Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise de sentimentos de tweets para identificar tendências de Bitcoins no mercado financeiro

Pedro Coutinho de Castro
Breno Rios Ferreira

Monografia apresentada como requisito parcial
para conclusão do Curso de Computação — Licenciatura

Orientador
Prof. Dr. Vinicius Ruela Pereira Borges

Brasília
2019

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise de sentimentos de tweets para identificar tendências de Bitcoins no mercado financeiro

Pedro Coutinho de Castro
Breno Rios Ferreira

Monografia apresentada como requisito parcial
para conclusão do Curso de Computação — Licenciatura

Prof. Dr. Vinicius Ruela Pereira Borges (Orientador)
CIC/UnB

Prof. Dr. Douglas Cedrim Oliveira
Instituto Federal Goiano - campus Rio Verde

Prof. Dr. Ricardo Barros Sampaio
Fundação Oswaldo Cruz

Prof. Dr. Thiago de Paulo Faleiros
Departamento de Ciência da Computação - Universidade de Brasília

Prof. Dr. Wilson Henrique Veneziano
Coordenador do Curso de Computação — Licenciatura

Brasília, 24 de junho de 2019

Dedicatória

Dedicamos este trabalho primeiramente Deus, pr ser essencial m nossas vidas, nosso guia, e s nossos pais e mãs.

Agradecimentos

Ao professor Vinicius, pelos textos traduzidos, orientação, seu grande desprendimento em ajudar-nos e aos nossos amigos Igor da Silva Bonomo e Beatriz Chiarelli que nos auxiliaram na realização deste trabalho.

Resumo

Predição de eventos relativos ao mercado financeiro é sempre uma tarefa de alta complexidade, visto que os fatores que podem ser responsabilizados por movimentos de mercado é bastante vasto e diverso. As criptomoedas, tecnologia inovadora que tem ganhado bastante adesão e visibilidade no comércio, ainda são cercadas de desconfianças. Aliado à popularização dessas criptomoedas, as redes sociais se tornaram uma fonte rica de informações, pois seus usuários publicam mensagens e opiniões relacionadas com diversos assuntos. Nesse sentido, alguns usuários e especialistas do domínio financeiro analisam essas mensagens das redes sociais relacionadas ao mercado financeiro e às criptomoedas para auxiliar nas tarefas de tomadas de decisão. Este projeto propõe desenvolver um método para identificar relações entre as mensagens da rede social *Twitter* e o valor de mercado do *Bitcoin*. O método proposto é composta por técnicas de pré-processamento e caracterização de textos, em conjunto com algoritmos de classificação, a fim de analisar a polaridade (positivo, negativo e neutro) de sentimentos presentes nos *tweets*. Modelos não-supervisionados são também empregados para extrair tópicos de conjuntos de *tweets* e para agrupá-los conforme as relações de similaridade, possibilitando a identificação de padrões. Experimentos foram realizados para validar o método proposto e consistem de três etapas: (i) comparar o volume total de *tweets* positivos, negativos e neutros, com o valor de mercado do *Bitcoin*; (ii) obter os tópicos relevantes presentes nos documentos, e assim, buscar relações entre os sentimentos predominantes e os tópicos; (iii) visualizar os agrupamentos formados pela aplicação da técnica *K-means*. O método proposto obteve resultados relevantes, mas levando em consideração o atual momento de crescente valorização da criptomoeda, sendo necessário testes adicionais para comprovar sua plena eficácia.

Palavras-chave: *Bitcoin*, Twitter, análise de sentimentos, aprendizado de máquina, visualização de textos

Abstract

The prediction of events related to the financial market is a complex task, since the responsible factors for market movements is diverse and vast. The growing popularity of cryptocurrencies are still seen with mistrust by investors and the financial market specialists. In the last years, social networks have become a powerful source of information as their users can post text messages and opinions related to various subjects. Specifically, several users such messages related to financial market and cryptocurrencies in order to support them on decision taking tasks. This work proposes a method based on sentiment analysis to identify relationships between the messages of the social network Twitter and the Bitcoin's market value. The proposed method is composed by preprocessing and text characterization techniques, along with supervised models for sentiment classification according to tweets' polarities (positive, negative and neutral). Unsupervised models are also employed to extract topics from tweets' sets and to cluster them by taking into account its similarity relations for identifying relevant patterns. Experiments were performed in three steps to validate the proposed method: (i) compare the total volume of positive, negative and neutral tweets, in relation to the Bitcoin's market value; (ii) obtain the most relevant topics in the documents, and thus to seek relations between the predominant sentiments and the most relevant topics; (iii) visualize the clusters formed by the application of K-means. The proposed method obtained relevant results, but emphasizing the long term of growing Bitcoin appreciation. Therefore, additional experiments are required by considering other Bitcoins values at the financial market.

Keywords: Bitcoin, Twitter, sentiment analysis, machine learning, text visualization

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

No início do processo civilizatório, o comércio era realizado prioritariamente através do escambo de mercadorias. Entretanto, tais trocas tinham valor abstrato e muito particular, fato que poderia resultar em impasses comerciais. Alternativamente, no século VII a.C., foram criadas as primeiras moedas, de ouro e prata, a fim de solucionar esse tipo de questão. Esse sistema evoluiu progressivamente ao vigente na atualidade, em que os bancos centralizam quase que a totalidade das operações financeiras. Além disso, os bancos são responsáveis pela segurança e confiabilidade de todo esse complexo e custoso sistema [2].

Como reflexo da dependência da atuação dos bancos intermediando as operações e os fortes avanços da tecnologia, as moedas digitais ou criptomoedas se consolidaram como forte alternativa ao sistema financeiro convencional [3]. O *Bitcoin* é a criptomoeda que ganhou maior notoriedade e apesar de não ser atrelada à política econômica de um governo, se tornou uma moeda confiável e consolidada no mundo [4]. Sua ascensão fez com que alguns países passassem a empregá-la no comércio e em operações de câmbio [5], além de ter estimulado o surgimento de dezenas de criptomoedas paralelas, conhecidas como *altcoins*. Tal fato ocorre em decorrência do código-fonte do *Bitcoin* estar disponível publicamente, sendo então amplamente utilizado como base para outros projetos[1].

O sucesso da criptomoeda *Bitcoin* se deve principalmente à tecnologia *blockchain*, que fornece segurança, privacidade e imutabilidade às operações financeiras realizadas. O *blockchain* pode ser definido como um banco de dados descentralizado e acessível em uma rede, cujos nós são responsáveis por registrar e validar todas as operações financeiras, denominadas transações. Por isso, tal rede ponto-a-ponto (*peer-to-peer*) dispensa a necessidade de um agente centralizador responsável para validar as transações. Como consequência, os custos operacionais demandados pelas instituições centralizadoras que

[1] *https://medium.com/@gblaender*

intermedeia as operações financeiras são reduzidos, tornando o uso dessa moeda bastante atrativo em diferentes cenário.

A popularização da internet resultou na produção em massa de conteúdo. Como consequência desse fenômeno, nascem as redes sociais que por sua capacidade de retro-alimentação, isto é, produzem e consomem conteúdo, têm revolucionado as formas de relacionamentos, em todos os níveis, como por exemplo: relacionamentos interpessoais, amizade, namoro e casamento; cliente-empresas, captação de clientes, *feedback* dos serviços e publicidade [6]. Não obstante, formar opiniões, dissipar informações, criar tendências, também são exemplos das inúmeras formas de atuação nesse ecossistema que trouxe o mundo a uma nova era.

Com tantas riquezas de informações disponíveis em domínio público, as opiniões dos usuários, publicadas em seus perfis nas redes sociais, despertaram a atenção de empresas para análise de satisfação, em detrimento dos tradicionais formulários de questões, utilizados até então. Outra ocorrência derivada das redes sociais é o engajamento em massa, popularmente denominada de "viralização", que detém o poder de alavancar rapidamente novas tecnologias, conteúdos, costumes, entre outros, podendo resultar em mudanças na dinâmica da sociedade [6].

As redes sociais são algumas das ferramentas mais atrativas na internet, uma vez que é possível compartilhar informações, promover a interação entre pessoas e disseminar ideias [7]. Devido a esses motivos [8], é interessante permitir que essas informações sejam catalogadas e analisadas de forma a entender os diferentes grupos de usuários presentes na rede. Além disso, pode-se direcionar a coleta desses dados ao seguir um perfil específico de usuários ou participar de grupos temáticos, o que viabiliza o estudo de opiniões que surgem nessas redes sociais.

O *Twitter*[2], Rede Social Online (RSO), uma das redes sociais mais populares, permite interação entre usuários, de maneira fácil, simples e objetiva, tendo se mostrado uma fonte rica em informações e opiniões de usuários. O *Twitter* permite que usuários escrevam mensagens textuais de até 280 caracteres em seus perfis, geralmente expressando suas opiniões e sentimentos em relação a temas ou assuntos específicos. Particularmente, a grande disponibilidade de *tweets* sobre o mercado financeiro e o desenvolvimento de ferramentas computacionais para coleta possibilitou a análise desses textos para identificar tendências, expressas pelos seus usuários, e a partir disso, prever o comportamento do mercado de valores a partir das mensagens presentes nas redes sociais [9].

Como a quantidade de mensagens textuais nas redes sociais cresce cada vez mais, tarefas de análise desses textos e as opiniões dos usuários tornam-se inviáveis se realizadas por um especialista humano. Nesse contexto, a área de análise de sentimentos, intersecção

[2] *https://twitter.com/*

das áreas de mineração de textos e processamento de linguagem natural, pode contribuir com métodos e técnicas para extrair automaticamente conhecimento relevante e implícito de textos. Na literatura, a análise de sentimentos foi empregada em textos de diversos domínios do conhecimento, como na identificação de locais de crimes [10], no auxílio à decisão de vencedores de concursos de televisão [11], na classificação de filmes [12], entre outros. Por isso, essa pesquisa considera a oportunidade de aplicar técnicas de análise de sentimentos em conjuntos de *tweets* relacionados ao *Bitcoin*.

No segmento de predição da bolsa, existem pesquisas que mostram a capacidade do *Twitter*, quando associada ao mercado financeiro, de relacionar factíveis indicadores do mercado à bolsa de valores [13]. O artigo publicado por *Bollen et al.* [14], propõe estudar o comportamento das bolsas de valores, por meio do humor expresso no *Twitter*. Entretanto, seus experimentos não apresentaram resultados contundentes, pois não foram obtidas informações sobre mecanismos causadores que possam conectar estados de humor ao comportamento da *Dow Jones Industrial Average* (*DIJA*).

Não está claro na literatura que a análise de sentimentos possa identificar perfis, específicos, de investidores de criptomoedas. Contudo, foi observado que o processo de coleta dos *tweets* é o momento oportuno para qualificar a busca desses dados de modo a selecionar possíveis consumidores de *Bitcoins*. Definir previamente um conjunto de palavras-chave é um meio viável para separar, com maior precisão, os textos relevantes de usuários das redes sociais [15]. Portanto, esse trabalho propõe analisar os *tweets* de forma ampla, sem a presença de um filtro complexo, considerando que a única palavra-chave utilizada foi "*Bitcoin*".

O objetivo principal desta pesquisa é verificar se existe relação entre os *tweets* que possuem termos associados ao *Bitcoin* e seu comportamento e valorização dessa criptomoeda no mercado de valores. Nesse contexto, a relação de causa pode ser descrita como o movimento coletivo, dos grupos presentes na rede, enquanto que a relação de efeito reflete o valor corrente da criptomoeda. O entendimento dessa relação de causa e efeito viabiliza a criação de um indicador, baseado em sentimentos expostos em redes sociais, que funcionaria de forma colaborativa com os indicadores, conhecidos e amplamente usados, do mercado financeiro, como por exemplo: oscilador estocástico, indicador bandas de *Bollinger* [16]. Como consequência, o método descrito neste trabalho pretende apresentar ao mercado financeiro uma nova proposta de ferramenta para o auxílio à tomada de decisões de compra e venda de *Bitcoins*.

Esta pesquisa propõe investigar as seguintes hipóteses:

- (i) "é possível realizar análise de sentimentos de *tweets* relacionadas ao *Bitcoin*?";
- (ii) "é possível identificar contas de usuários investidores de *Bitcoin* por meio da análise de sentimentos?".

Com a finalidade de responder a hipótese (i), foram realizadas as seguintes tarefas: a primeira consiste em reduzir o volume original do conjunto de dados textuais através da aplicação de técnicas de pré-processamento de texto e, subsequentemente, caracterizar os textos a fim de estruturá-los. Em seguida, classificá-los com suas respectivas polaridades (positivo, negativo e neutro). Esses procedimentos facilitam a aplicação de algoritmos supervisionados de aprendizado de máquina, para classificação dos *tweets*, de acordo com os sentimentos previamente definidos. Assim, podem-se quantificar as polaridades referentes às respectivas mensagens, para comparar o volume total de polaridades com o valor corrente de mercado do *bitcoin*.

Por outro lado, a hipótese (ii) pode ser respondida ao empregar técnicas de classificação, modelo de extração de tópicos (padrão recorrente de coocorrência de palavras [17]) e algoritmo de agrupamento nos conjuntos de *tweets*, que visam identificar perfis específicos de investidores e eventuais responsáveis pelo movimento do mercado. O retorno proporcionado por essas aplicações promove a visualização e a correlação entre os principais agentes, o volume de textos, de uma determinada polaridade; o movimento do valor de mercado, da moeda; e a, possível, palavra, ou conjunto de palavras, responsável por tais variações de preço. Finalmente os resultados são interpretados com objetivo de identificar e dar sentido a essas tendências.

O restante dessa monografia está estruturada da seguinte maneira:

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

Capítulo 2

Fundamentação Teórica

No decorrer dessa fundamentação teórica serão apresentados os principais conceitos, ferramentas e técnicas utilizadas para realização dessa pesquisa científica. A Seção 2.1 aborda a conceituação e uma breve formulação sobre o funcionamento do *Bitcoin*. A Seção 2.2 conceitua o processamento e a caracterização de textos, atestando a utilização e explicando as técnicas aplicadas. A Seção 2.3 descreve sobre a análise dos sentimentos textuais. A Seção 2.4 contextualiza sobre a redução de dimensionalidade, aborda técnicas como: PCA e *Truncated SVD*, e destaca sua importância. A Seção 2.5 desenvolve sobre modelagem e extração de tópicos, e exemplifica com o modelo utilizado nessa pesquisa.

2.1 *Bitcoin*

O *Bitcoin* é que uma moeda criptografada e foi difundida por meio da internet. Essa moeda digital funciona de forma descentralizada através da tecnologia de rede *peer-to-peer* para realizar suas transações, sem interferência de bancos ou instituições financeiras. Todas as transações de *bitcoins* são registradas e validadas por meio da tecnologia *blockchain*, que pode ser definido como um banco de dados distribuído de registros onde cada movimentação é verificada por consenso da maioria das entidades participantes da rede [18]. Ademais, o *blockchain* registra as transações na cadeia por meio de um processo conhecido como mineração, em que os nós autorizados da rede competem entre si para ganhar o direito de inserir um novo bloco de transações na cadeia ao resolver um problema matemático complexo (por exemplo, quebrar um *hash*). O nó vencedor também recebe uma recompensa em forma de *Bitcoins* [19].

A autoria do *Bitcoin* ainda é uma incógnita, porém a teoria mais difundida é que o seu criador teria sido *Satoshi Nakamoto* [20]. Por outro lado, especula-se que a autoria desta famosa criptomoeda supracitada foi um trabalho de não apenas uma pessoa, mas de um grupo de grandes mentes visionárias com intuito de revolucionar a história [18].

As moedas virtuais estão ganhando popularidade e evidência no mercado financeiro atual. Tal fato ocorre diante das inúmeras vantagens obtidas com a sua utilização, como a deflação, a ausência de taxação, a segurança e a simplicidade nas operações, tornando-as atrativas para o mercado [21].

2.1.1 Funcionamento do *Bitcoin*

É importante salientar que o controle da inflação é fundamental para a confiabilidade da moeda, pois traz estabilidade e previsibilidade, para os consumidores e investidores de criptomoedas [22]. Isso é possível pelo número limitado de moedas a serem mineradas, no caso do *Bitcoin*, 21 milhões de unidades, número que será alcançado apenas entre 2110-2140 [23].

O fluxograma mostrado na Figura 2.1 exemplifica o processo de mineração de *bitcoins*, possível a todo computador capaz de realizar o processo de mineração, que consiste em: resolver um problema matemático complexo, registrar no *blockchain* a solução do problema (*proof-of-work*), o registro de todas as operações com *Bitcoins* que ocorreram nos últimos 10 minutos e uma referência para o bloco imediatamente anterior [18].

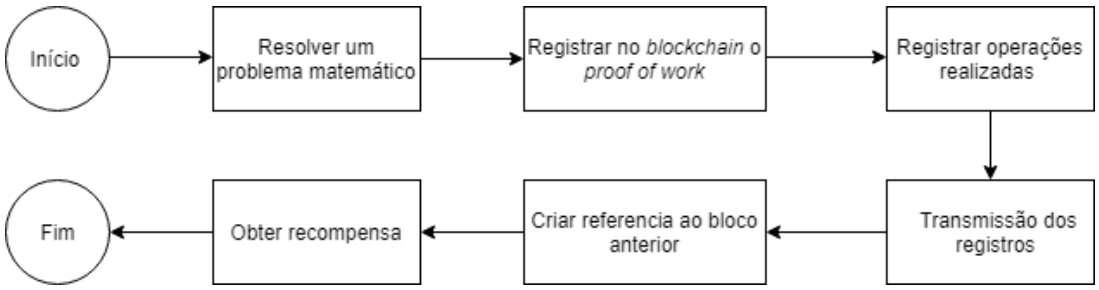

Figura 2.1: Fluxo da mineração de *bitcoins*.Fonte: própria (2019).

Diferentemente do processo de criação de *bitcoins*, que exige certa complexidade e um custo computacional alto para os interessados em registrar essas operações, o processo de comercialização de criptomoedas preza pela simplicidade, visto que para adquirir *bitcoins* basta o usuário ter uma carteira digital, contatar uma corretora que comercialize a moeda e, finalmente, realizar sua aquisição. Ao efetuar uma transação, os fundos de *Bitcoin* ficam atrelados ao código alfanumérico da carteira [18]. Da mesma forma, ocorrem transações interpessoais, independentemente da localidade, eliminando as barreiras burocráticas existentes para envio de dinheiro ao exterior.

Por outro lado, a ausência de taxação nas transações financeiras é decisiva para atrair a atenção do mercado e gerar grande preocupação nas instituições bancárias. Vale ressaltar que as transações envolvendo *bitcoins* são descentralizadas e não necessitam de instituições

com altos custos operacionais, como os bancos, para validar essas operações. Contudo, a falta de regulamentação traz certa insegurança jurídica para a moeda, como, por exemplo, na ocorrência de alguma fraude, uma vez que não há responsáveis, nem garantias legais pelas operações [23].

É importante salientar a questão da segurança das transações, que é peça precípua para promover confiabilidade ao mercado. As transações são validadas pelos nós da rede e, posteriormente, registradas em um dos blocos do *blockchain*. Porém, há que se ressaltar que existe a possibilidade, remota, de um computador conseguir descriptografar todos os blocos ou toda a cadeia, até a fonte (primeiro bloco), e com isso obter para si todos os *Bitcoins* que foram minerados [24].

