também apresentando a curva ROC (TANG; WANG; CHEN, 2011). Embora este estudo não contemple uma investigação sobre a causalidade entre ambas as métricas, isto será tratado em trabalhos futuros.

3.7 DISPONIBILIZAÇÃO DOS DADOS

Os dados gerados pela estruturação de cada *tweet* e os dados dos índices de predição foram armazenados em um banco de dados SQLite3 criando uma massa de dados históricos que foi utilizada para realizar a análise da performance preditiva dos índices gerados e poderá ser explorada em estudos futuros para possíveis aprimoramentos que podem ser implementados, entre eles a criação de um *score* para cada usuário com base na acurácia dos seus *tweets* em realizar a predição da volatilidade para as criptomoedas. O Modelo de Entidade-Relacionamento (MER) desenvolvido

Figura 3 – MER



Fonte: Do Autor

para armazenar os dados no banco de dados segue o esquema ilustrado na Figura 3.

3.8 TREINAMENTO DO MODELO

Para criar um modelo de inteligência artificial foi realizado o treinamento de um modelo utilizando bibliotecas de *deep learning* com a arquitetura BERT (DEVLIN *et al.*, 2018), considerada como o estado da arte para PLN. As bibliotecas utilizadas foram PyTorch, Tez e Transformers, todas elas disponíveis para a linguagem Python.

O treinamento do modelo consiste em um aprendizado de máquina supervisionado, necessitando de uma grande massa de dados para alimentar o processo de treinamento. Para isso, foi utilizado o *dataset* GoEmotions (DEMSZKY *et al.*, 2020) que possui 58.000 comentários publicados no Reddit e classificados em 28 diferentes categorias de emoções. O processo de classificação e formatação dos textos se deu de modo bem estruturado e documentado, trazendo confiabilidade ao *dataset*.

Após realizar o *download* do *dataset* é necessário realizar a separação do mesmo em três partes, sendo estas o treino, a validação e o teste. O *dataset* é carregado por uma função disponibilizada pela biblioteca datasets que retorna o *dataset* já estruturado em um formato próprio para o treinamento, sendo um objeto do tipo dicionário que possui três chaves que nos permitem acessar um seleção de textos de tamanho adequado para cada etapa do treinamento do modelo, sendo estas chaves *train*, *validation* e *test*. Cada *dataset* retornado possui um *id* para cada linha, o texto e a sua classificação de acordo com as emoções segundo a Tabela 3.

Após a separação do *dataset*, é criada uma função que irá gerar um vetor de classificação para cada frase marcando com o valor 1 (um) o sentimento presente no texto e com o valor 0 (zero) todos os sentimentos que não estão presentes. Desta forma é gerada uma matriz com cada linha correspondendo a um texto do *dataset* GoEmotions e a coluna a um sentimento dentre os 28. Tornando os dados em um formato próprio para ser processado por algoritmos de *machine learning*, com uma matriz para

cada *dataset* (treino, validação e teste) associada ao *dataset* correspondente.

Ao analisar a distribuição de cada classe de sentimentos no *dataset* de treino, percebemos que há uma super-representação da classe neutro, não havendo impacto no âmbito geral para este estudo, pois, na Figura 5, podemos verificar um histograma com a representação de cada classe. O índice de cada emoção pode ser verificado na Tabela 3.

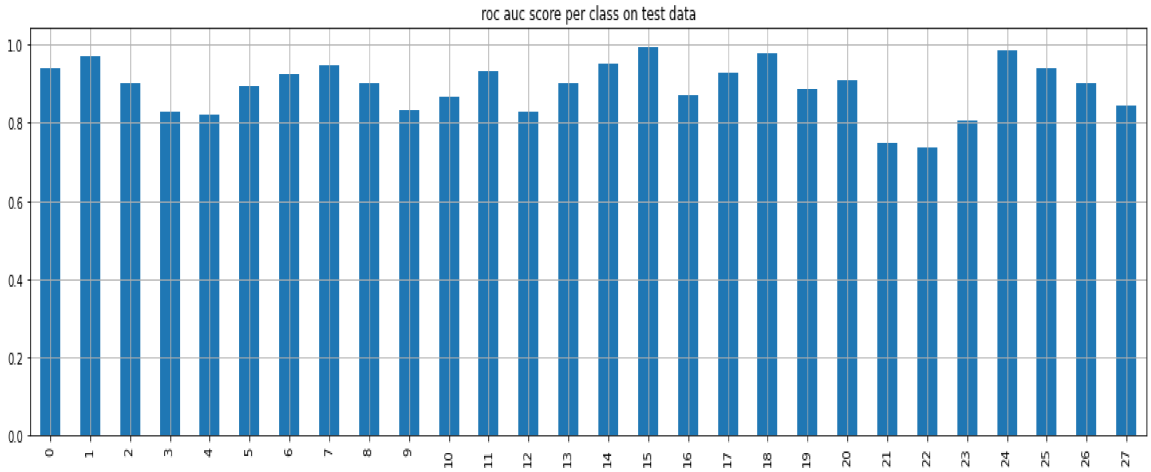
Para realizar o pré-processamento para o treinamento do modelo foi criada uma classe com métodos de suporte ao treinamento do modelo. A classe recebe a lista de textos e os *labels* correspondentes em sua inicialização e implementa internamente um tokenizador, implementando assim o modelo BERT. O tokenizador foi escolhido em função da melhor performance com o modelo SqueezeBert utilizado como *backbone* para o processo de treinamento. Os modelos *backbone* são uma forte linha base para o modelo a ser treinado atuando como extrator das *features* para os dados utilizados como base do treinamento e decodificados para facilitar o processo de treinamento (GAO *et al.*, 2019). O tokenizador converte o texto em tokens a cada iteração em que é chamado para que o algoritmo de *deep learning* possa processá-lo alimentando o treinamento do modelo para predição das emoções presentes nos textos.

