UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO

Enzo Yamamura

Sentimentos do Reddit como proxy para efeito demanda nas oscilações do Bitcoin com teste de poder preditivo via LSTM multivariado

São Carlos

2022

Enzo Yamamura

Sentimentos do Reddit como proxy para efeito demanda nas oscilações do Bitcoin com teste de poder preditivo via LSTM multivariado

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Jó Ueyama

Versão original

São Carlos

2022

Folha de aprovação em conformidade
com o padrão definido
pela Unidade.
No presente modelo consta como
folhadeaprovacao.pdf

Este trabalho é dedicado a minha família, meu porto seguro,
e a Luiz Carlos de Jesus Júnior, pelo suporte inestimável.

RESUMO

YAMAMURA, E. **Sentimentos do Reddit como proxy para efeito demanda nas oscilações do Bitcoin com teste de poder preditivo via LSTM multivariado**. 2022. 54p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

Neste trabalho buscamos averiguar se existe ganho informacional na predição de tendências do Bitcoin considerando sentimentos manifestados no fórum Reddit como indicativo de humor de mercado. Estabelecemos que há associação positiva com indícios de causalidade, com sentimentos do Reddit antecipando oscilações no Bitcoin, e finalizamos ajustando um modelo de Long Short Term Memory aos dados, obtendo uma previsão precisa nos dados de teste. Concluímos que o Reddit (e potencialmente outras mídias sociais), pelo menos para Bitcoin e com base no período estudado, pode agregar riqueza preditiva, efetivamente servindo como *proxy* para tendências e sentimentos online, dado seu formato irrestrito (que permite desde notícias, opiniões e até conteúdo humorístico) e a quantidade de usuários da plataforma. O que reforça a natureza do ativo, ao menos no período considerado, de investimento especulativo e não de reserva de valor, sendo portanto altamente suscetível a emoções ou percepções coletivas e movimentos de manada.

Palavras-chave: Reddit. NLP. Bitcoin. Séries Temporais. Redes Neurais Recorrentes. LSTM.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	Redes Neurais Artificiais
API	Interface de Programação de Aplicações
ARIMA	Modelo Autoregressivo Integrado com Médias Móveis
BTC	Bitcoin
CPU	Processador
CSS	Cascading Style Sheets
EMH	Hipótese de Mercado Eficiente
GME	Ação da GameStop
GPU	Placa de Vídeo
HTML	Linguagem de Marcação de Hipertexto
IA	Inteligência Artificial
LSTM	Long Short Term Memory
NLP	Processamento de Linguagem Natural
NNAR	Modelo Autoregressivo de Redes Neurais
RNN	Redes Neurais Recorrentes
UTC	Tempo Universal Coordenado
VADER	Valence Aware Dictionary and Sentiment Reasoner

SUMÁRIO

1 INTRODUÇÃO

A hipótese do mercado eficiente, tal como cunhada por FAMA (1991) e Fama et al. (1969), estabelece que o mercado de ativos financeiros reflete na íntegra toda informação disponível. Consequentemente, segundo Nguyen e Shirai (2015) *apud* FAMA (1991), toda mudança em preços seria decorrente de novas informações ou notícias e, como estas são impossíveis de antever, os preços de ativos deveriam seguir uma *random walk*, ou seja, uma oscilação aleatória, com o melhor preditor para o preço futuro sendo o preço atual. Dito isso, Walczak (2001) estabelece que qualquer oscilação em ativos financeiros não seria previsível com mais de 50% de acurácia.

Porém, segundo Kahneman e Tversky (2013), decisões financeiras não são unicamente tomadas apenas com base no valor e fundamentos, mas também em percepções de risco e emoções. Emoções que se tornaram amplamente disseminadas com o advento das mídias sociais, tal como Twitter, Reddit e portais de notícia. A premissa deste trabalho é de que existe uma relação causal entre sentimentos online e oscilação de ativos (mais especificamente Bitcoin) a ser investigada com intuito de averiguar se podemos melhorar estratégias de investimentos em criptomoedas com base no humor de mercado online ou se os sentimentos não antecipam, mas apenas reagem às oscilações.

Em carta aos investidores do primeiro trimestre de 2021[1] o Twitter apresentou os seguintes dados:

- 199 milhões de usuários.
- 500 milhões de tweets por dia.
- 1 em cada 5 adultos nos Estados Unidos utilizava a plataforma.
- Cada publicação tem um limite de 280 caracteres.

Por sua vez, a plataforma de fóruns Reddit[2]:

- mais de 430 milhões de usuários ativos mensalmente .
- 52 milhões de usuários ativos diariamente.
- 1 em cada 4 adultos americanos usam o fórum.

[1] <https://s22.q4cdn.com/826641620/files/doc_financials/2021/q1/Q1'21_InvestorFactSheet.pdf>. Acesso em: 12 Jul. 2021.

[2] <https://backlinko.com/reddit-users>. Acesso em 12 Jul. 2021.

- Cada publicação tem um limite de 40.000 caracteres (maior riqueza informacional que Twitter).
- Registrados 2 bilhões de comentários em 2020 (mesmo com filtros e regras anti *spam*).

Ambos os supracitados possuem APIs próprias e gratuitas para extração de informações com variados critérios de filtragem por tópico (*hashtags* no Twitter e *subreddits*, ou sub-fóruns, no Reddit) e por popularidade da publicação. Além do mais ambos historicamente já impactaram o mercado de ativos significantemente. São exemplos:

- Os tweets de Elon Musk[3] que, independentemente da seriedade do conteúdo, causam oscilações consideráveis nas criptomoedas.
- Salto da ação da Gamestop[4] de US$ 18 para mais de US$ 450 em janeiro de 2021, causado não por questões intrínsecas ou macroeconômicas mas sim por emoções: como forma de protesto aos grandes fundos de investimento operando vendidos (apostando na queda) nas ações da companhia, um grupo de investidores amadores se uniu no *subreddit /wallstreetbets* em movimento massivo de compra de ações da companhia (GME), elevando os preços de forma sem precedentes.

Fundos de investimento passaram a monitorar ambas as redes[5], considerando não só sentimentos mas também o volume de menções atreladas a cada ativo de interesse, de modo a compilar humores de mercado para ativos específicos.

Com os fatos elencados, é reiterada a importância de se considerar informações de humor de mercado das novas mídias para definição de estratégias de investimento. Estudos midiáticos[6] já têm indicado aumento no percentual de adultos que se informam principalmente via redes sociais (45% no Reino Unido em 2020), e o aumento é consistente ano a ano. Além disso, um estudo da Gartner de 2010[7] apontava que, dez anos trás, a maioria dos consumidores já contava com redes sociais para guiá-los na tomada de decisões de consumo, com a ascensão dos *influencers* impactando atividades de compra de até 74% da população já naquela época.

[3] <https://www.forbes.com/sites/ronshevlin/2021/02/21/how-elon-musk-moves-the-price-of-bitcoin-with-his-tw
>. Acesso em: 12 Jul. 2021.

[4] <https://www.forbes.com/sites/emilsayegh/2021/03/09/losing-touch-with-reality--a-gamestop-lesson/
>. Acesso em: 12 Jul. 2021.

[5] <https://www.cnn.com/2021/02/03/investing/wall-street-reddit-hedge-funds/index.html.
>. Acesso em: 12 Jul. 2021.

[6] <https://www.ofcom.org.uk/___data/assets/pdf_file/0013/201316/
news-consumption-2020-report.pdf.>. Acesso em: 12 Jul. 2021.

[7] <https://archive.nytimes.com/www.nytimes.com/external/readwriteweb/2010/07/26/
26readwriteweb-majority-of-consumers-use-social-networks-t-90514.html.>. Acesso em: 12 Jul. 2021.

Estudos anteriores considerando tanto canais oficiais como extra-oficiais de informação (redes sociais) na predição do preço de ativos, reportam acurácias direcionais acima de 50% (e até próximas a 90%) são possíveis, subvertendo a hipótese de EMH (*efficient markets hypothesis*) ou hipótese de mercados eficientes, como visto em Shah, Isah e Zulkernine (2018), Nguyen e Shirai (2015) e Sul, Dennis e Yuan (2017), o que atesta aplicabilidade e possível rentabilidade via análise de sentimentos para geração estratégias de compra e venda de ativos financeiros.

Neste trabalho buscamos compilar uma base de dados de comentários do Reddit computando seus respectivos sentimentos para investigar sua associação com a série histórica do preço do Bitcoin. O enfoque é averiguar se existe relação: a oscilação no Bitcoin causa oscilações nos sentimentos ou vice-versa? Caso o segundo se prove verdadeiro, prosseguiremos num esforço de agregar previsibilidade para o preço do Bitcoin via sentimentos manifestados no Reddit.

O objetivo final deste trabalho é bem resumido pelo trecho:

> ...humanos tomam decisões usando a maior quantidade de informações disponíveis. Isso geralmente toma vários minutos para que descubram novas informações e tomem a decisão. Um algoritmo é capaz de processar gigabytes de texto de múltiplas fontes em segundos. Potencialmente, poderíamos explorar este ínterim para criar uma estratégia de trading. (VELAY; DANIEL, 2018). (Tradução livre, própria)

Em linha com o trabalho de Velay e Daniel (2018), a motivação desta monografia é de definir a viabilidade de se desenvolver um algoritmo de investimento em criptomoedas que capture e processe sentimentos online para prever valores futuros destas com uma precisão ao menos marginalmente maior do que a aleatória (50%), explorando o lapso que um ser humano demora para processar informações e antecipando movimentos de manada na demanda comprando ou vendendo ativos. Como isto é apenas um estudo preliminar, nos limitamos a estudar relações de associação, causa e efeito, finalizando com um breve teste preditivo.

2 REVISÃO BIBLIOGRÁFICA

2.1 Fundamentação Teórica

2.1.1 *Machine Learning* - Aprendizado de máquina

Segundo o site da IBM[1]:

> Machine Learning é uma área de inteligência artificial (IA) e ciências computacionais que foca no uso de dados e algoritmos para imitar o jeito que humanos aprendem, gradativamente melhorando sua precisão. (Tradução livre)

Consiste no estudo de algoritmos computacionais que melhoram constantemente através de experiência e uso de dados (MITCHELL et al., 1997). Ajusta-se um modelo sobre dados históricos de modo que seja capaz de extrapolar o aprendido para dados inéditos(LORENA; GAMA; FACELI, 2000).

Segundo Lindholm et al. (2021), envolve aprendizado, interpretação e ação via construção de programas de computador que, em interface com dados, extraem informações, classificações, conclusões ou decisões. Difere de análise de dados pois é automatizado e os programas computacionais aprendem com os dados de forma dinâmica, efetivamente treinando.

Neste trabalho utilizaremos Aprendizado Supervisionado, tendo em vista o formato tabular dos dados extraídos. Aprendizado Supervisionado engloba modelos em que tanto *inputs* como *outputs* estão presentes possibilitando que, via otimização de uma função objetivo, algoritmos possam ser utilizados para prever informações de interesse a partir de dados novos (LINDHOLM et al., 2021). São tarefas de algoritmos de aprendizado supervisionado problemas de classificação e regressão.

Porém, os dados utilizados, apesar de rotulados e estruturados, são séries temporais. Logo, optamos por utilizar *Long Short Term Memory* (LSTM), criado por Hochreiter e Schmidhuber (1997), uma rede neural recorrente.

Redes Neurais Artificiais são algoritmos bioinspirados na forma como neurônios operam: com dentritos recebendo informação na forma de impulsos nervosos (sinapses) e axônios processando e propagando os impulsos para outros neurônios conectados. O diagrama abaixo ilustra o processo: providenciamos os dados na camada de entrada, que são processados nas camadas ocultas e temos os resultados (classificações) na camada de saída. Efetivamente, cada conexão representa um peso, e a cada iteração a informação é

[1] <https://www.ibm.com/cloud/learn/machine-learning>. Acesso em 16 Jul. 2021.

propagada até a camada de saída, na qual uma função de ativação transforma os dados no *output*.

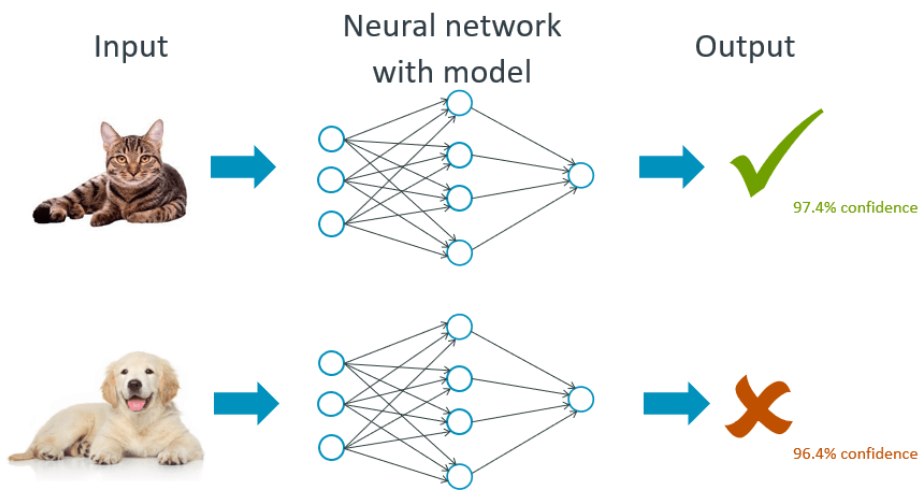

Figura 1 – Redes Neurais. Fonte: artigo do Medium[2].