Vale ressaltar que mesmo sendo uma tecnologia extremamente inovadora e com um potencial enorme, ainda precisa de adesão em massa em todos os níveis da sociedade, para atingir uma maior estabilidade cambial, dando fim a era do dinheiro físico e nos impulsionando diretamente para o futuro no âmbito tecnológico e financeiro [25].

2.1.2 Aplicações

Na literatura, diversos trabalhos propuseram aplicações com base no *blockchain* para outros domínios do conhecimento, dado o universo de possibilidades que essa tecnologia proporciona e os vários benefícios disponíveis.

Um exemplo de aplicação dessa tecnologia é a empresa *Everledger*, que faz uso dessa ferramenta como registro permanente de diamantes e histórico de transferências. Tal registro também contém a listagem de características únicas, para identificação, das pedras como: comprimento, largura, cor, peso, profundidade, entre outras [18].

No ramo de tráfego aéreo existe o conceito de *System Wide Information Management (SWIM)*, modelo usado para gerenciar as informações de tráfego aéreo, têm aplicações na Europa e nos Estados Unidos. O *SWIM Registry* é parte integrante desse modelo e nele são armazenados os dados de tráfego aéreo. No Brasil, já existem estudos que fazem a proposição desse modelo de registro usando o *blockchain* como alternativa às técnicas atualmente usadas, com a garantia da autenticidade, segurança, eficiência, integridade e confiabilidade desses registros [26].

2.2 Mineração de textos

A definição acerca do termo mineração de dados, dada sua multidisciplinaridade, pode ser definida como uma análise exploratória de dados, aplicação de algoritmos de aprendizado de máquina e reconhecimento de padrões, em busca de conhecimento [27].

Vista como uma especialização da área de mineração de dados, a mineração de textos objetiva extrair conhecimento relevante e útil em documentos de textos [28], como as relações de similaridade entre os documentos de textos, como também suas características globais e locais implícitas. Nesse sentido, diversas tarefas podem ser aplicadas, podendo ser citadas a classificação [29], regressão [29], agrupamento [30], reconhecimento de entidades nomeadas [31], simplificação de textos [32], detecção de anomalia [32] etc.

O processo de mineração de textos pode ser definido conforme as etapas ilustradas na Figura 2.2. Primeiramente ocorre a coleta e formação do(s) conjunto(s) de textos. O pré-processamento é aplicados aos textos originais para remover termos irrelevantes e redundantes. Em seguida, a caracterização dos textos determina os vetores de características associados aos textos do conjunto. Os algoritmos de aprendizado de máquina recebem tais vetores como entrada e a extração de informação determina os padrões relevantes e o conhecimento implícito nos textos. Por fim, a experimentação valida e avalia os modelos de aprendizado de máquina gerados, possibilitando a interpretação dos resultados pelos especialistas [1].

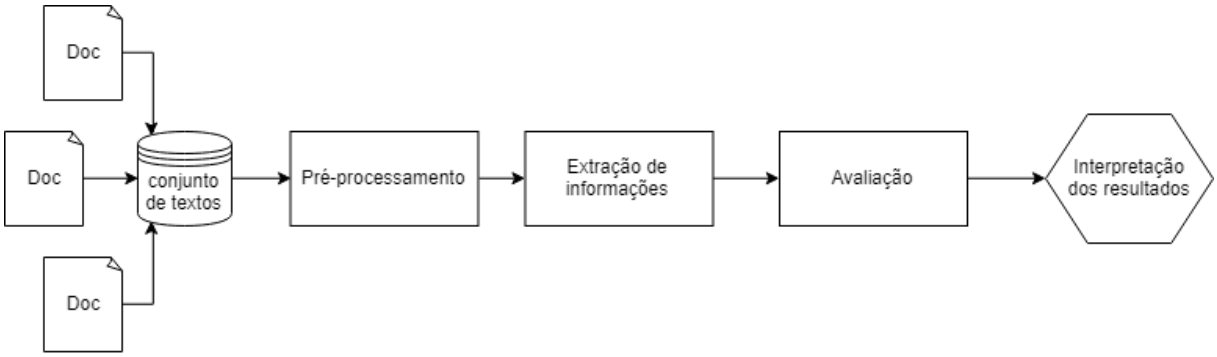

Figura 2.2: Etapas de um processo tradicional de mineração de textos [1]. Fonte: própria (2019).

As próximas subseções descrevem em detalhes cada uma dessas etapas.

2.2.1 Pré-processamento de textos

A qualidade dos textos coletados é uma etapa fundamental em processos de mineração de textos. Em sua forma original, textos podem conter ruídos, ausência de normas de linguagem, inconsistência e/ou redundância de informações, que afetam os algoritmos empregados em processos de mineração de textos, gerando resultados equivocados e não representativos [33]. Por isso, pré-processar o texto é um procedimento fundamental para evitar que resultados insatisfatórios em tarefas de análise textual sejam obtidos [34]. Isso ocorre pelo fato de existirem grandes quantidades de palavras responsáveis unicamente pela coesão textual, como: preposições, conjunções, artigos, advérbios, números, pronomes e pontuações. O pré-processamento realiza o tratamento e a limpeza dessas palavras, isso pode evitar as distorções dos resultados [35].

Após a obtenção do conjunto de documentos textuais, técnicas de pré-processamento são aplicadas para formatar e estruturar os textos, sem remover suas características naturais. A técnica de remoção de *stopwords* consiste em extrair termos, de baixa relevância, presentes nos textos [36]. Adicionalmente a isso, a técnica de stemização (*stemming*) reduz as palavras ao seu radical, visando tratar variações morfológicas das palavras, ao remover as desinências, afixos, e vogais temáticas. Consequentemente, os termos derivados de um mesmo radical serão contabilizados como um único termo. Em conjunto com as técnicas anteriormente descritas, *URL's* presentes no texto são removidas por não conterem significado relevante para o texto, o mesmo ocorre com características informais presentes em textos, não regidos pela norma culta da língua, como: *emojis*, espaços extras e caracteres especiais. O uso conjunto dessas técnicas resulta na redução do custo computacional, uma vez que o conjunto de textos apresentará um vocabulário reduzido [37].

2.2.2 Caracterização de textos

O volume de informações textuais presentes em redes sociais como o *Twitter* é grande e diverso, mesmo com limitação de 280 caracteres por mensagem (*tweet*). Isso justifica o interesse cada vez maior em analisar esses dados e obter informações úteis a partir dessa rede. Um aspecto comum aos *tweets* é a falta de padronização das sentenças, uma vez que os textos podem apresentar abreviações, *emojis*, *emoticons* e gírias.

Os algoritmos de aprendizado de máquina, que no processo de mineração de textos são empregados na etapa de extração de informações, são incapazes de manipular dados textuais ou um conjunto dos termos de textos. Portanto, tais algoritmos requerem que os dados de entrada estejam padronizados e representados numericamente sob a forma de vetores de características.

Para viabilizar tal processamento, deve-se extrair as características dos textos por meio da análise da frequência das palavras nos textos de uma coleção. Para esse propósito, técnicas de caracterização de textos foram propostas, sendo o *bag-of-words* e o *term frequency-inverse document frequency* (*TF-IDF*) as abordagens mais populares na literatura de mineração de textos [38]. Como resultado, cada *documento* é transformado em um vetor de características. É importante notar que esses vetores obtidos a partir de uma coleção de documentos possuem a mesma dimensão, isto é, a mesma quantidade de atributos, mesmo possuindo diferentes quantidades de palavras.

Bag-of-words

O modelo *bag-of-words* é uma representação de fácil abstração, entendimento e implementação, de um documento exibido como vetores de termos. Tal modelo é amplamente utilizado e é apropriado para tarefas de mineração de textos. A técnica *bag-of-words* recebe com entrada os dados textuais e os transforma uma representação numérica estruturada. Isso é realizado através de uma matriz (documento, termo) em que as linhas representam o conteúdo presente, ou não, no documento e as colunas, os termos catalogados no *bag-of-words*. Para exemplificar[1], têm-se as seguintes sentenças:

1. *"Aplicativos inteligentes criam processos de negócios inteligentes"*
2. *"Os robôs são aplicativos inteligentes"*
3. *"Eu faço inteligência de negócios"*

As sentenças acima não possuem o mesmo comprimento, o que exemplifica a forma não-estruturada dos textos. A Tabela 2.1 é uma matriz (documento, termo), a primeira linha consiste de todos os termos presentes nas sentenças 1, 2 e 3, já aplicadas técnicas de pré-processamento de textos. As representações numéricas reproduzem os termos presentes e a contagem de cada um, em cada documento.

Os vetores gerados pelo *bag-of-words* caracterizam-se pela alta dimensionalidade, que estará diretamente relacionada com a quantidade de palavras existente em uma coleção

[1]A Tabela 2.1 e os exemplos relativos a *bag-of-words* foram extraídos do endereço: *http://www.darrinbishop.com/blog/2017/10/text-analytics-document-term-matrix/*.

Tabela 2.1: Exemplo de *bag-of-words*.

	inteligen	**aplic**	**cri**	**negoci**	**process**	**robo**	**sao**	**eu**	**fa**
Doc1	2	1	1	1	1	0	0	0	0
Doc2	1	1	0	0	0	1	1	0	0
Doc3	1	0	0	1	0	0	0	1	1

de documentos. Além disso, essa representação possui limitações associadas ao grande número de palavras com características semelhantes [38].

Term Frequency-Inverse Document Frequency

A técnica, *TF-IDF*, quantifica a importância de uma palavra presente em um documento em relação a uma coleção de documentos ou conjunto de palavras. A técnica foi desenvolvida de modo a buscar equilíbrio entre o *term frequency* (*TF*) e o *inverse document frequency* (*IDF*), o que significa que a importância de uma palavra é definida, proporcionalmente, pelo número de vezes que a palavra aparece em um documento e a quantidade de ocorrências da mesma na coleção de documentos analisada [39]. Essa busca por equilíbrio visa diminuir peso de palavras comuns aos mais diversos textos pelo fato de não carregarem consigo um valor semântico relevante para identificar assuntos "chave" em um texto. Matematicamente, as funções anteriormente, citadas são definidas conforme a Equação 2.1 [40]:

$$TF(t) = \frac{Frequência\ do\ termo}{Total\ de\ termos} \tag{2.1}$$

Proposta inicialmente por, *Hans Peter Luhn (1957)* [41], a função *TF* define a importância da palavra pelo cálculo da razão entre número de ocorrências e a quantidade, de palavras presentes no documento. Contudo, essa solução pode acarretar resultados que não condizem com a realidade, como por exemplo:

> "A UnB é importante para o ramo de pesquisa científica"

A frase proposta possui palavras centrais para o entendimento do contexto: "UnB", "ramo", "pesquisa", "científica". Porém, as outras palavras "A", "é", "importante", "para" "o", "de", por serem ferramentas de coesão são comuns, e podem ser amplamente citados, em diversos documentos de categorias diferentes em relação ao assunto de interesse. A partir disso, o *IDF* foi criado de modo a contrabalancear a contabilização efetuada pela técnica *TF*, sendo representada pela Equação 2.2 [40]:

$$IDF(t) = \log_e \left(\frac{Total\ de\ documentos}{Total\ de\ documentos\ contendo\ o\ termo\ t} \right) \tag{2.2}$$

Desta forma, quanto mais raro for o termo, maior será seu valor na medida *IDF*, alcançando as palavras com legítima importância contextual para o conjunto de documentos. Logo, ao relacionar *TF* e *IDF*, obtém-se a medida *TF-IDF*, dada pela Equação 2.3 [40]:

$$TF\text{-}IDF = TF \times IDF \tag{2.3}$$

Para propósitos de exemplificação, a técnica *TF-IDF* pode ser descrita da seguinte maneira: supondo um banco de dados com descrições de milhares de moedas e uma busca por "moeda *bitcoin*". A técnica pode ser eficaz, pois a partir dos cálculos informados é aguardado que a palavra "moeda" tenha um valor menor (menor peso) que a palavra "*bitcoin*", já que a primeira palavra possui grande frequência, enquanto a segunda aparecerá em menos documentos.

Modelo espaço vetorial

Em busca de resultados mais precisos para a obtenção de documentos que respondam parcialmente a uma expressão de busca, o modelo espaço vetorial gera um conjunto de documentos ordenado pelo grau de similaridade de cada documento. Isso é feito através da associação de pesos dos termos de indexação com aqueles utilizados na expressão de busca [42]. A representação de um documento é feita por um vetor, onde cada elemento (palavra) representa o peso, ou relevância, do seu respectivo termo de indexação para o documento. Cada vetor irá representar a posição do documento em um espaço multidimensional, onde cada termo de indexação representa uma dimensão ou eixo e cada elemento do vetor (peso) é normalizado de forma a assumir valores entre zero e um [42]. O espaço vetorial contém N dimensões, a similaridade (sim) entre um documento d_j e uma expressão de busca q pode ser calculada com a seguinte Equação 2.3:

$$sim(d_j, q) = \frac{\sum_{i=1}^{N}(w_{i,j}) \times \sum_{i=1}^{N}(w_{i,q})}{\sqrt{\sum_{i=1}^{N}(w_{i,j}^2)} \times \sqrt{\sum_{i=1}^{N}(w_{i,q}^2)}} \tag{2.4}$$

Onde w_{ij} é o peso do i-ésimo termo do documento d_j e w_{iq} é o peso do i-ésimo termo da expressão de busca q.

A Figura 2.3 em conjunto com a Equação 2.5 exemplificam o cálculo definido na Equação 2.3:

$$sim(doc, busca) = \frac{(0.7 \times 0.8) + (0.6 \times 0.5)}{\sqrt{0.7^2 + 0.6^2 + 0.3^2} \times \sqrt{0.8^2 + 0.5^2}} \cong 0.94 \tag{2.5}$$

Figura 2.3: Exemplo de associação de peso sobre um documento e uma busca. Fonte: própria (2019).

2.2.3 Aprendizado de máquina

Diz-se que um programa de computador aprende a partir de uma experiência E referente a alguma classe de tarefas T e avaliação de desempenho P, se o desempenho nas tarefas T, for medido por P, aprimora com a experiência E [43].

É importante destacar que os sistemas de aprendizado de máquina são categorizados como: supervisionado, não supervisionado e por reforço. No aprendizado supervisionado é fornecido ao algoritmo, ou indutor, um conjunto de instâncias de treinamento para os quais o rótulo, categoria (por exemplo: positivo, negativo e neutro) da classe associada, é conhecido [44].

Um ramo do aprendizado supervisionado é o semi-supervisionado, método que combina dois tipos de dados, rotulados e não rotulados. Fato que reduz a necessidade de rotular manualmente instâncias. Essa tarefa é inviável para grandes conjuntos de dados textuais [40].

Já no aprendizado não supervisionado, o indutor analisa os dados fornecidos e tenta determinar se podem ser agrupados através das similaridades entre os dados, formando agrupamentos.

Aprendizado supervisionado

O funcionamento do aprendizado supervisionado pode ser descrito da seguinte forma: dado um conjunto de exemplos rotulados na forma ($\mathbf{x_i}$, y_i), em que $\mathbf{x_i}$ representa uma instância de dados (exemplo) e y_i denota sua categoria, pode produzir um classificador, também denominado modelo, preditor ou hipótese, capaz de predizer precisamente o rótulo de novos dados. Esse processo de indução de um classificador a partir de uma amostra de dados é denominado treinamento. O classificador obtido também pode ser visto como uma função f, onde recebe uma instância x e determina sua categoria [45].

Como consequência do aprendizado supervisionado, criou-se uma nova subcategoria, a classificação. Modelos de classificação são, basicamente, sistemas utilizados, desenvolvidos através de experimentação, cujo objetivo é rotular amostras através de métodos estatísticos e/ou redes neurais de processamento [46]. Para exemplificar esse procedi-

mento, podemos observar que a plataforma *Gmail* é capaz de separar os *e-mails* recebidos em principal, social, promoção e, até mesmo, *spam*, com elevados índices de acertos. Ao longo do desenvolvimento do projeto foram utilizados os seguintes classificadores: máquinas de vetores de suporte (*SVM*), redes neurais, *Naive Bayes*. Em conjunto com o modelo gerativo de extração de tópicos, alocação de *Dirichlet* latente (*LDA*), e também, a técnica de agrupamento (*K-means*).

Maquinas de vetores de suporte

As máquinas de vetores de suporte (*SVM*), são modelos de aprendizado de máquina supervisionado que em conjunto com algoritmos de aprendizado executam tarefas de classificação e análise de regressão [47]. Como forma de exemplificar considere que um conjunto de instâncias, X, compostas de dois tipos de dados, distintos, foram caracterizados e dispostos no quadrante (x, y). O conjunto de dados pode ter N características distintas, por exemplo: homem e mulher, cachorro e gato, entre outros. A Figura 2.4 apresenta círculos e triângulos para ilustrar esses dois grupos distintos.

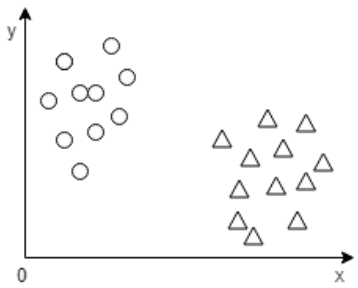

Figura 2.4: Conjunto de dados com duas características distintas. Fonte: própria (2019).

O *SVM* traça hiperplanos a fim de obter uma margem interna, livre de instâncias, que separe as *features*. Satisfeita essa condição, o *SVM* faz uso de vetores de suporte para calcular a maior margem de separação possível entre o hiperplano e as *features* mais próximas. Esse processo é conhecido como separação ótima obtém o hiperplano de maior margem, Figura 2.5.

Essa técnica parte do princípio de que se as classes são separáveis, então a solução que traz os melhores resultados é aquela que alcança a máxima separação entre as classes, ou seja, o hiperplano que possui a maior distância até os elementos mais próximos de cada classe.

A técnica computacional, *SVM*, permite rapidez na implantação em aplicações; modularidade no *design*, o que possibilita a combinação de *kernels* em aprendizados diferentes; excelente desempenho em grandes dimensões de dados e capacidade de generalização [48].

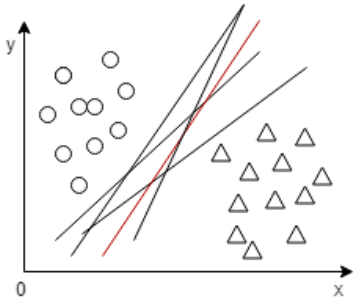

Figura 2.5: Hiperplanos usados pelo *SVM* para a realização da separação ótima. Fonte: própria (2019).

Redes neurais artificiais

As redes neurais são modelos matemáticos inspirados no sistema nervoso central. Essa inspiração se deve ao fato de que o cérebro pode ser comparado a um computador, não linear e paralelo; possui capacidade de organizar os neurônios de forma a realizar tarefas de reconhecimento de padrão, percepção e controle motor, mais rápido que qualquer computador digital [49].

Rede neural pode ser definida como um processador maciça e paralelamente distribuído, possui unidades simplificadas de processamento que naturalmente armazenam conhecimento e os disponibiliza para uso. Assim como o cérebro, as redes neurais artificiais adquirem conhecimento a partir do ambiente por um processo de aprendizagem [49].

Existem arquiteturas variadas de redes neurais artificiais, e serão apresentados dois modelos: *perceptron* simples e multicamadas.

