Revista Terra & Cultura: Cadernos de Ensino e Pesquisa

ISSN 2596-2809

DATA MINING, MACHINE LEARNING E BUSINESS INTELLIGENCE - UM ESTUDO DE CASO SOBRE CRIPTOMOEDAS

DATA MINING, MACHINE LEARNING, AND BUSINESS INTELLIGENCE - A CASE STUDY ON CRYPTOCURRENCIES

Felipe Israel Marinho [1]
Mario Henrique Akihiko da Costa Adaniya [2]

RESUMO

Este trabalho objetivou-se no uso de técnicas diversas como Data Mining, Machine Learning e técnicas gráficas utilizando-se de ferramentas de Business Intelligence para a extração, tratativa e a análise de dados provenientes de redes sociais, mais especificamente em grupos de mensagem do aplicativo Telegram, na intenção de validar se podemos verificar correlação entre os sentimentos que os usuários expressam nesta rede social e a flutuação de preços da criptomoeda Bitcoin. A intenção é apresentar uma análise comparativa completa entre os dados provenientes da rede social Telegram com o cenário de maior alta do preço desta criptomoeda, em dezembro de 2017. Primeiramente uma pesquisa exploratória com objetivo de introduzir diversos aspectos de todas as tecnologias utilizadas no desenvolvimento deste trabalho será apresentada. A aplicação da pesquisa se dará em dados reais provenientes da rede social Telegram. Ao final teremos um quadro comparativo utilizando modelos estatísticos para encontrarmos uma correlação entre as variáveis trabalhadas durante o desenvolvimento e a pesquisa, que são análise sentimental e os preços das moedas.

Palavras-chave: data mining; machine learning; business intelligence; criptomoedas.

ABSTRACT

This work aimed at the use of diverse techniques such as Data Mining, Machine Learning and graphic techniques using Business Intelligence tools for extracting, processing and analyzing data coming from social networks, more specifically in application message groups Telegram, in order to validate if we can verify correlation between the feelings that the users express in this social network and the price fluctuation of the Bitcoin cryptocurrency. The intention is to present a complete comparative analysis between the data coming from the Telegram social network with the scenario of higher price of this crypto-currency in December 2017. First, an exploratory research aimed at introducing several aspects of all technologies used in the development of this job will be displayed. The application of the survey will take place in real

[1] Discente do curso de Ciência da Computação do Centro Universitário Filadélfia.
[2] Docente do curso de Ciência da Computação do Centro Universitário Filadélfia.

data coming from the social network Telegram. At the end we will have a comparative table using statistical models to find a correlation between the variables worked during the development and the research, which are sentimental analysis and the prices of the currencies.

Keywords: data mining; machine learning; business intelligence; criptocurrency.

1 INTRODUÇÃO

Com a evolução da tecnologia, surgiram também muitas formas de extração de dados. Estes dados podem ser provenientes de smartphones, smartwatches, aparelhos de IoT como câmeras ou sensores diversos, ou até mesmo computadores pessoais. Essa diversidade de fontes de dados trouxe a necessidade de armazenamento em larga escala de forma distribuída e heterogênea. Nesse contexto, Data Mining e Machine Learning surgiram como uma forma inteligente e otimizada para extrair conhecimento de grandes quantidades de dados. Os estudos nessas áreas têm despertado cada vez mais interesse daqueles que desejam explorar o conhecimento contido em suas bases de dados.

Quando realizado e implementado corretamente, Data Mining e Machine Learning podem trazer muitas melhorias no processo de tomada de decisão. Lidar com grandes quantidades de dados usando apenas planilhas ou métodos empíricos torna-se inadequado para visualizar padrões e insights relevantes. É nesse contexto que a mineração de dados surge como uma resposta para as perguntas que surgem em grandes empresas ou até mesmo entre entusiastas. É possível utilizar a mineração de dados para entender o que as pessoas estão falando sobre um determinado assunto nas mídias sociais, compreender o comportamento dos clientes, identificar nichos de mercado, avaliar a opinião pública e muitas outras aplicações.

Este trabalho visa realizar um levantamento histórico sobre Data Mining, Machine Learning e o impacto da análise de dados provenientes de mídias sociais na flutuação de preços de criptomoedas. Além disso, será demonstrado o uso da análise de sentimento em mensagens de dezembro de 2017 e como métodos estatísticos podem ser utilizados para inferir se a opinião e o sentimento dos usuários em relação à criptomoeda Bitcoin flutuaram na mesma direção em que o preço da moeda variou.

Este trabalho tem como objetivo investigar como a análise de sentimentos dos usuários de mídias sociais pode auxiliar na compreensão da flutuação de preços da criptomoeda Bitcoin. Através dessa análise, pretendemos verificar se os sentimentos expressos pelos usuários em

postagens de grupos ou fóruns de mensagens podem ter algum impacto nos preços da criptomoeda, seja ele positivo, negativo ou inexistente.

2 METODOLOGIA

Segundo Prodanov e Freitas (2013), todo tipo de conhecimento ou acontecimento que foi observado ou constatado por alguém não deixava de ser conhecimento, mas o que diferenciava o conhecimento da ciência era o embasamento científico. A ciência requer evidências e provas, além da aplicação de metodologia, buscando estabelecer relações entre o fato estudado e outros fatos periféricos, independentemente da aparência.

Com base nisso, o presente trabalho utilizou como fontes de pesquisa livros, periódicos, relatórios, trabalhos de conclusão de curso, páginas da internet e outros tipos de documentos. Em relação à originalidade, as fontes de pesquisa utilizadas neste documento foram consideradas fontes primárias e secundárias:

- Primárias: referenciamos trabalhos que eram considerados bases primárias para muitas outras pesquisas, como por exemplo Nakamoto (2008).

- Secundárias: em nossa revisão bibliográfica, citamos várias fontes de trabalho secundárias.