Em Redes Neurais Artificiais é assumida independência entre neurônios e ausência de memória (ou seja, tanto as entradas quanto as classificações na saída acima são independentes entre si), o que não é ideal para situações em que a sequência importa, com dependência de resultados anteriores. Rumelhart, Hinton e Williams (1986) criam então o modelo de Redes Neurais Recorrentes, trazendo para Redes Neurais o método de *backpropagation*, em que o mesmo processo anterior ocorre mas o erro passa a ser computado para reajustar os pesos das ligações retroativamente, reiniciando o processo sucessivamente até a entrada mapear razoavelmente bem a saída, minimizando o gradiente descendente do erro. A figura abaixo do site da IBM ilustra bem a diferença, o diagrama da esquerda sendo uma Rede Neural Recorrente e o da direita uma Rede Neural Artificial:

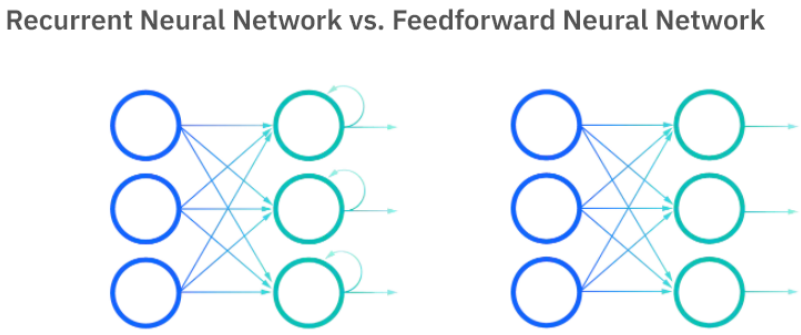

Figura 2 – RNN x ANNs. Fonte: site da IBM[3].

[2] <https://ahmetozlu.medium.com/long-short-term-memory-lstm-networks-in-a-nutshell-363cd470ccac>. Acesso em 24 Mar. 2022.

[3] <https://www.ibm.com/cloud/learn/recurrent-neural-networks>. Acesso em 24 Mar. 2022.

Por sua vez, Redes Neurais Recorrentes possuem pouca memória, falhando em em carregar informações ao longo de séries mais extensas. O algoritmo Long Short Term Memory (HOCHREITER; SCHMIDHUBER, 1997) é um tipo de Rede Neural Recorrente capaz de aprender dependências ao longo de extensos períodos de tempo. Possui portas de esquecimento que são treinadas de acordo com ganho informacional, filtrando quais informações passadas são propagadas ou descartadas iterativamente.

2.1.2 Análise de Sentimentos

Em 2017 estimava-se que 90% dos dados disponíveis no mundo haviam sido criados nos 2 anos precedentes (2015 e 2016)[4]. Grande parte dos quais são não estruturados.

Dados não estruturados são organizados em vários formatos, cujo propósito é a leitura, por humanos, dentro de um contexto cultural (GARCÍA; LUENGO; HERRERA, 2015). São fotos, imagens, áudios, publicações online, artigos, e-mails, dentre outros.

Esta exuberância de informação não estruturada levou ao surgimento da área de *natural language processing* (NLP), ou Processamento de Linguagem Natural. Trata-se de uma coleção de métodos para fazer com que dados não-estruturados possam ser interpretados computacionalmente (ABRAHAM et al., 2018).

Dentro de NLP existe o secto de análise de sentimento, cujo enfoque é extrair emoções e opiniões expressados via texto. De acordo com Mohapatra, Ahmed e Alencar (2019) existem duas alternativas mais usuais para análise de sentimento: baseada em aprendizado de máquina e em léxicos (ou dicionário). Enquanto a primeira utiliza técnicas de aprendizado de máquina para classificar sentimentos, a segunda utiliza um dicionário de sentimentos associado a palavras de opinião de modo a obter a polarização do texto, palavra a palavra.

No presente trabalho faremos uso do VADER (*Valence Aware Dictionary and Sentiment Reasoner*). Idealizado por Hutto e Gilbert (2015), trata-se de um léxico calibrado especificamente para mídias sociais, tal como Twitter, Facebook e Reddit. Além de detectar a polaridade palavra a palavra (positiva, neutra ou negativa), detecta também a intensidade dos sentimentos e a polarização composta do texto. Processa gírias, contrações, símbolos, estilos e até *emojis* utilizados nas redes sociais, sem a necessidade de tratamento extensivo ou tokenização, reduzindo grande parte do pré-processamento de textos de comentários online. Mais informações estão disponíveis no GitHub da ferramenta[5].

O dicionário foi comparado com 11 ferramentas de classificação de sentimentos alternativas ao longo de 4 domínios distintos, obtendo a melhor performance em todos os testes referentes a textos de mídias sociais (HUTTO; GILBERT, 2015).

[4] <https://www.iflscience.com/technology/how-much-data-does-the-world-generate-every-minute/>. Acesso em 16 Jul. 2021.

[5] <https://github.com/cjhutto/vaderSentiment>. Acesso em 24 mar. 2022.

2.2 Trabalhos Relacionados

2.2.1 Requisitos Esperados

No presente trabalho buscaremos inovar sobre a literatura existente, tomando como principal inspiração o artigo que originou o KryptoOracle (MOHAPATRA; AHMED; ALENCAR, 2019), uma ferramenta que integra análise de sentimento, aprendizado de máquina, arquitetura Spark (para obter dados em tempo real) e aprendizado online.

As principais contribuições deste trabalho serão:

- Observar se sentimentos de Reddit possuem relação com oscilação do Bitcoin.
- Verificar se há causalidade e se humores variam com as oscilações do Bitcoin ou vice-versa.
- Testar se um modelo LSTM multivariado para prever o preço do Bitcoin em função dos sentimentos do Reddit gera um bom ajuste.

2.2.2 Trabalhos Prévios

A motivação inicial da criação do Bitcoin, em 2008, pelo grupo de programadores sob o pseudônimo de Satoshi Nakamoto (NAKAMOTO, 2008), era de criar uma reserva de valor fiduciária sem a presença do intermediário, ou seja, da autoridade central de controle monetário. Com o intuito de contornar as percebidas vulnerabilidades causadas por variações monetárias atreladas à política e ideais, a moeda foi criada sem autoridade central com poder de impactar a oferta. Isto se tornou possível graças à tecnologia do *Blockchain*: um sistema descentralizado (*peer-to-peer*) com registro compartilhado, criptografado (portanto seguro) e descentralizado, contendo todos os registros históricos de transações. Tal base é criada em cima de blocos de operações em ordem cronológica, que são criptografados e então agrupados com blocos mais antigos, criando uma corrente de blocos (*Blockchain*), fazendo com que se torne tanto seguro como transparente (STENQVIST; LÖNNÖ, 2017). O valor de qualquer moeda depende da confiança do público, da aceitação e das expectativas envolvidas. Ou seja, com oferta fixa o Bitcoin depende apenas de:

- Demanda.
- Fatores exógenos: conflitos, regulamentações, manifestações públicas de agentes influentes e afins.

Em sua obra, Kahneman e Tversky (2013) estabelecem que decisões financeiras são significantemente impactadas por risco e emoções. Este tema também é explorado por outros autores, como Dolan (2002), reiterando que a tomada de decisões é altamente

impactada por emoções. Somando-se a isso, segundo Roberts (2017), desde o início o Bitcoin atraiu investidores amadores com mentalidade de aposta, fazendo com que se tornasse um ativo altamente volátil e particularmente especulativo (suscetível a emoções de mercado) desde sua criação, ao invés da reserva de valor liberal e independente que foi concebido para ser.

Ao longo dos anos, com a ascendência e massificação das mídias sociais, estas gradativamente se consolidaram dentre os principais meios de informação, com vários canais de notícia e celebridades aderindo às mesmas. Algumas notícias, por exemplo, acabam sendo veiculadas em primeira mão no Twitter antes de qualquer outro veículo informativo, tal como a queda de avião do US Airways no rio Hudson, em 2009[6]. Além disso, milhões de usuários diariamente fazem uso da plataforma para expressar suas opiniões. Por conta disso e do agrupamento de temas intrínseco (definido pelo limite de caracteres e *hashtags*), Abraham et al. (2018) apontam que a plataforma se tornou uma fonte exuberante de dados sobre os sentimentos da população e a evolução dos mesmos acerca de praticamente qualquer tópico. Ou seja, a vasta disponibilidade dos dados (com API própria e gratuita para extração), as publicações rotuladas, a objetividade imposta pela restrição de caracteres e a aderência global fazem da plataforma uma mina de ouro para dados de opinião em formato já semi estruturado e classificado em tópicos (KOULOUMPIS; WILSON; MOORE, 2011).

Em FAMA (1991) e Fama et al. (1969) estabelecem a hipótese do mercado eficiente. Ou seja, dado que toda a informação é plenamente disponível, qualquer oscilação nos ativos seria devido à notícias novas. Notícias novas são, por definição, aleatórias e imprevisíveis. Portanto, argumentam que um ativo financeiro só pode ser previsto com uma acurácia de, no máximo, 50% de acerto. Ou seja, para os autores existia aleatoriedade intrínseca e impossibilidade de antecipação dos movimentos de preço.

Todavia, com a emergência das novas redes sociais e a consolidação destas como meios de informação citada anteriormente, alguns autores como Shah, Isah e Zulkernine (2018), Nguyen e Shirai (2015), Sul, Dennis e Yuan (2017), Bollen, Mao e Zeng (2011), Abraham et al. (2018) e Stenqvist e Lönnö (2017), dentre outros, subvertem a máxima de Fama com modelos que buscam captar humores das novas fontes de informação online e estabelecem não só correlações mas predições superiores ao nível de aleatoridade (50%).

Nguyen e Shirai (2015) realizam análise de sentimentos em publicações do Twitter e obtém predição do preço de ativos selecionados com mais de 60% de acurácia. Embasando-se apenas em análise de sentimentos sobre notícias, em Shah, Isah e Zulkernine (2018), chegam a uma acurácia direcional de 70,59% na previsão de tendências de curto prazo para certos ativos. Bollen, Mao e Zeng (2011) fazem uso de redes neurais: tomando sentimentos do Twitter como dados de entrada, elaboram previsão do índice de DOW Jones, atingindo

[6] <https://www.brandwatch.com/blog/twitter-stats-and-statistics/>. Acesso em 24 Mar. 2022.

86,7% de precisão.

Existem também esforços no sentido de prever preços de Bitcoin via modelos univariados, ou seja, embasando-se apenas na série histórica do próprio ativo. Munim, Shakil e Alon (2019) fazem uso de modelo auto-regressivo integrado de média móvel (ARIMA), um dos mais populares na literatura para predições envolvendo séries temporais, e do modelo auto-regressivo de redes neurais (NNAR).

Tendo em vista que o escopo deste trabalho não é de contribuir para a área de Análise de Sentimento ou Séries Temporais, mas de aplicá-las para estabelecer relações entre sentimentos de redes sociais e variações no preço do Bitcoin, tomamos inspiração principalmente dos seguintes artigos: Mohapatra, Ahmed e Alencar (2019), Abraham et al. (2018) e Stenqvist e Lönnö (2017), detalhadas a seguir.

Stenqvist e Lönnö (2017) coletam dados via API do Twitter, sobre os quais realizam tratamentos para remover publicações dúbias e irrelevantes, usando o VADER para classificar a polarização das mesmas. Então coletam dados de variações por minuto do Bitcoin via CoinDesk[7], construindo tabela relacional com as polarizações médias por período de tempo e as respectivas oscilações no Bitcoin, sobre as quais aplicam uma previsão simples (*naive prediction*), em que a variação do preço do Bitcoin é testada contra a direção correspondente da classificação de polarização das publicações. Via tais procedimentos, conseguem uma acurácia de 83%, mas argumentam que os achados são embasados em amostra por demais limitada para tecer conclusões.

Abraham et al. (2018) utilizam abordagem similar, porém para auferir relações não só com o Bitcoin mas também com a segunda criptomoeda de maior adesão: Ethereum. Também utilizam dados de publicações coletados via API do Twitter com integração via pacote Tweepy do Python mas suplementam as informações com dados do Google Trends. Argumentam que a plataforma era então utilizada em mais de 70% das buscas online, portanto servindo como *proxy* para o interesse público geral e, portanto, fatores macroeconômicos. Também empregam o VADER, argumentando se tratar do melhor dicionário de sentimentos disponível especificamente calibrado para redes sociais. Porém, observam que as polarizações tendem à neutralidade, o que dificulta o estabelecimento de relação clara. Para circular este obstáculo, empregam dados de volume de publicações no Twitter, que extraem do portal <www.bitinfocharts.com>. Ao testarem as correlações de Pearson e Spearman, percebem forte relação linear entre o preço do Bitcoin, os dados provenientes do Google Trends e o volume de *tweets* (publicações do Twitter). Portanto, optam por uma regressão linear múltipla como algoritmo de aprendizado. Concluem que a maior parte dos estudos prévios teria encontrado causalidade espúria, devido a considerarem unicamente períodos de ascensão da criptomoeda. Também observam que, a despeito da queda nos preços, os *tweets* sobre o Bitcoin tendem a ter viés positivo, devido

[7] <https://www.coindesk.com/>. Acesso em 17 Jul. 2021.

a grande parte dos que se manifestam positivamente sobre a criptomoeda a verem como reserva de valor e não como investimento especulativo. Também sugerem que modelos mais complexos (não lineares) de aprendizado de máquina podem incrementar a previsibilidade do Bitcoin via *tweets*.

Mohapatra, Ahmed e Alencar (2019) tomam inspiração dos dois artigos anteriores e transformam o problema de predição de preços de Bitcoin via sentimentos do Twitter em um problema de *Big Data*. Fazendo uso de arquitetura Spark, implementam aprendizado em tempo real. Também utilizam VADER como dicionário para detecção de sentimentos em publicações coletadas do Twitter via API e pacote Twython do Python. Extraem dados de variação de Bitcoin através da API do portal Cryptocompare. O diferencial deste trabalho é que adicionam ponderação por nível de influência da postagem, via contagem de seguidores, curtidas e republicações. Após os tratamentos preliminares de dados, compilam as variações de Bitcoin juntamente com os sentimentos compostos e ponderados por influência, obtendo dados em formato tabular. Considerando este formato estruturado, utilizam o algoritmo de aprendizado supervisionado tido como estado da arte: XGBoost. Por fim, implementam uma arquitetura em Spark: o modelo XGBoost é inicialmente treinado com séries históricas do Twitter analisadas via VADER e dados históricos de Bitcoin e, de acordo com erros, acertos e dados inéditos, o modelo é constantemente retreinado em sistema de Online Learning, gerando um modelo perpetuamente atualizado e que aprende em tempo real. Concluem estabelecendo que melhorias poderiam ser feitas gerando visualizações em tempo real e testando outros algoritmos de aprendizado pois, por limitações do Spark, os algoritmos mais modernos de *deep learning* não eram então implementáveis.