Perceptron simples

Para elucidar o modelo de *perceptron* simples, e assim obter entendimento sobre o funcionamento do modelo. O neurônio é composto de sinais de entrada x_1, x_2, x_m; pesos sinápticos $w_{\kappa_1}, w_{\kappa_2}, w_{\kappa_m}$ onde o κ_m é o neurônio; um somatório dos produtos do sinais de entrada com os pesos sinápticos; u_κ é o retorno desse somatório [49], formalmente definido como na Equação 2.6:

$$u_\kappa = \sum_{m=1}^{n} w_{\kappa_m} x_m \tag{2.6}$$

Em seguida temos o *Bias*, b_κ, uma constante que auxilia o modelo a se adaptar aos dados fornecidos; uma função de ativação $\phi(\cdot)$, que resulta na saída y_κ [49], formalmente descrita como na Equação 2.7:

$$y_\kappa = \phi(u_\kappa + b_\kappa) \tag{2.7}$$

O modelo completo é ilustrado pela Figura 2.6. Inicialmente o modelo apresenta os sinais de entrada que fluem através dos pesos sinápticos em seguida é aplicada a junção aditiva assim como foi definido na Equação 2.6. Posteriormente é também aplicada a função de ativação, isso resulta na saída Y_k, definida na Equação 2.7.

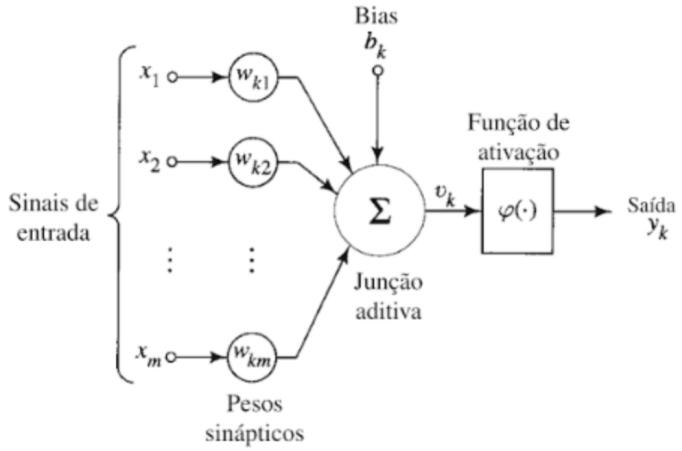

Figura 2.6: Modelo não linear de um neurônio. Fonte: própria (2019).

Perceptron multicamadas

O algoritmo *perceptron* multicamadas ou *multi-layer perceptron* (*MLP*), tem com característica uma ou mais camadas ocultas de neurônios. Obtém, a partir de treinamento, uma função $f(\cdot) : R^n \rightarrow R^o$ através do conjunto de textos, onde n é o número de dimensões de entrada e o, de saída. Dado um conjunto de características $X = x_1, x_2, x_n$ e um alvo y, o *MLP* pode realizar tarefas de classificação ou regressão, por meio de um aproximador não linear.

A Figura 2.7 exemplifica o *MLP* com uma camada oculta. A camada de entrada é constituída por neurônios, x_1, x_2, x_n, recursos de entrada. Cada neurônio da camada oculta (segunda camada),a_1, a_2, a_k, realiza uma soma ponderada dos recursos de entrada, x_n, com os pesos sinápticos, w_n. Em seguida é aplicada sobre a_1, a_2, a_k a função de ativação $f(\cdot)$ que envia esses valores para a camada de saída, que os transforma em valores de saída [50].

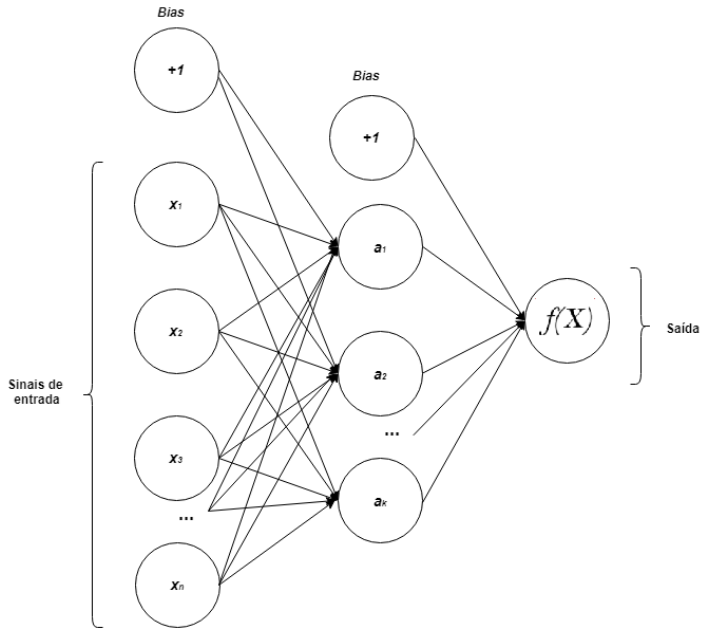

Figura 2.7: *Perceptron* multicamadas com uma camada oculta. Fonte: própria (2019).

O *perceptron* multicamadas tem capacidade de aprender modelos não lineares, porém possui a desvantagem de ter muitos hiperparâmetros a serem calibrados, por exemplo: função de ativação, *momentum*, número de neurônios nas camadas ocultas, entre outros [50].

Naive Bayes

Esse modelo probabilístico é bastante simplificado, funciona de forma a analisar cada palavra independente, por exemplo:

- "meu cachorro gosta de brincar"
- "Cachorro pode ser agressivo"
- "Ela ama cachorro"

O modelo, *Naive Bayes*, classifica as *features* (instâncias presentes no espaço vetorial) de teste, a partir do conjunto de dados rotulados, por meio da aplicação do calculo de probabilidade, mostrado na Equação 2.8. É realizado o treinamento do classificador onde se determina a probabilidade de uma *feature* possuir determinada polaridade. Porém, se a palavra a ser rotulada não for parte do conjunto de treino, o modelo aplica como padrão a probabilidade da classe de maior frequência no documento.

$$p(C_\kappa \mid x) = \frac{p(C_\kappa)p(x \mid C_\kappa)}{p(x)} \tag{2.8}$$

Tabela 2.2: Exemplo de *features* de treino, rotuladas.

Feature	**Polaridade**
cachorro	positivo
gosta	positivo
brincar	positivo
cachorro	negativo
pode	positivo
ser	positivo
agressivo	negativo
ama	positivo
cachorro	positivo

A Tabela 2.2, exemplifica como as *features* de treino podem ser listadas.

Aprendizado não supervisionado

A função do aprendizado não supervisionado é identificar a organização dos padrões presentes nos dados por meio de agrupamentos (*clusters*). Tal fato proporciona observar similaridades e dissimilaridades entre os padrões presentes nos conjuntos de dados analisados, e assim, extrair conclusões sobre os dados. É possível definir agrupamento como: dado um conjunto $X = x_1, x_2, x_n$ que representa uma coleção com n documentos, uma partição P $= G_1, G_2, G_k$ com k grupos, tal que [51]:

- $G_i \neq \emptyset$, para todo $i \in 1, 2, k$;
- $G_1 \cup G_2 \cup ... \cup G_k = X$, a união de todos os grupos é igual ao conjunto de dados original;
- $G_i \cap C_j$ para todo $i \neq j$, não há interseção entre as *features* de dois grupos diferentes.

Um agrupamento, ou grupo, é uma coleção de *features* que são similares, entre si, e dissimilares, a objetos presentes em outros grupos. Um exemplo de algoritmo que realiza tarefas de agrupamento é o *k-means*.

K-means

O funcionamento da técnica *k-means* consiste na fixação aleatória de k centroides, tal que o valor de k representa a quantidade de particionamento dos dados, ou seja, k representa o numero de grupos a serem formados [51]. Em seguida associa-se cada *feature* (elementos do conjunto de dados) ao centroide mais próximo. Feito isso, os centroides são recalculados baseados nas *features* classificadas. Esse processo é repetido até que não ocorra mais alterações nos grupos [51].

O algoritmo *k-means* pode ser descrito como na Figura 2.8 [51]:

Algoritmo 1: O algoritmo k-means

Entrada:
$X = \{x_1, x_2, ..., x_n\}$: conjunto de documentos
k: número de grupos

Saída:
$P = \{G_1, G_2, ..., G_k\}$: partição com k grupos

```
1 selecionar aleatoriamente k documentos como centroides
  iniciais;
2 repita
3     para cada documento x ∈ X faça
4         computar a (dis)similaridade de x para cada
          centroide C ;
5         atribuir x ao centroide mais próximo ;
6     fim
7     recomputar o centroide de cada grupo;
8 até atingir um critério de parada;
```

Figura 2.8: Algoritmo de particionamento do *k-means*.

2.2.4 Avaliação de performance de classificadores

Após a exposição de alguns classificadores, a próxima etapa é avaliar a performance do processo de classificação.

Matriz de confusão e *F1-Score*

A matriz de confusão é bastante utilizada para essa tarefa e o seu funcionamento é de fácil compreensão. Considere uma hipótese de uma matriz 2x2 (VR, VP), valores reais (VR) e valores preditos (VP). A matriz de confusão se baseia em um problema de classificação binária, cujas categorias são denominadas positiva e negativa, como na Figura 2.9.

Obtêm-se os seguintes resultados: verdadeiro positivo, falso positivo, falso negativo e verdadeiro negativo. Observa-se que os verdadeiros positivos são instâncias classificadas

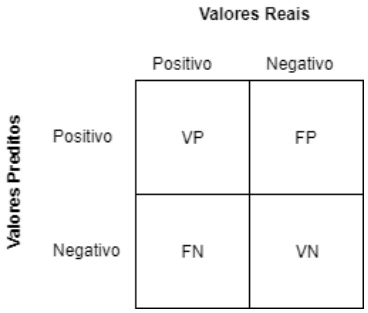

Figura 2.9: Matriz de Confusão. Fonte: própria (2019).

como positivas, mas que são originalmente positivas; já os falsos positivos são instâncias classificadas como positivas, porém são originalmente negativas; isso se repete para os falsos negativos; e os verdadeiros negativos são equivalentes aos verdadeiros positivos [33].

A matriz de confusão traz consigo três conceitos que são fundamentais: *Accuracy*, *Precision*, *Recall* e o *F1-Score*. *Accuracy* é a razão entre as previsões verdadeiras (acertadas), com a soma de todas a previsões [15], Equação 2.9:

$$accuracy = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.9}$$

Já o *recall* [15] tem o objetivo, distinto, de mensurar a qualidade das predições, ou seja, identificar o quão bom o classificador é em obter previsões, positivas ou negativas, de forma correta. Considerando que o objeto de avaliação sejam as previsões positivas, o cálculo é feito da razão entre os verdadeiros positivos sobre a soma dos verdadeiros positivos com os falsos negativos, Equação 2.10:

$$recall = \frac{VP}{VP + FN} \tag{2.10}$$

O *precision* [15], é o cálculo da proporção de identificações corretas. Isso é realizado, tendo como base as instâncias positivas, através da razão entre os verdadeiros positivos sobre a soma dos verdadeiros positivos e os falsos positivos, Equação 2.11:

$$precision = \frac{VP}{VP + FP} \tag{2.11}$$

A medida, *F1-Score*, [52] é definida pela média harmônica entre o *recall* e o *precision*, Equação 2.12:

$$F1\text{-}Score = 2 \times \left(\frac{precision \times recall}{precision + recall} \right) \tag{2.12}$$

A pontuação *F1-Score1* alcança o melhor valor, ou seja, *recall* e *precision* perfeitas, com um valor 1 e a pior, seria com valor 0. O *F1-Score1* é importante para fornecer uma medida realista do desempenho de testes [53].

2.3 Análise de sentimentos

Com o advento da internet e os avanços tecnológicos subsequentes, foram criados incontáveis meios para entreter a humanidade. A tentativa que se mostrou mais acertada foi, até o momento, a criação das redes sociais. A adesão em massa alcançada nas redes tem gerado, como consequência, grandes massas de dados [54].

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

O nome "análise de sentimentos" é de alguma forma, autoexplicativo, porém não obstante a isso é possível definir que: são técnicas cujo objetivo é extrair automaticamente informações subjetivas de textos escritos em linguagem natural [55]. Tais definições podem ser descritas de N formas, porém neste trabalho foram usadas três: positivo, negativo e neutro.

A aplicação de técnicas de análise de sentimentos viabiliza a transformação de textos em efetivo conhecimento, e consequentemente, geram inteligência. Esse conceito é amplamente estudado em disciplinas como sistemas de informação, onde é desenvolvida a capacidade de diferenciar dados, informação e conhecimento.

Inicialmente, o conceito de dados é definido da seguinte forma: São códigos que constituem a matéria prima da informação, ou seja, é o conteúdo que ainda não apresenta relevância. Os dados representam um ou mais significados de um sistema que isoladamente não podem transmitir uma mensagem ou representar algum conhecimento [56].

O resultado do processamento de dados são as informações [56]. As informações têm significado e podem contribuir no processo de tomada de decisões.

Por outro lado o conhecimento é o ato ou efeito de abstrair ideia ou noção de alguma coisa, como por exemplo: conhecimento das leis; conhecimento de um fato (obter informação); conhecimento de um documento; conhecimento da estrutura e função de determinados sistemas. O saber, a instrução ou domínio científico estão relacionados com o conhecimento [57].

Esta etapa, de análise de sentimentos, desperta o questionamento: "Qual é a importância de saber o que pensam as pessoas?" ("*What other people think*") [58]. Essa

questão é bastante significativa e vai de encontro com as possibilidades geradas a partir da aplicação de técnicas de análise de sentimentos. Sendo assim, se faz necessário apontar exemplos, de possíveis aplicações da técnica, como: avaliar a receptividade do consumidor, perante produtos e serviços oferecidos (*feedback*), mapear áreas de risco, observar novas necessidades de mercado e finalmente, prever eventos significativos.

Especificamente, o exercício da "adivinhação" é uma tarefa muito relativa e complexa, o que não significa absoluto impedimento a tentativa. Os "documentos" utilizados, providos pelo *microblogging, Twitter*, detém caráter dinâmico mesmo frente a sua limitação, de 280 caracteres. Essa rede tem a característica de carregar impressões quase imediatas, dos seus usuários, em resposta a acontecimentos recentes. A análise realizada sobre mensagens de texto, curtas e praticamente instantâneas, reflete a real impressão deixada junto ao consumidor, isso potencializa a obtenção de sentimentos característicos.

A continuidade do processo de análise de sentimentos se faz dependente da capacidade que um algoritmo tem de realizar tarefas de rotulação do texto. Existem alguns métodos para a obtenção dessas polaridades, que posteriormente servirão de rótulos para o conjunto de dados, por exemplo: análise da polaridade do documento, como um todo; das sentenças, que compõem o documento; ou das características e atributos, dos objetos presentes no documento.

Neste projeto, utiliza-se a tarefa de rotulação dos textos, a um método de aprendizado de máquina supervisionado. Essa metodologia tem alcançado bons resultados considerando a abundância de trabalhos relacionados ao assunto, na literatura.

2.4 Redução de dimensionalidade

A palavra dimensionalidade está relacionada ao número de características de uma representação de padrões, ou seja, a dimensão do espaço de características. A alta dimensionalidade dos dados pode afetar o desempenho de algoritmos de aprendizado de máquina, em um processo denominado *maldição da dimensionalidade* [59], em que as dimensões se tornam irrelevantes ao serem comparadas uma outras, principalmente no cálculo de dissimilaridade entre instâncias multidimensionais. A redução de dimensionalidade pode minimizar esse fenômeno e se baseia em duas abordagens: seleção de atributos e transformação [33]. A seleção de atributos consiste em selecionar um sub-conjunto dos atributos conforme algum critério de relevância de atributos em consonância com a tarefa de aprendizado de máquina adotada. Apesar dessa abordagem preservar os valores originais dos atributos, a tarefa de determinar o critério adequado para selecionar os atributos mais relevantes é considerada difícil.

A outra abordagem se baseia na projeção de características, que transforma os atributos originais em um conjunto reduzido de novos atributos. Essa abordagem compreende uma função f, que recebe um conjunto $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ de instâncias de dados de dimensionalidade m e realiza um mapeamento para um conjunto $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ de dimensionalidade p, em que $p < m$ [60]. A função de mapeamento deve ser definida de acordo com algum critério de preservação das relações de similaridade e de vizinhança dos dados e deve produzir um espaço de baixa dimensão, que preserve ao máximo seus padrões e estruturas originais.

2.4.1 *PCA e TruncatedSVD*

Uma das técnicas de redução de dimensionalidade mais populares é a Análise de Componentes Principais (*PCA - Principal Component Analysis*) [61]. O *PCA* realiza uma transformação ortogonal nos dados definidos no espaço multidimensional para um espaço de baixa dimensão em que a variância dos dados é maximizada. O *PCA* calcula a matriz de covariância dos atributos e realiza sua decomposição espectral, obtendo autovetores e autovalores. Os autovetores associados aos maiores autovalores são fundamentais para definir as componentes principais, que retém grande parte da variação dos dados e representam o espaço de baixa dimensão. No entanto, o *PCA* apresenta limitações ao processar matrizes muito esparsas, característica comum aos espaços multidimensionais definidos por vetores *TF-IDF*, diminuindo sua performance e por vezes inviabilizando o processo de redução de dimensionalidade.

O transformador *Truncated SVD* (*LSA*) realiza redução da dimensionalidade linear através da decomposição do valor singular truncado (*SVD*). Diferentemente do *PCA*, o *Truncated SVD* não centraliza os dados antes de calcular a decomposição do valor singular. Ou seja, esse transformador pode processar, eficientemente, matrizes esparsas [50].

O uso do *SVD* completo é bastante incomum, normalmente é necessária uma decomposição unitária completa do espaço nulo da matriz. Em vez disso, geralmente é suficiente e mais eficiente, para armazenamento, calcular uma versão reduzida do *SVD*. Para uma matriz $m \times n$ M de classificação r, o *Truncated SVD* pode ser definido como na Equação 2.13 [62].

$$\dot{M} = U_t \Sigma_t V_t^* \tag{2.13}$$

Apenas os vetores da coluna t dos vetores de linha U e t de V^* correspondentes aos valores t maiores Σ_t são calculados. O resto da matriz é descartado. Isso pode ser muito mais rápido e mais econômico do que o compacto *SVD* se $t \ll r$. A matriz U_t é assim $m \times t$, Σ_t é $t \times t$ diagonal e V_t^* é $t \times n$ [62].

A matriz obtida pelo *Truncated SVD* deixa de ser uma decomposição exata da matriz original M, porém a matriz aproximada $\dot{M}$ está se aproxima bastante de M [62].

Neste projeto, a técnica de redução de dimensionalidade será empregada em um conjunto de textos para diminuir os esparsos, remover recursos redundantes, aperfeiçoar os resultados da técnica de agrupamento, e viabilizar a visualização em duas dimensões das instâncias.

2.4.2 Visualização dos textos

Como este projeto requer identificar padrões e relações de similaridade entre os *tweets* sobre *Bitcoins*, uma alternativa viável e interessante consiste em empregar processos de visualização exploratória, que realiza a descoberta de conhecimento implícito e relevante baseado no uso de representações gráficas dos dados e de recursos de interatividade [63]. A visualização exploratória pode ser vista como um processo de geração de hipóteses, em que os especialistas humanos utilizam visualizações para validar as hipóteses inicialmente estabelecidas sobre o domínio dos dados. As representações gráficas geradas pelas visualizações permitem que os especialistas humanos interpretem e identifiquem padrões nesses dados, fazendo com que novas hipóteses sejam elaboradas.