Os hiperparâmetros adotados na inicialização da classe de treinamento possuem um *dropout* de 0,3 e uma transformação linear com o valor de 768 para o tamanho da amostra de *input* e o número de labels (28) como tamanho do *output*. Como função de otimização foi utilizado o AdamW *optimizer* que é uma função de otimização que possui uma melhor performance para processamentos *batch* de larga escala, se mostrando vantajoso quando aplicado para o processamento de redes neurais com arquitetura BERT (YOU *et al.*, 2019). Para a função de *loss* foi definida a função Binária de *Cross Entropy* para cada neurônio e a função de *foward* é definida para orquestrar o treinamento do modelo a cada *epoch*.

O treinamento do modelo foi realizado sem a utilização de GPU, com 8 *epochs*, *batch size* de 64 e o número de 10 *jobs* finalizando o processo de treinamento em aproximadamente 30 horas.

Verificando a performance do modelo treinado foi realizado o *plot* de um histograma com a AUC (*Area Under the Curve*) da curva ROC (*Receiver Operating Characteristic*) de cada classe na avaliação do *dataset* de teste representado na Figura 4. Todas as classes de emoções obtiveram uma precisão acima de 80%, com exceção das classes Orgulho e Realização tendo apenas a primeira um impacto que pode ser considerado significativo para este estudo. Pois o impacto que a classe da emoção proporciona é relativo ao seu valor apresentado na Tabela 3. Quanto maior o valor absoluto do sentimento, maior o impacto que este causará, pois o sentimento é considerado um indicativo da dinâmica no valor do ativo. Desta forma, os sentimentos com valores mais próximos à 10 e a -10 são considerados mais relevantes para este

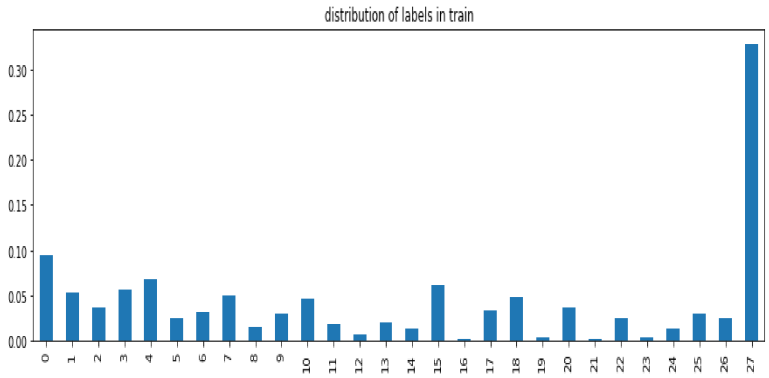
Figura 4 – Performance do Modelo



Fonte: Do Autor

trabalho.

Figura 5 – Distribuição entre as Classes



Fonte: Do Autor

Neste ponto, com o modelo gerado, é realizado o *load* do mesmo por uma função do Tensorflow e este então é utilizado para identificação das emoções presentes nos *tweets* e posterior correspondência aos pesos que serão utilizados na equação de geração do *score* de predição da variação dos valores das criptomoedas conforme descrito na seção 3.5.

3.9 RESULTADOS

Com toda a arquitetura desenvolvida o algoritmo foi executado gerando as predições para a volatilidade de cada criptomoeda (Bitcoin, Ethereum e Binance Coin). Com os valores armazenados no banco de dados, foi selecionada uma amostra de 57 predições (19 para cada criptomoeda) das quais foram utilizadas para aplicação das métricas e verificação da performance de predição. Para realizar a comparação entre os índices gerados e os valores de mercado foi desenvolvido um algoritmo capaz

de obter os dados da *exchange* Binance buscando pelo histórico de cotações para cada criptomoeda. Para realizar a conexão com a API da *exchange* foi identificada a biblioteca Python ccxt que fornece funções que facilitam a comunicação com a API de diversas *exchanges*, dentre elas a Binance. Os dados fornecidos pela *Exchange* estão em formato OHLCV que significam *Open*, *High*, *Low*, *Close* e *Volume*, correspondendo as informações de valor de abertura, maior valor, menor valor, valor de fechamento e volume para cada janela de tempo. Estes dados são muito utilizados no mercado financeiro sendo geralmente utilizados para formar gráficos no estilo *candlestick*. A janela de tempo definida para o formato OHLCV foi de uma hora.