Em linha com os estudos anteriores, buscaremos identificar se existe a mesma relação para sentimentos do Reddit (ou se são apenas reativos ao Bitcoin). Caso exista, seguiremos com a sugestão de Mohapatra, Ahmed e Alencar (2019) e Abraham et al. (2018) e ajustaremos um modelo de *deep learning* com LSTM multivariado, averiguando se sentimentos do Reddit como *proxy* de humor de mercado ajudam a prever as oscilações no preço de fechamento do Bitcoin no conjunto teste.

2.2.3 Tabela comparativa

	Dados	Tratamentos	Análise de Sentimentos	Algoritmo	Atualização Tempo Real	Data Viz
Nguyen e Shirai (2015)	-Yahoo Finance: 18 ações consideradas - Yahoo Finance Message Board: postagens referentes às ações consideradas	- Remoção de Stop Words - Lematização via Stanford CoreNLP	- SVM com kernel linear como modelo de classificação - Latent Dirichlet Allocation (LDA) para obter tópicos ocultos - JST para obter tópicos ocultos - Associação tópico - sentimento	- SVM para classificar acurácia da classificação do sentimento x oscilação positiva / negativa das ações	Não	Não
Stenqvist e Lönnö (2017)	- Coindesk (Bitcoin) - Twitter API (publicações sobre BTC)	- Remoção de publicações duplicadas - Remoção de hashtags, palavras recorrentes, bigramas e trigramas	VADER	- Naive Predictor: sentimento classificado como positivo = subida no preço do bitcoin negativo = queda	Não	Não
Abraham et al. (2018)	- Dados Ethereum e Bitcoin: fonte não informada - Google Trends (referente a BTC e ETH) - Twitter API (publicações sobre BTC e ETH) - Volume de Tweets (bitinfocharts.com)	- RegEx para remover símbolos irrelevantes - Remoção de letras maiúsculas - Preprocessamento (lematização, tokenização)	VADER	Regressão Linear Múltipla	Não	Não
Mohapatra, Ahmed e Alencar (2019)	- Bitcoin: Cryptocompare API - Twitter API (publicações sobre BTC) + Twhython	- Remoção de links, imagens, vídeos, hashtags - Mantiveram ID, texto, username e número de seguidores - Score da publicação leva em conta não só a polarização mas também a contagem de seguidores, número de likes e número de retweets: considera a influência	VADER	XGBoost	Arquitetura Spark para atualizar dados em tempo real via integrações com APIs do Twitter e Cryptocompare, inserindo dinâmica de aprendizado online perpétuo	Não
O que propomos	- Bitcoin: Base do Kaggle - Base de Comentários do Reddit (via BigQuery)	- Remoção de formatação de texto via Regex - Identificação do idioma dos comentários via Fast Detect - Exclusão de comentários não ingleses p/ VADER - Mater pontuação composta recebida pelo comentário, data e Subreddit - Score do comentário do Reddit = *upvotes - downvotes*	VADER	LSTM	Próximos passos	Próximos passos

Tabela 1 – Tabela Comparativa. Elaboração própria.

3 DESENVOLVIMENTO

3.1 Detalhamento do problema

Nesta seção definimos o problema a ser atacado e pormenores. Todos os passos realizados estão em códigos e extensamente documentados no GitHub[1] deste projeto (rodapé). Vale ressaltar que a visualização dos gráficos só é possível via NBViewer[2] por conta da escolha do pacote de visualizações Plotly.

Antes de mais nada, esquematizamos as etapas do trabalho abaixo:

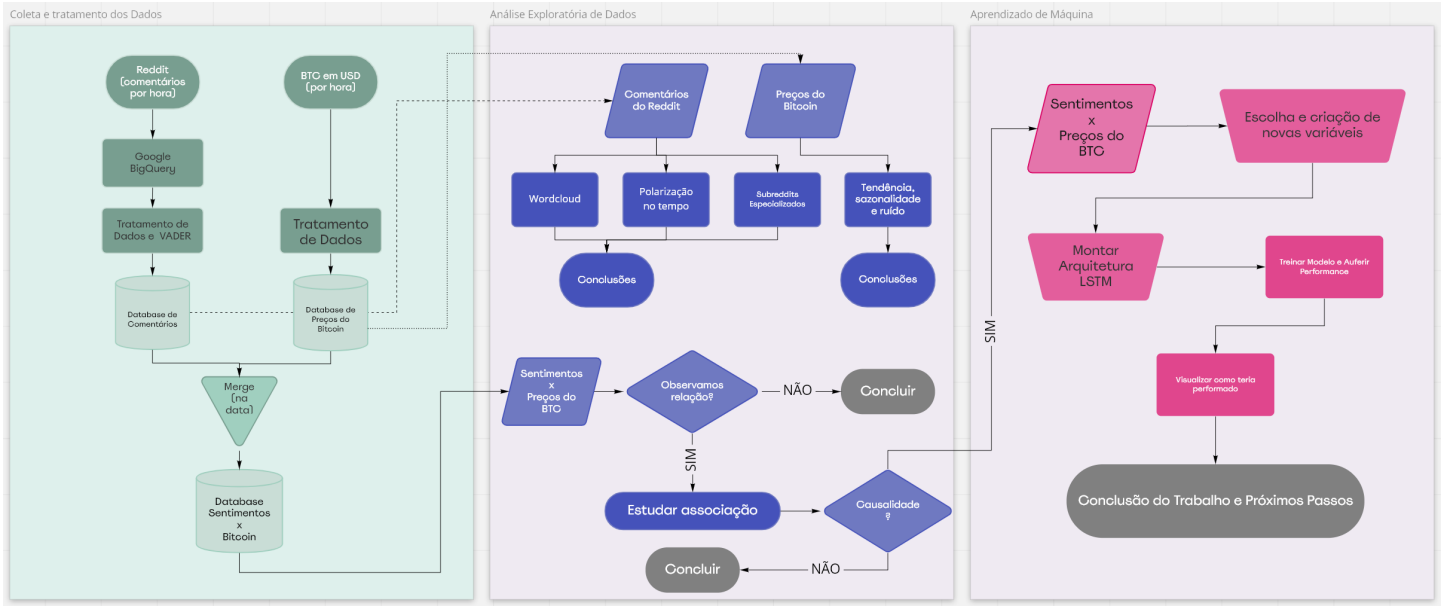

Figura 3 – Fluxograma deste Trabalho. Elaboração própria.

Em suma:

- Extraímos os dados do Reddit e do Bitcoin em dólares.
- Comentários do Reddit: formato é tratado, aplicamos VADER (para obter sentimentos) e agrupamos sentimentos ponderados por *score* por dia.
- Bitcoin: apenas agrupamos por dia.
- Analisamos a base de sentimentos do Reddit e do Bitcoin separadamente.
- Investigamos presença de relação entre ambas as bases.

[1] <https://github.com/Yamamuen/MBA_thesis>. Acesso em 24 Mar. 2022.
[2] <https://nbviewer.org/>. Acesso em 24 Mar. 2022.

- Analisamos como se associam e verificamos se existe causalidade.

- Encerramos com aplicação de um LSTM para averiguar se usar sentimentos do Reddit como *proxy* para variações na demanda traz ganhos preditivos.

3.2 Coleta dos dados

3.2.1 Dados do Reddit

Começamos esta seção com agradecimentos especiais ao usuário e arquivista do Reddit *Stuck_In_the_Matrix*, seu único nome público conhecido é Jason. Jason desenvolveu o pacote Pushift, que permite que consultas virtualmente ilimitadas de *web scrapping* sejam feitas no Reddit, contornando as limitações da própria API do Reddit, PRAW. A documentação do Pushift se encontra disponível em seu GitHub[3]. Além disso, passou a usar a biblioteca para arquivar um volume massivo de dados do fórum ao longo dos anos[4]. Outro usuário, Felipe Hoffa, foi responsável por carregar estes dados compilados por Jason dentro da estrutura de nuvem do Google BigQuery, disponibilizando os dados publicamente na tabela *fh-bigquery*[5].

A tabela acima possui 1,7 bilhão de comentários em diferentes níveis de agregação entre 2015 e 2019. Optamos por usá-la pois a construção de uma ferramenta de *web scrapping* própria tomaria muito tempo e não é o escopo deste curso. Visando completude, consideramos comentários em todo Reddit, independentemente do subreddit (sub-fórum) e do tipo (respostas à publicações ou publicações em si). Como o intuito é considerar apenas sentimentos relacionados ao Bitcoin, filtramos apenas comentários que continham "Bitcoin" ou "BTC". A consulta em SQL usada no BigQuery em cima da tabela *fh-bigquery* foi:

```
SELECT
  subreddit,
  created_utc,
  body,
  score
FROM
  [reddit-btc-analysis:comments.reddit_btc_comments]
WHERE
  (LOWER(body) LIKE '% bitcoin%'
    OR LOWER(body) LIKE '% bitcoin %'
    OR LOWER(body) LIKE '% btc.%'
    OR LOWER(body) LIKE '% btc %')
```

Figura 4 – Consulta realizada no Google BigQuery. Elaboração própria.

[3] <https://github.com/pushshift/api>. Acesso em 24 Mar. 2022.
[4] <https://pushshift.io/>. Acesso em 24 Mar. 2022.
[5] <https://www.reddit.com/r/bigquery/comments/3cej2b/17_billion_reddit_comments_loaded_on_bigquery/>. Acesso em 24 Mar. 2022.

Acima consideramos só os comentários transformados em letras minúsculas para fins de filtragem. Vale o adendo de que o filtro considera espaços específicos em " btc " e " btc." pois algumas linguagens possuem a combinação destas letras em outras palavras. Selecionamos dados sobre o subreddit do comentário, data de criação em UTC (Tempo Universal Coordenado), corpo da mensagem e Score (votos positivos - negativos). Não trouxemos mais dados pois existe limite de 1 terabyte para consulta gratuita na ferramenta.

Então a tabela acima foi salva como base de dados no BigQuery e exportada como extensões CSV em vários lotes de pouco mais de 1 milhão de comentários cada um. Quando reagregados, obtivemos uma base de 5.137.769 de comentários contendo menções ao Bitcoin entre 01-01-2015 e 31-12-2019.

3.2.2 Dados do Bitcoin

Existem diversos portais com APIs próprias e gratuitas para extração de dados de Bitcoin. Ao longo do desenvolvimento deste trabalho, desenvolvemos um código para interagir com a API do CryptoCompare trazendo o valor do Bitcoin em dólares por hora[6]. Porém, devido à demora e à restrição de consultas gratuitas, optamos por seguir com uma base pronta do Kaggle, cobrindo um período mais amplo, entre 2012 e 2021. A base de dados foi compilada e carregada na plataforma pelo usuário Zielak, trazendo dados do preço em dólares de abertura, fechamento, alta, baixa, ponderado e volume negociado do Bitcoin em intervalos de 1 minuto. Mais detalhes sobre a base estão no endereço do rodapé[7]. Aqui simplesmente baixamos o arquivo em extensão CSV.

3.3 Tratamento e Preparação dos Dados

Devido ao volume das bases e ao extensivo processamento necessário, detalhamos cada parte separadamente a seguir.

3.3.1 Tratamento - Dados do Reddit

Concatenando os dados extraídos do Reddit via Google BigQuery da tabela *fh-bigquery* obtivemos um total de 5.137.769 comentários contendo "Bitcoin"ou "BTC". A tabela gerada possui as seguintes colunas:

- Subreddit: o nome do sub-fórum onde foi feito o comentário.
- Body: o corpo do comentário, o texto.

[6] <https://github.com/Yamamuen/btc_pred/blob/main/cryptocompare.py>. Acesso em 24 Mar. 2022.

[7] <https://www.kaggle.com/datasets/mczielinski/bitcoin-historical-data>. Acesso em 24 Mar. 2022.

- Created UTC: data de criação em segundos de acordo com UTC (Tempo Universal Coordenado).
- Score do Comentário: é igual ao número de votos positivos subtraído dos votos negativos.

Daqui em diante utilizamos o pacote Swifter[8] para suplementar os tratamentos com processamento paralelo (automaticamente usa todos as unidades de processamento lógico do computador).

Primeiramente convertemos a data de criação de UTC para o formato de *Timestamp*, com data, hora e segundo. Em seguida tratamos o texto removendo formatações (HTML e CSS), espaços extra, negrito, itálico e removendo links para sites. A despeito do VADER trabalhar bem com contrações, gírias, emojis e outros estilos online, julgamos válido este tratamento preliminar. Também removemos 6.578 linhas exatamente duplicadas.

Sobre os comentários tratados aplicamos a ferramenta de Processamento de Linguagem Natural Fast Text[9] treinada em cima de uma base de 176 idiomas, de modo a identificar os comentários na língua inglesa, uma vez que o VADER só funciona em textos em inglês. Abaixo temos o resultado da análise e percebemos que, em relação ao total, o número de comentários em outros idiomas é relativamente baixo. Filtramos a base para preservar apenas comentários em inglês.

	english	others
contagem	5077054	60715

Tabela 2 – Idiomas da base. Elaboração própria.