Processos de visualização exploratória requerem o emprego de técnicas de visualização da informação, que recebem como entrada um conjunto de dados multidimensional e geram como saída metáforas visuais (denominadas *layouts*). A ideia chave da visualização é aproveitar a capacidade do sistema de percepção humano em tarefas de interpretação de imagens, uma vez que, podem-se identificar padrões e estruturas relevantes nos dados de maneira mais eficiente quando comparada com uma simples inspeção visual nos dados brutos [64].

Diversas técnicas de visualização da informação foram propostas, em que o gráfico de dispersão, as coordenadas paralelas e o mapa de calor são as mais populares [65]. A Figura 2.10 é um exemplo de gráfico de coordenadas paralelas que permite comparar a característica de várias observações individuais em um conjunto de variáveis numéricas, representada por linhas no gráfico. O conjunto de dados Iris apresenta 150 espécies de flores, caracterizadas por 4 atributos (comprimento da Sépala, largura da Sépala, comprimento da Pétala e largura da Pétala) e categorizadas em três espécies (*Setosa*, *Versicolor* e *Virginica*). Em ambos os *layouts*, as cores estão associadas às espécies das flores.

No entanto, essas técnicas apresentam limitações ao gerar *layouts* intuitivos quando a dimensionalidade dos dados é alta. Por isso, diversas técnicas de visualização baseadas em algoritmos de redução de dimensionalidade tem sido propostas nos últimos anos [66]. Essas abordagens são baseadas no posicionamento de pontos no espaço visual [67], em que

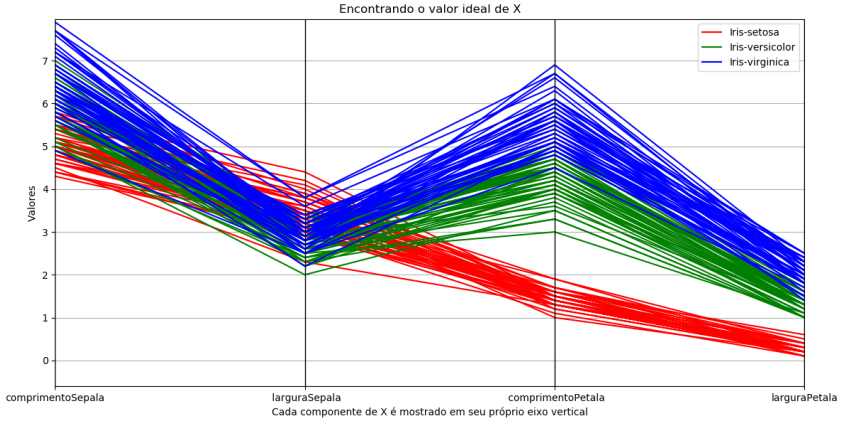

Figura 2.10: Layout obtido pela técnica de visualização coordenadas paralelas.

cada símbolo visual está associado a uma instância do conjunto de textos e a cor de cada ponto representa uma categoria. A visualização posiciona os pontos no *layout* conforme as relações de similaridade das instâncias no espaço original, isto é, dados similares ficam próximos no *layout*, enquanto que dados dissimilares ficam mais distantes. Portanto, essas visualizações não demandam a definição de eixos coordenados, logo, a análise de padrões nos dados ocorre pela densidade e geometria dos grupos formados no *layout*.

A Figura 2.11 ilustra o *layout* produzido pela técnica de visualização baseada no PCA utilizando o conjunto de dados Iris. De acordo com o posicionamento dos pontos, percebe-se que duas espécies de flores apresentam formas mais similares.

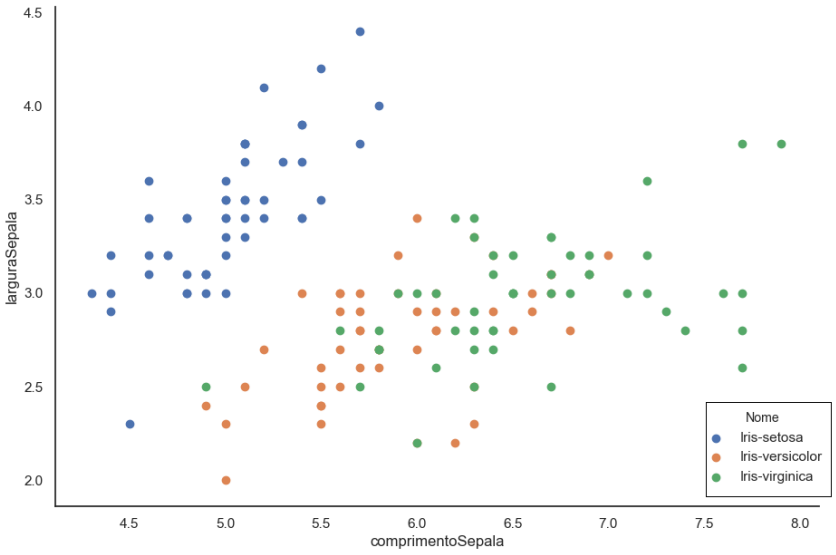

Figura 2.11: *Layout* produzido por uma visualização baseada em PCA utilizando o conjunto de dados iris.

Neste projeto, um dos objetivos é visualizar conjuntos de *tweets* para aprimorar a compreensão das análises comparativas entre a variação do mercado e o volume polaridades obtidas dos *tweets*, responder às hipóteses pré-definidas e desenvolver novos questionamentos sobre o método proposto.

2.5 Modelagem e extração de tópicos

Dada uma coleção de documentos, a modelagem de tópicos é um modelo gerador, que visa extrair os principais assuntos abordados pela coleção. Os documentos podem ser gerados através da utilização de distribuições sobre tópicos. Os tópicos são um conjunto de palavras que possuem frequência em documentos que estão semanticamente relacionados, são formados por uma distribuição probabilística de termos [68].

O *Latent Dirichlet Allocation* (*LDA*) é um algoritmo cuja função é modelar os tópicos presentes no documento. Considera que cada documento é composto por um conjunto de tópicos, cabe ao processo identificar palavras e agrupá-las com tópicos que possuem relacionamento com as mesmas, a Tabela 2.3 ilustra esse fenômeno.

Tabela 2.3: Demonstração dos tópicos relevantes.

Tópicos	comer	dormir	brincar	miar	latir
Tópico 1	0.1	0.3	0.2	0.4	0.0
Tópico 2	0.2	0.1	0.4	0.0	0.3

Esse modelo Bayesiano é completo e embasado na geração de tópicos como distribuições de *Dirichlet*, capaz de classificar documentos não conhecidos utilizando informações fornecidas previamente [69]. Essa capacidade de inferir tópicos deriva do seguinte processo gerativo [70]:

1. Escolha $N \sim Poisson(\xi)$
2. Escolha $\theta \sim Dir(\alpha)$
3. Para cada uma das N palavras w_n
    - Escolha o tópico $z_n \sim$ Multinomial(θ)
    - Escolha a palavra para $p(w_n|z_n,\beta)$, a probabilidade multinomial condicionada ao tópico z_n.

A Tabela 2.3 exemplifica um conjunto de documentos em que as principais palavras são: "comer", "dormir", "brincar", "miar" e "latir", inferidas a partir da distribuição de probabilidade (*Dirichlet*) sobre as palavras de cada documento. A observação dos tópicos permite perceber que há grande possibilidade do "Tópico 1" estar relacionado ao animal gato e o "Tópico 2" ao cachorro, dado que os maiores valores são de palavras características dos respectivos animais.

O ajuste dos parâmetros para a aplicação do *LDA* influencia no resultado do algoritmo, pois um alto valor de α, que relaciona a distribuição documento-termo, pode significar que cada documento terá uma maior mistura de tópicos. Em contraposição, o valor baixo provavelmente indicará uma mistura de poucos tópicos. A escolha do parâmetro β, relacionado à distribuição tópico-palavra, com um alto valor significa que cada tópico pode resultar em uma mistura de várias palavras. Enquanto que um valor baixo pode indicar que o tópico será formado por poucas palavras.

Capítulo 3

Revisão de Literatura

Este capítulo tem grande influência na elaboração dessa pesquisa científica, pois nele estão relatadas as pesquisas primárias, projetos inspiradores e que ajudaram na construção deste trabalho. A Seção 3.1 retrata resumos de artigos que empregam projetos semelhantes a este. A Seção 3.2 informa sobre a tese explorada, explicitando a forma de abordagem.

3.1 Trabalhos relacionados

Em vista do desenvolvimento de tecnologias linguísticas e as mídias sociais, que fornecem possibilidades poderosas para investigar o humor dos usuários e os estados psicológicos das pessoas. Neste artigo, é discutida a possibilidade de melhorar a precisão das previsões dos indicadores do mercado de ações através dos estados psicológicos dos usuários do *Twitter*.

A análise foi feita em um conjunto de 755 milhões de *tweets*, coletados no período, de 13/02/2013 a 29/09/2013, que observa o sentido das palavras para avaliar a presença de oito emoções básicas.

O resultado da pesquisa indicou que a adição de informações do Twitter não nos permite aumentar significativamente a precisão e que usando o algoritmo *Support Vector Machine* foram obtidos melhores resultados para prever indicadores da *DIJA (Dow Jones Industrial Average)* [71].

Pesquisa consiste em trabalhar com o proposito de tentar prever movimentos do mercado de ações, utilizando especialmente o *Twitter*, já que essa mídia social tem muita representação da opinião pública sobre eventos atuais. Especificamente, em como as mudanças nos preços das ações de uma empresa, os aumentos e quedas, estão correlacionadas com as opiniões públicas expressas em *tweets* sobre essa empresa. A metodologia usou de técnicas de caracterização textual, *Word2vec* e *Ngram*, e princípios do aprendizado de máquina supervisionado, para realizar tarefas de classificar os sentimentos nos *tweets*. Com a finalidade de observar a correlação entre os movimentos do mercado de ações de uma

empresa e os sentimentos presentes nos *tweets*. Uma das teses da pesquisa tentou evidenciar que notícias positivas e *tweets*, a respeito de uma empresa, definitivamente, podem encorajar as pessoas a investirem nas ações daquela empresa, resultando na valorização de suas ações. A conclusão da pesquisa, mostra que existe correlação entre valorizações e desvalorizações dos preços das ações com os sentimentos do público nos *tweets* [72].

Pesquisa fundamentada na economia comportamental, na qual busca indícios de que as emoções e sentimentos derivados de *tweets* podem estar correlacionados ao valor da *Dow Jones Industrial Average* (DJIA).

A metodologia desenvolvida consistiu da análise de *tweets* diários no período de 28 de fevereiro a 19 de dezembro de 2008. Foram extraídos dos textos, sentimentos (positivo ou negativo) e as dimensões de humor (calma, alerta, confiante, vital, amável e feliz). A partir disso, foi investigada a hipótese seguinte: "Estados públicos de humor são elementos preditivos relacionados às mudanças nos valores de fechamento do *DJIA*?". Tais investigações foram pautadas através da aplicação da análise de causalidade de *Granger* e da rede neural *Fuzzy*.

Os resultados das análises se mostraram expressivos, principalmente, em relação ao humor público dividido em seis dimensões, pois suas mudanças correspondiam diretamente nos valores da *DIJA*, ocorridos após 3 ou 4 dias. Além disso, os testes indicam que há margem para evolução, a fim de encontrar melhorias nos modelos anteriormente aplicados [14].

Este trabalho que foi desenvolvido na PUC-MG se propõe a desenvolver uma análise da homofilia política entre usuários do *Twitter* durante a eleição presidencial dos Estados Unidos em 2016. Foi realizado um estudo durante um período do ano eleitoral a partir de dados coletados no *Twitter*: *tweets*, perfis de usuários e suas redes de contatos. Com objetivo de analisar as tendências em relação às intenções de votos sobre os principais candidatos: *Donald Trump* e *Hillary Clinton*. O pleito foi caracterizado por uma disputa acirrada, principalmente, após as primárias partidárias que resultou na polarização entre representantes do partido republicano e, democrata, fato gerou uma série de embates políticos e ideológicos.

Vale ressaltar que as redes sociais são plataformas populares e democráticas, têm assumido papel primordial na exposição pública de opiniões, promoção de debates e disseminação de informação entre pessoas. É fato que a política tem sido um tema amplamente mencionado, com isso, os autores identificaram a oportunidade de analisar a homofilia política entre usuários do *Twitter*, tendência das pessoas possuírem características e comportamento semelhantes ao de seus pares.

A metodologia utilizada foi à coleta de dados do *Twitter*, o cálculo do sentimento dos *tweets*, os critérios de identificação dos sujeitos de cada *tweet* e o cálculo da homofilia

entre usuários. Durante 122 dias foram coletados 3.6 milhões de *tweets* de 18.450 usuários diferentes do *Twitter*, que utilizou como parametrização pessoas que comentavam sobre a eleição presidencial americana. A partir disso foi utilizada uma ferramenta para extrair o sentimento das mensagens, verificando se um usuário tem sentimento favorável (positivo), desfavorável (negativo) ou neutro em relação ao candidato. Os resultados obtidos indicaram que a homofilia está presente nos dados analisados e que houve bastante manifestação negativa em relação aos candidatos, principalmente para o candidato *Trump*. Também houve sentimento positivo, mas em menor intensidade [73].

Pesquisa realizada nas Universidades do Estado do Amazonas (UEA) e Federal do Amazona (UFAM), que através da observação da ascensão das redes sociais, viu a oportunidade de se extrair informações relevantes através de métodos de aprendizagem de máquina. O *Twitter*, uma rede social online com um grande número de postagens diárias, tornou-se uma importante fonte de informações sobre eventos diversos, consequentemente, a plataforma foi escolhida como fonte de dados da pesquisa. Em contrapartida, essas informações são caracterizadas pela difícil compreensão, uma vez que há uma diversidade contextual e um custo elevado para processar os dados. Nesse contexto, o trabalho propôs uma caracterização de informações relevantes sobre eventos. Foram avaliadas técnicas de aprendizagem de máquina não supervisionadas para detecção de tópicos com o intuito de analisar uma nova abordagem que permita extrair informações e mostrar a viabilidade de se utilizar o *Twitter* para descobrir relatos relevantes de um evento.

A metodologia abordada parte da escolha de uma técnica de aprendizagem de máquina (*K-means, Non-negative Matrix Factorization*) para extração de tópicos, análise de duas abordagens de pré-processamento para textos, buscando eliminar ruídos existentes nos dados, e fazendo agrupamentos dos dados, a fim de obter um valor de grupos que melhor represente o conjunto de dados. O resultado obtido através de dados coletados sobre a Operação Lava Jato da Polícia Federal no ano de 2016 se mostrou assertivo, pois os tópicos resultantes eram semelhantes com as principais notícias que estavam sendo veiculadas pela mídia [74].

Esse artigo tem o objetivo de dissertar sobre classificação de filmes com base em suas legendas e informações capturadas de redes sociais. Utilizando a computação afetiva, mais especificamente: análise de sentimentos e reconhecimento de emoções, o trabalho busca auxiliar usuários a encontrar filmes de interesse, já que a quantidade de filmes disponíveis são muito grandes.

A proposta do trabalho foi atribuída em quatro fases: extração dos dados, seleção dos dados, transformação dos dados e seleção de atributos. Ao utilizar dicionário de palavras do *LIWC*, método para estudar fatores emocionais, cognitivos e estruturais, foi possível classificar a qualidade de filmes a partir de sua legenda. Além disso, foi utilizada

a plataforma *WEKA*, que possui uma série de algoritmos relacionados à aprendizagem de máquina, para mineração de texto com o intuito de distinguir se os filmes são bons ou ruins [12].

3.2 Desafios

Durante a produção desse trabalho uma vasta gama de artigos foi objeto de estudo e com isso se tornaram fontes de inspiração e desenvolvimento do projeto, porém, esse foi um momento de observação devido à existência de poucos trabalhos relacionando criptomoedas e análise de sentimentos. Isso se deve pela complexidade de trabalhar com *tweets*, uma rica fonte de dados, mas repleta de peculiaridades linguísticas, limitações de caracteres e na captação dos *tweets*, entre outros. Apesar de todas as dificuldades, é uma ferramenta de estudo muito válida e com a evolução das pesquisas na área será possível buscar resultados ainda mais expressivos na predição de criptomoedas. Além disso, a elaboração do algoritmo dessa pesquisa pode se tornar uma ferramenta para investimento na bolsa de valores, principalmente no mercado de *Bitcoin*, um diferencial ou uma alternativa para os principais instrumentos de aplicações financeiras, como: sites de informações, oscilação na bolsa, histórico da moeda e indicadores estatísticos.

Capítulo 4

Método Proposto

O Capítulo 4, tem como objetivo detalhar toda a metodologia executada no projeto e suas devidas análises e alguns detalhes da implementação realizada na linguagem de programação *Python*. A Seção 4.1 aborda sobre o conjunto de textos extraído com suas respectivas parametrizações e as ferramentas utilizadas para captação e classificação. A Seção 4.2 refere-se às técnicas de pré-processamento empregadas e suas categorizações. A Seção 4.3 descreve a estratégia de classificação das mensagens de acordo com seu sentimento. A Seção 4.4 se baseia na frequência de tópicos e detalha essa caracterização através do algoritmo *Latent Dirichlet Allocation*. A Seção 4.5 descreve agrupamentos e identificação de padrões. A Seção 4.6 descreve por meio de análise visual um estudo comparativo entre os *tweets* sobre *bitcoins* e as variações dessa criptomoeda no mercado financeiro.

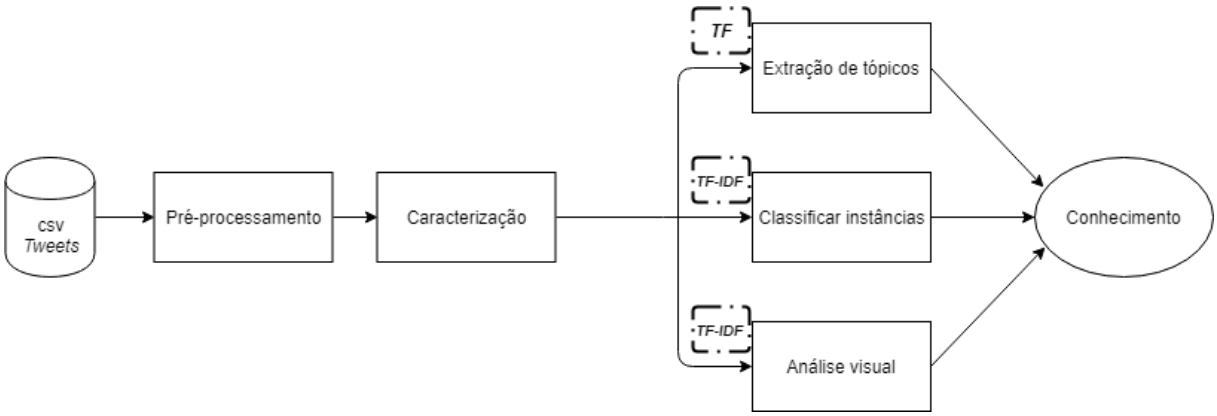

Figura 4.1: Fluxo de atividades do projeto. Fonte: própria (2019).

4.1 Conjuntos de textos

Conseguir um amplo e qualificado conjunto de textos é uma tarefa difícil para validar técnicas de mineração de textos. Inicialmente foi programado o desenvolvimento de um algoritmo em *Python* em conjunto com a *API* fornecida pelo *Twitter* para obtenção dos textos em tempo real, porém a *API* apresentou limitações por ser uma versão *Standard*. Isso custou diversas reconfigurações dos métodos do próprio *Twitter*, como a alteração na quantidade de caracteres, 140 para 280 e que o número de *tweets* a serem retornados por página são de no máximo 100.