Fontelles (2009) mostrou que para cada tipo de população existia um tipo de pesquisa a ser aplicada, e essa classificação determinou o tipo de pesquisa que foi realizada. Neste trabalho, em relação ao processo de pesquisa na revisão bibliográfica, foi realizada uma pesquisa do tipo exploratória, com o objetivo de introduzir e familiarizar o leitor com os temas que seriam abordados. Na parte do estudo de caso, foi realizada uma pesquisa descritiva, na qual tratamos e descrevemos a relação entre algumas variáveis que serviram de base para nossa pesquisa. Além disso, no estudo de caso, utilizamos informações provenientes da rede social chamada Telegram, por meio da qual fizemos a análise de sentimentos do público e das notícias em geral para entender o clima em determinados períodos. A escolha do Telegram como plataforma de análise deveu-se ao fato de que muitas pesquisas eram realizadas em outras plataformas, mas não foram encontrados muitos resultados sobre extração de dados do Telegram.

Para apresentar os resultados encontrados ao final da pesquisa, utilizamos abordagens

quantitativas e qualitativas:

- Quantitativa: relacionamos algumas variáveis e, por meio de métodos estatísticos, apresentamos a relação numérica entre as mensagens positivas e negativas nas redes sociais e a direção na qual os preços da criptomoeda Bitcoin flutuam.

- Qualitativa: com base nos dados analisados estatisticamente e nos resultados quantitativos da pesquisa, apresentamos percepções e análises dos dados obtidos.

3 REVISÃO BIBLIOGRÁFICA

Da fertilização in vitro à seleção dos fazendeiros de quais vacas manter para si e quais vender para os abatedouros, o Data Mining (DM) é um meio de solucionar problemas que têm grandes quantidades de dados e variáveis e que, muitas vezes, acabam sendo difíceis de processar pelo raciocínio humano. Estamos envoltos em problemas para os quais a utilização de DM pode trazer melhorias significativas em suas soluções. Para que seja possível trabalhar com esse tipo de análise, se faz necessária uma grande capacidade de armazenamento e de processamento.

Atualmente, a computação nas nuvens tornou o armazenamento e o processamento de dados algo trivial. Dados que provavelmente se perderiam em discos rígidos 10 anos atrás, hoje fazem parte da base de informação que algumas empresas possuem de seus clientes e usuários, armazenados na nuvem. Por se tratar do uso de métodos estatísticos e matemáticos, o DM, também chamada como Mineração de Dados, possui uma história que remonta a períodos anteriores mesmo ao computador como conhecemos, trazendo à tona estudos como o Teorema de Bayes, do século XVIII, e a análise de regressão, do século XIX, ambos métodos que buscam padrões em dados (SAUNDERS *et al*., 2018).

Segundo Fayyad *et al*. (1996), DM é o processo não-trivial de identificar padrões válidos, úteis, novos e compreensíveis nos dados. Quando tratamos de padrão, falamos sobre uma forma de agrupamentos de um subconjunto de dados específico, ou até mesmo um modelo aplicado a este subconjunto. Por não-trivial, Fayyad *et al.* (1996) nos diz que o processo não se trata apenas de simples cálculos de média ou valores predefinidos e já conhecidos, mas que também será necessário pesquisa sobre os dados ou inferências dos padrões a respeito dos dados tratados.

Fayyad *et al.* (1996) traz uma das mais antigas e mais citadas definições de DM, sendo assim referência importante no que diz respeito ao assunto. Este, apresenta o conceito de Knowledge Discovery in Databases (KDD), que se tornou a base de todo o conhecimento acerca da Mineração de Dados. Baseado na literatura em inglês (WEISS; INDURKHYA, 2010; FRIEDMAN; KAMBER, 1997, por exemplo), Fayyad *et al.* (1996) definiu o conceito de DM pelas palavras acima, porém no seu artigo original, utilizava aquelas palavras para referir-se ao que chamou de KDD. Até meados dos anos 1995, muitos pesquisadores consideravam o KDD e a DM como sinônimos segundo Chen e Zhang (1996). Apenas em 1995, em uma importante conferência, foi criada a distinção entre os dois conceitos de acordo com Adriaans e Zantinge (1996). Assim, KDD passou a englobar todo o procedimento de coleta à interpretação dos dados, e DM referia-se especificamente a uma das etapas do processo de KDD.

Entretanto, segundo Hand *et al.* (1998), DM se trata de um meio de descobrir estruturas interessantes, inesperadas e valorosas dentro de grandes bases de dados. Hand *et al*. (1998) mostra o relacionamento direto de DM com a estatística, porém voltado para as ferramentas utilizadas no processo. Hand *et al.* (1998) também expõe a necessidade de grandes quantidades de dados, porém pontua o fato de que é preciso atentar-se não somente ao armazenamento ou à leitura desses dados, mas a problemas mais fundamentais, como a maneira de determinar a representatividade dos dados, ou de analisar os dados em um período de tempo razoável, entre outras questões.

Dessa maneira, é possível perceber um amadurecimento no campo do DM, da época de publicação de Fayyad *et al.* (1996) a Hand *et al.* (1998). Torna-se claro que o DM é uma das etapas dentro de um processo maior chamado KDD, tendo nesse contexto o papel de ser o passo responsável pela aplicação de algoritmos a dados previamente filtrados e limpos, buscando, nesses dados, padrões que possam ter algum significado. Ainda que, muitas vezes, essas tratativas sejam confundidas com pesquisas estatísticas, por serem formas de pesquisa exploratória, a diferença se dá pela quantidade de dados. Enquanto profissionais de Estatística trabalham com pequenas quantidades de dados, cientistas de dados trabalham com até mesmo bilhões de registros de informação.

O uso desse tipo de tecnologia se estende a muitos campos, Fayyad *et al*. (1996) cita alguns exemplos de áreas em que, já em 1996, o DM poderia ser utilizado: em campanhas de marketing, por exemplo, detectando padrões de comportamento dos clientes. Ao falar de DM

nesse contexto, frequentemente surge o exemplo: uma rede de supermercados encontrou um padrão nos dados coletados de seus clientes: aqueles que compram fraldas na sexta-feira tendem a comprar também cerveja. Uma pessoa comum pode ter dificuldades para encontrar esse padrão em uma grande base de dados, mas o DM expõe essa relação e, com esse tipo de conhecimento, a rede de supermercados pode explorar esse padrão, deixando os dois produtos em prateleiras próximas.