3.9.1 Análise da Correlação

Para realizar a comparação, é esperado que o efeito das redes sociais no mercado das criptomoedas (partindo do pressuposto de que há um efeito neste sentido, já que este trabalho não investigou relações de causalidade) possa ocorrer após algum tempo, sendo então executado o teste para diferentes janelas de tempo. Foram então determinados os intervalos de tempo de 0, 1, 2, 3, 6, 12, 24 e 48 horas para a aplicação das métricas e verificada a correlação dos *scores* gerados com a volatilidade de cada criptomoeda como mostrado na Tabela 4. Deste modo, em sua execução normal, o algoritmo geraria predições em *loop*, tendo a capacidade de gerar uma predição para cada criptomoeda a cada dois minutos, essas predições serão comparadas com os valores de mercado em cada uma das janelas de tempo descritas.

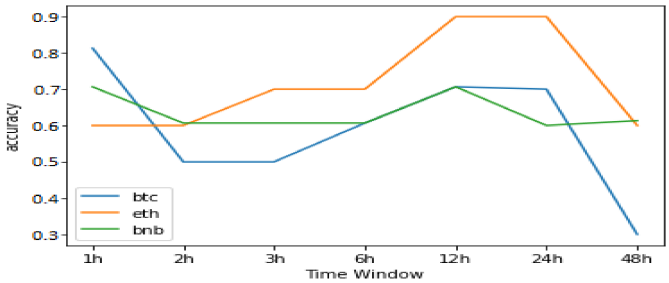
Todas as etapas seguidas para a realização da análise dos resultados podem ser observado na Figura 2. Para realizar a correlação entre os valores do *score* gerado e os valores de mercado das criptomoedas foi necessário realizar o tratamento nos dados. Primeiramente foi criada uma lista com 12 valores, com uma janela de tempo de uma hora, com os valores de mercado contendo apenas o valor correspondente ao momento atual do mercado, podendo, neste caso, ser utilizado os valores dos parâmetros *Open* ou *Close* do formato OHLCV, foi realizada a escolha do valor *Open* para ser utilizado com referência. Para cada criptomoeda foram geradas 19 predições sendo que as predições são geradas em *loop* constante, não havendo um horário determinado para ocorrer. Desta forma os 19 *scores* foram organizados em uma janela de tempo de 12 horas (conforme o horário que estes haviam sido gerados), sendo feita a média dos *scores* quando havia mais de um *score* dentro da mesma janela de uma hora. Por fim, chegou-se à uma única lista de 12 valores de *score* e uma lista de 12 valores históricos de mercado para cada intervalo de tempo (1, 2, 3, 6, 12, 24 e 48), conforme descrito na Tabela 4. Com as duas listas geradas, foi utilizada a função *pandas.DataFrame.corr* da biblioteca Pandas para gerar o grau de correlação entre os valores dos *scores* gerados e os valores históricos das criptomoedas extraídos da *Exchange*.

Tabela 4 – Datasets

Criptomoeda	Janela de tempo	Correlação
Bitcoin	Imediato	0.2374
	1 hora	0.3961
	2 horas	0.3647
	3 horas	0.2190
	6 horas	-0.5401
	12 horas	-0.0154
	24 horas	-0.1541
	48 horas	0
Ethereum	Imediato	-0.5382
	1 hora	-0.5484
	2 horas	-0.0356
	3 horas	0.3210
	6 horas	-0.0511
	12 horas	-0.7175
	24 horas	-0.7609
	48 horas	0
Binance Coin	Imediato	0.0514
	1 hora	-0.2430
	2 horas	-0.6223
	3 horas	-0.5531
	6 horas	0.4022
	12 horas	0.0982
	24 horas	0.1298
	48 horas	0