Vale ressaltar que um *tweet* é composto por atributos nomeados de metadados como: autor do *tweet* (*userID*), contador de *retweets*, geolocalização, horário de publicação, linguagem, *hashtags* e o texto. Esta pesquisa faz uso apenas do texto dos *tweets* para classificação das polaridades, os outros elementos são descartados. Vale destacar o atributo *hashtag* é um elemento utilizado na construção dos *rankings* de assuntos mais comentados (*tranding topics*), do twitter, e pode conter informações semânticas do texto.

O primeiro conjunto de *tweets* utilizado neste trabalho foi obtido de duas formas distintas, a primeira delas contendo aproximadamente 50 mil[1] instâncias (*tweets*) rotuladas por especialistas humanos em positivo, negativo e neutro. Esses *tweets* foram obtidos no dia 28/03/2018, entretanto mesmo sendo um conjunto farto, esse fato restringiria as análises a um único dia, o que motivou a busca por uma nova fonte de *tweets*. Deste conjunto de dados foram consideradas a informação textual dos *tweets* e sua categoria de sentimento.

Para complementar a confiabilidade da análise, alguns *tweets* foram coletados utilizando uma ferramenta online[2], que consiste de um algoritmo que permite a busca e *download* dos 100 *tweets* mais recentes por requisição. A partir disso foi desenvolvido um *script* na linguagem *Python* que automatiza esse processo, que realiza requisições a cada 2 minutos no site, obtendo assim um conjunto de cerca de 500 mil *tweets*, referentes a um período 25/03/2019 até 09/04/2019. Esses *tweets* foram classificados por um modelo de classificação treinado a partir do primeiro *dataset* de 50 mil instâncias. Tal fato viabilizou aprofundar as análises dos *tweets* e assim identificar, possíveis, padrões existentes nos textos.

4.2 Pré-processamento e Caracterização de Texto

Dentre as diversas técnicas de pré-processamento de texto existentes na literatura, foram aplicadas no trabalho:

[1]Endereço do conjunto de textos com 50mil *tweets*: *https://data.world/mercal/btc-tweets-sentiment/*.
[2]Endereço para a captação de *tweets*: *https://twitter-sentiment-csv.herokuapp.com*.

Remoção de *Stopwords*: *Stopwords* (palavras vazias) importantes para o entendimento de frases, mas a nível computacional ocupa espaço em memória e aumenta o tempo de processamento. Por isso, é interessante remover esses caracteres e palavras para a construção do *corpus*, conjunto estruturado de termos. O processo de remoção de *stopwords* é feito para língua inglesa em cada *tweet*.

Remoção de URL: As *URLs* não tem papel importante na identificação da polaridade (sentimento) dos *tweets*, logo se tornam desnecessárias para os processamentos posteriores.

Stemização: Consiste no processo de reduzir a palavra ao seu radical diminuindo assim o tamanho do vetor resultante obtido após a aplicação do *bag of words*, ou seja, o tamanho do vocabulário a ser usado. Isso resulta em ganhos de performance e economia de espaço em memória.

Remoção de espaço extra: Como o público que usa o *Twitter* tem uma linguagem que foge à norma culta, é muito comum encontrar *tweets* com múltiplos espaçamentos e isso também é tratado no algoritmo proposto.

A figura 4.2 mostra as etapas realizadas na metodologia desse trabalho para a realização do pré-processamento dos textos.

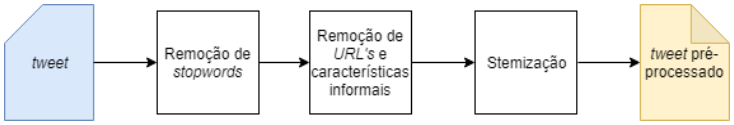

Figura 4.2: Etapas do pré-processamento do *tweet*. Fonte: própria (2019).

Nas figuras 4.3 e 4.4 são mostrados exemplos de mensagens do *Twitter*, a primeira é um *tweet* original e a segunda é a mesma mensagem, após o pré-processamento com o uso das técnicas de remoção de *stopwords*, remoção de URL's, remoção de espaços extras e stemização.

Figura 4.3: Exemplo de *tweet*. Fonte: https://twitter.com/breninhorf.

Figura 4.4: Exemplo de *tweet*, após o pré-processamento. Fonte: própria (2019).

4.3 Classificação: previsão de sentimentos

A escolha do classificador mais adequado à tarefa de análise de sentimentos nos *tweets* é de suma importância para essa pesquisa, pois o classificador escolhido deve ser capaz de rotular, automaticamente, novas instâncias do conjunto de textos relacionando-os com o conteúdo treinado e garantir a melhor taxa possível na assertividade dos rótulos, de acordo com seu sentimento [46]. Na busca pelo classificador mais eficiente para análise do conjunto de textos, em cada *tweet* coletado, foi implementado para cada um dos classificadores um algoritmo *GRIDSearchCV* [75]. Esse tipo de algoritmo efetua diversas simulações, variando nos parâmetros de cada classificador em intervalos pré definidos (Ex.: C: 1-100). O *GRIDSearchCV* possibilita testar todas as combinações possíveis e/ou desejáveis, tem o objetivo claro de encontrar os melhores parâmetros de cada classificador. Essa "calibragem" proposta pelo algoritmo permite o aprimoramento dos resultados de cada classificador e em seguida a partir de um comparativo de resultados, identificar a melhor ferramenta para então dar continuidade ao projeto, por meio do classificador que obteve o melhor *F1-score*.

4.4 Extração de Tópicos

Como tentativa de obter definições concretas sobre o que efetivamente motivou a variação dos valores diários do *bitcoin*, foi usado o *Latent Dirichlet Allocation* (*LDA*), ferramenta que permite a extrair tópicos relevantes dos documentos. O objetivo é vincular uma variação, positiva ou negativa, a um tópico específico e seu conjunto de palavras relacionadas que, em tese, motivaram aquela reação no gráfico.

Para realização do *LDA*, empregando a biblioteca *Scikit Learn*, foi utilizada uma *bag-of-words*, contabilizado as palavras para aplicação do modelo probabilístico, estabelecido os paramêtros de modo de aprendizado online, para analisar as coleções de documentos, incluindo aquelas que chegam em um fluxo. Também implementado o número de tópicos

10, padrão da biblioteca e 30, baseado em teste realizado na literatura [76]. Definido o número de termos gerados igual a 30 para análise de uma possível relevância dos termos na cotação do *bitcoin*.

Posteriormente a identificação dos tópicos relevantes, é realizada uma comparação com os principais termos presentes nos grupos obtidos pela técnica de agrupamento (*k-means*), para avaliar se existe relação entre ambas.

4.5 Agrupamento: identificação de padrões

A técnica de agrupamento foi usada com o intuito de aglomerar termos semelhantes que possam ser relevantes para predição do valor do *bitcoin*. O algoritmo *k-means* realiza a segmentação dos textos, e também, mostra os textos mais próximos e expressivos através do cálculo da sua distância em relação aos centroides, o k em *k-means*, locais que representam o centro do agrupamento. Aliado a isso foi aplicado o algoritmo *elbow curve*, técnica que auxilia a obtenção da quantidade ideal de grupos presentes no conjunto de textos [77].

A primeira etapa para a realização dos agrupamentos foi a utilização do algoritmo *elbow curve*, que é uma técnica proposta para determinar o valor de k, onde varia número de centroides gradativamente, até não obter mudança significativa na variância, encontrando um número ideal. A segunda etapa ocorre para reduzir a dimensionalidade dos textos, empregando a decomposição de valores singulares, *TruncatedSVD*, com o objetivo de reduzir o número de linhas, e preservar a similaridade entre as colunas. Esse método é oportuno para gerar um novo espaço vetorial com uma possível melhora na qualidade dos resultados na continuidade do processo de mineração de textos.

A última etapa consiste na aplicação do algoritmo *k-means*, onde a primeira iteração é calcular a distância média de todos os pontos que estão atrelados ao centroide e então mudar a posição dos centroides até encontrar palavras bem agrupadas [78].

4.6 Análise visual

Depois de cumpridas as etapas de obtenção do conjunto de textos, inicia-se a análise dos *tweets* obtidos por meio de análise visual de textos.

Diariamente são registradas todas as movimentações relacionadas às criptomoedas por sites especializados e, muitas delas, representadas através de gráficos. Isso possibilita o monitoramento das flutuações da moeda em diversos intervalos de tempo. Neste projeto foram usados intervalos de 24h, desde a abertura do mercado, 00h01min, até seu fechamento, 00h, nas datas referentes à captura dos textos.

As análises foram feitas de forma a estabelecer um paralelo entre quantidade total de *tweets*, referente a cada polaridade; o movimento, do valor da moeda. Aliado a isso, os textos foram agrupados, e posteriormente, foram extraídos os principais tópicos presentes nos conjuntos de textos, possibilitando assim a observação dos eventuais causadores de variações no mercado.

Os textos selecionados foram separados, levando em consideração suas respectivas datas, em diferentes grupos de análise: diários, semanais e mensais. Cada grupo é segregado de forma a separar a quantidade original de textos, em *tweets*, de acordo com a sua polaridade produz-se o primeiro conjunto de gráficos de barras, determinando assim o volume original de textos positivos, negativos e neutros. Em seguida, o valor do *bitcoin* é coletado na abertura e fechamento[3], do mercado financeiro, com intuito de identificar a variação, ascendente, descendente ou estática da moeda, em um determinado período de tempo. E assim, é finalizada a primeira análise, cuja comparação é realizada entre o volume de *tweets* e o valor da moeda.

A segunda etapa consiste em identificar e relacionar elementos que possam ser responsabilizados pela alteração do valor de mercado da criptomoeda, por exemplo: Atualmente, 25/05/2019, a aprovação, ou não, do projeto e emenda constitucional (PEC), da previdência social, no Brasil. Esse é um assunto que tornou a economia nacional, absolutamente, volátil, e toda notícia, boa ou ruim, relacionada a esse assunto tem efeitos diretos no mercado financeiro. Ou seja, a presença do tópico "PREVIDÊNCIA" aliado a um volume muito grande de textos, polarizados como positivo, deveria ser relacionado a um crescimento de mercado na bolsa de valores brasileira. Por fim, são definidos grupos de instâncias a que pertencem os tópicos, a fim de dar uma maior amplitude a identificação desses potenciais agentes de mutação do mercado, ou seja, ampliar o vocabulário de palavras que possam vir a causar a variação do valor da moeda.

[3]Endereço de coleta das informações de valor de mercado do *bitcoin*: *https://br.tradingview.com/symbols/BTCBRL*.

Capítulo 5

Resultados Experimentais

No decorrer deste capítulo são apresentados os resultados obtidos por intermédio de experimentos. A Seção 5.1 faz referência aos classificadores e técnicas aplicadas, suas diversas parametrizações e discorrendo sobre o porquê de suas escolhas. A Seção 5.2 exibe os resultados encontrados de cada classificador empregado traçando o nexo causal com os sentimentos capturados. A Seção 5.3 mostra através da técnica *LDA* os termos mais relevantes encontrados na nossa base de textos. A Seção 5.4 discorre sobre a técnica *K-means*, descrevendo os grupos, os centroides e os padrões encontrados nos agrupamentos e os padrões internos aos grupos. A Seção 5.5 faz uma breve análise sobre as implementações e enaltece as vantagens do método proposto com base nos resultados obtidos.

5.1 Experimentação dos classificadores e *K-Means*

Cada parâmetro presente em um classificador foi obtido após a execução do algoritmo *GRIDSearchCV*. O objetivo é efetuar testes em sequência com múltiplos valores a fim de buscar a melhor calibragem possível para o classificador dentro de um conjunto de possíveis parâmetros pré-determinados. Foi usado o conjunto de texto do dia 28/03/2018 para a realização dos experimentos de *GRIDSearchCV*.

Na sequência serão expostos os resultados que definiram a escolha do classificador do projeto e sua respectiva calibragem.

5.1.1 Máquinas de vetores de suporte - *SVM*

Para a aplicação do *GRIDSearchCV* no *SVM* (implementação em *python* usada da biblioteca *sklearn*[50]) foram usadas duas funções *kernel*, não linear (*Radial basis function - RBF* ou *Kernel Gaussiano*) e linear, a fim de comparar metodologias opostas e, assim, encontrar a melhor calibragem de parâmetros para o conjunto de textos. Ambos os testes

tiveram em comum a variação de C (parâmetro de penalidade do termo de erro) entre 0 e 100, porém diferentemente do *kernel* linear, o *RBF* teve o parâmetro γ (*gamma*) variando entre 0,001 e 0,0001.

Em um dos conjuntos de textos, uma série de valores são testados em diferentes parâmetros, com o uso do *GridSearchCV* para o *kernel* linear, que viabilizou a melhor configuração para precisão e, *recall*. A partir das configurações encontradas, a calibragem que obteve melhor média *F1-score* foi selecionada.

A Figura 5.1 mostra a variação dos resultados de classificação do *SVM* aplicado com *kernel* linear e o hiperparâmetro C obtendo valores de 0 a 100.

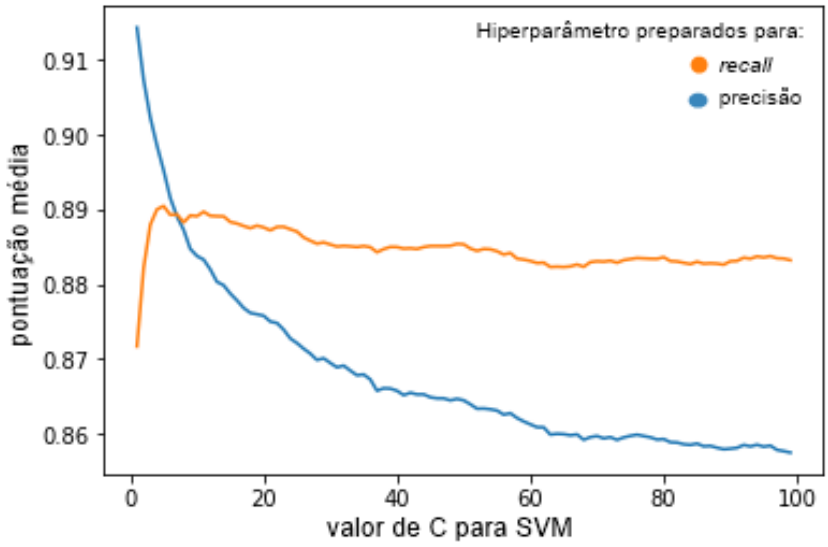

Figura 5.1: Variação do parâmetro C.

A Tabela 5.1 mostra a melhor média *F1-score*, obtidas dentre a variação mostrada na Figura 5.1. O parâmetro, $C = 1$, foi o que obteve os melhores resultado quando aplicado junto ao *kernel* linear.

Tabela 5.1: *SVM - Kernel: Linear,* $C = 1$.

	precision	***recall***	***f1-score***
negativo	0.91	0.74	0.81
neutro	0.88	0.97	0.92
positivo	0.96	0.91	0.93

O *kernel RBF* possui também o atributo γ a ser variado, com isso foi definido uma lista que contém valores de 0.005 até 0.01, a serem atribuídos ao γ, com espaços de 0.05 entre cada valor de teste, durante a realização do *GRIDSearchCV*, a fim de obter o melhor resultado do atributo γ [79]. A Figura 5.2 mostra os resultados de pontuação média em função da variação do hiperparâmetro, γ.

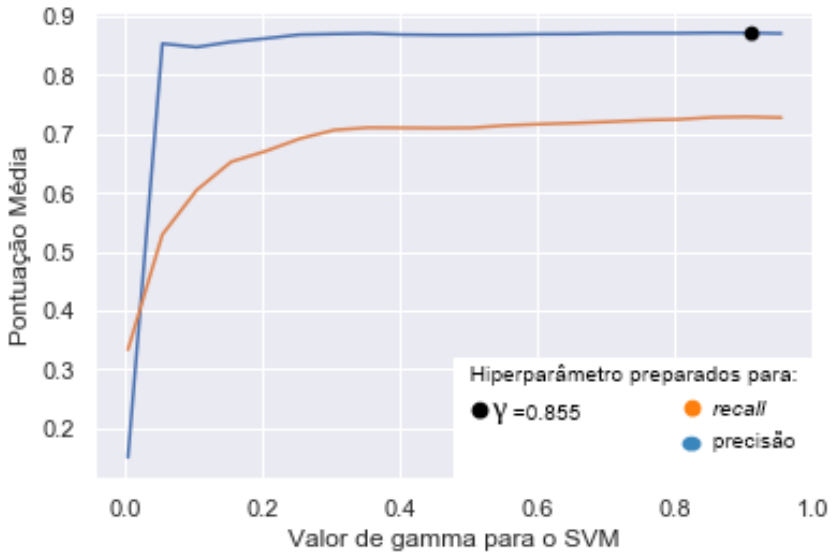

Figura 5.2: Gráfico da variação de γ: 0 até 1.0.

A Tabela 5.2 mostra que $\gamma = 0.855$, obteve a melhor média *F1-Score*.

A partir da obtenção do γ, considerado ideal para o conjunto de textos analisados neste trabalho, o objetivo passa a ser encontrar o melhor valor para o atributo C, agora para o *kernel RBF*. O algoritmo *GRIDSearchCV* é aplicado para variar o C em valores de 0 a 100, a fim de ter uma boa margem de análise.

A Tabela 5.3 mostra a melhor avaliação da classificação realizada com o *SVM*, com a utilização do *Kernel RBF* (*Radial basis function*). O parâmetro, $C = 14$, foi identificado como melhor valor para uso conjunto com o *kernel RBF* e $\gamma = 0.855$.

Tabela 5.2: *Melhor resultado para $\gamma = 0.855$.*

	precision	***recall***	***f1-score***
negativo	1.00	0.47	0.64
neutro	0.73	0.94	0.82
positivo	0.91	0.77	0.84

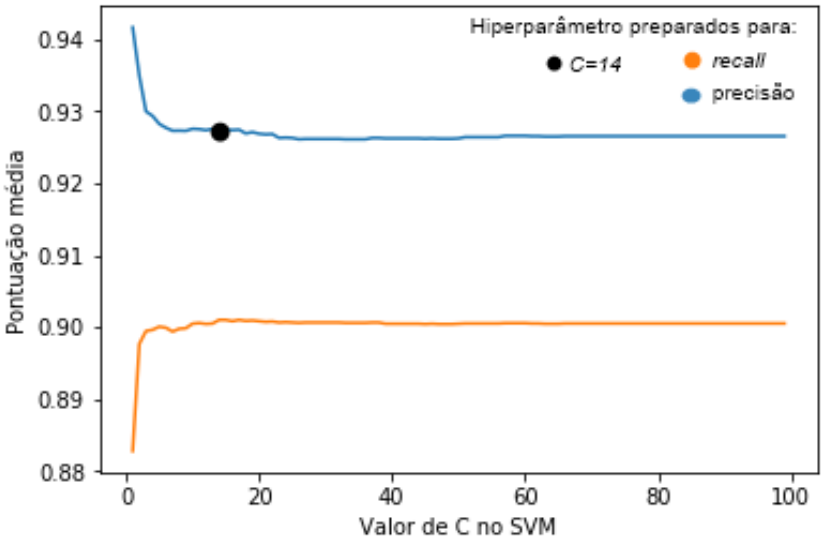

Figura 5.3: Variação do C: 0-100 e $\gamma = 0.855$.