Outro exemplo é na detecção de fraudes no uso de cartões de crédito, empresas como Visa e Mastercard verificam a todo o momento transações de milhares de seus clientes, na busca de padrões quebrados (como compras com valores elevados, em regiões distantes das comuns ao cliente, entre outros). O mercado financeiro, mais reservado na divulgação dos tipos de análise ou trabalhos que realiza, utiliza o DM com o objetivo de prever a flutuação dos preços, e já em 1993 os sistemas automatizados da LBS Capital Management superavam grande parte dos maiores investidores do mundo no que diz respeito a lucros (HALL *et al*., 1996).

As etapas do processo KDD

Uma das primeiras pesquisas a ter grande impacto no mundo da estatística, e que atualmente é amplamente utilizada em KDD, foi escrita em 1763 por Thomas Bayes. Trabalhando com conceitos da teoria das probabilidades, Bayes postulou em seu teorema a probabilidade da ocorrência de determinado evento dependente de outro (BAYES, 1763).

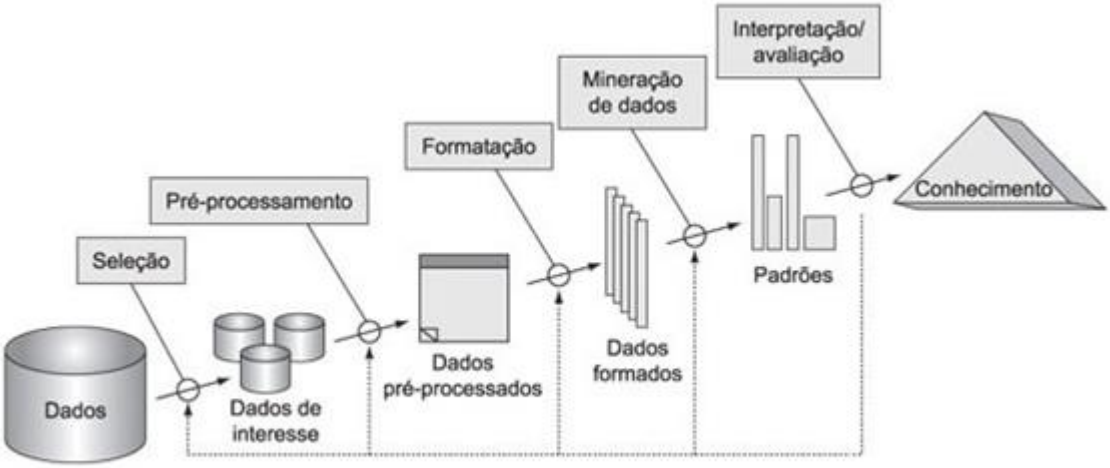
O termo KDD (Knowledge Discovery in Databases, ou Extração de Conhecimento de Banco de Dados) foi cunhado em 1989 por Gregory Piatetsky-Shapiro, em um workshop criado por ele e outros, chamado KDD-1989. O evento, que inicialmente contava com 69 participantes, tornou-se uma importante conferência internacional, com ouvintes e palestrantes do mundo todo (KDD-SHAPIRO, 1989). Contudo, os conceitos nos quais KDD se baseia não surgiram junto com o termo: em 1960, estatísticos, matemáticos e economistas já utilizavam técnicas de análise de dados. Para se referir às práticas de busca de dados, eram utilizados termos como Data Fishing ou Data Dredging (Pesca ou Dragagem de dados). Nota-se que os termos utilizados para se referir à busca de dados relacionam-se com atividades de extração como a pesca e a mineração, além da dragagem que é uma técnica de desassoreamento de cursos d'água. Dessa maneira, fica implícita a noção de que o processo consiste na separação de alguns

ISSN 2596-2809

elementos relevantes dentro de um universo maior de outros elementos descartáveis (SMITH, 2002).

O processo do KDD surge como uma maneira de organizar e padronizar o processo de captura e tratamento de grandes quantidades de dados até a aplicação de algoritmos para detecção de padrões. Abaixo, um diagrama mostra todas as etapas do processo de KDD ou DM:

Figura 1 - Processo do KDD.



Fonte: Teófilo (2015)

De acordo com a Figura 1, as primeiras etapas no processo de KDD (seleção de dados, pré-processamento e formatação) são de extrema importância para as etapas que se seguem, pois são responsáveis pela limpeza dos dados, removendo informações desnecessárias como palavras de parada e palavras repetidas para que os algoritmos possam realizar uma melhor aprendizagem, além da verificação de padrões nos dados. A essas etapas damos o nome de análise de dados. Diversos algoritmos podem ser aplicados a esses dados filtrados e pré-analisados para a geração do conhecimento, como Árvore de Decisão, k-Médias, SVM, kNN e Naive Bayes (SANTOS, 2016).

Até hoje, muitas pesquisas são realizadas com base nas etapas de análise de dados do processo de KDD, e o campo da Psicologia contribuiu extensivamente nesta fase, como em Spasic *et al*. (2012) e sua pesquisa na busca de padrões em mensagens suicidas, e Fátima et al. (2011) com o uso de psicologia do trabalho e DM para redução de absenteísmo nas empresas. A etapa de análise de dados assemelha-se ao que muitos profissionais de diversas áreas realizavam em planilhas eletrônicas muitos anos atrás: tinham uma visão superficial do aspecto

geral dos dados e procuravam por padrões a serem seguidos. A próxima etapa (aplicação de algoritmos e levantamento de conhecimento) é atualmente o diferencial, pois por meio desses dados podemos adquirir novos conhecimentos de forma automatizada (TANG *et al.*, 2008).

Pré-processamento, descoberta de padrões e avaliação do conhecimento aprendido, cada um desses termos refere-se a um agrupamento de tarefas no processo de KDD. Enquanto no pré-processamento trabalha-se com a integração de diversas fontes de dados, sua limpeza e transformação para um formato apropriado à aplicação posterior, a fase de descoberta de padrões é a aplicação do DM nos dados previamente trabalhados, lembrando que para cada tipo de dado e para cada resultado final, um algoritmo em específico pode ser necessário. Já na etapa de avaliação, de posse dos dados filtrados e já com o algoritmo de DM aplicado, avalia-se a aprendizagem adquirida de forma visual, por meio de gráficos, tabelas ou outras maneiras que facilitem a interpretação dos dados a partir de esquemas visuais (APPEL *et al.*, 2010).

