

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

GABRIEL CARVALHO DE SOUZA

Orientador: Prof. Dr. Guilherme Tavares de Assis

Coorientador: Prof. Dr. Israel José dos Santos Felipe

**DESENVOLVIMENTO E VALIDAÇÃO DE UMA PLATAFORMA DE
AUXÍLIO, BASEADA EM *TWEETS*, À TOMADA DE DECISÃO DO
INVESTIDOR BRASILEIRO NO MERCADO FINANCEIRO**

Ouro Preto, MG
2021

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

GABRIEL CARVALHO DE SOUZA

**DESENVOLVIMENTO E VALIDAÇÃO DE UMA PLATAFORMA DE AUXÍLIO,
BASEADA EM *TWEETS*, À TOMADA DE DECISÃO DO INVESTIDOR BRASILEIRO
NO MERCADO FINANCEIRO**

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Guilherme Tavares de Assis

Coorientador: Prof. Dr. Israel José dos Santos Felipe

Ouro Preto, MG
2021



FOLHA DE APROVAÇÃO

Gabriel Carvalho de Souza

Desenvolvimento e validação de uma plataforma de auxílio, baseada em tweets, à tomada de decisão do investidor brasileiro no mercado financeiro

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 18 de Agosto de 2021.

Membros da banca

Guilherme Tavares de Assis (Orientador) - Doutor - Universidade Federal de Ouro Preto
Israel José dos Santos Felipe (Coorientador) - Doutor - Universidade Federal de Ouro Preto
Jadson Castro Gertrudes (Examinador) - Doutor - Universidade Federal de Ouro Preto
Raoni de Oliveira Inácio (Examinador) - Doutor - Universidade Federal de Ouro Preto

Guilherme Tavares de Assis, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 18/08/2021.



Documento assinado eletronicamente por **Guilherme Tavares de Assis, PROFESSOR DE MAGISTERIO SUPERIOR**, em 18/08/2021, às 16:57, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0206597** e o código CRC **13C2E9D6**.

Resumo

No atual momento, a análise de sentimentos em redes sociais possui fundamental importância para o monitoramento de informações essenciais e voltadas ao planejamento e à gestão de grandes negócios. Na ótica do mercado financeiro, tal análise pode tornar-se uma fonte valiosa de informação, uma vez que auxiliar investidores, trazendo notícias e repercussões diárias sobre seus ativos ou investimentos, e relacionar o sentimento expresso em tais notícias com os preços dos ativos podem ser de grande relevância em uma tomada de decisão. Contudo, aplicar a análise de sentimentos sobre notícias disponíveis, de forma manual, é humanamente ineficiente e inviável. Desta forma, ferramentas que automatizam a análise de sentimentos no contexto do mercado financeiro, por meio de processos de extração dos dados necessários e análise propriamente dita dos resultados obtidos, podem ser potencialmente contributivas. Assim, este trabalho possui, como objetivo geral, propor, desenvolver e validar uma plataforma que auxilie o investidor brasileiro no processo de compra e venda de ações mediante a divulgação de *tweets* e suas polaridades. Para tanto, a plataforma faz uso da rede social *Twitter* e encontra-se dividida em três módulos, a saber: (a) módulo responsável pela coleta, processamento e classificação dos *tweets*; (b) módulo responsável pelo recebimento, organização e apresentação dos dados; e (c) módulo responsável pela comunicação entre os dois outros módulos. Experimentações práticas, considerando o foco associado à contribuição da plataforma para os usuários, geraram resultados satisfatórios quanto às análises realizadas, identificando, em alguns momentos, padrões onde o sentimento expresso pelos *tweets* acompanhava o movimento dos preços dos ativos. Os experimentos com foco associado ao desempenho da arquitetura proposta também geraram resultados satisfatórios quanto ao tempo de execução das etapas envolvidas nos processos internos dos módulos e no tráfego eficiente e confiável de mensagens entre os módulos.

Palavras-chave: *Twitter*. Mineração de dados. Análise de sentimentos. Mercado financeiro. Auxílio a tomada de decisão.