Tabela 5.3: *SVM - Kernel: RBF*, $C = 14$, $\gamma = 0.855$.

	precision	***recall***	***f1-score***
negativo	0.93	0.81	0.87
neutro	0.91	0.96	0.94
positivo	0.95	0.93	0.94

5.1.2 *Multinomial Naive Bayes - MNB*

Por sua simplicidade, o processo de classificação do *MNB* foi executado sem a aplicação do algoritmo *GRIDSearchCV*, e retornou os resultados de classificação descritos na Tabela 5.4.

5.1.3 Redes Neurais

***Perceptron* Simples**

O *perceptron* simples, é outro modelo que não exige a aplicação do *GRIDSearchCV*. Os resultados de classificação do método estão na Tabela 5.5.

Tabela 5.4: *Multinomial Naive Bayes*.

	precision	***recall***	***f1-score***
negativo	0.85	0.45	0.58
neutro	0.83	0.91	0.87
positivo	0.86	0.84	0.85

Tabela 5.5: *Percetron* Simples.

	precision	***recall***	***f1-score***
negativo	0.72	0.76	0.74
neutro	0.92	0.92	0.92
positivo	0.93	0.92	0.93

Tabela 5.6: *Redes Neurais Multicamadas.*

	precision	***recall***	***f1-score***
negativo	0.00	0.00	0.00
neutro	0.74	0.90	0.82
positivo	0.77	0.79	0.78

Perceptron* multicamadas - *MLP

Ao contrário do modelo *perceptron* simples, o multicamadas possui diversos parâmetros a serem calibrados, sendo o *GRIDSearchCV* importante para a obtenção da melhor calibragem desse modelo. Os atributos testados foram: *momentum*, *activation* (função de ativação), *learning rate* (taxa de aprendizado), *hidden layer sizes* (número de neurônios nas camadas ocultas), *solver* (solução para otimização de peso). Foram testadas todas as combinações possíveis dos seguintes conjuntos de atributos: função de ativação, *logistic e tanh*; taxa de aprendizado, *invscaling, adaptive e constant*; solução para otimização de peso, *sgd, adam e lbfgs*; *momentum* (quando o *solver* for *sgd*), 0.1 ao 0.9. Também foi variada a quantidade de neurônios presentes em cada camada, inicialmente com tuplas de 1 a 100 neurônios na primeira camada e a segunda camada vazia (1 a 100,), a tupla padrão do classificador no *Sklearn* é (100,). Após esse teste foi observado que a variação dos resultados entre camadas com um neurônio de diferença, por exemplo: (2,) e (3,), é irrelevante. Portanto foram definidos espaços de 10 neurônios para a sequência dos testes do modelo com duas camadas. Foi seguido um padrão de tuplas, $(x, \frac{x}{2})$, por exemplo: $(10, 5), (20, 10)$ até $(100, 50)$.

Desta maneira, foram obtidos os melhores parâmetros para o conjunto de textos do projeto: *activation: tanh*, *hidden layer sizes*: (20,10), *learning rate: invscaling*, *solver: lbfgs*. Os resultados de classificação do modelo *MLP*, são descritos na Tabela 5.6.

5.2 Resultados da classificação

Nesta Seção 5.2 são mostrados os resultados do processo de classificação dos *tweets* referentes a períodos distintos de tempo. Juntamente ao valor corrente do *Bitcoin*, no

mercado, extraídos do site *Trending View*[1] no mesmo período, seguindo como referência o fuso horário UTC±00:00. Objetivando identificar indícios de que, quando as mensagens são em sua maioria positivas, a moeda tende a valorizar. Os resultados serão mostrados nas Figuras 5.4 a 5.7.

5.2.1 Análise Diária

A Figura 5.4 apresenta uma comparação dos sentimentos dos *tweets* em relação à cotação do *Bitcoin* no dia 28/03/2018. Ao analisar a Figura 5.4 foi observado que o dia em questão foi estável em relação ao dia anterior, e o valor da moeda se manteve em R$ 27.000,00. Em paralelo a essa análise, os sentimentos coletados dos *tweets* desse dia se mostraram equilibrados entre positivos e neutros, com cerca de 45% de mensagens classificadas como positivas, 43% neutras e 12% negativas, em 50.864 *tweets*.

Vale resaltar que mesmo com a predominância de *tweets* com polaridades positivas no dia em questão, o valor do *bitcoin* não obteve crescimento, superioridade das polaridades está relacionada a variação da cotação da criptomoeda ao longo do dia, que obteve vários momentos de inflexão.

[1]Endereço de coleta das informações de valor de mercado do *bitcoin*: *https://br.tradingview.com/symbols/BTCBRL*.

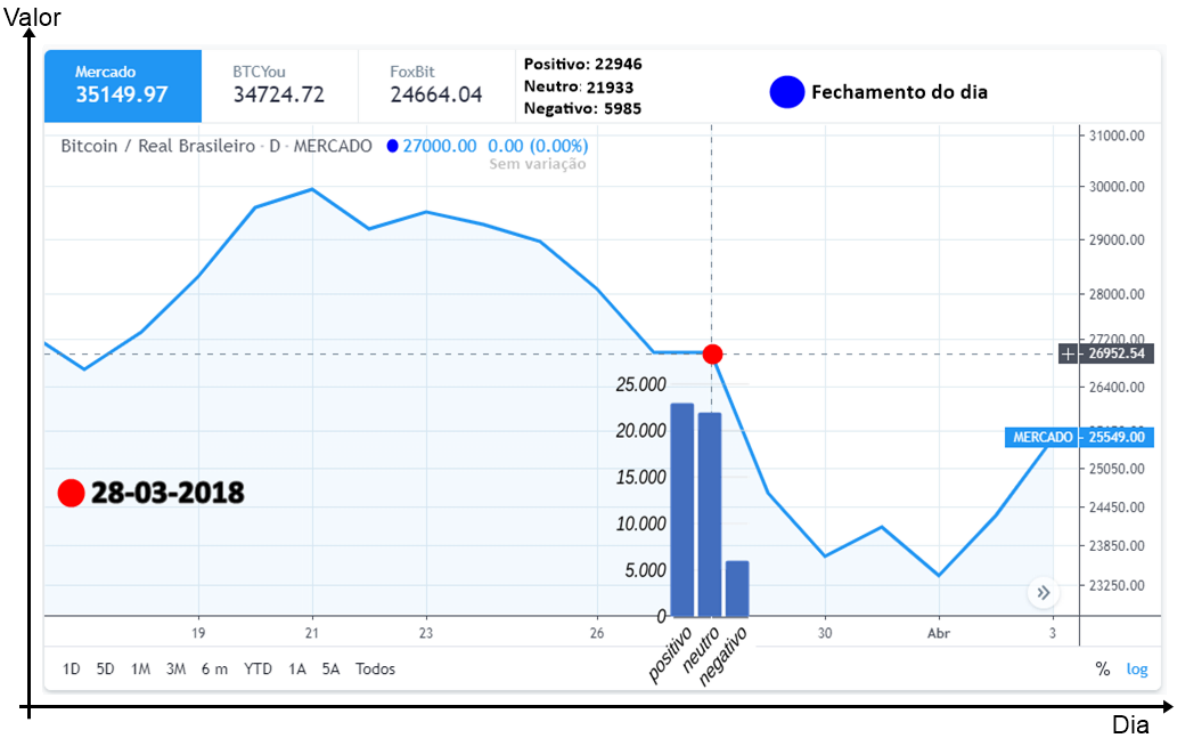

Figura 5.4: Volume de sentimentos em relação ao valor corrente do *bitcoin*, no dia 28/03/2018.

5.2.2 Análise Semanal

Na análise semanal, 25/03/2019 a 31/03/2019, representado na Figura 5.5, houve uma ascensão no valor do *Bitcoin*, de aproximadamente 7%. Tal crescimento que fez com que a moeda chegasse ao valor de R$ 16.493,00 no final do dia 31/03/2019. As polaridades dos *tweets* coletados na semana, assim como na análise diária, mostrou um equilíbrio entre positivos e neutros, com cerca de 40% de mensagens classificadas como positivas, 48% neutras e 12% negativas, em 19.2040 *tweets*.

A próxima análise semanal, 01/04/2019 a 07/04/2019, ilustrada na Figura 5.6. Período evidenciado por uma notável subida no valor da moeda, de R$ 16.480,00 para R$ 20.202,00, aproximadamente 23%. Na observação das polaridades, obteve-se 11.624 *tweets*, onde 36% foram classificados como positivas, 53% neutras e 11% negativas.

Nesse momento ainda não é possível identificar os fatores responsáveis pelo movimento do valor de mercado do *bitcoin*, porém nas etapas de extração de tópicos relevantes (*LDA*) e dos principais termos por grupo (*K-means*), será possível identificar os fatores responsáveis por essas variações.

Figura 5.5: Volume de sentimentos em relação ao valor corrente do *bitcoin*, entre os dias 25/03/2019 e 31/03/2019.

5.2.3 Análise março ∼ abril

Assim como nas análises semanais, a análise do período março ∼ abril, 25/03/2019 a 09/04/2019, aconteceu em um momento de crescimento do valor da moeda, explícito no aumento percentual de quase 33%, com o valor de fechamento, 09/04/2019, R$ 20.498,00.

A análise das polaridades, assim como nas outras análises, indicou superioridade nos números de mensagens classificadas como positivas e neutras em relação às negativas, cerca de 39% positivas, 50% neutras e 11% negativas, em 510882 *tweets*. Informações descritas estão presentes na Figura 5.7.

5.3 Extração de tópicos

A aplicação do modelo probabilístico proporciona extrair termos relevantes em conjuntos textuais. Em vista disso, o modelo foi aplicado em *tweets* diários, semanais e no período de março a abril de 2019, com o intuito de enriquecer a análise dos resultados do processo de classificação de *tweets*. Isso é feito ao relacionar os tópicos extraídos ao conjunto de polaridades obtidas no processo de classificação dos *tweets*.

Figura 5.6: Volume de sentimentos em relação ao valor corrente do *bitcoin*, entre os dias 01/04/2019 e 07/04/2019.

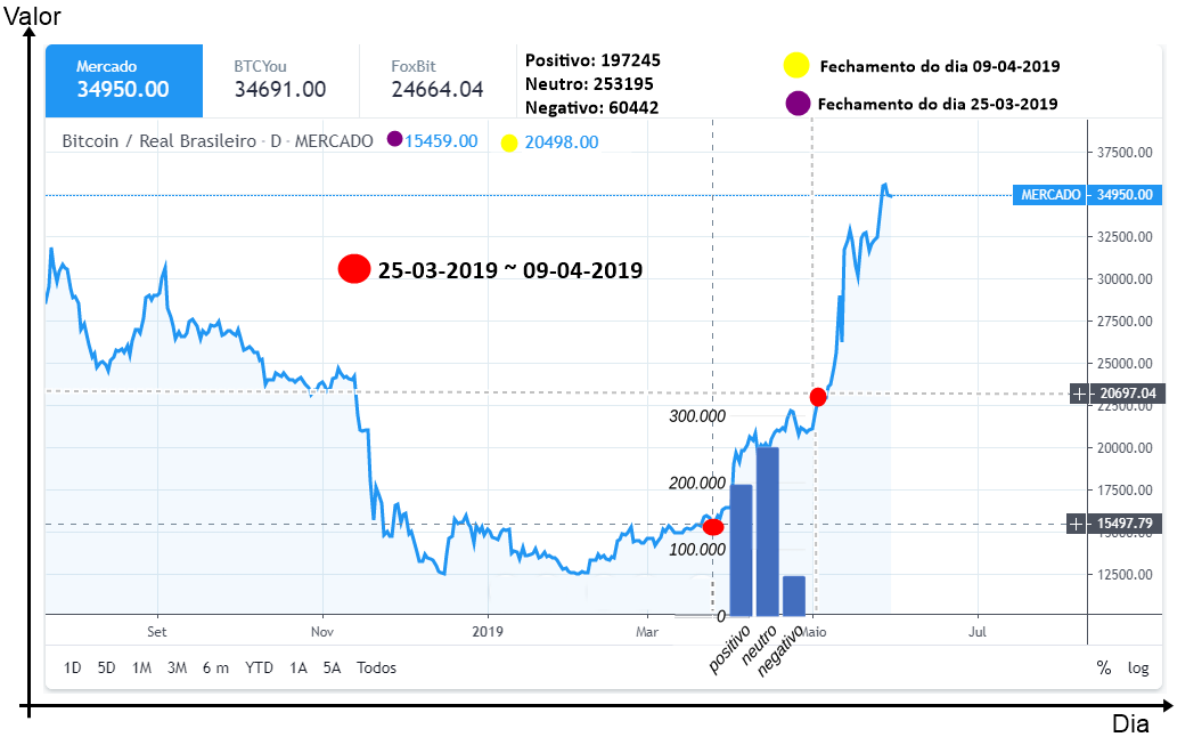

Figura 5.7: Visualização do volume de sentimentos coletados nos textos e da variação do valor corrente da moeda, março ∼ abril.

5.3.1 Análise Diária

Na Figura 5.8, são mostrados os 30 termos mais relevantes do dia 28/03/2018, onde palavras como "*bitcoin*", criptomoeda; "*blockchain*", tecnologia base de funcionamento do *bitcoin*; "*airdrop*", presentes concedidos a usuários da criptomoeda que realizaram tarefas ou simplesmente se cadastraram na lista de bonificações oficial da criptomoeda. Tais palavras possuem absoluta relação com as criptomoedas, explicando a maior frequência dessas palavras nos *tweets*.

Figura 5.8: Termos mais relevantes coletados no dia 28/03/2018.

5.3.2 Análise Semanal

As análises são divididas em semanas 1 e 2, respectivamente. Na Figura 5.9, são mostrados os 30 termos mais relevantes da semana 1, 25/04/2019 a 31/03/2019. As palavras "*btc*", abreviação para *bitcoin* e "*blockchain*", "*airdrop*". Mostram uma repetição dos três termos mais frequência na análise diária. Porém, é possível observar a forte presença de novas criptomoedas, como: *Ethereum*, *Cardano*, *Stellar* e *Litecoin*. Além disso, pode-se notar

a presença do termo "*free*", que pode ter relação com o termo "*airdrop*", e indicar uma busca por bonificações dentro das criptomoedas mencionadas.

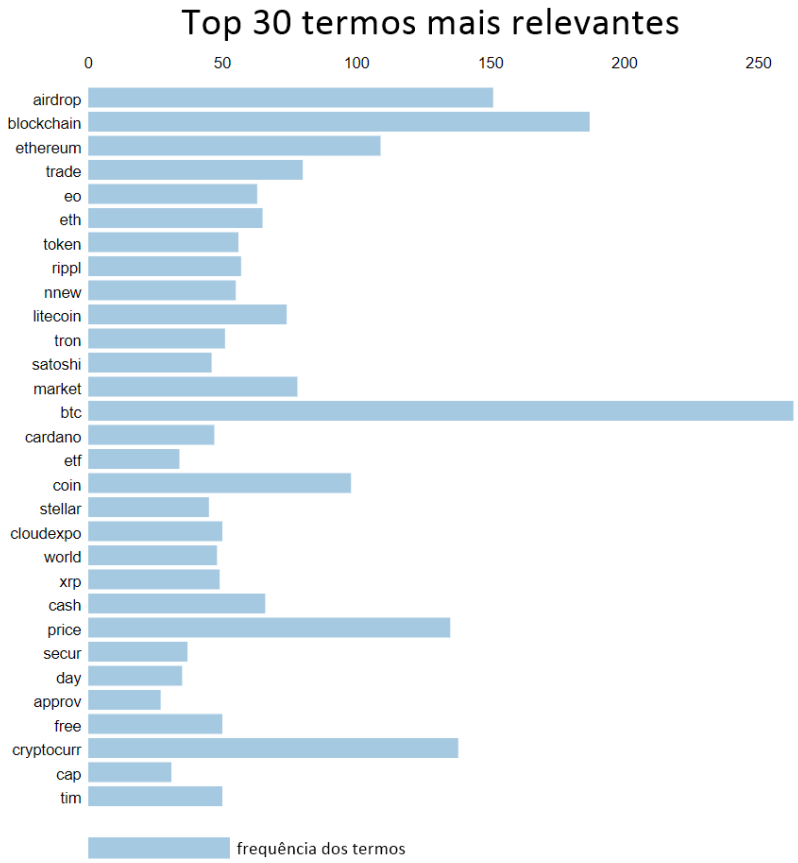

Figura 5.9: Termos mais relevantes coletados entre os dias 25/04/2019 e 31/03/2019.

A Figura 5.10, ilustra os 30 termos mais relevantes da semana 2 (01/04/2019 a 07/04/2019). Diferentemente da semana 1, quatro palavras foram mais frequentes: "*btc*"; "*leverag*", tem como significado alavancagem; "*primexbt*", corretora de criptomoedas; "*crypto*", termo que pode ser uma criptomoeda, carteira de investimento, entre outros. Foi observada também, a presença dos termos "*leverag*", "*buy*" (comprar) e "*price*" (preço), vão de encontro com o momento de grande crescimento do valor da moeda vivenciado no mesmo período, evidenciado na análise da Figura 5.6.

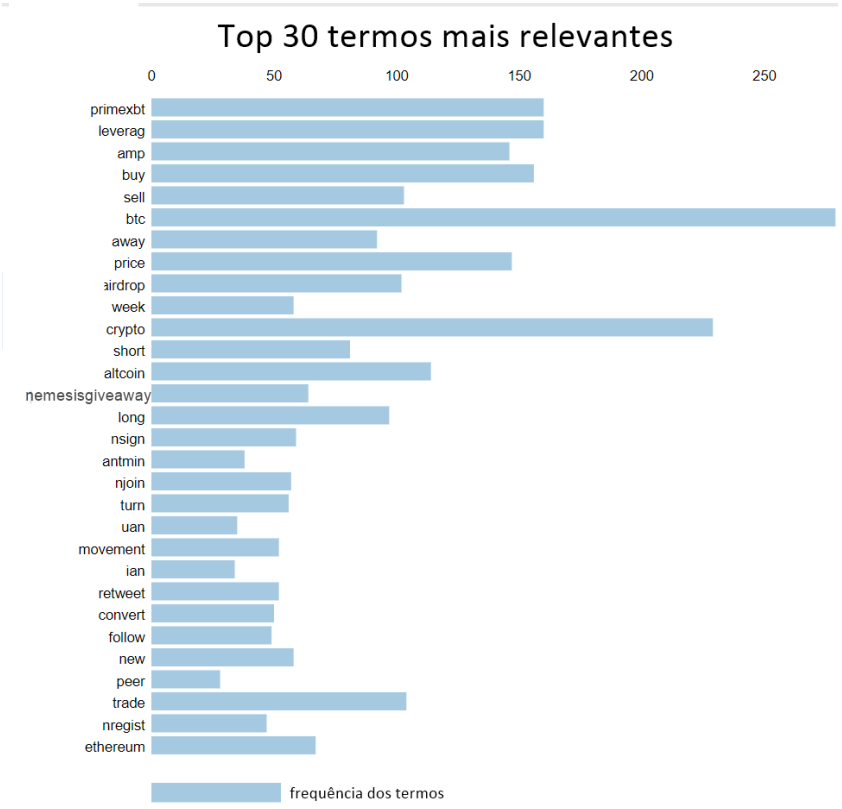

Figura 5.10: Termos mais relevantes coletados entre os dias 01/04/2019 e 07/04/2019.