As pesquisas de Appel *et al.* (2010) e Santos (2016) trazem definições do processo de DM muito próximas das definições de KDD criadas por Fayyad *et al.* (1996). A pesquisa de Appel *et al*. (2010) utiliza uma divisão de etapas criadas por Rezende (2003), na qual o KDD é dividido em 3 grupos de atividades; já Santos (2016) utiliza a mesma divisão criada por Fayyad *et al.* (1996), na qual o processo todo é composto por 4 grupos de atividades.

Todos os passos do KDD são necessários para uma correta extração de conhecimento a partir de uma base de dados (HORTA, 2014). Horta (2014) divide o processo em 3 agrupamentos de atividades: descrição e visualização, associação e clustering (ou agrupamento) e classificação e estimação (ou predição), seguindo assim o modelo proposto por Chye *et al.* (2004) num modelo próximo ao de Rezende (2003). A fase de descrição e visualização auxilia na compreensão dos dados tratados, principalmente quando tratamos de grandes quantidades de dados. Na associação, o objetivo é encontrar relações entre os dados, enquanto no clustering eles são agrupados. Enquanto etapas finais, a classificação e a estimação tratam da predição e rotulação dos dados (KOHAVI *et al*., 1997).

Horta (2014) esclarece sua visão a respeito do processo de KDD trazendo uma pesquisa do tipo descritiva e quantitativa, aplicando a resolução de problemas com insolvência apoiada em dados provenientes de empresas como o Serasa ou a Bolsa de Valores de São Paulo. Diferentemente dos métodos convencionais que utilizam-se de indicadores provenientes de pesquisas.

ISSN 2596-2809

Redes Sociais

Por mais que as primeiras redes de computadores (ARPANET) tenham surgido em meados dos anos 60, foi apenas a partir dos anos 80 que as formas de comunicação começaram a se popularizar, devido à popularização dos computadores pessoais. A rede IRC, uma das redes e aplicações pioneiras no quesito comunicação, foi criada por volta de 1988 e continua sendo utilizada até hoje. A primeira aplicação que realizava o envio e recebimento de mensagens de e-mail foi criada em 1971, por um programador chamado Ray Tomlinson, onde ele utilizou-se da rede ARPANET para realizar essa troca de mensagens entre pessoas conectadas à mesma rede.

A primeira rede social de que se tem notícia a utilizar os protocolos padrões da Internet atualmente foi uma rede de nome SixDegrees.com, um sistema onde você poderia criar um perfil e trocar mensagens com outras pessoas, porém aqueles que não estavam cadastrados na rede eram convidados a fazer parte dela. No seu auge operacional, o SixDegrees.com chegou a ter mais de 3 milhões e meio de usuários registrados em seu site (STWEBDESIGNER, 2016).

Segundo Boyd e Ellison (2008), redes sociais são sistemas online que permitem aos seus usuários criar um perfil, seja ele público ou privado, junto a uma lista de outros usuários com os quais ele compartilha interesses em comum, e onde ele seja capaz de visualizar e navegar por sua lista de conexões. Ainda segundo os autores, um dos parâmetros que tornaram as redes sociais algo único é o fato de que, através delas, podemos tornar visíveis nossas conexões sociais. Muitas conexões que talvez nem fossem possíveis no mundo real, devido à distância ou círculos diferentes, hoje são possíveis graças ao uso delas.

Atualmente, grande parte do nosso tempo é gasto interagindo em redes sociais. As cinco redes sociais mais utilizadas no mundo em termos de usuários ativos por mês são: Facebook, YouTube, WhatsApp, Facebook Messenger e WeChat (STATISTA, 2018).

Uma das maiores quebras de paradigma da Web 2.0 foi o surgimento do Facebook. Ainda que muitas outras plataformas tenham sido desenvolvidas para aproximar usuários com gostos parecidos, nenhuma delas chegou perto da revolução que o Facebook trouxe. Foi a primeira vez que, em escala global, as barreiras foram quebradas para trazer indivíduos de todo o planeta para dentro de uma plataforma, unindo todos os aspectos de suas vidas, reencontrando velhos amigos ou fazendo novas amizades através de jogos online fornecidos na própria

plataforma. Ele também foi um diferencial no que diz respeito à liberdade de expressão. Muitos jovens egípcios utilizam a plataforma por ser uma das poucas formas de liberdade de expressão que eles possuem. Além disso, essa quebra de barreiras permite conectar pessoas que antes poderiam não se encontrar no mundo real, facilitando assim o intercâmbio cultural, a aprendizagem de novas línguas, entre outros aspectos (DARWISH, 2011).

Do recorde de maior número de comparecimento para votação nas eleições americanas de 2008 (onde Barack Obama e John McCain disputavam a presidência dos Estados Unidos) ao apoio prestado à população do estado de Santa Catarina, também em 2008, quando houve uma temporada incessante de chuvas, ambos os eventos foram marcados por algo em comum: o advento da Comunicação Mediada pelo Computador (CMC). Essa nova comunicação permitiu não apenas que os indivíduos se comunicam, mas também amplificou a capacidade de conexão entre eles de forma mais rápida, dinâmica e colaborativa, criando assim um novo ambiente de rede na Internet, conectando não apenas computadores, mas também pessoas: as redes sociais mediadas por computador (RECUERO, 2011).

Mesmo com toda essa mudança de paradigma na comunicação, outro importante aspecto das redes sociais diz respeito à nossa privacidade. Gastamos grande parte do nosso tempo compartilhando informações, como fotos, vídeos, opiniões, entre outros. No entanto, é importante considerar a privacidade das informações compartilhadas. Um grande número de usuários já compartilhou informações sobre compromissos futuros ou condições de saúde sem controle algum de quem pode visualizar essas informações. A solução encontrada para esse problema foi o desenvolvimento de um sistema de recomendação e auxílio à configuração de privacidade baseado no perfil dos usuários (GHAZINOUR et al., 2016).