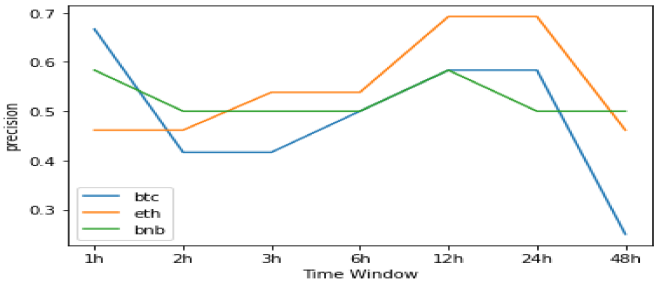
Como é possível observar, a correlação entre os índices de predição gerados e os valores de volatilidade das criptomoedas é muito baixa, não atingindo uma correlação de 80% em nenhum momento, isso demonstra que o índice não reflete com perfeição da dinâmica do mercado cripto. A aferição da correlação foi realizada com ambos os valores em escalas diferentes, o que não interfere no resultado obtido para o grau de correlação. Ainda assim, foi realizada uma comparação direta entre os valores produzidos com os valores históricos do mercado, multiplicando o valor gerado pelo índice por um valor escalar para que ambos os valores estejam na mesma escala. Essa comparação revelou valores muito discrepantes, demonstrando a dinâmica da relação entre eles já representada pelos valores de correlação. Desta forma, ficou claro que o sistema não possui um grau preditivo suficiente para determinar valores absolutos, contudo o índice ainda pode ter bons resultados preditivos com o desenvolvimento de melhorias no sistema.

Figura 6 – Acurácia



Fonte: Do Autor

Figura 7 – Precisão



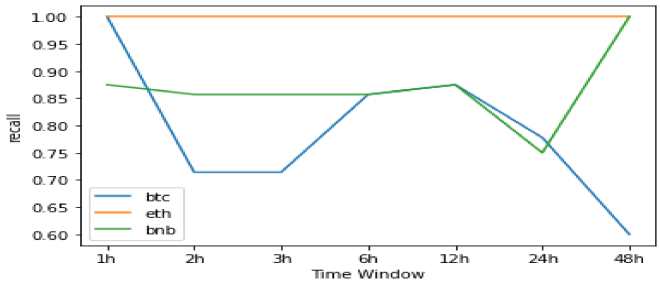
Fonte: Do Autor

3.9.2 Análise da predição de valorização ou desvalorização

Uma outra abordagem realizada foi a aferição de métricas de performance considerando apenas a direção apontada pelo índice. Desta forma, para gerar as métricas de acurácia, precisão *recall* e *f1-score*, foi considerada apenas a direção apontada pelo índice gerado, sendo os valores positivos um indicativo de valorização e os valores negativos um indicativo de desvalorização da criptomoeda. Com os valores transformados em indicações de valorizações ou desvalorizações, foi possível criar uma matriz de confusão para cada criptomoeda e cada janela de tempo, segundo o modelo exemplificado na Tabela 1 apresentada no Capítulo 2.

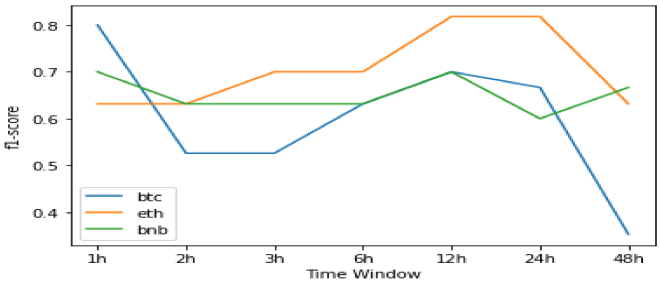
Como é possível verificar, a matriz de confusão nos fornece os dados de verdadeiros positivos (VP), falsos positivos (FP), verdadeiros negativos (VN) e falsos negativos (FN) para cada predição gerada em cada uma das oito janelas de tempo. Desta forma é possível realizar o cálculo da acurácia para cada janela de tempo conforme mostrado na Figura 6. A acurácia nos mostra uma performance de predição que, para as três criptomoedas, demonstra um aumento significativo com janelas de tempo de 12 e 24 horas, mostrando também um claro declínio após esse período. Este mesmo movimento também é percebido nas métricas de precisão demonstrado na Figura 7, *recall* demonstrado na Figura 8 e *f1-score* demonstrado na Figura 9. É importante destacar que os valores obtidos foram significativamente altos, em especial quando consideramos a capacidade de predição de desvalorização para a criptomoeda Ethereum.

Figura 8 – Recall



Fonte: Do Autor

Figura 9 – F1-Score



Fonte: Do Autor

A melhor performance preditiva com intervalos de tempo de 12 e 24 horas chama a atenção para a possibilidade da influência das redes sociais no mercado cripto se dá após algumas horas. É provável também que o potencial preditivo das redes sociais oscile em diferentes janelas de tempo, apresentando uma dinâmica própria. Esta possibilidade poderá ser investigada em um trabalho futuro. Para que outras possibilidades sejam investigadas futuramente, os resultados foram salvos no banco de dados, possibilitando a aplicação de outras métricas e a geração de outros *insights*.