5.3.3 Análise Março ∼ Abril

Na Figura 5.11, são apresentados os 30 termos mais relevantes entre um período específico de março e abril, 25/03/2019 a 09/04/2019. O maior período, contínuo, de captação de textos para a pesquisa. Como já esperado, os termos mais frequentes foram os que fazem alusão à moeda, "*bitocoin*", "*btc*", e "*crypto*". Conforme as outras análises, alguns termos podem ser representativos com a movimentação da moeda no mesmo período, como: "*buy*", "*bonus*", "*trade*" (comércio).

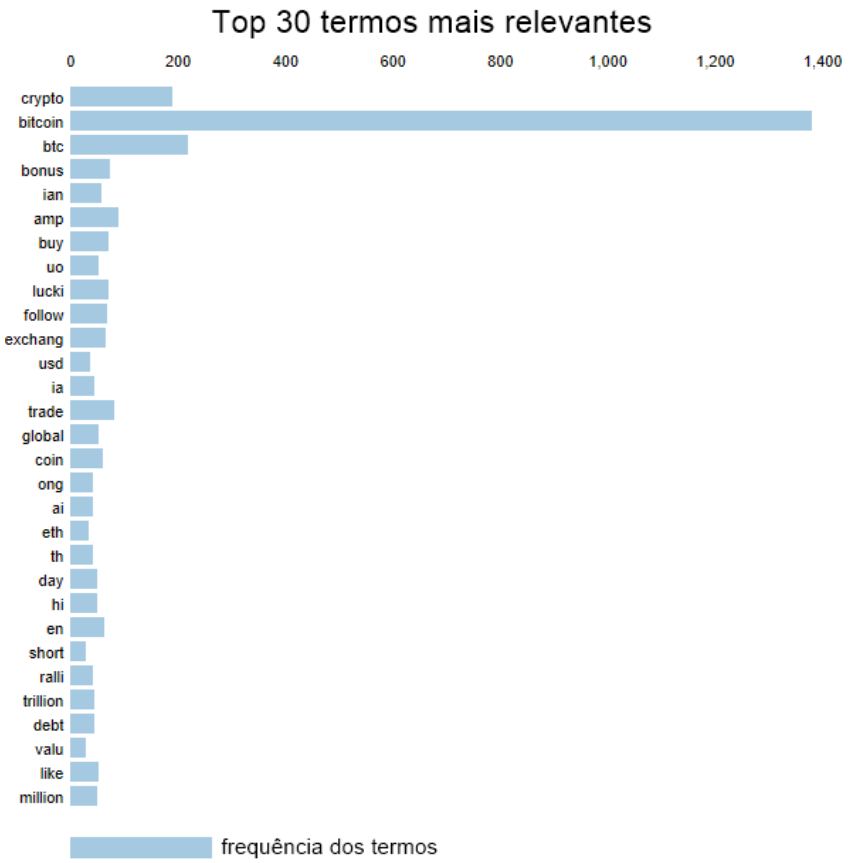

Figura 5.11: Termos mais relevantes coletados no intervalo mensal de Março para Abril.

5.4 Agrupamento (*k-means*)

As análises de agrupamento foram aplicadas nos textos referentes aos seus respectivos dias, sem a extensão de períodos de tempo como nas outras técnicas. Isso ocorre devido a capacidade, reduzida, de processamento dos computadores que executaram as técnicas.

5.4.1 Curva cotovelo

A técnica curva cotovelo (*elbow curve*) foi aplicada a fim de encontrar o número ideal de centroides para o conjunto de textos analisados. A Figura 5.12, ilustra um gráfico com a presença de "cotovelos", descrevendo possíveis valores para o número total de centroides, k, são eles: 4, 5 e 8. Porém, ao observar baixa diferença do valor $k = 3$ em relação ao $k = 4$, foi aplicado o agrupamento também para $k = 3$, a fim de observar relações com os sentimentos presentes nos textos, ou seja, tentar formar grupos que tenham características positivas, negativas ou neutras.

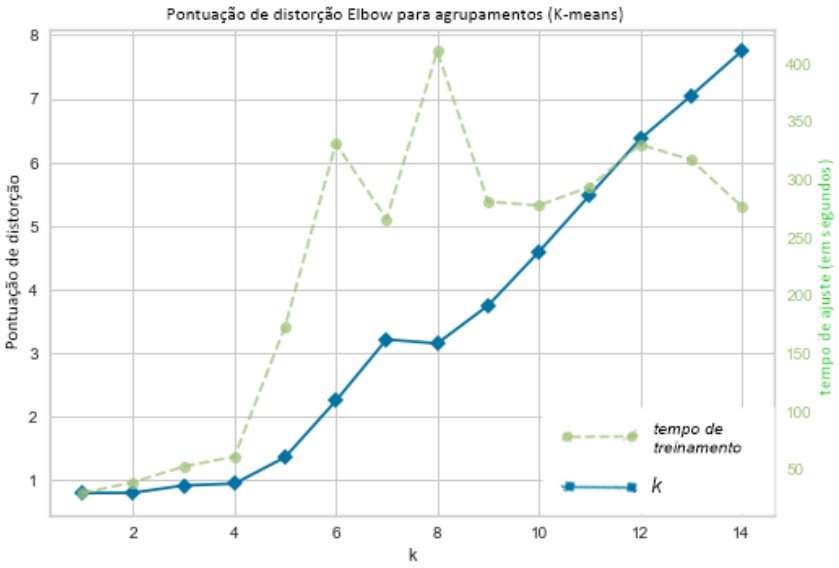

Figura 5.12: Resultado da curva cotovelo.

5.4.2 Análise de agrupamento

As Figuras 5.13 a 5.15, mostram exemplos selecionados para demonstrar as diferentes aplicações dos centroides indicados pela técnica *elbow curve*. Os dados desse dia específico foram utilizados por evidenciarem grupos mais bem definidos para visualização.

Foi utilizada a biblioteca *sklearn* para gerar os gráficos dos agrupamentos, e dentre os vários parâmetros, foi definido o *k-means++*, método que favorece a convergência dos centroides gerados. O *max_iter*, quantidade máxima de vezes de execução do algoritmo, limitado à até 300 iterações (valor padrão). O número de grupos e centroides não foram definidos, porém nas Figuras 5.13 a 5.15 serão mostrados experimentos que realizam a variação dos mesmos para qualificar a análise. Também realizada a utilização de técnicas de redução de dimensionalidade, *LSA* e *TruncatedSVD*, para os valores obtidos através do *TF-IDF* e, assim, gerar os gráficos em duas dimensões.

Para melhor entendimento dos gráficos de agrupamentos, os pontos vermelhos representam os centroides, e os demais pontos são as *features*, onde cada cor ilustra um grupo.

A visualização dos gráficos permitirá inferir que quando as *features* possuem uma distância pequena entre si, nos grupos que são pertencentes, os textos possuem uma grande taxa de similaridade, em contraposição, quando a distância é grande, maior a taxa de dissimilaridade. E quando há uma distância pequena entre grupos, os textos podem possuir um grau de pertinência associado a mais de um grupo.

O particionamento em três grupos ($k = 3$) foi executado com objetivo de obter grupos que possam se relacionar com as polaridades usadas no trabalho. A Figura 5.13 mostra que o agrupamento é bem definido, e permite a identificação dos grupos.

A determinação desse mapeamento das instâncias do espaço de alta dimensionalidade para os pontos no espaço bidimensional (que define o *layout*) foi obtida pelo uso da técnica *Truncated SVD*. Em linhas gerais, a técnica de visualização produziu uma imagem a partir do espaço bidimensional resultante da redução de dimensionalidade do espaço de características definido pelos vetores *TF-IDF* da coleção de textos.

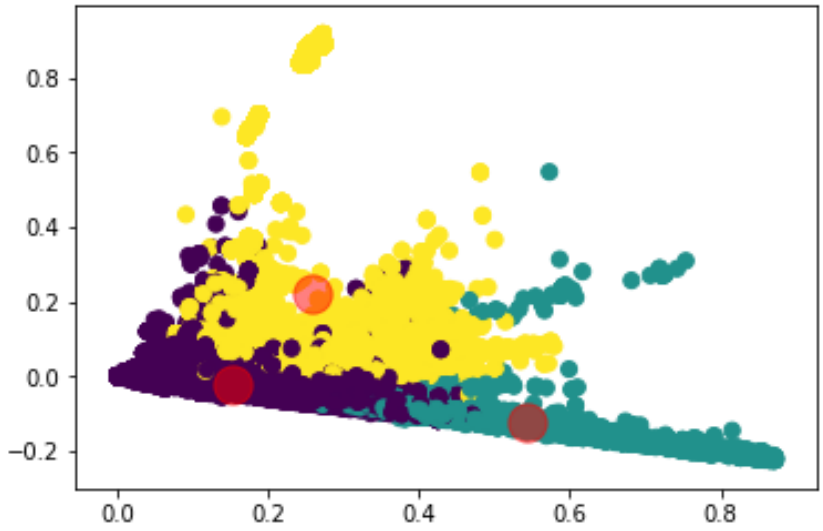

Figura 5.13: Agrupamento dos textos, em três grupos, para os *tweets* coletados no dia 27/03/2018 .

A Tabela 5.7 explicita os termos com maior importância em cada grupo, obtidos através de um algoritmo que converte cada centroide em uma lista decrescente das colunas maiores valoradas (principais termos) pelo *TF-IDF*. Vale destacar que assim com na técnica *LDA*, percebemos um protagonismo das palavras: *"btc", "bitcoin", "blockchain" e "airdrop"*.

Tabela 5.7: Principais termos obtidos dos centros dos agrupamentos, $k = 3$.

grupo 1	**grupo 2**	**grupo 3**
bitcoin	bitcoin	de
crypto	btc	en
get	crypto	hi
like	cryptocurr	ong
new	blockchain	uo
use	price	ia
trade	bitcoin btc	iang
amp	ethereum	ian
blockchain	btc bitcoin	la
exchang	airdrop	ong uo
one	nnew	bitcoin
market	nnew airdrop	que
buy	airdrop nnew airdrop	de bitcoin
peopl	airdrop nnew	ia iang
follow	crypto blockchain	iang ong

O particionamento em quatro grupos ($k = 4$) é possível observar uma clara intersecção entre dois dos centroides presentes. Fato que confirma a tendência observada no processo de execução da curva cotovelo para o particionamento das *features* em três grupos.

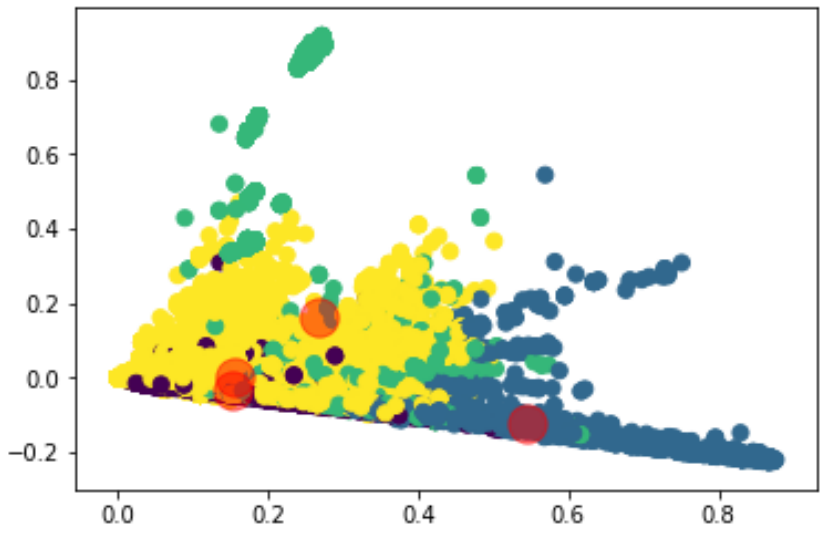

Figura 5.14: Agrupamento dos textos para $k = 4$, *tweets* coletados no dia 27/03/18.

Tabela 5.8: Principais termos obtidos dos centros dos agrupamento, $k = 4$.

grupo 1	**grupo 2**	**grupo 3**	**grupo 4**
cryptocurr	bitcoin	bitcoin btc	btc
bitcoin	crypto	btc	bitcoin
bitcoin cryptocurr	blockchain	bitcoin	btc bitcoin
hi	de	could	price
ong	get	turn	usd
uo	like	back time	eth
ia	trade	time put invest	hour
iang	market	could turn back	btc eth
ian	amp	turn back	btc usd
crypto	use	turn back time	xrp
ethereum	price	back time put	crypto
ong uo	one	put invest money	eur
blockchain	airdrop	invest money bitcoin	bitcoin price
ia iang	exchang	put invest	ltc
iang ong	new	time put	updat

É possível observar, na Tabela 5.8, que há diferença no agrupamento e definição dos textos mais importantes. As palavras citadas anteriormente aparecem em menor volume e novos textos ganham notoriedade, por exemplo: *"invest money bitcoin", "put invest bitcoin", "time put invest", "turn back time"*, entre outros.

No processo de particionamento em cinco grupos ($k = 5$), visto na Figura 5.15, é possível notar que três grupos bem definidos, conforme a similaridade de seus *tweets* constituintes, enquanto que dois centroides estão praticamente sobrepostos.

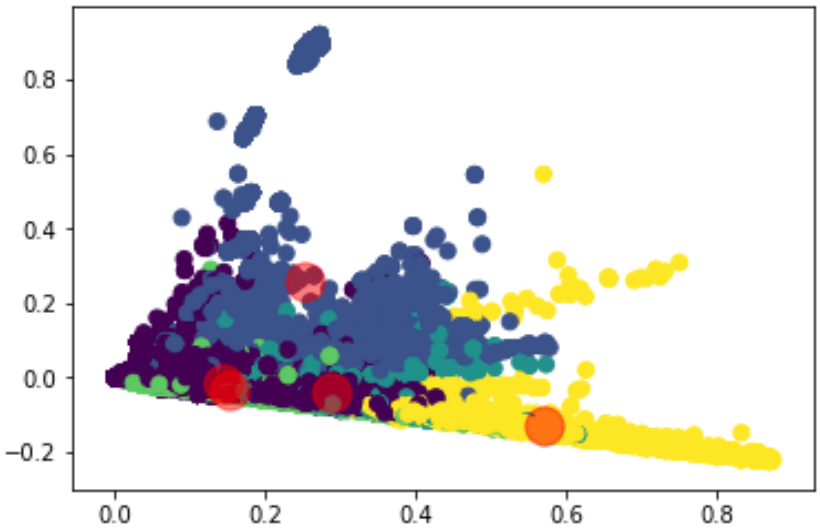

Figura 5.15: Agrupamento dos textos para $k = 5$, *tweets* coletados no dia 27/03/18.

Observado na Tabela 5.9 a presença de textos bastante coesos dentro de cada grupo. O grupo 1 tem o predomínio de termos relacionados a *cryptocurrency*, ou seja, comércio de criptomoedas. O grupo 2, possui termos referentes as ações possíveis com criptomoedas, *"buy", "mine", "use", "get"*. O grupo 3, apresenta como os principais termos, o *bitcoin* e as criptomoedas derivadas dele, *"ethereum", "ripple", "litecoin", "xrp" e "crypto"*. Já os grupos 4 e 5 não apresentam termos diferentes dos vistos anteriormente.

Tabela 5.9: Principais termos obtidos dos centros dos agrupamento, $k = 5$.

grupo 1	**grupo 2**	**grupo 3**	**grupo 4**	**grupo 5**
cryptocurr	bitcoin	crypto	airdrop nnew airdrop	btc
bitcoin cryptocurr	get	ethereum	airdrop nnew	bitcoin
bitcoin	mine	bitcoin	nnew airdrop	price
crypto	de	bitcoin ethereum	nnew	bitcoin btc
blockchain	buy	bitcoin crypto	airdrop	new
cryptocurr bitcoin	use	xrp	blockchain airdrop	follow
ethereum	time	eo	blockchain airdrop nnew	blockchain
btc	like	btc	bitcoin satoshi	de
crypto cryptocurr	cash	rippl	satoshi	like
cryptocurr crypto	money	blockchain	crypto blockchain airdrop	hi
price	blockchain	litecoin	satoshi crypto blockchain	btc bitcoin
market	think	crypto bitcoin	satoshi crypto	usd
altcoin	good	eth	bitcoin satoshi crypto	en
eth	peopl	price	crypto blockchain	market
use	free	top	blockchain	eth

5.5 Discussão

Tarefas de classificação

Para o desenvolvimento do projeto, foram previamente escolhidas metodologias distintas para classificar os *tweets*. Os parâmetros dos classificadores *SVM* e *Multilayer* Perceptron foram obtidos por meio do recurso *GRIDSearchCV* (*SVM*, *perceptron* multicamadas). A escolha dos melhores hiperparâmetros para os modelos proporcionou o melhor resultado de classificação dentre todos os testes realizados. Como visto nos resultados experimentais, o melhor classificador foi o *SVM*, configurando com um *kernel* não linear *RBF*.

Aprimorar os algoritmos de classificação pode acarretar em um aumento perceptível na assertividade (média *F1-Score* próxima a 100) do classificador, com isso obtém-se maior confiança nos resultados de classificação. Tal fato permite analisar os padrões presentes nos textos com maior precisão e confiabilidade. Os resultados da classificação têm interferência direta nos resultados obtidos na análise visual, pois são realizadas comparações entre os o volume de textos classificados em relação ao valor corrente da moeda.

Análise visual

A experimentação de atrelar o volume original de sentimentos com a variação do valor da moeda revelou respostas concordantes com o momento de ascensão da cotação do *Bitcoin* no mercado, visto que a quantidade de elementos com polaridades positivas foram

superiores as negativas, em todos os momentos. Vale mencionar que o momento, de expressiva valorização da moeda, pode ter influência nas análises.

Existem algumas vantagens no emprego de análise visual sobre os *tweets* como observar padrões implícitos em agrupamentos, estabelecer comparações, criação de novas hipóteses, entre outras. Neste projeto foram identificadas relações entre o volume de *tweets*, sobre *bitcoin* de uma determinada polaridade e o respectivo valor de mercado. Esses fatores demonstram a riqueza de possibilidades presentes na análise visual de textos.

Extração de tópicos

Na utilização do modelo probabilístico *LDA*, os termos extraídos, em geral, não produziram direcionamentos condizentes com os principais eventos veiculados pelas mídias sociais (*Twitter*) e/ou tradicionais (portais de notícias). Os termos mais frequentes, em sua maioria, não são relevantes por se tratarem de termos comuns as maioria das mensagens relacionadas a *Bitcoin*, como por exemplo: *bitcoin, btc, airdrop*. Esse fato que evidencia a necessidade de um ajuste no pré-processamento de textos a fim de qualificar a análise dos tópicos extraídos.

Agrupamento via *K-means*

Um processo de agrupamento de dados pode ser considerado ideal se as distâncias entre os grupos são maximizadas e se as distâncias entre os elementos de um grupo são minimizadas. A técnica *elbow curve* foi aplicada com o objetivo de encontrar o valor ideal de K (número de particionamentos) para agrupar os conjuntos de *tweets*.

Para o conjunto de *tweets* usado os possíveis valores foram 4, 5 e 8. Entretanto, pelo caráter (análise de sentimentos positivos, negativos e neutros) do trabalho e a equivalência dos resultados da curva cotovelo entre os valores de $k = 3$ e $k = 4$, foi adicionado o experimento para $k = 3$. Esse experimento permitiu a observação de três grupos bem definidos, ou seja, próximos do ideal, entretanto os termos mais relevantes presentes nos agrupamentos não foram esclarecedores, apresentando os mesmos problemas do modelo *LDA*, explicado anteriormente.