Quando tratamos de rede social, uma das formas de comunicação que mais cresce nos últimos tempos, é importante levar em consideração o uso das redes sociais no ambiente de trabalho. Muitas pesquisas são realizadas nesse campo, e algumas sugerem o uso de redes sociais internas, da própria empresa, para compartilhamento de conteúdos relacionados à empresa, como novas vagas, imagens e vídeos que agreguem valor ao dia do colaborador. Exemplos de empresas que trabalham dessa forma são a IBM e a Atos. Dentro de um fluxo organizacional, é importante que se tenha definido corretamente como liberar o uso das redes sociais para que sejam alcançados os objetivos estratégicos da empresa. Muitas empresas consideram o uso de redes sociais um desperdício de tempo e produtividade. O uso de redes

sociais internas da empresa é considerado valioso apenas para empresas com grande quantidade de funcionários, pois isso fomenta o lado social e colaborativo dentro do próprio ambiente de trabalho (YOKOYAMA, 2014).

Com isso, concluímos a importância das redes sociais, desde seu surgimento até suas necessidades atuais. É claro que, ao lidarmos com a pluralidade humana, enfrentaremos muitas necessidades diferentes. Por exemplo, no que diz respeito à privacidade, muitas pessoas sentem a necessidade de serem públicas, o que não as torna um problema, mas apenas demonstra que essas pessoas têm uma preferência diferente em relação à privacidade. No entanto, é importante levar em conta que uma pessoa que valoriza a privacidade evitará problemas ao usar redes sociais no ambiente de trabalho, ao passo que uma pessoa que mantém todas as suas postagens públicas e publica algo durante o horário de trabalho não está focada em suas atividades.

O uso de KDD nas redes sociais

O KDD aplicado a redes sociais foi um dos principais assuntos em destaque no ano de 2018. Após o escândalo de vazamento de dados pessoais de mais de 50 milhões de perfis do Facebook para a empresa Cambridge Analytica, as pessoas começaram a questionar o poder das redes sociais sobre nossos dados. Embora esse aspecto negativo seja uma realidade, o KDD também pode ser utilizado de forma benéfica, melhorando o sistema de propagandas, notificações personalizadas e recomendação de páginas e perfis relevantes (CADWALLADR, 2018).

A aplicação de KDD em redes sociais, como o Twitter, para a classificação sentimental de textos curtos, como tweets, tem sido uma área de pesquisa. Algoritmos, como o de Supervisão Distante, que utilizam emojis para categorizar sentimentos, têm alcançado até 83% de precisão (GO *et al*., 2009). Esses algoritmos processam o texto removendo emojis, nome de usuários, links e letras repetidas, deixando o texto pronto para classificação. Testes com diferentes algoritmos, como Naive Bayes, MaxEnt e SVM, atingiram em média 80% de precisão (GO *et al*., 2009).

Outro estudo abordou a análise sentimental e mineração de opiniões em textos de crítica e análise de filmes. Nesse caso, as mensagens foram categorizadas com base na quantidade de estrelas ou notas numéricas atribuídas aos reviews. Com o uso de algoritmos como Naive

Bayes, MaxEnt e SVM, obteve-se uma média de 80% de precisão na classificação de comentários positivos e negativos (PANG *et al.*, 2008).

Já na área de predição do mercado financeiro, Huang *et al*. (2017) capturaram tweets de empresas de tecnologia listadas na bolsa e as categorizaram de acordo com a polaridade dos tweets. Com base nos resultados, conseguiram prever corretamente a direção dos preços em média em 60% dos casos (HUANG *et al*., 2017).

O uso do KDD em redes sociais pode beneficiar diversas áreas e tem ganhado destaque nas grandes corporações. A análise de dados pode auxiliar na compreensão da opinião pública sobre empresas e também ser útil para análises financeiras. O Brasil está iniciando sua jornada nesse campo e promete acompanhar o crescimento dessa tecnologia nos próximos anos (FELIX *et al.*, 1998).

Criptomoedas e Bitcoin

Mesmo após 10 anos da criação da primeira criptomoeda, a identidade de seu criador, Satoshi Nakamoto, permanece desconhecida. Em outubro de 2008, Nakamoto lançou o artigo que marcou o início da tecnologia do Bitcoin, intitulado "Bitcoin: A peer-to-peer Electronic Cash System". O primeiro bloco da criptomoeda, conhecido como Bloco Gênesis, foi minerado em 3 de janeiro de 2009, iniciando assim a blockchain, o banco de dados descentralizado que registra todas as transações da rede (BTCHISTORY, 2018).

O Bitcoin é um sistema monetário eletrônico descentralizado, projetado para funcionar sem a necessidade de uma autoridade central. Ele possui um limite total de 21 milhões de moedas, o que o torna imune à inflação. As transações são validadas pela rede Bitcoin por meio de um algoritmo de consenso distribuído chamado Prova de Trabalho. Esse algoritmo resolve um complicado enigma matemático, e o nó que o resolve primeiro é encarregado de criar um novo bloco contendo as transações válidas e a hash do bloco anterior (REID, 2011).

Ao longo do tempo, a mineração de Bitcoins evoluiu rapidamente. Inicialmente realizada apenas pelas CPUs dos usuários, logo foram utilizadas GPUs para aumentar o poder de processamento. Com o crescimento da competição, surgiu o conceito de mineração em piscinas, onde os usuários unem seus recursos para aumentar as chances de encontrar um bloco

válido. As recompensas pela mineração podem ser divididas proporcionalmente entre os participantes ou serem pagas de forma fixa por processamento (PPS) (ROSENFELD, 2011).

A mineração de Bitcoins é uma corrida que acontece a cada 10 minutos, onde o poder de processamento determina a probabilidade de resolver o enigma e receber a recompensa. A mineração em piscinas se tornou uma prática comum para maximizar os lucros e garantir uma participação ativa na rede (NAKAMOTO, 2008).