Abstract

At present, the analysis of feelings in social networks is of fundamental importance for monitoring essential information focused on the planning and management of large businesses. Such analysis can become a valuable source of information from the financial market perspective since it helps investors bring daily news and repercussions about their assets or investments. Relating the sentiment expressed in such news to dosing prices can be of great relevance in making a decision. However, applying the analysis of disagreements on available news by hand is humanly inefficient and unfeasible. In this way, tools that automate the sentiment analysis in the financial market context can be potentially profitable by extracting the necessary data and correctly analyzing the results. Thus, this work has as a general objective, to develop and validate a platform that helps the Brazilian investor in buying and selling shares through the disclosure of their polarities. Thus, this work has, as a general objective, to propose, develop and validate a platform that helps Brazilian investors in the process of buying and selling shares by disclosing their tweets and their polarities. three modules, namely: (a) module responsible for the collection, processing and classification of tweets; (b) module responsible for receiving, organizing and presenting data; and (c) module responsible for communication between the two other modules. Practical experiments, considering the focus associated with the platform's contribution to users, generated satisfactory results in terms of the analyzes carried out, identifying, at times, patterns where the sentiment expressed by the tweets accompanied the movement of asset prices. The experiments focused on the performance of the proposed architecture also generated satisfactory results regarding the execution time of the steps involved in the internal processes of the modules and the efficient and reliable traffic of messages between modules.

Keywords: Twitter. Data mining. Sentiment analysis. Financial market. Assistance in decision making.

Lista de Ilustrações

Figura 2.1 – Classificação do mercado financeiro.	7
Figura 2.2 – Descoberta de conhecimento em base de dados.	8
Figura 2.3 – Arquitetura do <i>framework Scrapy</i>	9
Figura 2.4 – Tarefas desempenhadas por técnicas de mineração de dados.	11
Figura 2.5 – Arquitetura do <i>framework RabbitMQ</i>	13
Figura 3.1 – Arquitetura de funcionamento da plataforma proposta.	23
Figura 3.2 – Fragmento do <i>script</i> que realiza a extração dos <i>tweets</i> a partir de regras Xpath	25
Figura 3.3 – Fragmento do <i>script</i> que realiza o pré-processamento dos <i>tweets</i>	26
Figura 3.4 – Funcionamento do terceiro módulo	30
Figura 3.5 – Fórmula do retorno linear de um ativo	31
Figura 3.6 – Primeira etapa do módulo de comunicação	32
Figura 3.7 – Segunda etapa do módulo de comunicação	32
Figura 3.8 – Tela inicial da plataforma	33
Figura 3.9 – Tela de apresentação dos resultados obtidos	34
Figura 4.1 – Caso de teste 1	39
Figura 4.2 – Caso de teste 2	40
Figura 4.3 – Caso de teste 3	41

Lista de Tabelas

Tabela 2.1 – Comparativo de Trabalhos Relacionados	20
Tabela 3.1 – Manual de funcionamento da plataforma proposta	24
Tabela 4.1 – Parametrização dos casos de teste	35
Tabela 4.2 – Resultados da performance da plataforma	37

Lista de Abreviaturas e Siglas

UFOP	Universidade Federal de Ouro Preto
API	Application Programming Interface
SVM	Support Vector Machine
VSM	Vector Space Model
DJIA	Dow Jones Industrial Average
SGBD	Sistema de Gerenciamento de Banco de Dados
HTML	HyperText Markup Language,
PLN	Processamento de Linguagem Natural
NLTK	Natural Language Toolkit
HTTP	Hyper Text Transference Protocol
AMQP	Advanced Message Queuing Protocol
RAM	Random Access Memory
RT	Retweet
TF-IDF	<i>Term Frequency – Inverse Document Frequency</i>
IBGE	Instituto Brasileiro de Geografia
IBOV	Ibovespa

Sumário

1	Introdução	1
1.1	Justificativa	3
1.2	Objetivos Geral e Específicos	4
1.3	Organização do Trabalho	5
2	Revisão Bibliográfica	6
2.1	Fundamentação Teórica	6
2.1.1	Mercado financeiro	6
2.1.2	Gerência de dados na <i>Web</i>	8
2.1.2.1	Coleta de dados	9
2.1.2.2	Extração de dados	10
2.1.2.3	Mineração de dados	10
2.1.3	Arquitetura de microserviços	12
2.1.4	<i>Twitter</i>	13
2.1.5	Análise de sentimentos	14
2.2	Trabalhos Relacionados	15
3	Desenvolvimento	22
3.1	Arquitetura de funcionamento	22
3.1.1	Coleta e processamento dos dados	23
3.1.2	Recebimento, organização e apresentação de dados	29
3.1.3	Comunicação entre módulos	31
3.2	Interface e parametrização da plataforma	33
4	Experimentos Computacionais	35
4.1	Descrição dos experimentos	35
4.2	Análise dos resultados obtidos	36
5	Considerações Finais	42
5.1	Conclusão	42
5.2	Trabalho futuro	42
	Referências	44