Então aplicamos o Valence Aware Dictionary and Sentiment Reasoner (VADER) nos comentários ingleses tratados, extraindo a polarização composta, que retorna um agregado ponderado do texto de acordo com cada polarização das palavras contidas. Esta polarização composta varia entre -1 (extremamente negativa) até +1 (extremamente positiva). Mesmo utilizando processamento paralelo e recrutando todos os processadores lógicos, esta etapa tomou 15 horas e 35 minutos devido ao tamanho da base. Alternativamente, usar uma solução que recrutasse todos os processadores da GPU ao invés da CPU talvez tivesse acelerado o processo. Obtemos uma base no formato da tabela 3. Filtrando as polarizações máximas, tanto positivas quanto negativas (+/-1), notamos que são 3 casos, todos *spams*, então optamos por removê-los.

[8] <https://github.com/jmcarpenter2/swifter>. Acesso em 24 Mar. 2022.
[9] <https://fasttext.cc/docs/en/language-identification.html>. Acesso em 24 Mar. 2022.

subreddit	score	Created Date	Comment	Sentiment
millionairemakers	2.0	2015-01-01 00:00:43	Starting out is tough. Strict Regulations in t...	0.8163
changetip	0.0	2015-01-01 00:01:22	To quit looking at the price of btc every 15 m...	0.0000
sportsbook	2.0	2015-01-01 00:01:22	That's the reason I dislike betting using bitc...	0.1441
Bitcoin	2.0	2015-01-01 00:01:28	Yup, you probably already installed bitcoin-se...	0.5927
Bitcoin	0.0	2015-01-01 00:01:46	I really like Coinbase for its iPhone app. I h...	0.9657

Tabela 3 – Excerto da base pós tratamento. Elaboração própria.

Como cada *score* sinaliza validação por outros usuários, consideramos esta variável como *proxy* para influência e apoio, de forma similar ao "efeito influencer" do trabalho de Mohapatra, Ahmed e Alencar (2019), porém no Reddit. Portanto, optamos por aplicar o MinMaxScaler[10], normalizando os *scores* entre 0 e 1. Eis o racional: dado que existem pontuações extremas com milhares de votos positivos ou negativos e grande maioria sem nenhuma interação ($score = 0$), decidimos dar o peso mínimo de 0 para o comentário com maior número de reprovações e de 1 para o comentário com maior número de votos positivos, evitando assim a desconsideração da grande maioria de comentários, que não possuem nenhum voto ($score = 0$) e que, portanto, teriam peso igual a 0. Dessa forma a composição com a polarização do sentimento é direta, bastando multiplicar o *score* normalizado ([0, 1]) com a polarização ([-1, 1]), devidamente amplificando ou reduzindo o peso da polarização daquele comentário no dia de acordo com aprovação ou reprovação dos demais usuários.

3.3.2 Tratamento - Dados do Bitcoin

A base baixada do Kaggle possuí 4.857.376 entradas. Novamente transformamos a data, porém desta vez de segundos em Unix UTC para data, hora e segundo da operação. Obtemos o seguinte formato:

Timestamp	Open	High	Low	Close	Volume_(BTC)	Volume_(Currency)	Weighted_Price	Date
1617148560	58714.31	58714.31	58686.00	58686.00	1.384487	81259.372187	58692.753339	2021-03-30 23:56:00
1617148620	58683.97	58693.43	58683.97	58685.81	7.294848	428158.146640	58693.226508	2021-03-30 23:57:00
1617148680	58693.43	58723.84	58693.43	58723.84	1.705682	100117.070370	58696.198496	2021-03-30 23:58:00
1617148740	58742.18	58770.38	58742.18	58760.59	0.720415	42332.958633	58761.866202	2021-03-30 23:59:00
1617148800	58767.75	58778.18	58755.97	58778.18	2.712831	159417.751000	58764.349363	2021-03-31 00:00:00

Tabela 4 – Excerto da base pós tratamento. Elaboração própria.

Por fim, como consideraremos variações diárias do Bitcoin, precisamos agrupar estes dados, sem perda informacional, por dia. Para tanto:

- Agrupamos a estampa temporal por dia.
- Open: consideramos a primeira entrada desta coluna por dia.

[10] <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>. Acesso em 24 Mar. 2022.

- Close: consideramos a última entrada desta coluna por dia.
- High: consideramos o valor máximo desta coluna por dia.
- Low: consideramos o valor mínimo desta coluna por dia.
- Volume_(BTC): apenas somamos o volume transacionado por dia.
- Volume_(Currency): apenas somamos o volume transacionado por dia.

De tal forma conseguimos traduzir a informação, originalmente minuto a minuto, no equivalente diário exato. Por fim, filtramos para o mesmo período da base do Reddit, compreendendo os dias entre 01-01-2015 e 31-12-2019.

3.4 Análise de Dados

Nesta seção investigaremos os dados tratados com o objetivo de averiguar se existe relação ao longo do tempo entre a variação de sentimentos acerca do Bitcoin no Reddit e a oscilação da criptomoeda em si.

3.4.1 Análise de Dados de Sentimentos do Reddit

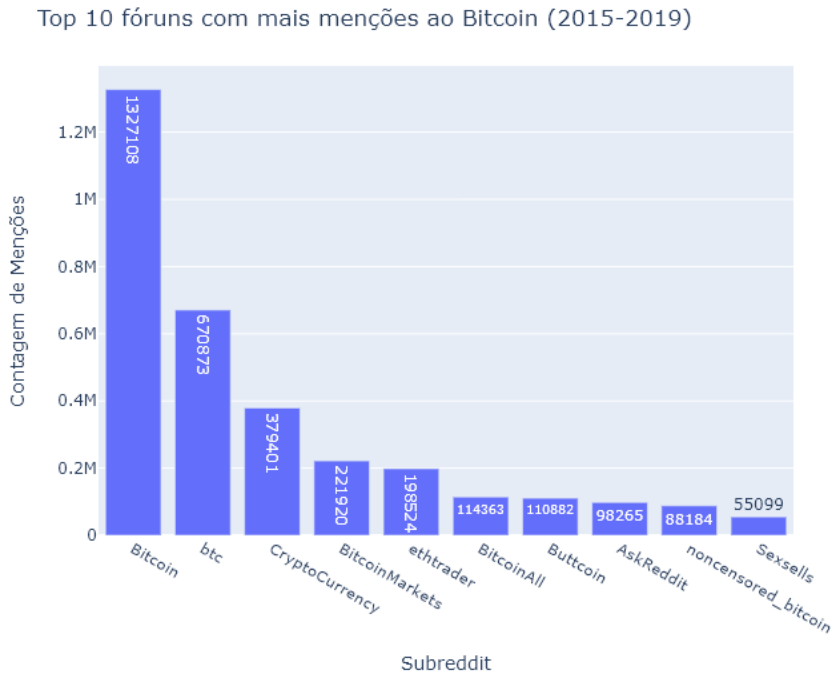

Figura 5 – Top 10 Subreddits com menções ao BTC. Elaboração própria.

Iniciamos a análise conferindo quais são os principais subreddits com menções ao Bitcoin entre 2015 e 2019. Vale ressaltar que o acesso e o conteúdo do Reddit são irrestritos

e, por conta disso, há um número considerável de menções ao Bitcoin (gráfico da figura 5) em fóruns com nomes peculiares. Os 2 últimos ("noncensored_bitcoin" e "Sexsells") dentre os top 10 surpreendem, mas se deve à importância do Bitcoin para pessoas trabalhadoras sexuais online[11], que recebem na moeda e, portanto, têm interesse na mesma como reserva de valor.

Com a democratização do acesso à tecnologia e investimentos em criptomoedas, existem poucas barreiras à entrada. Logo, se restringíssemos as análises apenas aos fóruns "sérios" corremos o risco de perder riqueza informacional sobre como pessoas reais e nichos que impactam a demanda do Bitcoin estavam antecipando ou reagindo às oscilações. É sabido que fóruns online são o lar de *trolls*[12] mas, ainda assim, com a finalidade de captar o humor geral de mercado, não vamos descartar informações de nenhum subreddit com base em seu nome, público ou seriedade. Dentre os top 4 fóruns com menções temos os principais especializados: Bitcoin, btc, CryptoCurrency e BitcoinMarkets. Do total dos comentários ao longo do período considerado, 51,3% se concentrou nestes.

De acordo com a documentação do VADER[13], sentimentos compostos são positivos quando maiores que 0,05, neutros entre -0,05 e +0,05 e negativos quando menores que -0,05. Investigamos a existência de viés nos 10 principais fóruns com menções:

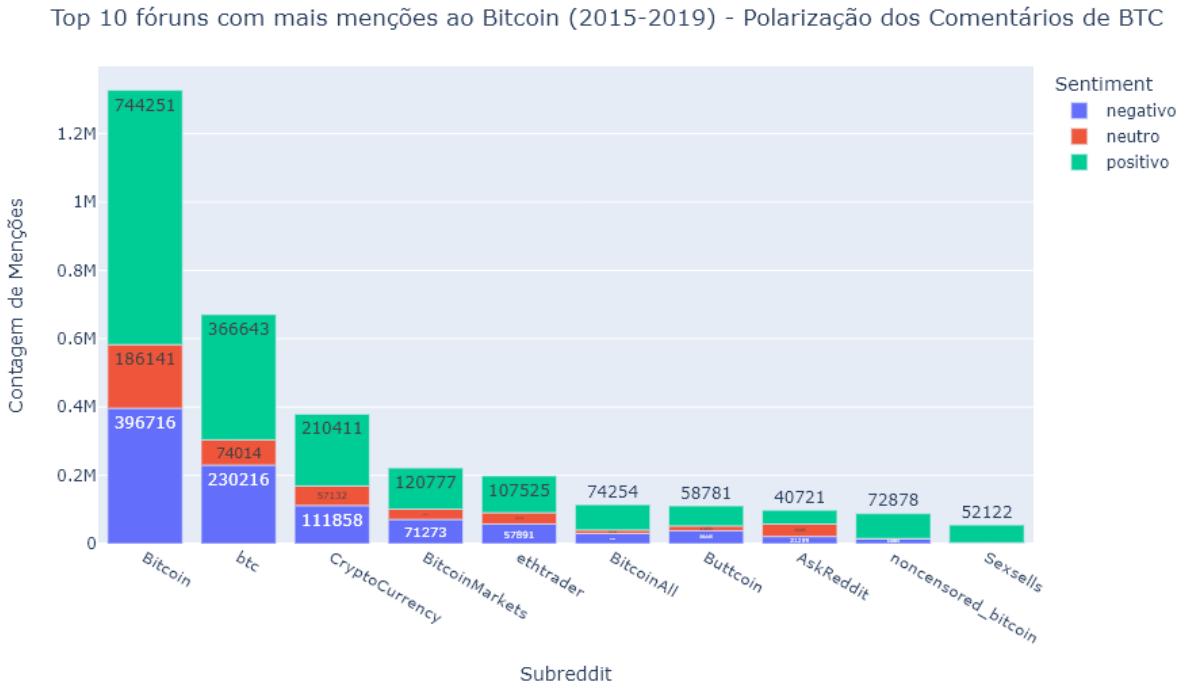

Figura 6 – Top 10 Subreddits e Polarização. Elaboração própria.

11 <https://www.cnbc.com/2022/02/05/bitcoin-a-lifeline-for-sex-workers-like-ex-nurse-making-1point3-mi
html>. Acesso em 24 Mar. 2022.

12 Em suma pessoas que aplicam trotes por brincadeira na Internet.

13 <https://github.com/cjhutto/vaderSentiment>. Acesso em 24 Mar. 2022.

Percebemos que os 2 últimos tem um viés mais positivo, o que se deve ao supracitado: trabalhadores sexuais digitais que recebem em Bitcoin tendem a manifestar sentimentos mais positivos pois não veem a moeda como aposta ou investimento, mas como reserva de valor. Os outros parecem mais balanceados, com certo viés otimista, porém Bitcoin e btc saltam aos olhos com sentimentos positivos predominando consideravelmente. Isso pode ter relação com o público que é atraído por estes fóruns ou até mesmo com o período considerado, em que o preço do Bitcoin foi, predominantemente, ascendente.

Passamos então por uma breve análise de quais palavras são mais frequentes em comentários com menções ao Bitcoin via Word Cloud, em todo Reddit temos:

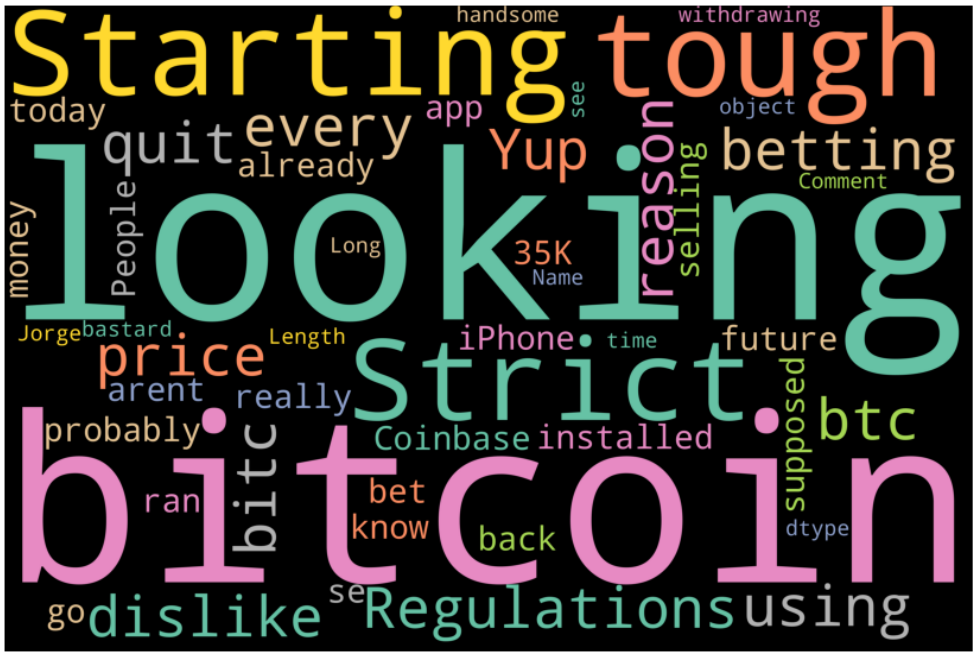

Figura 7 – Word Cloud, palavras mais frequentes. Elaboração própria.