Uma das formas técnicas para obter Bitcoins é a mineração, mas não é a única. Em março de 2010, surgiu a primeira exchange de compra e venda de Bitcoins, a BitcoinMarket.com, que operou por um ano até ser interrompida devido a fraudes nas transações. Em maio do mesmo ano, ocorreu a famosa troca de 10 mil moedas de Bitcoin por pizzas. Desde então, as criptomoedas valorizaram significativamente, com o Bitcoin mantendo uma Linha de Tendência de Alta até 2018, quando alcançou o pico de cerca de 20 mil dólares (MERCHANT, 2013).

O crescimento das criptomoedas deu origem às exchanges, empresas que facilitam a negociação de criptomoedas por outros ativos. Essas plataformas centralizadas requerem validação da identidade do usuário e possibilitam a compra, venda e troca de criptomoedas. No entanto, elas rompem com o conceito de descentralização proposto por Nakamoto, pois as transações internas não são registradas na blockchain, tornando-as entidades terceiras e fiduciárias (DECKER *et al*., 2015).

Um desafio enfrentado pela rede Bitcoin é a limitação no número de transações que pode processar em um curto período. Enquanto a Visa suporta 24 mil transações por segundo, o Bitcoin suporta apenas 7. Isso levou ao surgimento de outras criptomoedas que buscam atender a maior demanda de transações, como o Ripple, com capacidade de 1.500 transações por segundo, seguido pelo BitcoinCash, Litecoin, Dash e Ethereum (AMORÓS, 2018).

Em dezembro de 2017, durante o aumento recorde do valor do Bitcoin, a rede enfrentou problemas com cerca de 200 mil transações presas e demorou semanas para confirmá-las. Outra questão importante é o espaço em memória necessário para manter uma cópia da Blockchain, que chega a 180 gigabytes com o uso da carteira oficial Bitcoin Core. Porém, carteiras lightweight surgiram para resolver esse problema, permitindo a implantação em dispositivos com menor capacidade de processamento (NAKAMOTO, 2008).

ISSN 2596-2809

As criptomoedas, incluindo o Bitcoin, ainda estão em constante desenvolvimento e evolução. Novas moedas e tecnologias são lançadas regularmente, e há uma lista crescente de mais de 2.000 criptomoedas no mercado. Com um valor de mercado próximo a 1 trilhão de reais e potencial para crescer ainda mais, as criptomoedas ganham cada vez mais destaque nos meios de comunicação e nos negócios (CHEZ, 2018).

Machine Learning

Os usuários da rede não apenas consomem conteúdo, mas também o produzem, gerando dados repletos de opiniões, desejos e experiências que podem influenciar outros usuários ou interessar a empresas para aprimorar produtos. A análise de sentimentos nesses textos é uma poderosa ferramenta para transformar o conteúdo em dados analisáveis, conforme mencionado por Ávila (2017).

O conceito de Machine Learning refere-se ao aprendizado por máquinas, que, ao simular o aprendizado humano, possibilita um ganho de tempo significativo. Esse aprendizado permite a construção de algoritmos capazes de fazer previsões precisas com base em dados amostrais. Os três principais tipos de algoritmos de Machine Learning são: supervisados, não supervisionados e semi supervisionados.

Nos algoritmos de aprendizagem supervisionada, o algoritmo utiliza informações previamente classificadas para inferir novas informações sem classificação. Exemplos incluem árvores de decisão, Naïves Bayes e máquinas de vetores de suporte (MOHRI *et al.*, 2012). Em contraste, os algoritmos de aprendizagem não supervisionada não recebem informações classificadas, apenas dados, e devem inferir suas próprias conclusões. Exemplos são algoritmos de agrupamento, detecção de anomalias e redes neurais (NEGRETTO, 2016). Já os algoritmos de aprendizagem semi supervisionada combinam características dos dois métodos anteriores, aplicando-se quando parte dos dados é classificada, mas outra parte não. Exemplos são algoritmos de Maximização de Expectativa, treinamento e máquinas de vetores de suporte transdutivo (NEGRETTO, 2016).

Essas técnicas são utilizadas para compreender rapidamente uma variedade de assuntos com base em dados amostrais não estruturados. As informações na web são produzidas de

forma livre e diversificada, mas ao aplicar Machine Learning , os dados tornam-se mais compreensíveis e estruturados.

Kauer (2016) destaca que Machine Learning baseia-se na construção de algoritmos capazes de fazer previsões sobre dados, produzindo modelos a partir de instâncias de dados de entrada. Esses modelos consistem em hipóteses formuladas sobre os dados, capazes de prever informações que não estão diretamente presentes nas instâncias. Cada instância é representada por um vetor de atributos, e a classificação ocorre quando um atributo (rótulo) prevê dados com base nos outros atributos.

Essas técnicas têm sido amplamente utilizadas para a análise de sentimentos, classificação e agrupamento de dados, superando as limitações de conhecimento e velocidade do processamento humano. A área de Machine Learning permite que os computadores trabalhem com inteligência artificial, atendendo às exigências dos consumidores por produtos tecnológicos cada vez mais inteligentes e eficientes.

A utilização de modelos de Machine Learning na análise de sentimentos possibilita a estruturação de dados não estruturados da web, tornando-os compreensíveis quantitativa e qualitativamente. Isso é essencial para o desenvolvimento de produtos e serviços de alta qualidade, a fim de atender às exigências e expectativas dos usuários da internet (KAUER, 2016).

Portanto, o uso de mecanismos de previsão de informações na análise de textos da web é uma ferramenta eficaz para a construção de conjuntos de dados relevantes para diversas áreas, contribuindo para a constante evolução da tecnologia e atendendo às necessidades crescentes dos usuários da rede.

5 ESTUDO DE CASO

Neste capítulo será demonstrado o estudo de caso que foi aplicado nesta pesquisa. Nele, através de métodos estatísticos e técnicas de computação como extração de dados, análise sentimental, entre outras, será validado a hipótese seguinte: existe uma ligação entre os sentimentos expressos pelos usuários de redes sociais e a movimentação dos preços da criptomoeda Bitcoin?