1 Introdução

A sociedade está cada vez mais dependente do mundo digital: o acesso à *Internet*, via computadores e *smartphones*, trouxe mais comodidade e entretenimento para as pessoas, gerou muitos empregos e ampliou exponencialmente a quantidade de informação difundida pelo mundo, conforme aponta Muller (2016). A vista disto, uma enorme e diversificada massa de informação é publicada e vista por milhões de pessoas a todo momento. De acordo com o relatório divulgado pela Comscore (2015), estima-se que 45% da população gasta em média 650 horas criando e absorvendo conteúdo virtual. Tal relatório afirma que 90% do tempo gasto dá-se em redes sociais, como *Facebook*¹, *Twitter*², *Instagram*³ e *Pinterest*⁴, e não em portais de notícias de uma forma geral.

Por permitir que os usuários não só consumam as informações, mas possam também interagir com o conteúdo publicado, as redes sociais atraíram a atenção do mundo inteiro. De acordo com Sulz (2019), o que se observa é que a interconectividade dos usuários, por funcionalidades dispostas em tempo real pelas redes sociais, aumenta o contato direto do público com a notícia ou publicação de uma forma geral, aumentando o valor agregado a essas comunidades. Muitos enxergam as redes sociais como uma fonte valiosa de dados, pois expressam as opiniões, intensões, preocupações e desejos de seus usuários; ademais, se explorados, estes dados possibilitam descobrir relações com vários fenômenos, econômicos, políticos e culturais (RANCO et al., 2015).

Com o movimento crescente da indústria, fazer previsões com uma base de dados proveniente de redes sociais tornou-se cada vez mais comum (NETO; ARRUDA, 2018). Conforme aponta Team (2020), o volume de dados produzido diariamente nestes ambientes virtuais é intangível, sendo caracterizado, segundo o autor, como uma mina de ouro para as empresas que souberem analisá-los. Devido a esse acúmulo de dados, processos focados em *Big Data*⁵, como mineração de dados, aprendizado de máquina e análise de sentimentos, são essenciais na tentativa de granular estes dados e estudá-los, de modo a obter informações úteis e efetivas no auxílio à tomada de decisão, além de *insights* valiosos sobre o comportamento humano e empresarial (DISTRITO, 2020).

Dentre as redes sociais existentes, o *Twitter* destaca-se quanto à quantidade de informação exibida. Segundo Statusbrew (2018), a rede social possui 326 milhões de usuários ativos e mais de 500 milhões de *tweets* diários. Por se tratar de uma rede social com uma comunicação rápida e

¹ <https://www.facebook.com/>

² <https://twitter.com/>

³ <https://www.instagram.com/>

⁴ <https://pinterest.com/>

⁵ *Big Data* é uma área que estuda como tratar e analisar um conjunto de dados grande demais para ser analisado por sistemas tradicionais.

em escala global, a análise manual desse volume de dados é impraticável segundo [Aguiar \(2012\)](#). Conforme apontado pelo autor, coletar e submeter publicações do *Twitter*, a um tratamento de dados rápido e automatizado, produzem resultados que podem ser aplicados a diversas áreas.

Em meio a tanta informação propagada no *Twitter*, uma quantidade considerável está relacionada ao mercado financeiro. Publicações que exibem em seu conteúdo temas como: retorno das ações, volatilidade dos ativos, motivações financeiras, empresas que compõem a bolsa, pronunciamentos estratégicos, dentre outros, provocam efeitos anormais nas carteiras de investimento, pelo fato de, potencialmente, influenciarem na tomada de decisão de investidores ([SOUZA et al., 2015](#)). Segundo [Dondio \(2012\)](#), a rede social, a partir de seus grupos de usuários, gera dados que, gradualmente, vem adquirindo credibilidade como fonte válida para a análise do mercado de ações.

No entanto, devido ao comportamento volátil e incerto do mercado e à assimetria de informação encontrada nas redes sociais ([REGAL et al., 2019](#)), um conjunto de dilemas pode ser encontrado pelo investidor no momento da tomada de decisão: quais ativos comprar ou vender, em qual quantidade e qual o momento certo de fazê-lo. Para os autores, além desses fatores, são escassas as ferramentas disponíveis que possam trazer, de forma confiável, informações relevantes para apoiar o investidor na tomada de decisão.