Temos a palavra "procurando", "estrito", "difícil", "apostando", "começando" e "regulamentações" em proeminência. Algumas destas possivelmente refletem fatores exógenos, como as regulamentações cada vez mais estritas acerca da criptomoeda. Chama atenção também o número de pessoas começando e procurando por algo, mas uma breve investigação nos subreddits mostra que muitos buscam ajuda nos fóruns, com publicações e guias inteiros dedicados a iniciantes, além de fóruns especializados nisso, como o /BitcoinBeginners.

Já nos fóruns especializados (top 4, figura 8) percebemos menções a conflitos, ao Coinbase (plataforma online de compra e venda), ao cenário atual, a medo, aplicativos, previsões, reservas e iPhone. Várias plataformas de criptomoedas oferecem aplicações para iOS e, considerando que a população dos países de língua inglesa majoritariamente utiliza

dispositivos iOS[14] a presença de iPhone neste gráfico é justificada.

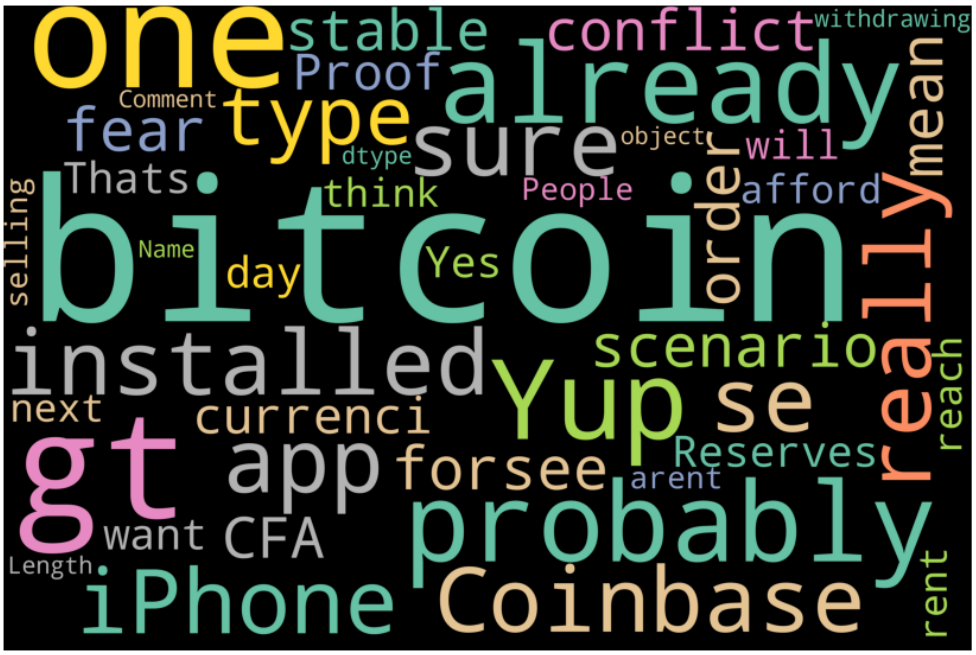

Figura 8 – Word Cloud, Top 4 especializados. Elaboração própria.

Decidimos dividir os comentários em negativos, neutros e positivos e repetir a análise em todo Reddit, nas figuras 9, 10 e 11, respectivamente.

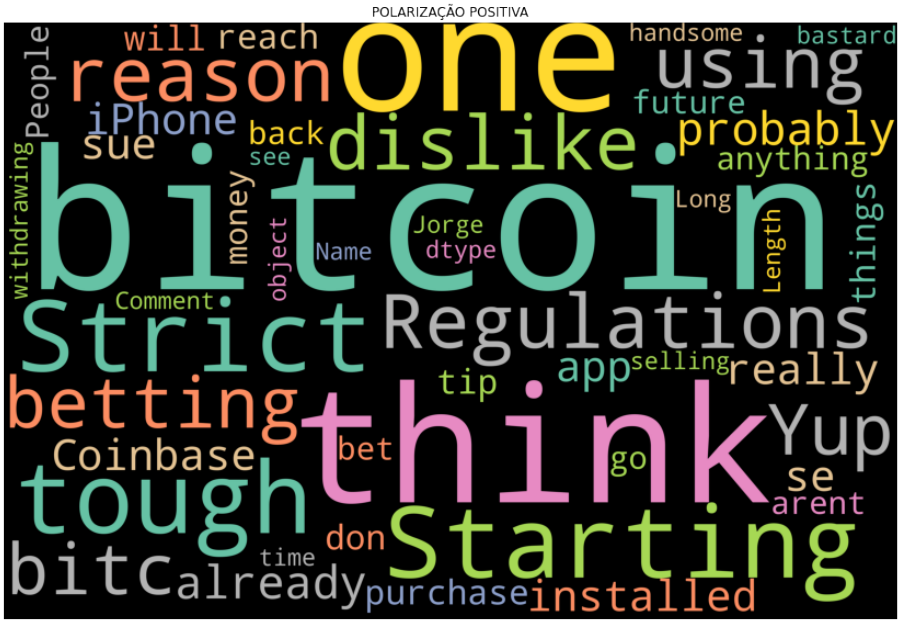

Figura 9 – Word Cloud, Comentários Positivos. Elaboração própria.

[14] <https://deviceatlas.com/blog/android-v-ios-market-share>. Acesso em 24 Mar. 2022.

Figura 10 – Word Cloud, Comentários Neutros. Elaboração própria.

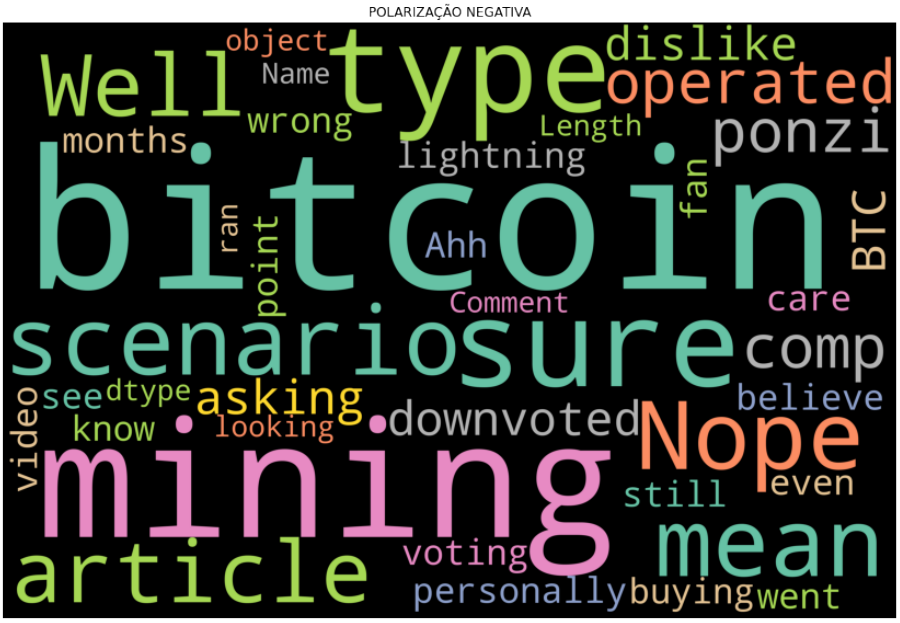

Figura 11 – Word Cloud, Comentários Negativos. Elaboração própria.

No Word Cloud de comentários positivos (figura 9) chamam atenção: "não gostar", "difícil", "estrito" e "regulamentações", dentre outros. Mas vale reforçar que o VADER analisa o comentário inteiro e não palavras singulares. A palavra "pensar" também se destaca, sinalizando que possivelmente grande parte dos comentários positivos são opiniões.

Já no de comentários neutros (figura 10) percebemos várias palavras igualmente recorrentes e nada muito conclusivo.

Por fim, na figura 11 temos as palavras mais recorrentes nos comentários negativos: temos grande quantidade de referências à "mineração" de Bitcoin, "cenários", "ponzi", "artigos" e à palavra "certeza". As demais palavras mais recorrentes tem aparente conotação negativa. Neste caso as associações negativas são mais flagrantes, como a presença de menções a esquemas Ponzi, possivelmente pessoas que enxergam a criptomoeda como um esquema de pirâmide e pessoas expressando descontentamento com outros comentários, sinalizando que deram votos negativos (*downvoted*) ou *dislikes*.

Buscando entender como os sentimentos ponderados pela pontuação e agregados por dia se comportaram ao longo do tempo:

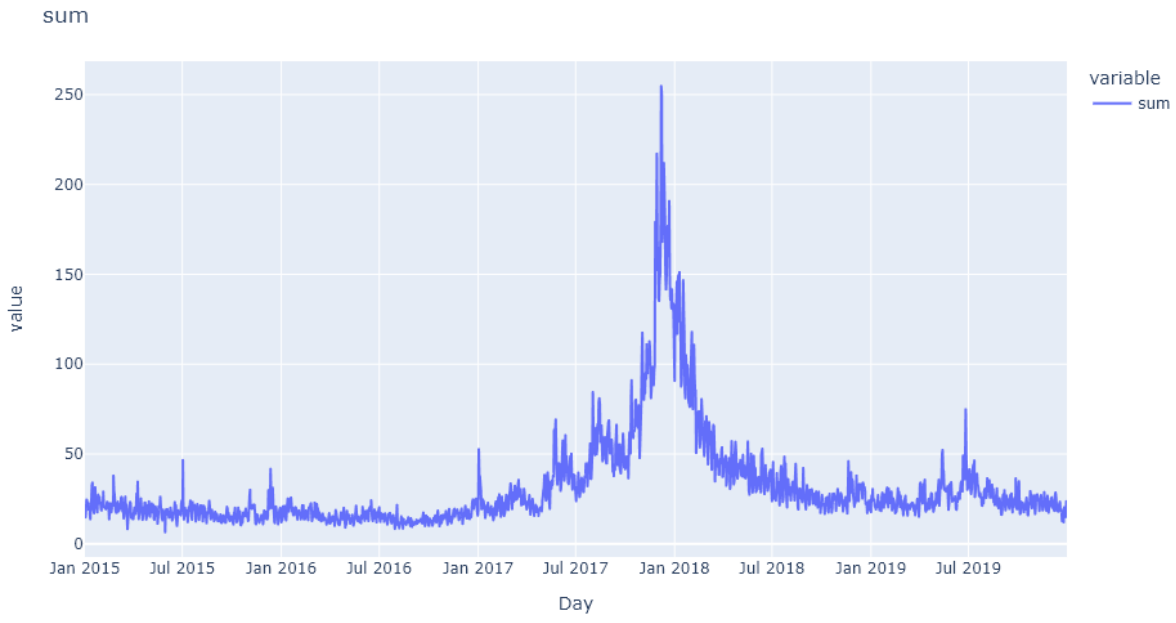

Figura 12 – Soma de Sentimentos Ponderados no Tempo. Elaboração própria.

Nota-se que em Janeiro de 2018 houve algo que impactou os sentimentos negativamente após uma ascensão forte em 2017. A queda na polarização somada por todo Reddit começa em dezembro de 2017, um mês antes do que foi chamado "*crash* do Bitcoin", quando na segunda quinzena de Janeiro verificou-se a queda de 25% na criptomoeda.

Nesta etapa foi verificado que as séries temporais de sentimentos agregados era

idêntica quando considerados os comentários de todo Reddit ou os comentários dos Top 4 fóruns (especializados). Portanto, de modo a preservar riqueza informacional, optamos por dar seguimento no estudo com a base inteira.

3.4.2 Análise de Dados do Preço Bitcoin

Iniciamos com uma análise da série histórica do preço do Bitcoin observando seu comportamento no tempo.

Figura 13 – Oscilação do Bitcoin por dia. Elaboração própria.

Novamente destaca-se o pico ao longo de 2017 seguido pelo vale em janeiro de 2018. Elencamos os seguintes acontecimentos que ocorreram no período[15]:

- Ascensão histórica em 2017 com onda de otimismo.
- Queda especulativa no final de 2017.
- Queda por rumores de banimento na Coréia do Sul no começo de Janeiro de 2018.
- No final de janeiro de 2018, hackers invadiram o Coincheck, maior mercado de balcão de criptomoedas do Japão, provocando queda no preço de todas as criptomoedas.

Percebe-se que os 2 primeiros são efeitos via demanda e, portanto, passíveis de serem capturados via humor de mercado, enquanto os últimos dois itens podem ser considerados

15 <https://en.wikipedia.org/wiki/Cryptocurrency_bubble>. Acesso em 24 Mar. 2022.

fenômenos exógenos ou até mesmo *black swams* sendo, portanto, imprevisíveis. Porém, levantamos a hipótese de que monitorar fóruns com opiniões globais em tempo real talvez gere informações de forma mais tempestiva do que acompanhando portais oficiais de imprensa, em particular para ativos de natureza especulativa altamente impactada pela demanda como é o caso para criptomoedas no cenário atual.