Os resultados obtidos demonstram uma capacidade preditiva do modelo que chega a 55% de precisão para o Bitcoin e a Binance Coin e 69% para o Ethereum. O modelo demonstrou uma melhor performance em prever a desvaloriação das criptomoedas chegando a obter 100% de *recall* para o Ethereum e 85% para o Bitcoin e Binance Coin. Todos estes valores são referentes a janela de tempo de 12h, onde o modelo demonstrou ter a melhor capacidade preditiva. Embora estes resultados sejam razoáveis, eles ainda apontam para a necessidade de melhor desenvolvimento da capacidade preditiva do modelo, melhorando partes do processo ou combinando o modelo com outros modelos de predição que utilizem outras abordagens.

3.10 DISCUSSÃO

O tema da predição do valor de ativos no mercado financeiro é certamente um tema complexo e já vastamente explorado por diversas abordagens. Contudo, este

estudo trouxe oportunidade de explorar diversas ferramentas tecnológicas no processo de examinar uma abordagem viável e que ainda possui oportunidades à serem exploradas de formas inovadoras, utilizando tecnologias ainda recentes no mercado. Para as redes sociais foi possível investigar formas de acesso ao seu conteúdo, sendo identificada uma biblioteca que facilita o acesso a API do Twitter e fornecendo funções que permitem realizar buscas exploratórias no conteúdo da plataforma. As criptomoedas em si não tiveram a sua tecnologia explorada neste trabalho, contudo a obtenção dos dados da *exchange* Binance permitiu a identificação da biblioteca ccxt que permite uma fácil conexão com a APIs das principais *exchanges*, assim explorando algumas das tecnologias empregadas no mercado de criptomoedas. Por fim, o processo de treinamento de um modelo de IA e todo o tratamento dos dados das redes sociais gerou uma valiosa oportunidade para explorar as novas tecnologias de PLN com dados de redes sociais e aplicando análise de sentimentos no textos extraídos.

O estudo realizado neste trabalho demonstrou a possibilidade de uma fácil conexão com as redes sociais para a análise do seu conteúdo e a aplicação de técnicas avançadas de análise de sentimentos, podendo ser aplicadas em outros contextos. Para a teoria econômica da Sabedoria das Massas (GOLUB; JACKSON, 2010), este estudo apresenta uma pequena contribuição executando mais uma vez na prática os seus conceitos. Deste modo, pequenas contribuições são oferecidas em diversos temas diferentes.

Embora, ao fim de todo o processo, este trabalho não gere resultados conclusivos, as métricas aplicadas revelam um forte potencial preditivo das redes sociais. Desta forma, este índice pode não ser aplicado diretamente para a predição da volatilidade das criptomoedas, mas pode funcionar como uma variável para integrar a outras fórmulas possíveis, sendo combinado com outras métricas do mercado financeiro. Sendo assim, este trabalho se mostra como uma contribuição para um estudo maior sobre a influencia das redes sociais no mercado de criptomoedas.

4 CONCLUSÃO

Neste trabalho foram avaliadas abordagens de aplicação de PLN para predição dos valores de volatilidade das criptomoedas utilizando como fonte de dados a análise de sentimentos de textos extraídos da rede social Twitter. Para isso, foi levantado o estado da arte na utilização de textos extraídos da *web* para cálculo da predição do valor de ativos no mercado financeiro, utilizando como base uma revisão sistemática do tema (NASSIRTOUSSI *et al.*, 2014) e acrescendo a esta artigos e publicações mais recentes relacionadas ao tema. Nas obras consultadas não foi identificada nenhum outro estudo que utilizasse 28 categorias de sentimentos para a análise de sentimentos dos textos, sendo a maior parte dos estudos concentrada apenas na identificação dos sentimentos como positivo, negativo e neutro.

Para realizar o processamento dos dados do Twitter foi desenvolvido um *pipeline* que extrai os dados do Twitter utilizando a biblioteca tweepy para realizar a conexão com a plataforma da rede social. Os dados foram pré-processados utilizando a biblioteca spacy para realizar a formatação do texto e a remoção de *stop words* e os *emoticons* foram removidos com a biblioteca Emoji. Para a análise de sentimentos, foi treinado um modelo utilizando o *dataset* GoEmotions e o *framework* BERT com a arquitetura Transformers, gerando um modelo capaz de reconhecer sentimentos categorizando-os em 28 emoções. Com o *pipeline* desenvolvido e capaz de processar os textos e identificar as emoções presentes associadas as criptomoedas, foi desenvolvida uma tabela para conversão dos sentimentos identificados em valores e uma função para realização do cálculo de predição utilizando a média dos sentimentos identificados e valores dos metadados fornecidos pelo Twitter.

Para aferir a performance do algoritmo foram obtidas informações sobre cotações das criptomoedas monitoradas (Bitcoin, Ethereum e Binance Coin) na *exchange* Binance utilizando a biblioteca ccxt. Foi então gerada uma matriz de confusão para cada valor de predição nas janelas de tempo de 0, 1, 2, 3, 6, 12, 24 e 48 horas, sendo em seguida empregadas as métricas de correlação, acurácia, precisão, *recall* e *f1-score* para cada uma das janelas.