Não se sabe ao certo qual é o modelo do mercado de ações; compreender o seu comportamento e suas tendências futuras são grandes desafios ([ATSALAKIS; VALAVANIS, 2009](#)). Portanto, como forma de trazer ao investidor um apoio na sua tomada de decisão, a técnica de previsão, segundo [Regal et al. \(2019\)](#), foi desenvolvida: esta técnica busca prever o valor de uma variável no tempo usando, como entrada, valores desta mesma variável em conjunto com outros dados, em janela de tempos anteriores. Para [Melo \(2012\)](#), esta técnica está relacionada a compreender um passo adiante o comportamento do mercado. O autor aponta que características cíclicas nos movimentos dos preços dos ativos existem e que correlacionar estes preços com outros dados, como notícias ou avaliações sobre as empresas, é uma grande oportunidade de maximizar ganhos e evitar perdas.

Muitos trabalhos tentam relacionar o *Twitter* ao mercado de ações, como é visto em ([BOLLEN; MAO; ZENG, 2011](#)). Segundo os autores, uma análise criteriosa na vasta quantidade de dados encontrados nessa rede social pode trazer respostas que levem a alguma vantagem no mundo dos investimentos. [Bollen, Mao e Pepe \(2011\)](#), em uma abordagem de análise do *Twitter*, encontraram relações entre os indicadores de humor e o *Dow Jones Industrial Average (DJIA)*⁶. Já [Souza et al. \(2015\)](#) mostraram que o sentimento do *Twitter* para cinco empresas de varejo tem relação estatisticamente significativa com o retorno das ações e a volatilidade. Em estudo feito por [Santos, Laender e Pereira \(2015\)](#), concluiu-se que o volume de transações financeiras na bolsa de valores, em um determinado espaço de tempo, correlacionam para 66% dos ativos mencionados

⁶ DJIA é um índice de ações que representa a evolução da cotação de 30 ações em empresas norte-americanas líderes de mercado.

no *Twitter*.

Este capítulo encontra-se organizado como se segue. A Seção 1.1 apresenta a justificativa para a realização deste trabalho. A Seção 1.2 descreve os objetivos geral e específicos. A Seção 1.3 discorre sobre a organização do restante do trabalho monográfico.

1.1 Justificativa

O processo de formação dos preços dos ativos negociados no mercado financeiro e a tomada de decisão, seja de compra ou venda, estão ligados a diversos fatores, dentre eles, a relação e o sentimento do próprio investidor, segundo [Brown e Cliff \(2004\)](#). Para os autores, esta relação é forte; eles mencionam que, um alto nível de otimismo, faz com que os preços atuais sejam elevados, principalmente quando a análise é realizada com dados de pessoas institucionais, onde há uma maior evidência de previsão.

A ideia de acompanhar as redes sociais à procura de oportunidades no mercado financeiro é bastante estudada ([KOPSCHITZ, 2011](#)). Ser capaz de monitorar uma rede social constantemente, com a intenção de prever as tendências do mercado e o seu comportamento, pode dar uma vantagem a qualquer investidor. Em estudo feito por [Bollen, Mao e Zeng \(2011\)](#), baseado na economia comportamental, foi identificada uma precisão de 87.6% nos resultados correlacionados com o valor da DJIA, ao avaliar e traduzir para a ótica financeira as emoções transmitidas pelos usuários por meio de mensagens no *Twitter*.

Em notícia publicada por [d'Agosto \(2012\)](#), ações americanas subiram 0,4% em apenas 20 minutos; a causa foi atribuída a um comentário no *Twitter* feito por Bill Gross, fundador da Pimco, uma empresa global de gerenciamento de investimentos. Em 2013, a agência de notícias *The Associated Press* publicou uma *fake news* citando o então presidente dos Estados Unidos, Barack Obama ([ELBOGHDADY, 2013](#)). A notícia, segundo o site, dizia que o presidente havia se ferido por bombas que explodiram na Casa Branca. A publicação manteve-se irrefutável por 3 minutos, o suficiente para apagar com 136 bilhões de dólares do mercado acionário. Em 2019, o banco J.P. Morgan Chase mencionou a criação de um índice para avaliar os impactos dos *tweets* do presidente americano Donald Trump no mercado financeiro, uma vez que fora relatado a influência destes na movimentação do mercado e na volatilidade nas taxas de juros dos Estados Unidos ([NEWBURGER, 2019](#)).