Decompomos a série temporal em variação diária percentual e medidas de tendência e volatilidade:

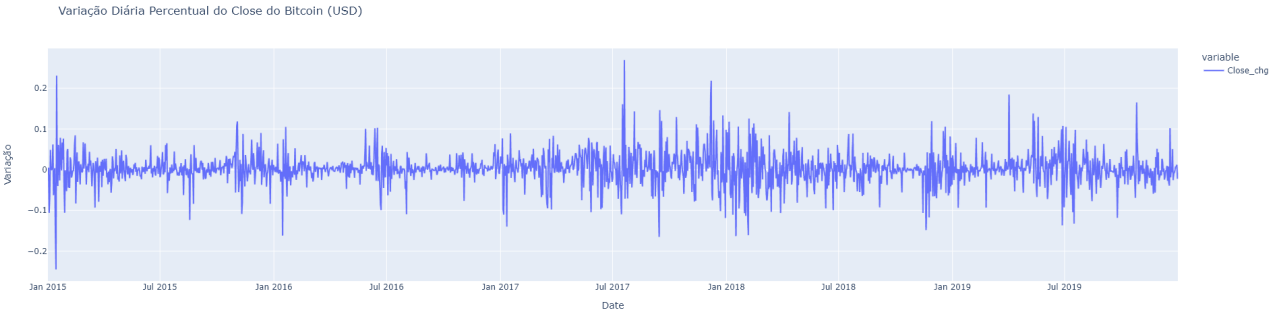

Figura 14 – Oscilação Percentual do Bitcoin por dia. Elaboração própria.

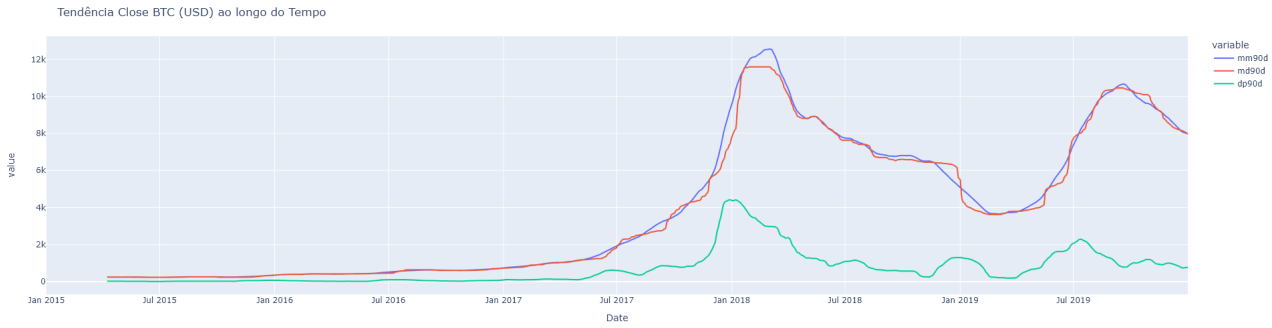

Figura 15 – Tendências do Bitcoin ao longo do tempo. Elaboração própria.

Fica evidente que nem a média nem o desvio padrão são constantes ao longo do tempo, sinalizando não-estacionariedade e, portanto, impossibilidade de usar modelos de predição de séries temporais mais simples como ARMA que não estacionarizam a série. Percebe-se que a frequência de 120 observações (ou aproximadamente 4 meses) com sazonalidade aditiva gera resíduos menos dispersos e sazonalidade mais demarcada na figura 17, que parece ter um certo padrão repetido a cada 4 meses. Observar-se-á que os resíduos explodem nos períodos correspondentes às instabilidades supracitadas.

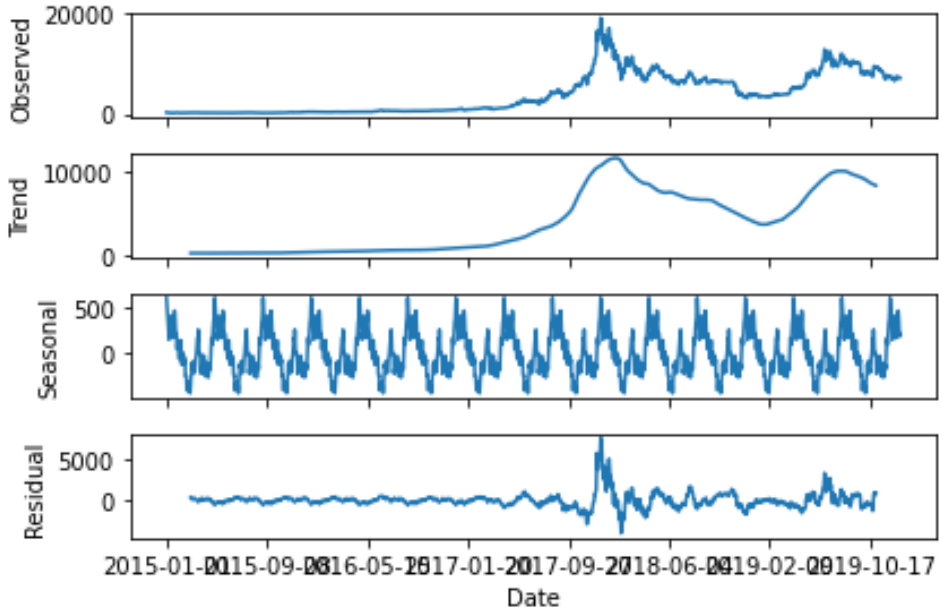

Figura 16 – Decomposição do Close do Bitcoin. Elaboração própria.

3.4.3 Análise Relacional entre Bitcoin e Sentimentos

Passamos então a considerar a variação conjunta entre as séries do Bitcoin em dólares e sentimentos agregados do Reddit. Para tanto, juntamos ambas em uma só base de acordo com o dia. Agrupadas por dia, ambas possuem 1.826 observações entre 01/01/2015 e 31/12/2019.

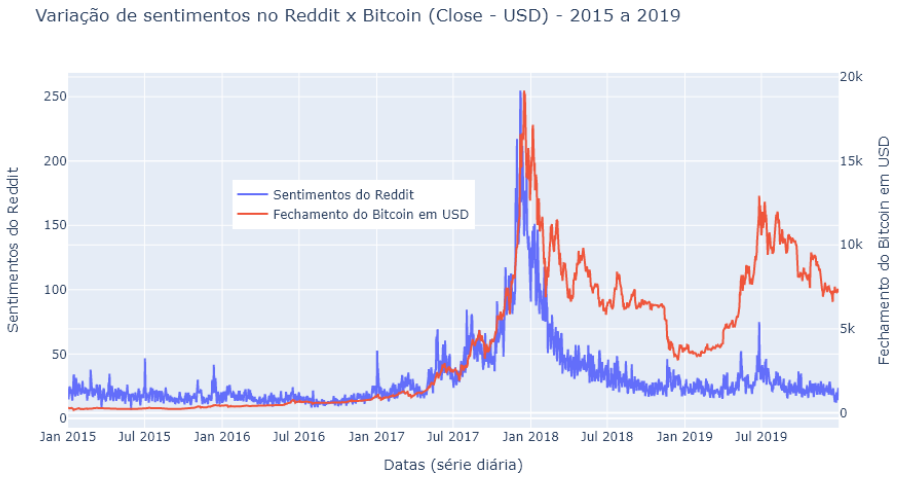

Figura 17 – Bitcoin e Sentimentos do Reddit. Elaboração própria.

É notável não só a correspondência os picos mas também como os sentimentos parecem antecipar o movimento de queda do Bitcoin que persistiu até janeiro de 2019.

Apesar do fechamento do Bitcoin ainda oscilar subindo e descendo pós janeiro de 2018, os sentimentos continuaram em queda vertiginosa, sem voltar aos níveis de 2017. Podemos aventar que após o *crash* de 2018 os usuários do Reddit ficaram, em geral, menos otimistas, ou que simplesmente o volume de menções caiu pois passaram a diversificar os investimentos com outras criptomoedas nascentes. Também percebemos picos nos sentimentos diários do Reddit ao longo de 2017 seguidos pela tendência sem precedentes de alta no preço da criptomoeda. Sabe-se que nem todas as pessoas que compram e vendem criptomoeda utilizam o Reddit, mas considerando que muitas, apesar de não interagirem nos fóruns, se informam pelo veículo (como mencionado nas estatísticas da introdução), pode-se inferir que trata-se de uma *proxy* que potencialmente reflete e (ou) resume os humores de mercado.

Partimos para uma investigação de distribuição dos sentimentos do Reddit no período correspondente ao maior pico e vale históricos acima e temos, na figura 18, que apesar da distribuição se manter relativamente constante, uma mudança acentuada no volume total de menções positivas é observada no final de dezembro de 2017, antecedendo o *crash*. Vale notar que o volume de comentários, e não somente seu teor, parece ter um efeito considerável na oscilação.

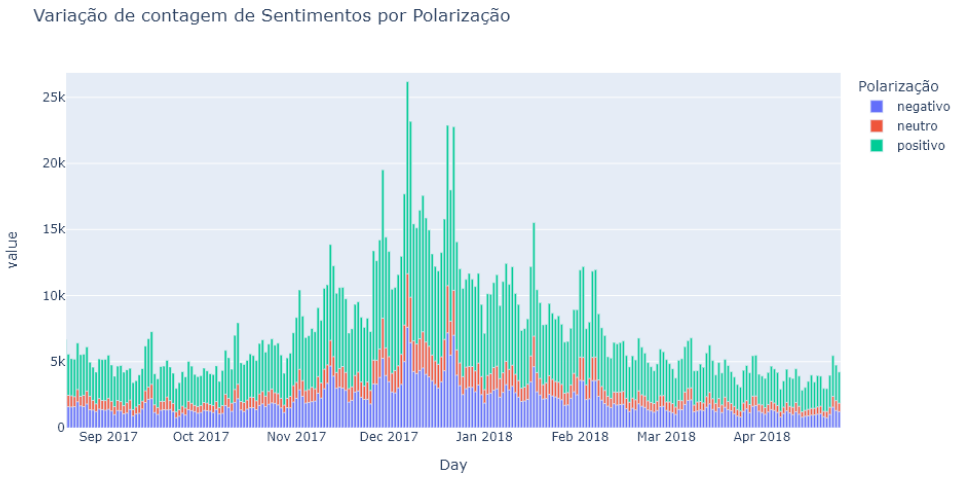

Figura 18 – Contagem de Polarização de Comentários do Reddit. Elaboração própria.

Para averiguar relações de causa efeito, repetimos a análise defasando os sentimentos agregados do Reddit em 2 semanas, obtendo a figura 19. A série de sentimentos fica ainda mais ajustada às oscilações do Bitcoin, sinalizando que talvez seja possível utilizar sentimentos do Reddit como indicativo de como a demanda está ou vai se comportar, impactando o Bitcoin. Possivelmente esse atraso ocorre devido ao tempo entre constatação dos sentimentos e efetiva decisão de compra por um ser humano, causado pela demora no processamento do humor de mercado ou até mesmo por hesitação, postergando o efeito manada.

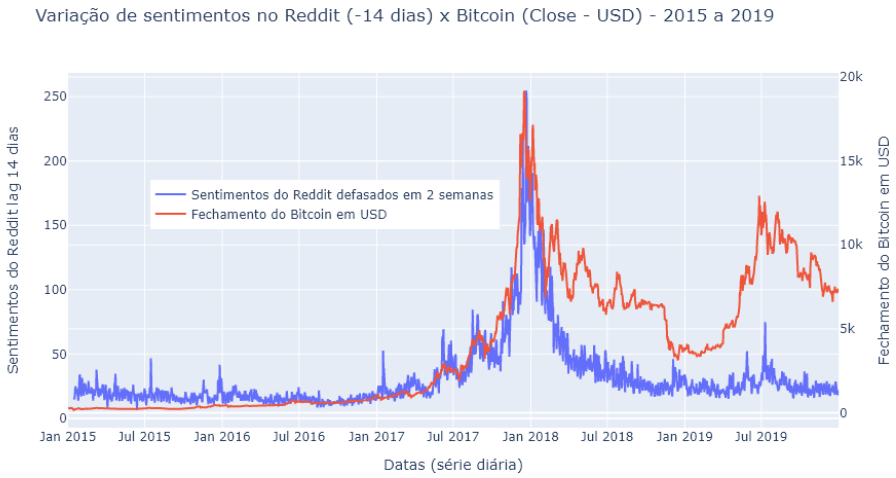

Figura 19 – Sentimentos defasados e Close de Bitcoin. Elaboração própria.

Em seguida, estudamos a associação entre Close de Bitcoin em dólares e sentimentos agregados do Reddit (figura 20), obtendo uma associação positiva: em dias com sentimentos positivos ou volumes altos o preço do Bitcoin tende a fechar em alta. Aqui vale ressaltar que o Close é o último preço registrado do Bitcoin em dólares no dia, ou seja, sentimentos agregados dentro do mesmo dia são associados a fechamentos maiores.

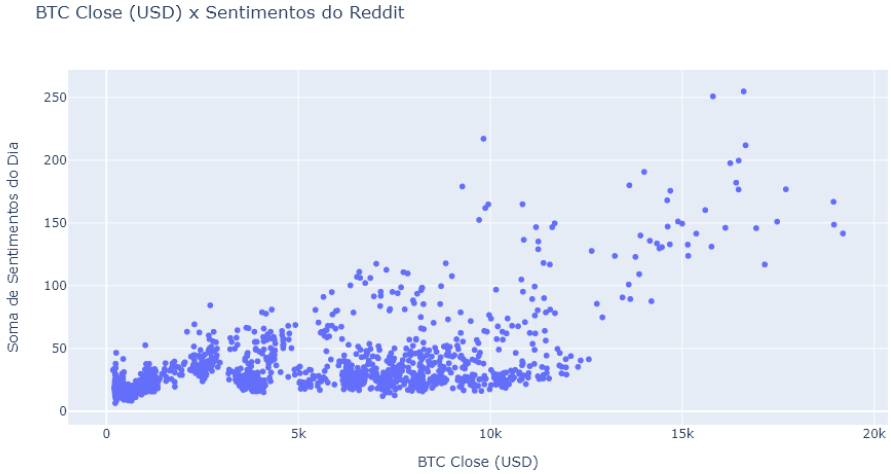

Figura 20 – Dispersão: Sentimento x Close Bitcoin. Elaboração própria.

Por sua vez, com sentimentos defasados em 2 semanas obtemos a relação da figura 21, também consideravelmente positiva e ligeiramente menos dispersa. Ou seja, sentimentos do Reddit do próprio dia ou em períodos anteriores possuem correlação linear positiva não irrisória com o preço de fechamento do Bitcoin em dólares.