4.1 CONTRIBUIÇÕES

Este trabalho contribui com a exploração da análise de sentimentos em textos de redes sociais para sua utilização na predição da volatilidade das criptomoedas, realizando a identificação de sentimentos complexos segundo a taxonomia proposta por Ekman (EKMAN, 1992). Algumas das tecnologias envolvidas neste trabalho são recentes e ainda possuem um vasto campo para serem exploradas, como é o caso das criptomoedas e da aplicação da arquitetura BERT em processos de PLN. Desta forma, este estudo contribui com as investigações para a melhor compreensão sobre

estas tecnologias utilizadas em conjunto. Além disso, o comportamento do mercado de criptoativos e a utilização massiva das redes sociais são fenômenos recentes e a sua análise ainda traz ganhos para a compreensão destes fenômenos.

Além das contribuições geradas pelo processo de desenvolvimento da aplicação, há ainda os resultados que demonstram um potencial preditivo das redes sociais que pode ser potencializado realizando otimizações no processo e na ferramenta desenvolvidos. Por fim, a análise dos resultados oferece uma oportunidade de revisão e melhoria do trabalho para outros interessados na atividade ou para trabalhos futuros.

4.2 TRABALHOS FUTUROS

Ao longo deste trabalho foram mencionados diversas oportunidades de trabalhos futuros que este estudo pode gerar. Em relação as redes sociais, podemos citar a inclusão de textos provenientes de outras fontes como Reddit e Quora e a utilização de textos em outros idiomas como o espanhol e o português. Sobre a análise dos resultados, é possível utilizar como fonte de dados para aferição da performance das predições os valores de mercado das criptomoedas fornecidos por outras *exchenges* como a Kraken e a Bitfinex, a aplicação de outras métricas de aferição da performance preditiva, utilizando ferramentas fornecidas pela econometria e a investigação da relação de causalidade entre as redes sociais e o mercado de criptomoedas. Em relação as janelas de tempo utilizadas para a comparação, é possível realizar uma investigação para identificar se a melhor performance preditiva sempre ocorre após 12h e 24h ou se o intervalo de tempo com a melhor performance muda. Sendo este o caso, é possível investigar também se existe algum ciclo de sazonalidade para os intevalos de tempo com a melhor performance preditivas.

Para os valores apresentados na Tabela 3, um trabalho futuro seria utilizar um algoritmo de *machine learning* capaz de calcular o valor para pesos que apresentem a melhor performance preditiva. Pois, como mencionado na seção de *Cálculo do índice de predição*, os valores foram estabelecidos apenas como uma configuração inicial, não havendo um critério bem estruturado para que estes tenham a melhor performance possível. Além disso, um sistema de pesos para a reputação dos usuários também pode ser desenvolvido, armazenando os *ids* das contas e criando uma variável de reputação onde o valor é incrementado conforme o usuário contribui mais para predições assertivas e decrementando conforme ele erra. Desta forma a própria Equação Equação (1) pode ser revisada sendo acrescentada a variável de reputação assim como outras variáveis de metadados providos pelas plataformas de redes sociais.

Para o treinamento do modelo, em um trabalho futuro pode-se explorar a possibilidade de utilizar uma técnica de *transfer learning* para aprimorar o modelo desenvolvido com textos voltados para o mercado financeiro. Esta técnica é utilizada quando se possui uma pequena massa de dados para treinamento de um modelo e um outro

modelo que já atende a necessidade geral, mas pode ter sua performance otimizada em alguma situações específicas. Neste caso seria realizada a coleta e classificação dos sentimentos de uma massa de dados voltada para o mercado de criptomoedas sendo esta massa de dados utilizada para retreinar o modelo atual utilizando *transfer learning* e assim buscar uma melhora de performance na identificação de sentimentos relacionados ao mercado cripto.

Como pode ser observado, existem muitos desdobramentos possíveis para este trabalho. Algumas das oportunidades de estudo mencionadas nesta seção poderiam gerar trabalhos muito extensos devido a sua complexidade. Isto se deve ao fato de que o tema abordado neste trabalho ser vasto e frutífero, ainda carecendo de novos estudos e investigações.

REFERÊNCIAS

AGARWAL, Mansi; SAXENA, Abhishek. An overview of natural language processing. **International Journal for Research in Applied Science and Engineering Technology (IJRASET)**, v. 7, n. 5, p. 2811–2813, 2019.

ANTONOPOULOS, Andreas M. **Mastering Bitcoin: unlocking digital cryptocurrencies**. [*S.l.*]: "O'Reilly Media, Inc.", 2014.

ASLAM, Naila; RUSTAM, Furqan; LEE, Ernesto; WASHINGTON, Patrick Bernard; ASHRAF, Imran. Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets using Ensemble LSTM-GRU Model. **IEEE Access**, IEEE, 2022.