ISSN 2596-2809

Levantamento de passos a serem seguidos

A aplicação do estudo de caso que será apresentada neste trabalho será baseada nas seguintes etapas:

- Extração de dados de preços do Bitcoin no período de 01/12/2017 até 31/12/2017.
- Tratamento de dados de preços pois toda base de dados histórica disponível contém informações minuto a minuto dos preços, por isso será necessário certa tratativa nos dados.
- Utilizar ferramentas de Business Intelligence para impressão de gráficos a respeito da tratativa de preços.
- Extração de dados de mensagens de usuários no período de 01/12/2017 até 31/12/2017.
- Realização de análise sentimental nas mensagens de usuários.
- Extração de dados de mensagens de notícias no período de 01/12/2017 até 31/12/2017.
- Realização de análise sentimental nas mensagens de notícias.
- Utilizar ferramentas de Business Intelligence para impressão de gráficos a respeito da análise sentimental.
- Realizar análise de correlação entre análise sentimental e preços.
- Demonstrar resultados.

6 RESULTADO OBTIDO

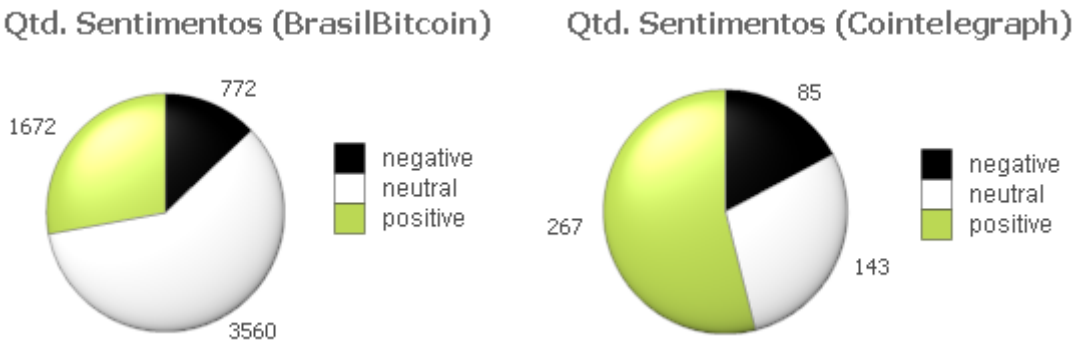
Para verificação das hipóteses levantadas iremos utilizar o teste de qui quadrado. O teste qui quadrado é utilizado com a intenção de verificar se existe uma dependência entre duas variáveis qualitativas (no estudo de caso as variáveis qualitativas foram subir e descer para os preços e positivo, neutro e negativo para os comentários dos usuários).

A análise estatística dos dados utilizando o método do Qui quadrado vai auxiliar na evidenciação de correlação ou não das duas variáveis levantadas, sendo elas a direção de flutuação do preço do Bitcoin e o sentimento expressado, pelos usuários e suas mensagens e pelas notícias que circularam na rede no período analisado.

ISSN 2596-2809

Inicialmente vamos trabalhar com os dados do grupo BrasilBitcoin, onde foram analisadas as mensagens trocadas entre os usuários de forma mais descontraída. Informações sobre quantidade de mensagens podem ser encontradas na Figura 2.

Figura 2 - Quantidade de mensagens baseadas em sua polaridade nos 2 grupos pesquisados



Fonte: próprio autor.

Sendo assim, o quadro comparativo entre sentimento e as mensagens é apresentado abaixo:

Direção de Preço / Sentimento	caiu	subiu	Total Geral
neutro	24	34	58
positivo	3	1	4
Total Geral	27	35	62
Probabilidade do preço	0,42622	0,57377	

7 CONCLUSÃO

Esta pesquisa verificou a correlação entre os sentimentos expressos pelos usuários da criptomoeda Bitcoin e a flutuação de preços de todo o mês de dezembro de 2017, na intenção de verificar se em momentos de muitos comentários e notícias positivas o preço da criptomoeda subia e se em momentos de muitos comentários e notícias negativas o preço caia. O estudo de caso apresentado nesta pesquisa utilizou-se de técnicas de extração de dados em grupos de

conversa da rede social Telegram, análise sentimental de texto e parte de processamento e análise de dados utilizando ferramentas de Business Intelligence.

Através das técnicas empregadas foi realizado a transformação de dados do valor da criptomoeda Bitcoin, de 1 em 1 minuto para períodos de 12 horas, extraímos mensagens de usuários da rede social Telegram, aplicamos algoritmos de análise sentimental nas mensagens e realizamos um cruzamento de todas essas informações com o método estatístico Qui Quadrado de Pearson. Com a análise final concluímos que os resultados (tanto das mensagens de grupo de usuários quanto de bot de notícias) nos retornaram uma baixa associação entre os as variáveis estudadas.

De acordo com a metodologia aplicada e os resultados obtidos, podemos verificar que a nossa hipótese de ligação entre a variação dos preços da criptomoeda Bitcoin e o sentimento expresso pelos usuários nas redes sociais é baixa pois está abaixo dos níveis aceitáveis. Isto significa que utilizando a metodologia acima não conseguimos encontrar ligação entre os dois fatos estudados.

Como trabalhos futuros fica indicado a implantação ou melhoria de algoritmos para Processamento de Linguagem Natural (PLN) diretamente na língua portuguesa, também uma verificação do impacto negativo nos dados finais quando se precisa realizar a tradução de mensagens. Outro ponto interessante seria a utilização de outras fontes de dados, visto que os grupos de conversação de usuário no Telegram podem nos dar um panorama muito enviesado a respeito da criptomoeda Bitcoin.

REFERÊNCIAS

ADRIAANS, Pieter; ZANTINGE, Dolf. **Data Mining**. [*S.l.*]: Addison-Wesley Professional, 1996.

APPEL, Ana Paula. **Métodos para o pré-processamento e mineração de grandes volumes de dados multidimensionais e redes complexas**. 2010. Tese (Doutorado em Métodos para o pré-processamento e mineração de grandes volumes de dados multidimensionais e redes complexas) - USP, São Carlos, 2010.