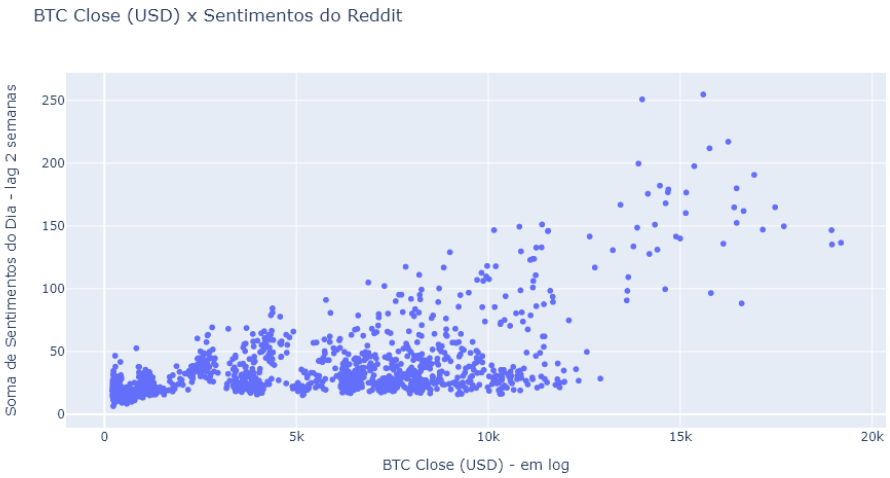

Figura 21 – Dispersão: Sentimento - 14 dias x Close Bitcoin. Elaboração própria.

Enfatizamos que a variável de sentimentos do Reddit é uma agregação diária que soma a pontuação de todos os comentários ponderados pelos respectivos *scores*, então de certa forma considerar volumes é tautológico. Ainda assim decidimos suplementar a visualização anterior com o volume correspondente a cada ponto e há indícios de que o volume de comentários também tende a ser maior em fechamentos maiores de Bitcoin (figura 22).

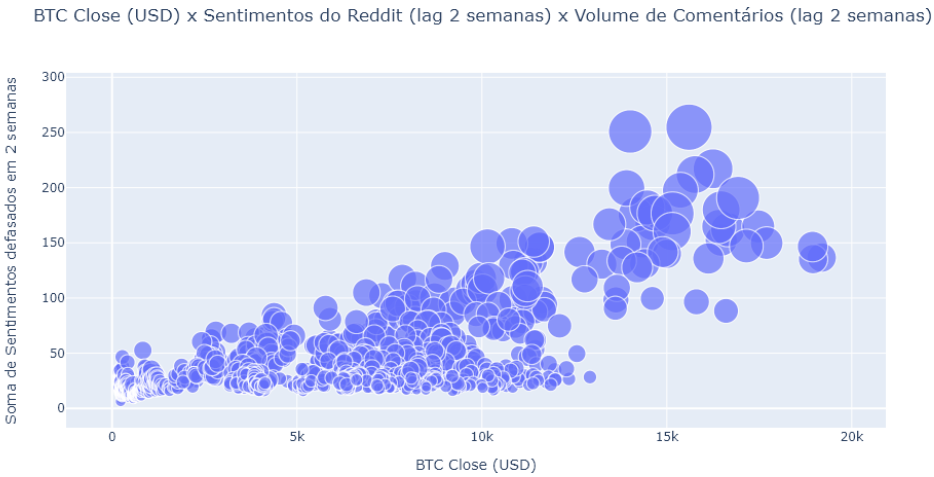

Figura 22 – Gráfico de Bolhas: Sentimento - 14 dias x Close Bitcoin x Volume de Comentários - 14 dias. Elaboração própria.

Por fim, embasando-nos nos resultados acima, investigamos a correlação linear de Pearson entre fechamento do Bitcoin com seus próprios valores passados e com sentimentos defasados em semanas múltiplas de 2.

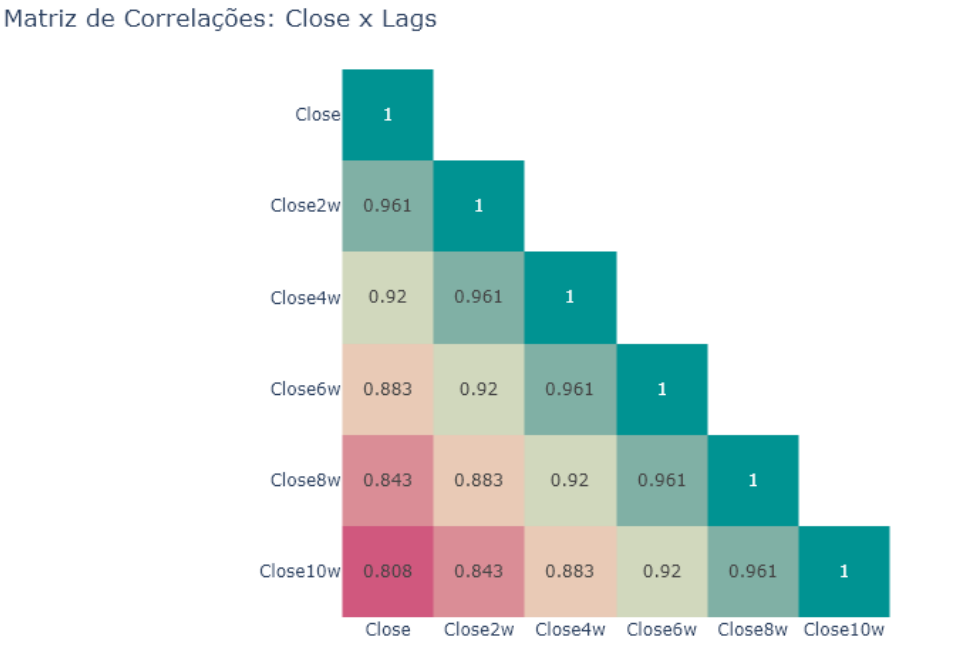

Figura 23 – Matriz de Correlação: Fechamento contra seus valores defasados. Elaboração própria.

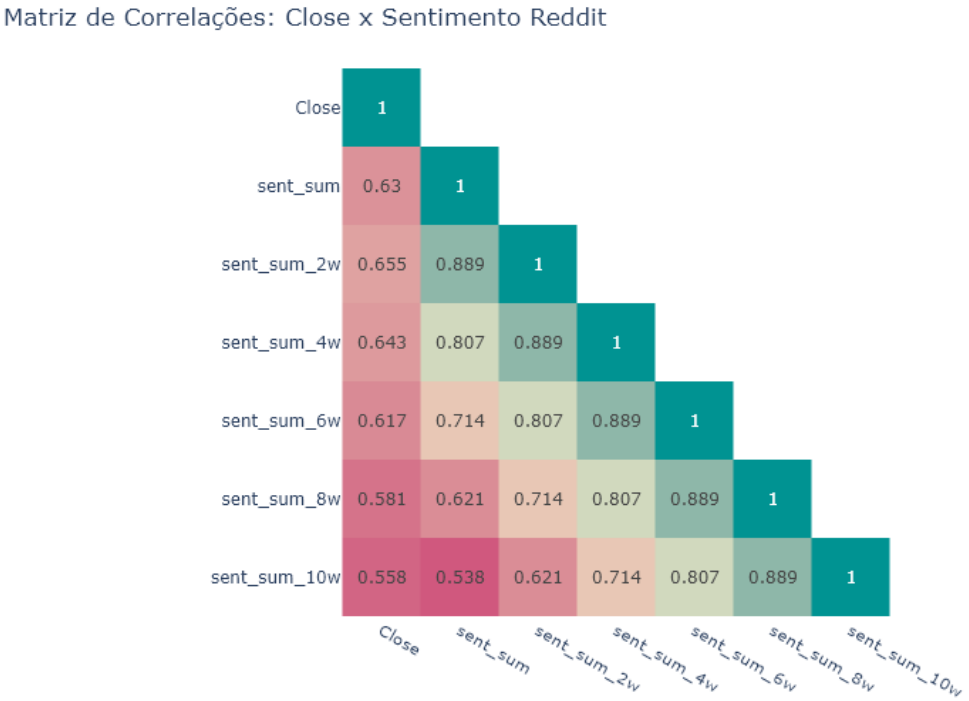

Figura 24 – Matriz de Correlação: Fechamento contra sentimentos defasados. Elaboração própria.

Pelos correlogramas acima conseguimos observar que o preço de fechamento do Bitcoin se trata de uma série com dependência temporal e memória (possivelmente auto-regressiva) e que sentimentos do Reddit defasados possivelmente podem contribuir com poder preditivo, dados os indícios de associação positiva alta. Ademais, as visualizações desta seção indicam que as oscilações no Reddit dão sinais de antecipar as do Bitcoin em certa medida.

Concluímos a análise de dados selecionando algumas variáveis para um esforço seminal de modelar a relação e identificar potencial preditivo na seção subsequente:

- Consideraremos a variável fechamento do Bitcoin como dependente, ou explicada.
- Consideraremos sentimentos diários agregados do Reddit como independente, ou explicativa. Optamos por não considerar volumes pois já estão incutidos nesta agregação diária.

3.5 Aprendizado de Máquina

Nesta etapa utilizaremos os dados e conclusões da seção anterior para ajustar um modelo de previsão do preço de fechamento do Bitcoin em dólares em função de sentimentos do Reddit.

O método utilizado será o Long Short Term Memory que, como descrito na revisão conceitual, é uma Rede Neural Recorrente adequada para séries com dependência temporal longa, onde não só a sequência das observações é considerada como também períodos distantes no tempo. O algoritmo automaticamente decide, via portas de esquecimento, quais informações e tempos passados serão ou não perpetuados para próximas rodadas de otimização de acordo com importâncias constatadas e atualizadas iterativamente.

Ademais, o LSTM também permite ajustar modelos multivariados à séries temporais, não possuindo a restrição de previsão futura condicionada unicamente a valores passados da própria variável explicada. Logo, estimaremos o preço de fechamento do Bitcoin em dólares com sentimentos do Reddit, com treinamento condicionado a como ambas as séries variam e sua relação ao longo do tempo.

A arquitetura LSTM se resume a:

- Entrada com dimensão de linhas equivalentes ao tamanho da janela escolhida e colunas ao número de atributos (1).
- Camada LSTM com 128 neurônios.
- Camada de ativação Leaky ReLU com 128 neurônios.
- Camada LSTM pós ativação com 128 neurônios.

- Camada de Dropout (0,3) para tentar prevenir *overfitting*.
- Saída Densa com apenas um neurônio: é um problema de predição, queremos estimar apenas um valor: close de Bitcoin.

Utilizamos uma camada de Leaky ReLU para evitar o problema de ReLUs normais: *Dead ReLU*, uma problema de gradiente de fuga em que, quando o modelo converge muito rápido, muitos neurônios acabam sendo desativados, minguando a capacidade do modelo. Ao invés de desativar os neurônios, a Leaky ReLU os mantém ativos com gradientes positivos ínfimos.

Resta definir:

- Janela temporal: quantos dias serão treinados e posteriormente utilizados para prever o próximo ponto?
- Tamanho do Batch: quantas amostras de N observações, com N = dias na Janela, serão usadas para prever o fechamento antes de cada *backpropagation* ser ativado, atualizando os novos pesos nas conexões entre neurônios?
- Epochs: quantas iterações para o experimento? Aqui fixamos em 200 epochs.

Sobre a janela temporal, dado que, em nossa análise preliminar, identificamos relações fortes de preço de fechamento do Bitcoin com sentimentos defasados semana a semana, decidimos testar várias janelas semanais: 1, 2, 4 e 10 semanas ou, respectivamente, [7, 14, 28, 70] dias.

Por sua vez, o tamanho do Batch foi escolhido de acordo com o padrão utilizado na literatura: múltiplos de 2, sendo os mais populares para bases pequenas 32 e 64. Trata-se de regra de dedo testada e aprovada para garantir boa generalização e eficiência computacional, não achamos nenhum artigo ou prova definitiva do porquê estes são os valores usualmente utilizados, apesar de vários artigos fazerem a recomendação[16]. Seguimos então testando os seguintes tamanhos de Batch: [4, 8, 32, 64, 256].

Para o treinamento segundo a arquitetura acima, utilizamos o otimizador Adam com taxa de aprendizado inicial de 0,001 e definimos a função perda a ser minimizada como o Erro Médio Absoluto, que simplesmente computa a média da diferença entre observado e predito em termos absolutos (optamos por esta função custo pois acreditamos que neste caso faz sentido a penalização estar na unidade e magnitude da diferença com o valor observado do Bitcoin). Com o intuito de evitar que o erro a ser minimizado divirja drasticamente ao longo de iterações mais tardias, inserimos uma função de decaimento

[16] <http://deeplearningbook.com.br/o-efeito-do-batch-size-no-treinamento-de-redes-neurais-artificiais/>. Acesso em 25 Mar. 2022.

exponencial de -0.1 no *callback* do ajuste. Por fim, como mais uma medida para evitar *overfitting*, definimos uma parada precoce: se após 3 iterações o erro no conjunto de validação não reduzir, interrompemos as Epochs.

Dividimos o conjunto treino como 80% das observações (de 2015 a 2018) e os últimos 20% (2019) como conjunto teste. Note que a ordem neste caso é de extrema importância, e no código aplicamos condições para assegurar que a ordem dos dados seja respeitada. Treinamos o modelo com os hiperparâmetros acima e validamos com o conjunto teste (para definir a parada, mas o conjunto teste em si não é utilizado para o treinamento, evitando vazamento de dados), computando as métricas de performance no conjunto teste para cada configuração, o que nos retorna a tabela 5.