BAHJA, Mohammed. Natural language processing applications in business. *In*: E-BUSINESS-HIGHER Education and Intelligence Applications. [*S.l.*]: IntechOpen, 2020.

BENITEZ-ANDRADES, José Alberto; GONZÁLEZ-JIMÉNEZ, Álvaro; LÓPEZ-BREA, Álvaro; AVELEIRA-MATA, Jose; ALIJA-PÉREZ, José-Manuel; GARCIA-ORDÁS, Marıa Teresa. Detecting racism and xenophobia using deep learning models on Twitter data: CNN, LSTM and BERT. **PeerJ Computer Science**, PeerJ Inc., v. 8, e906, 2022.

BRIN, Sergey; PAGE, Lawrence. The anatomy of a large-scale hypertextual web search engine. **Computer networks and ISDN systems**, Elsevier, v. 30, n. 1-7, p. 107–117, 1998.

BURNIE, Andrew; YILMAZ, Emine. An analysis of the change in discussions on social media with bitcoin price. *In*: PROCEEDINGS of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. [*S.l.*: *s.n.*], 2019. P. 889–892.

CHATRATH, Arjun; MIAO, Hong; RAMCHANDER, Sanjay; VILLUPURAM, Sriram. Currency jumps, cojumps and the role of macro news. **Journal of International Money and Finance**, Elsevier, v. 40, p. 42–62, 2014.

CHOPRA, Abhimanyu; PRASHAR, Abhinav; SAIN, Chandresh. Natural language processing. **International journal of technology enhancements and emerging engineering research**, Citeseer, v. 1, n. 4, p. 131–134, 2013.

COINMARKETCAP. **Today's Cryptocurrency Prices by Market Cap**. [*S.l.*: *s.n.*], 2022. `https://coinmarketcap.com/`. Accessed: 2022-03-24.

DEMSZKY, Dorottya; MOVSHOVITZ-ATTIAS, Dana; KO, Jeongwoo; COWEN, Alan; NEMADE, Gaurav; RAVI, Sujith. GoEmotions: A dataset of fine-grained emotions. **arXiv preprint arXiv:2005.00547**, 2020.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

EKMAN, Paul. An argument for basic emotions. **Cognition & emotion**, Taylor & Francis, v. 6, n. 3-4, p. 169–200, 1992.

FELDMAN, Ronen. Techniques and applications for sentiment analysis. **Communications of the ACM**, ACM New York, NY, USA, v. 56, n. 4, p. 82–89, 2013.

GAO, Shang-Hua; CHENG, Ming-Ming; ZHAO, Kai; ZHANG, Xin-Yu; YANG, Ming-Hsuan; TORR, Philip. Res2net: A new multi-scale backbone architecture. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 43, n. 2, p. 652–662, 2019.

GIANSTEFANI, Ilaria; LONGO, Luigi; RICCABONI, Massimo. The echo chamber effect resounds on financial markets: A social media alert system for meme stocks. **arXiv preprint arXiv:2203.13790**, 2022.

GOLUB, Benjamin; JACKSON, Matthew O. Naive learning in social networks and the wisdom of crowds. **American Economic Journal: Microeconomics**, v. 2, n. 1, p. 112–49, 2010.

GROSS-KLUSSMANN, Axel; KÖNIG, Stephan; EBNER, Markus. Buzzwords build momentum: Global financial Twitter sentiment and the aggregate stock market. **Expert Systems with Applications**, Elsevier, v. 136, p. 171–186, 2019.

JIN, Fang; SELF, Nathan; SARAF, Parang; BUTLER, Patrick; WANG, Wei; RAMAKRISHNAN, Naren. Forex-foreteller: Currency trend modeling using news articles. *In*: PROCEEDINGS of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. [*S.l.*: *s.n.*], 2013. P. 1470–1473.

KANE, Aditya; PATANKAR, Shantanu; KHOSE, Sahil; KIRTANE, Neeraja. Transformer based ensemble for emotion detection. **arXiv preprint arXiv:2203.11899**, 2022.

LIN, Pingping; LUO, Xudong. A survey of the applications of sentiment analysis. **International Journal of Computer and Information Engineering**, v. 14, n. 10, p. 334–346, 2020.

LIU, Ling; WU, Jing; LI, Ping; LI, Qing. A social-media-based approach to predicting stock comovement. **Expert Systems with Applications**, Elsevier, v. 42, n. 8, p. 3893–3901, 2015.

MOKNI, Khaled; BOUTESKA, Ahmed; NAKHLI, Mohamed Sahbi. Investor sentiment and Bitcoin relationship: A quantile-based analysis. **The North American Journal of Economics and Finance**, Elsevier, v. 60, p. 101657, 2022.