BATISTA, Alexandre da Costa. **Análise da aplicação de algoritmos de data mining em bases de dados de vendas de produtos**. 2009. 44 p. Trabalho de Conclusão de Curso (Graduação em Engenharia da Computação) - Escola Politécnica de Pernambuco, Universidade de Pernambuco, Pernambuco, 2009.

BAYES, Thomas. An Essay towards Solving a Problem in the Doctrine of Chances. **Philosophical Transactions of the Royal Society**, v. 35, p. 370-418, 1763.

CADWALLADR, Carole; GRAHAM-HARRISON, Emma. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. **The Guardian**, v. 17, 2018.

CHEN, Ming Syan; HAN, Jiawei; YU, Philip S. Data mining: An overview from a database perspective. **IEEE Transactions on Knowledge and Data Engineering,** v. 8, n. 6, p. 866-883, 1996.

CHYE, K.H.; CHIN, T.W.; PENG, G.C. Credit scoring using data mining techniques. **Singapore Management Review**, v. 26, p. 25-47, Jan 2004.

DARWISH, Ashraf; LAKHTARIA, Kamaljit I. The impact of the new Web 2.0 technologies in communication, development, and revolutions of societies. **Journal of advances in information technology**, v. 2, n. 4, p. 204-216, 2011.

DEALEXANDRE, F. A. **Algoritmos de aprendizado semi-supervisionado baseados em grafos aplicados na bioinformática**. 2016. Dissertação (Mestrado em Ciência da Computação) - UNESP, São José do Rio Preto, 2016.

FÁTIMA D. D. D. A contribuição da psicologia do trabalho para a redução do absenteísmo utilizando data mining. *In*: SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 43., 2011, Ubatuba. **Anais** [...]. Ubatuba, 2011.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, p. 37-54, 1996.

FELIX, Luis Carlos Molina. **Data mining no processo de extração de conhecimento de bases de dados**. 1998. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Universidade de São Paulo, São Paulo, 1998.

FRIEDMAN, Jerome H. Data mining and statistics: What's the connection. In: Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics. 1997.

GO, Alec; BHAYANI, Richa; HUANG, Lei. Twitter sentiment classification using distant supervision. 2009.

GRACCA, Andreia Sofia Dinis. Business Intelligence markup language e reporting. 2017.

HAND, David J. Data Mining: Statistics and More? **The American Statistician**, v. 52, n. 2, p. 112-118, 1998.

HALL, J.; MANI, G.; BARR, D. Computational Intelligence in Financial Engineering. *In*: IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering (CIFEr), March 1996.

HEAVEN, Douglas. Your data, safe at last?. New Scientist, v. 238, n. 3179, p. 22-23, 2018.

HORTA, Rui Américo Mathiasi; ALVES, Francisco José dos Santos; CARVALHO, Frederico Antônio Azevedo de. Seleção de atributos na previsão de insolvência: aplicação e avaliação usando dados brasileiros recentes. RAM. **Revista de Administração Mackenzie**, v. 15, n. 02, p. 125-151, 2014.

HUANG, Dashan; JIANG, Fuwei; TU, Jun; ZHOU, Guofu. Forecasting stock returns in good and bad times: The role of market states. 2017.

JOACHIMS, Thorsten. Text categorization with support vector machines: Learning with many relevant features.**European conference on machine learning**, p. 137-142, 1998.

KACPRZAK, Eduardo; CESAR-JR, Roberto M.; FERREIRA, Eduardo C. Aspectos da participação do Brasil no desafio internacional de mineração de dados KDD CUP 99. *In*: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 25., 2005. **Anais** [...]. 2005.

KAUER, Anderson Uilian. **Análise de sentimentos baseada em aspectos e atribuições de polaridade**. 2016.

KOHAVI, Ron; JOHN, George H. Wrappers for feature subset selection. **Artificial Intelligence**, v. 97, n. 1, p. 273-324, 1997.

NEGRETTO, Diego Henrique. **Análise de sentimento para textos curtos**. 2017. Dissertação (Mestrado em Modelagem matemática da Informação) - Fundação Getulio Vargas, Escola de Matemática Aplicada, 2017.

PANG, Bo; LEE, Lillian. Opinion mining and sentiment analysis. **Foundations and Trends® in Information Retrieval,** v. 2, n. 1-2, p. 1-135, 2008.

PIATETSKY-SHAPIRO, Gregory. KDD-89 Workshop. **KDnuggets Analytics Big Data Data Mining and Data Science,** Aug 1989.

PURKAIT, Soumyadeep. Bitcoin | Kaggle. Kaggle, 2018.

REZENDE, Solange Oliveira. **Sistemas Inteligentes**: Fundamentos e Aplicações. São Paulo: Manole, 2003.

RODRIGUES, Hélio L. S.; FERREIRA BRAGA, José Remo. Visualização da informação como ferramenta de apoio ao tratamento de dados empresariais. **Colloquium Exactarum**, v. 9, n. 2, p. 114-130, 2017.

SANTOS, Bruno *et al.* Data Mining: Uma abordagem teórica e suas aplicações. **Revista Espacios**, v. 37, n. 05, p. 23-29, nov. 2016.

SAUNDERS, Asena Atilla. **The History of Data Mining**. [*S.l.*]: Exastax, 2018.

SMITH, George Davey; EBRAHIM, Shah. Data dredging, bias, or confounding: They can all get you into the BMJ and the Friday papers. **BMJ**, v. 325, n. 7378, p. 1437-1438, 2002.

SPASIC, Irena *et al.* A Naïve Bayes Approach to Classifying Topics in Suicide Notes. **Biomedical Informatics Insights**, v. 5s1, p. BII.S8945, 2012.

TANG, H. A Simple Approach of Data Mining in Excel. *In*: INTERNATIONAL CONFERENCE ON WIRELESS COMMUNICATIONS, NETWORKING AND MOBILE COMPUTING, 4., 2008, China. **Proceedings** [...]. China: IEEE, 2008.

WEISS, Gary M.; DAVISON, Brian D. **Data Mining**. [*S.l.*]: Csrea, 2010.