Configuração	MSE	MAE	MAPE
Batches:4, Janela:7	0.036964	0.002086	0.036964
Batches:8, Janela:7	0.031455	0.001775	0.031455
Batches:32, Janela:7	0.049986	0.003773	0.049986
Batches:64, Janela:7	0.050934	0.003374	0.050934
Batches:256, Janela:7	0.030376	0.001684	0.030376
Batches:4, Janela:14	0.178074	0.045094	0.178074
Batches:8, Janela:14	0.029148	0.001342	0.029148
Batches:32, Janela:14	0.044508	0.003423	0.044508
Batches:64, Janela:14	0.104148	0.014979	0.104148
Batches:256, Janela:14	0.034186	0.002424	0.034186
Batches:4, Janela:28	0.030221	0.001490	0.030221
Batches:8, Janela:28	0.028412	0.001287	0.028412
Batches:32, Janela:28	0.111030	0.014240	0.111030
Batches:64, Janela:28	0.144233	0.024926	0.144233
Batches:256, Janela:28	0.055596	0.005234	0.055596
Batches:4, Janela:70	0.034192	0.001824	0.034192
Batches:8, Janela:70	0.037845	0.002122	0.037845
Batches:32, Janela:70	0.057765	0.005061	0.057765
Batches:64, Janela:70	0.225526	0.056159	0.225526
Batches:256, Janela:70	0.060865	0.005480	0.060865

Tabela 5 – Performance de diferentes configurações. Elaboração própria.

Na tabela também trazemos outras medidas de performance: MSE ou Erro Quadrático Médio (média dos desvios entre observado e predito ao quadrado) e MAPE ou Erro Absoluto Percentual Médio (equivalente ao erro absoluto médio sobre o observado). Porém,

como já dito, optamos por priorizar o MAE (Erro Absoluto Médio), então ordenando por MAE, MSE e então MAPE obtemos as 3 configurações que melhor performaram no conjunto de teste:

Configuração	MSE	MAE	MAPE
Batches:8, Janela:28	0.028412	0.001287	0.028412
Batches:8, Janela:14	0.029148	0.001342	0.029148
Batches:4, Janela:28	0.030221	0.001490	0.030221

Tabela 6 – Top 3 configurações. Elaboração própria.

Percebe-se que o ajuste gerou modelos que performaram consideravelmente bem no período de teste, dados os valores baixos das métricas. Então utilizamos a configuração que melhor performou no conjunto teste, de janela temporal de 28 dias (ou 4 semanas) com Batch de tamanho 8 e trazemos, graficamente, como as métricas se comportaram a cada iteração e como o modelo teria previsto o ano de 2019.

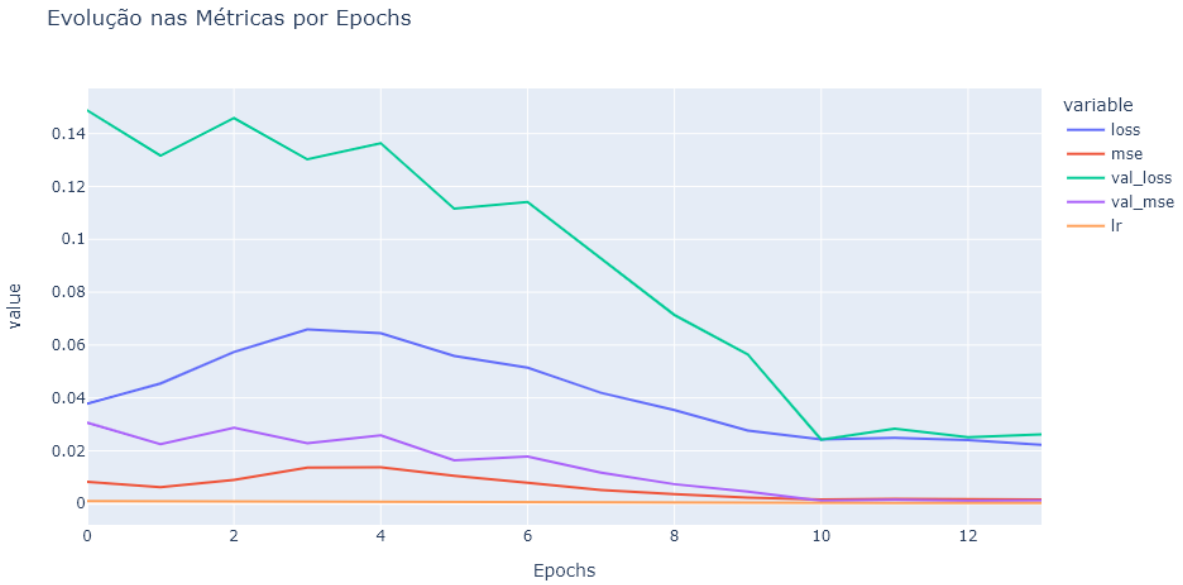

Figura 25 – Métricas ao longo das Iterações. Elaboração própria.

Figura 26 – Previsão no conjunto teste (2019). Elaboração própria.

A linha vermelha é a previsão no conjunto teste enquanto que a azul é a série observada em 2019. O modelo treinado consegue, com 28 observações prévias, prever bem o próximo close do Bitcoin. Tendo em vista que o ajuste observado acima parece "bom demais", vale testar, futuramente, a performance do modelo ajustado em outros períodos para conferir se a performance se repete ou investigar presença de *overfitting* (comum em LSTMs). Mesmo garantindo que não há vazamento de dados, implementando medidas para reduzir chance de sobre-ajuste e treinando apenas com sentimentos agregados do Reddit, causa estranheza o quanto o modelo acerta, com erro percentual médio (MAPE) de 2.8% em 2019.

4 CONCLUSÃO

4.1 Conclusão

O intuito inicial deste estudo era de averiguar a existência de relação causal entre oscilação de sentimentos do Reddit e do preço em dólares do Bitcoin. Inicialmente identificamos como as menções se comportaram, ficando clara a existência de viés positivo nos fóruns, e também trouxemos quais palavras são mais comumente associadas a textos positivos ou negativos em relação ao Bitcoin. Algumas palavras faziam sentido, indicando fatores exógenos como regulamentações, conflitos e afins, porém muitas não trouxeram nada conclusivo, o que era esperado, dado que o VADER polariza considerando o corpo textual como um todo e não as palavras singulares.

A associação entre sentimentos e Bitcoin foi observada via correlação linear de Pearson forte entre ambas e suas respectivas defasagens. Indícios de sentimentos causando as oscilações puderam ser observados graficamente: defasando a série de sentimentos em 2 semanas observamos melhor correspondência com as oscilações do Bitcoin, com picos de sentimentos diários precedendo os movimentos mais drásticos na série histórica da moeda. Também observamos relação entre volume de comentários e preço de fechamento do Bitcoin.

Por fim, utilizamos um algoritmo LSTM multivariado para predizer o preço do Bitcoin via sentimentos agregados do Reddit. Obtivemos erros por demais pequenos e um ajuste muito alto, então o potencial preditivo deve ser interpretado com cuidado e ceticismo. Também vale a reflexão de que durante o período considerado de fato grande parte dos investidores de criptomoedas eram amadores com forte presença no Reddit (o que pode justificar seu impacto no preço do Bitcoin), porém, mais recentemente, grandes grupos financeiros passaram a negociar criptomoedas, o que possivelmente reduz a capacidade de acerto do modelo em dados mais recentes.

Porém, o objetivo era provar existência de relação causal, o que, mesmo em agregações maiores (diárias), fica evidenciado. Dada a oferta fixa de Bitcoin, o preço deste é impactado unicamente pela demanda e fatores exógenos não previsíveis, logo, tendo em mente o formato do Reddit com um limite elevado de caracteres por comentários, conteúdo irrestrito (que possibilita não somente publicação de comentários, mas de artigos, notícias e opiniões) e sistema de validação por votos, este pode, potencialmente, servir de *proxy* não apenas para angariar humor de mercado mas também para obter de forma mais tempestiva polarizações decorrentes de notícias ou boatos relacionados a fatores exógenos ao redor do mundo antes mesmo destes serem divulgados por veículos de imprensa oficiais.

4.2 Próximos Passos

Dada a entrada de novos tipos de investidores no mercado de criptomoedas citada anteriormente e novas dinâmicas subsequentes, é válido testar o modelo em períodos mais recentes e auferir sua performance. Além disso, outros controles devem ser testados para evitar *overfit*, dada a suspeita levantada.

Possivelmente um estudo alternativo que meça quanto de Bitcoin é negociado reagindo a sentimentos online traga mais *insights* para a predição de efeitos na demanda.

Outro ponto de relevo é a polarização enviesada que obtivemos ao aplicar o VADER nos comentários do Reddit. Dado que o VADER é especificamente calibrado para mídias digitais e para reconhecer estilos, intensidades e símbolos, talvez os tratamentos preliminares foram por demais extensos, gerando classificações imprecisas. Nessa linha, uma filtragem de *spams* também poderia melhorar o modelo. Alternativamente, outras formas de agregar e considerar sentimentos (ou periodicidades mais granulares), além da soma ponderada diária, hão trazer novos achados.

Neste trabalho, devido à questões de prazo, nos restringimos à consulta de bases prontas para as duas fontes, porém, como mencionado nos capítulos introdutórios, existem várias APIs e possibilidades de integração. Semelhante ao feito por Mohapatra, Ahmed e Alencar (2019), poderíamos trazer dados em tempo real com uma arquitetura de *streaming* de dados de diferentes fontes online (como Twitter, Reddit e portais de notícia) via algum sistema de pub/sub como Kafka alimentando uma arquitetura de Apache Spark para realizar os tratamentos e treinamento em tempo real, já retornando as predições segundo o modelo selecionado e melhorando a cada iteração via aprendizado online. Virtualmente este mesmo fluxo poderia ser generalizado para qualquer criptomoeda ou ativo de interesse, com o adendo de que decisões de investimento em outros ativos, como ações, tendem a ser mais pautadas em considerações fundamentalistas. No futuro este modelo pode ser convertido em um algoritmo para compra e venda de criptomoedas em tempo real, capitalizando em cima da defasagem entre deliberação humana e tomada de decisão para auferir lucros.

REFERÊNCIAS

ABRAHAM, J. et al. Cryptocurrency price prediction using tweet volumes and sentiment analysis. **SMU Data Science Review**, v. 1, n. 3, p. 1, 2018.

BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. **Journal of computational science**, Elsevier, v. 2, n. 1, p. 1–8, 2011.

DOLAN, R. J. Emotion, cognition, and behavior. **science**, American Association for the Advancement of Science, v. 298, n. 5596, p. 1191–1194, 2002.

FAMA, E. F. Efficient capital markets: Ii. **The Journal of Finance**, v. 46, n. 5, p. 1575–1617, 1991. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1991.tb04636.x>.

FAMA, E. F. et al. The adjustment of stock prices to new information. **International Economic Review**, [Economics Department of the University of Pennsylvania, Wiley, Institute of Social and Economic Research, Osaka University], v. 10, n. 1, p. 1–21, 1969. ISSN 00206598, 14682354. Disponível em: <http://www.jstor.org/stable/2525569>.

GARCÍA, S.; LUENGO, J.; HERRERA, F. **Data preprocessing in data mining**. [S.l.]: Springer, 2015. v. 72.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

HUTTO, C.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: . [S.l.: s.n.], 2015.

KAHNEMAN, D.; TVERSKY, A. Choices, values, and frames. In: **Handbook of the fundamentals of financial decision making: Part I**. [S.l.]: World Scientific, 2013. p. 269–278.

KOULOUMPIS, E.; WILSON, T.; MOORE, J. Twitter sentiment analysis: The good the bad and the omg! In: **Fifth International AAAI conference on weblogs and social media**. [S.l.: s.n.], 2011.

LINDHOLM, A. et al. **Machine Learning - A First Course for Engineers and Scientists**. [s.n.], 2021. Disponível em: <https://smlbook.org>.

LORENA, A. C.; GAMA, J.; FACELI, K. **Inteligência Artificial: Uma abordagem de aprendizado de máquina**. [S.l.]: Grupo Gen-LTC, 2000.

MITCHELL, T. M. et al. Machine learning. McGraw-hill New York, 1997.

MOHAPATRA, S.; AHMED, N.; ALENCAR, P. Kryptooracle: A real-time cryptocurrency price prediction platform using twitter sentiments. In: . [S.l.: s.n.], 2019. p. 5544–5551.

MUNIM, Z. H.; SHAKIL, M. H.; ALON, I. Next-day bitcoin price forecast. **Journal of Risk and Financial Management**, Multidisciplinary Digital Publishing Institute, v. 12, n. 2, p. 103, 2019.

NAKAMOTO, S. Bitcoin: A peer-to-peer electronic cash system. **Decentralized Business Review**, p. 21260, 2008.

NGUYEN, T. H.; SHIRAI, K. Topic modeling based sentiment analysis on social media for stock market prediction. In: **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**. [S.l.: s.n.], 2015. p. 1354–1364.

ROBERTS, J. J. Big bitcoin crashes: What we learned. **Fortune, available at: http://fortune. com/2017/09/18/bitcoin-crash-history/(18 September)**, 2017.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **nature**, Nature Publishing Group, v. 323, n. 6088, p. 533–536, 1986.

SHAH, D.; ISAH, H.; ZULKERNINE, F. Predicting the effects of news sentiments on the stock market. In: IEEE. **2018 IEEE International Conference on Big Data (Big Data)**. [S.l.], 2018. p. 4705–4708.

STENQVIST, E.; LÖNNÖ, J. **Predicting Bitcoin price fluctuation with Twitter sentiment analysis**. 2017.

SUL, H. K.; DENNIS, A. R.; YUAN, L. Trading on twitter: Using social media sentiment to predict stock returns. **Decision Sciences**, Wiley Online Library, v. 48, n. 3, p. 454–488, 2017.

VELAY, M.; DANIEL, F. Stock chart pattern recognition with deep learning. **arXiv preprint arXiv:1808.00418**, 2018.

WALCZAK, S. An empirical analysis of data requirements for financial forecasting with neural networks. **Journal of management information systems**, Taylor & Francis, v. 17, n. 4, p. 203–222, 2001.