NAKAMOTO, Satoshi; BITCOIN, A. A peer-to-peer electronic cash system. **Bitcoin.–URL: https://bitcoin. org/bitcoin. pdf**, v. 4, 2008.

NASSIRTOUSSI, Arman Khadjeh; AGHABOZORGI, Saeed; WAH, Teh Ying; NGO, David Chek Ling. Text mining for market prediction: A systematic review. **Expert Systems with Applications**, Elsevier, v. 41, n. 16, p. 7653–7670, 2014.

ORTIZ-OSPINA, Esteban; ROSER, Max. The rise of social media. **Our world in data**, 2023.

ÖZKAYNAR, Kürşad. Marketing strategies of banks in the period of Metaverse, Block-chain, and Cryptocurrency in the context of consumer behavior theories. **International Journal of Insurance and Finance**, v. 2, n. 1, p. 1–12, 2022.

RAHEMAN, Ali; KOLONIN, Anton; FRIDKINS, Igors; ANSARI, Ikram; VISHWAS, Mukul. Social Media Sentiment Analysis for Cryptocurrency Market Prediction. **arXiv preprint arXiv:2204.10185**, 2022.

ROESSLEIN, Joshua. tweepy Documentation. **Online] http://tweepy. readthedocs. io/en/v3**, v. 5, 2009.

RUDKOWSKY, Elena; HASELMAYER, Martin; WASTIAN, Matthias; JENNY, Marcelo; EMRICH, Štefan; SEDLMAIR, Michael. More than bags of words: Sentiment analysis

with word embeddings. **Communication Methods and Measures**, Taylor & Francis, v. 12, n. 2-3, p. 140–157, 2018.

SEONG, Nohyoon; NAM, Kihwan. Predicting stock movements based on financial news with segmentation. **Expert Systems with Applications**, Elsevier, v. 164, p. 113988, 2021.

SPACY. **spaCy - Industrial-Strength Natural Language Processing**. [*S.l.*: *s.n.*], 2023. `https://spacy.io/`. Accessed: 2023-04-01.

SRINIVASA-DESIKAN, Bhargav. **Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras**. [*S.l.*]: Packt Publishing Ltd, 2018.

STATISTA. **Most popular social networks worldwide as of January 2022, ranked by number of monthly active users**. [*S.l.*: *s.n.*], 2022. `https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/`. Accessed: 2022-03-24.

SUSANTO, Perengki; HOQUE, Mohammad Enamul; SHAH, Najeeb Ullah; CANDRA, Andel Hopi; HASHIM, Nik Mohd Hazrul Nik; ABDULLAH, Nor Liza. Entrepreneurial orientation and performance of SMEs: the roles of marketing capabilities and social media usage. **Journal of Entrepreneurship in Emerging Economies**, Emerald Publishing Limited, v. 15, n. 2, p. 379–403, 2023.

TANG, Ke; WANG, Rui; CHEN, Tianshi. Towards maximizing the area under the ROC curve for multi-class classification problems. *In*: 1. PROCEEDINGS of the AAAI Conference on Artificial Intelligence. [*S.l.*: *s.n.*], 2011. v. 25, p. 483–488.

TRAN, Trang. Predicting Digital Asset Prices using Natural Language Processing: a survey. **arXiv preprint arXiv:2212.00726**, 2022.

VU, Tien Thanh; CHANG, Shu; HA, Quang Thuy; COLLIER, Nigel. An experiment in integrating sentiment features for tech stock prediction in twitter. *In*: PROCEEDINGS of the workshop on information extraction and entity analytics on social media data. [*S.l.*: *s.n.*], 2012. P. 23–38.

WALTON, Joseph. Cryptocurrency public policy analysis. **Available at SSRN 2708302**, 2014.

WĄTOREK, Marcin; DROŻDŻ, Stanisław; KWAPIEŃ, Jarosław; MINATI, Ludovico; OŚWIĘCIMKA, Paweł; STANUSZEK, Marek. Multiscale characteristics of the emerging global cryptocurrency market. **Physics Reports**, Elsevier, v. 901, p. 1–82, 2021.

YACOUBY, Reda; AXMAN, Dustin. Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models. *In*: PROCEEDINGS of the first workshop on evaluation and comparison of NLP systems. [*S.l.*: *s.n.*], 2020. P. 79–91.

YOU, Yang *et al.* Large batch optimization for deep learning: Training bert in 76 minutes. **arXiv preprint arXiv:1904.00962**, 2019.

YU, Yang; DUAN, Wenjing; CAO, Qing. The impact of social and conventional media on firm equity value: A sentiment analysis approach. **Decision support systems**, Elsevier, v. 55, n. 4, p. 919–926, 2013.

ZHANG, Li; ZHANG, Liang; XIAO, Keli; LIU, Qi. Forecasting price shocks with social attention and sentiment analysis. *In*: IEEE. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). [*S.l.*: *s.n.*], 2016. P. 559–566.