Vale ressaltar, que o agrupamento realizado com $k = 5$, apesar de os grupos não estarem bem definidos visualmente, apresentou ótimos resultados de termos relevantes presentes. Os termos possuem relação coesa entre si, o que possibilita entender as tendências referentes a cada grupo. Além disso, a similaridade dos *tweets* dificultou a formação de agrupamentos bem definidos referentes aos outros dias. Além disso, a alta complexidade computacional exigida pela técnica inviabilizou análises de longo prazo.

Capítulo 6

Conclusão

Este projeto estudou a aplicação de técnicas de análise de sentimentos para verificar se existem relações entre os sentimentos de *tweets* relacionados com *Bitcoins* e a valorização dessa criptomoeda no mercado financeiro. A principal motivação se deve pelo fato do *Bitcoin* ter se popularizado nos últimos anos, como também pela grande quantidade de informações e opiniões disponibilizadas pela rede social *Twitter*.

A Seção 6.1 descreve as considerações finais do estudo, juntamente com a validação das hipóteses e dos objetivos previamente propostos. A Seção 6.2 demonstra as principais limitações encontradas durante o desenvolvimento deste projeto. A Seção 6.3 tem como finalidade indicar projeções futuras proporcionadas pelo desenvolvimento dessa pesquisa.

6.1 Considerações finais

A realização desta pesquisa consistiu no desenvolvimento de uma metodologia baseada em análise de sentimentos com o intuito de identificar padrões e relações entre a valorização do *Bitcoin* no mercado financeiro, como também as opiniões e as manifestações dos usuários na rede social *Twitter*. A metodologia consistiu de etapas do processo convencional de mineração de textos, como pré-processamento, caracterização, extração de informações por meio de aprendizado de máquina e a experimentação para validar e avaliar as técnicas consideradas. Os modelos de aprendizado de máquina empregaram técnicas de classificação de *tweets* de acordo com a polaridade e abordagens não-supervisionadas, como a extração de tópicos via *Latent Dirichlet Allocation* e visualização de textos em conjunto com o algoritmo *K-Means*.

Os resultados experimentais na classificação de sentimentos de *tweets* possibilitaram observar que as polaridades obtidas no período relativo ao conjunto de textos analisados seguem, em geral, o movimento do mercado. As análises mostraram uma tendência contínua de valorização, em decorrência do forte crescimento do valor de mercado da moeda no

período, acompanhada de uma grande quantidade de *tweets* classificados como positivos. Além disso, foi observado que em raros momentos ocorreu a desvalorização da moeda paralelamente à baixa quantidade de *tweets* de polaridade negativa. Esses fatores validam a hipótese (i) "é possível realizar análise de sentimentos de *tweets* relacionadas ao *Bitcoin*?". No entanto, vale ressaltar que o *Bitcoin* está em um longo período de valorização, o que não deixa claro se a análise de sentimentos dos *tweets* em relação ao movimento do mercado de criptomoedas reflete diretamente o valor corrente da moeda. Por outro lado as técnicas de análise de sentimentos empregadas neste trabalho não possibilitaram identificar perfis específicos de investidores, questionamento da hipótese (ii) "é possível identificar perfis de investidores de *Bitcoin* por meio da análise de sentimentos?".

Os resultados apresentados pela modelagem de tópicos utilizando o *LDA*, neste trabalho, não foram esclarecedores. Isso se deve ao fato de os tópicos consistirem em sua maioria de termos comuns a *tweets* relacionados a *Bitcoin*. Tal fato evidencia que o emprego do *LDA* nessa pesquisa depende de um aprimoramento na etapa de pré-processamento de textos para remover termos naturalmente comuns e pouco influentes. Por isso, deve-se empregar a modelagem de tópicos juntamente com outras estratégias de aprendizado de máquina que possibilitem buscar termos realmente influenciadores de tendências de valorização ou desvalorização do valor corrente do *Bitcoin*.

A abordagem de agrupamento por meio do algoritmo *K-Means*, possui muitos fatores que dificultam a sua execução, porém o algoritmo obteve sucesso na tarefa de encontrar padrões condizentes com os sentimentos dos usuários de *Bitcoin* contribuindo para dar sentido à aplicação técnica de visualização que ilustra as comparações realizadas entre o volume total de *tweets* de uma determinada polaridade e o valor corrente da moeda.

Finalmente, a metodologia determinou indícios de relações entre o valor de mercado do *Bitcoin* e os sentimentos expressos nos *tweets*. Esse fato é um agente motivador para continuidade das análises por um período mais extenso, a fim de observar momentos de forte inflexão dos gráficos em todos os sentidos, visto que as análises desse trabalho ocorreram durante um momento de contínua valorização.

6.2 Limitações

A produção de um projeto com escala, relativamente grande, como esse traz limitações difíceis de serem transpassadas. Obtenção de um volume considerável de dados, processar esses dados de forma a buscar resultados relevantes, são alguns exemplos. Anteriormente, foram explicitadas as restrições impostas pela ferramenta de obtenção dos *tweets*, *API* do *twitter*, porém depois de ultrapassadas a barreira de obtenção desses dados é necessário processar esses dados em computadores pessoais, com capacidade de processamento infe-

rior a desejada, consequentemente, testes complexos como o *GRIDSearchCV* ou análises de conjuntos de dados, semanais e mensais, também explicitados no corpo do projeto, detiveram longos períodos de duração.

6.3 Trabalhos futuros

Em decorrência dos trabalhos de pesquisa desenvolvidos durante a elaboração e desenvolvimento do projeto foram observadas diversas possibilidades de aplicação das técnicas estudadas na obtenção de resultados em aplicações distintas das trabalhadas anteriormente. Identificar empresas em ascensão, qualidade de serviços prestados, *feedback* de consumidores nas mais diversas áreas, são trabalhos promissores e não possuem muitas divergências em relação ao projeto atual.

Referências

[1] Gupta, Vishal, Gurpreet S Lehal *et al.*: *A survey of text mining techniques and applications*. Journal of emerging technologies in web intelligence, 1(1):60–76, 2009. ix, 8

[2] Carvalho, Francisco Prancacio Araújo de, João Batista Lopes e Janaína Martins Vasconcelos: *Reflexões econômicas: dinheiro, economia e sociedade*. econômico, 16(31):45, 2014. 1

[3] Orrell, David e Roman Chlupatỳ: *The evolution of money*. Columbia University Press, 2016. 1

[4] Böhme, Rainer, Nicolas Christin, Benjamin Edelman e Tyler Moore: *Bitcoin: Economics, technology, and governance*. Journal of Economic Perspectives, 29(2):213–38, 2015. 1

[5] Chokun, Jonas: *Who accepts bitcoins as payment*. List of Companies, Stores, Shops Retrieved from https://99bitcoins.com/who-accepts-bitcoins-payment-companies-stores-take-bitcoins/, 2013. 1

[6] De Vries, Lisette, Sonja Gensler e Peter SH Leeflang: *Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing*. Journal of interactive marketing, 26(2):83–91, 2012. 2

[7] Tomaél, Maria Inês, Adriana Rosecler Alcará e Ivone Guerreiro Di Chiara: *Das redes sociais à inovação*. Ciência da informação, 34(2), 2005. 2

[8] Wasserman, Stanley: *Advances in social network analysis: Research in the social and behavioral sciences*. Sage, 1994. 2

[9] Kwak, Haewoon, Changhyun Lee, Hosung Park e Sue Moon: *What is twitter, a social network or a news media?* Em *Proceedings of the 19th international conference on World wide web*, páginas 591–600. AcM, 2010. 2

[10] Clarindo, João Paulo, Fábio Coutinho e André Lage Freitas: *Detecção de casos de violência patrimonial a partir do twitter*. 3

[11] Figueredo, Igleson F, Leandro B Marinho e Leonardo A Santos: *Investigando a influência de tweets em programas de votação popular no brasil*. 3

[12] Tavares, Rian e Gustavo Paiva Guedes: *Classificação de filmes: uma abordagem utilizando o liwc*. Em *6º Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2017)*, volume 6. SBC, 2017. 3, 31

[13] Zhang, Xue, Hauke Fuehres e Peter A Gloor: *Predicting stock market indicators through twitter "i hope it is not as bad as i fear"*. Procedia-Social and Behavioral Sciences, 26:55–62, 2011. 3

[14] Bollen, Johan, Huina Mao e Xiaojun Zeng: *Twitter mood predicts the stock market*. Journal of computational science, 2(1):1–8, 2011. 3, 29

[15] Lima, Milson Louseiro *et al.*: *Um modelo para predição de bolsa de valores baseado em mineração de opinião*. 2016. 3, 20

[16] Kannan, K Senthamarai, P Sailapathi Sekar, M Mohamed Sathik e P Arumugam: *Financial stock market forecast using data mining techniques*. Em *Proceedings of the International Multiconference of Engineers and computer scientists*, volume 1, página 4, 2010. 3

[17] Tight, Malcolm e Jeroen Huisman: *Theory and method in higher education research*. Emerald Group Publishing, 2015. 4

[18] Crosby, Michael, Pradan Pattanayak, Sanjeev Verma, Vignesh Kalyanaraman *et al.*: *Blockchain technology: Beyond bitcoin*. Applied Innovation, 2(6-10):71, 2016. 5, 6, 7

[19] Nakamoto, Satoshi *et al.*: *Bitcoin: A peer-to-peer electronic cash system*. 2008. 5

[20] Bradbury, Danny: *The problem with bitcoin*. Computer Fraud & Security, 2013(11):5–8, 2013. 5

[21] Beck, Roman: *Beyond bitcoin: The rise of blockchain world*. Computer, 51(2):54–58, 2018. 6

[22] Ulrich, Fernando: *Bitcoin: a moeda na era digital*. LVM Editora, 2017. 6

[23] Bryans, Danton: *Bitcoin and money laundering: mining for an effective solution*. Ind. LJ, 89:441, 2014. 6, 7

[24] Conti, Mauro, E Sandeep Kumar, Chhagan Lal e Sushmita Ruj: *A survey on security and privacy issues of bitcoin*. IEEE Communications Surveys & Tutorials, 20(4):3416–3452, 2018. 7

[25] Luther, William J: *Bitcoin and the future of digital payments*. The Independent Review, 20(3):397–404, 2016. 7

[26] *Desenvolvimento do registro swim para gerenciamento de tráfego aéreo com o suporte blockchain*. Em *2018 21ª Conferência Internacional sobre Sistemas Inteligentes de Transporte (ITSC)*. 7

[27] Fayyad, Usama, Gregory Piatetsky-Shapiro e Padhraic Smyth: *From data mining to knowledge discovery in databases*. AI magazine, 17(3):37–37, 1996. 7

[28] Barion, Eliana Cristina Nogueira e Decio Lago: *Mineração de textos.* Revista de Ciências Exatas e Tecnologia, 3(3):123–140, 2015. 8

[29] Sebastiani, Fabrizio: *Machine learning in automated text categorization.* ACM computing surveys (CSUR), 34(1):1–47, 2002. 8

[30] Larose, Daniel T e Chantal D Larose: *Discovering knowledge in data: an introduction to data mining.* John Wiley & Sons, 2014. 8

[31] Liu, Huan e Hiroshi Motoda: *Feature extraction, construction and selection: A data mining perspective*, volume 453. Springer Science & Business Media, 1998. 8

[32] Feldman, Ronen e James Sanger: *The text mining handbook: advanced approaches in analyzing unstructured data.* Cambridge university press, 2007. 8

[33] Tan, Pang Ning: *Introduction to data mining.* Pearson Education India, 2018. 9, 20, 22

[34] Alexander, Patricia A e Tamara L Jetton: *The role of importance and interest in the processing of text.* Educational psychology review, 8(1):89–121, 1996. 9

[35] Aggarwal, Charu C e ChengXiang Zhai: *Mining text data.* Springer Science & Business Media, 2012. 9

[36] Silva, Catarina e Bernardete Ribeiro: *The importance of stop word removal on recall values in text categorization.* Em *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 3, páginas 1661–1666. IEEE, 2003. 9

[37] Lovins, Julie Beth: *Development of a stemming algorithm.* Mech. Translat. & Comp. Linguistics, 11(1-2):22–31, 1968. 9

[38] Matsubara, Edson Takashi, Claudia Aparecida Martins e Maria Carolina Monard: *Pretext: Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words.* Techinical Report, 209:4, 2003. 10, 11

[39] Ramos, Juan *et al.*: *Using tf-idf to determine word relevance in document queries.* Em *Proceedings of the first instructional conference on machine learning*, volume 242, páginas 133–142. Piscataway, NJ, 2003. 11

[40] Medeiros, Hichemm Khalyd R. V.: *Um método semi-supervisionado para análise de sentimentos em textos da rede social twitter utilizando aprendizado ativo.* página 63, 2018. 11, 12, 13

[41] Luhn, Hans Peter: *A statistical approach to mechanized encoding and searching of literary information.* IBM Journal of research and development, 1(4):309–317, 1957. 11

[42] Lee, Dik L, Huei Chuang e Kent Seamons: *Document ranking and the vector-space model.* IEEE software, 14(2):67–75, 1997. 12

[43] *Machine learning.* Anual revisão de ciência da computação, 4(1):417–433. 13

[44] Monard, Maria Carolina e José Augusto Baranauskas: *Conceitos sobre aprendizado de máquina*. Sistemas inteligentes-Fundamentos e aplicações, 1(1):32, 2003. 13

[45] Lorena, Ana Carolina e André CPLF de Carvalho: *Uma introdução às support vector machines*. Revista de Informática Teórica e Aplicada, 14(2):43–67, 2007. 13

[46] Ikonomakis, M, Sotiris Kotsiantis e V Tampakas: *Text classification using machine learning techniques.* WSEAS transactions on computers, 4(8):966–974, 2005. 13, 35

[47] Cortes, Corinna e Vladimir Vapnik: *Support-vector networks*. Machine Learning, 20(3):273–297, Sep 1995, ISSN 1573-0565. https://doi.org/10.1007/BF00994018. 14

[48] Auria, Laura e Rouslan A Moro: *Support vector machines (svm) as a technique for solvency analysis*. 2008. 14

[49] Haykin, Simon: *Redes neurais: princípios e prática*. Bookman Editora, 2007. 15

[50] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot e E. Duchesnay: *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12:2825–2830, 2011. 16, 17, 23, 38

[51] Rezende, Solange O, Ricardo M Marcacini e Maria F Moura: *O uso da mineração de textos para extração e organização não supervisionada de conhecimento.* Embrapa Informática Agropecuária-Artigo em periódico indexado (ALICE), 2011. 18, 19

[52] Joachims, Thorsten: *A support vector method for multivariate performance measures*. Em *Proceedings of the 22nd international conference on Machine learning*, páginas 377–384. ACM, 2005. 20

[53] Zhang, Dell, Jun Wang e Xiaoxue Zhao: *Estimating the uncertainty of average f1 scores*. 2015. 21

[54] Gantz, John e David Reinsel: *The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east*. IDC iView: IDC Analyze the future, 2007(2012):1–16, 2012. 21

[55] Benevenuto, Fabrício, Filipe Ribeiro e Matheus Araújo: *Métodos para análise de sentimentos em mídias sociais*. Em *Brazilian Symposium on Multimedia and the Web (Webmedia), Manaus, Brasil*, 2015. 21

[56] Valentim, Marta Lígia Pomim *et al.*: *Inteligência competitiva em organizações: dado, informação e conhecimento*. DataGramaZero, Rio de Janeiro, 3(4):1–13, 2002. 21

[57] Silva, Heide Miranda da: *Sociedade da informação*, 2019. http://www.profcordella.com.br/unisanta/textos/tgs21_dados_info_conhec.htm, acesso em 2019. 21

[58] Pang, Bo, Lillian Lee *et al.*: *Opinion mining and sentiment analysis*. Foundations and Trends® in Information Retrieval, 2(1–2):1–135, 2008. 21

[59] Lee, John A e Michel Verleysen: *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007. 22

[60] Joia, Paulo, Danilo Coimbra, Jose A Cuminato, Fernando V Paulovich e Luis G Nonato: *Local affine multidimensional projection*. IEEE Transactions on Visualization and Computer Graphics, 17(12):2563–2571, 2011. 23

[61] Jolliffe, Ian: *Principal component analysis*. Springer, 2011. 23

[62] *Decomposições de conceito para dados de texto esparso grandes usando clustering*. Machine learning, 42(1-2). 23, 24

[63] Keim, Daniel A: *Information visualization and visual data mining*. IEEE transactions on Visualization and Computer Graphics, 8(1):1–8, 2002. 24

[64] Ware, Colin: *Information visualization: perception for design*. Elsevier, 2012. 24

[65] Card, Stuart, JD Mackinlay e B Shneiderman: *Information visualization*. Human-computer interaction: Design issues, solutions, and applications, 181, 2009. 24

[66] Nonato, Luis Gustavo e Michael Aupetit: *Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment*. IEEE transactions on visualization and computer graphics, 2018. 24

[67] Paulovich, Fernando V, Maria Cristina F Oliveira e Rosane Minghim: *The projection explorer: A flexible tool for projection-based multidimensional visualization*. Em *XX Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2007)*, páginas 27–36. IEEE, 2007. 24

[68] Blei, David M, John D Lafferty *et al.*: *A correlated topic model of science*. The Annals of Applied Statistics, 1(1):17–35, 2007. 26

[69] Faleiros, Thiago de Paulo: *Propagação em grafos bipartidos para extração de tópicos em fluxo de documentos textuais*. Tese de Doutoramento, Universidade de São Paulo, 2016. 27

[70] *Alocação dirichlet latente*. Journal of machine Learning research, 3(jan):993–1022. 27

[71] Porshnev, Alexander, Ilya Redkin e Alexey Shevchenko: *Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis*. Em *2013 IEEE 13th International Conference on Data Mining Workshops*, páginas 440–444. IEEE, 2013. 28

[72] Pagolu, Venkata Sasank, Kamal Nayan Reddy, Ganapati Panda e Babita Majhi: *Sentiment analysis of twitter data for predicting stock market movements*. Em *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*, páginas 1345–1350. IEEE, 2016. 29

[73] Caetano, Josemar Alves, Hélder Seixas Lima, Mateus Freira dos Santos e Humberto Torres Marques-Neto: *Utilizando análise de sentimentos para definição da homofilia política dos usuários do twitter durante a eleição presidencial americana de 2016*. Em *6º Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2017)*, volume 6. SBC, 2017. 30

[74] Souza, Bruno A, Thais G Almeida, Alice A Menezes, Carlos MS Figueiredo, Fabíola G Nakamura e Eduardo F Nakamura: *Uma abordagem para detecção de tópicos relevantes em redes sociais online*. Em *6º Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2017)*, volume 6. SBC, 2017. 30

[75] Bergstra, James e Yoshua Bengio: *Random search for hyper-parameter optimization*. Journal of Machine Learning Research, 13(Feb):281–305, 2012. 35

[76] Handa, Hisashi, Hisao Ishibuchi, Yew Soon Ong e Kay Chen Tan: *Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems*, volume 1. Springer, 2014. 36

[77] Kodinariya, Trupti M e Prashant R Makwana: *Review on determining number of cluster in k-means clustering*. International Journal, 1(6):90–95, 2013. 36

[78] Huang, Anna: *Similarity measures for text document clustering*. Em *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, páginas 9–56, 2008. 36

[79] Lima, Rodrigo Lucio de: *Avaliação do algorítimo svm na detecção de comportamentos suspeitos em cenas de vídeo*. B.S. thesis, Universidade Tecnológica Federal do Paraná, 2014. 40