Partindo da premissa de que o sentimento humano, expresso por meio de mensagens no *Twitter*, tem impacto significativo nos índices financeiros, diversos estudos têm desenvolvido estratégias para analisar a massa de dados existente e encontrar indícios de tal influência. Segundo [Ferreira e Ungaretti \(2021\)](#), o número de investidores brasileiros passou de 1.681.033 em 2019 para 3.229.318 em 2020. Com um aumento de quase 92% de pessoas envolvidas com o mercado, é fato que o volume de informação relacionada a esse tema tráfegada na *Web* também tenha aumentado; isso motiva grandes oportunidades para o surgimento de pesquisas nesta área.

A mineração de dados, aplicada em uma rede social de expressão como o *Twitter*, com o propósito de associação ao mercado financeiro, traz grandes impactos a alguns setores. A sociedade é impactada, uma vez que a comunicação via *Web* transformou-se em algo popular; a interpretação das opiniões descritas nela já é feita de forma intuitiva, porém nenhum valor informativo é extraído. Segundo [Falcão \(2020\)](#), a automatização do processo de mineração de dados e de análise de sentimentos pode auxiliar na capacidade de interpretação das pessoas, bem como a sua tomada de decisão. O sentimento do investidor é um dos aspectos das finanças comportamentais, definidas por [Bondt et al. \(2008\)](#), como o estudo de como a psicologia impacta nas decisões dos investidores.

A academia também é impactada neste processo, visto que a maior parte dos estudos feitos nesta área apresenta um trabalho manual em seu processo de classificação dos textos que, em muitas vezes, limita a escalabilidade das pesquisas, quando comparado com um processo automatizado, influenciando até em seus resultados ([PALMER, 2010](#)); ademais, um foco para o mercado financeiro brasileiro pode ser também um fator motivacional para que mais estudos aprofundem-se na área. Por fim, os agentes do mercado financeiro são também impactados, pois auxiliar na tomada de decisão do investidor de certa forma o deixa mais confiante e consciente para o mercado, reduzindo a ideia de racionalidade, já que investimentos podem passar a ser analisados com base em emoções junto às estatísticas.

Diante dessas ponderações e dos estudos realizados por ([NETO; ARRUDA, 2018](#)) e ([LIMA et al., 2016](#)), uma plataforma *Web*, que pudesse analisar os sentimentos presentes nas redes sociais e relacioná-los com os ativos da bolsa de valores brasileira, teria grande relevância para investidores.

1.2 Objetivos Geral e Específicos

O objetivo geral desse trabalho consiste em desenvolver e validar uma plataforma que auxilia o processo de compra e venda de ações mediante a divulgação de *tweets* e suas polaridades. Esta plataforma utiliza o *Twitter* como base de dados e engloba, em sua proposta, a análise e a exibição de informações relativas aos ativos negociados no mercado financeiro brasileiro.

Os objetivos específicos alcançados neste trabalho foram:

- promover a coleta automática e ilimitada de publicações sobre o mercado financeiro no *Twitter*;
- possibilitar a tomada de decisão de investidores por meio de indicadores de sentimento para determinados ativos a partir de postagens na rede social *Twitter* em tempo real;
- fornecer comparativo sobre trabalhos relacionados com o tema proposto.

1.3 Organização do Trabalho

O restante deste trabalho monográfico encontra-se organizado como se segue. No Capítulo 2, é apresentado o referencial bibliográfico, relacionado ao tema proposto e necessário para o entendimento deste trabalho. No Capítulo 3, é descrita a plataforma proposta neste trabalho, envolvendo suas características, arquitetura de funcionamento e interface. No Capítulo 4, os experimentos realizados são apresentados junto aos resultados obtidos. E por fim, no Capítulo 5, são apresentadas as conclusões deste trabalho e perspectivas de trabalho futuro.

2 Revisão Bibliográfica

Este capítulo destina-se à revisão bibliográfica; como forma de sustentar o presente trabalho, auxilia na definição dos objetivos e na delimitação do problema abordado. As subseções estão organizadas como se segue. A Seção 2.1 apresenta a fundamentação teórica. A Seção 2.2 discorre sobre trabalhos relacionados.

2.1 Fundamentação Teórica

Esta seção tem, como objetivo, apresentar conceitos relevantes para a fundamentação e a construção da proposta deste trabalho, bem como contribuir para um melhor entendimento da metodologia aplicada. Os assuntos abordados estão dispostos da seguinte forma: a Subseção 2.1.1 descreve o mercado financeiro, a Subseção 2.1.2 fala sobre o gerenciamento de dados na *Web*, a Subseção 2.1.3 discorre sobre a arquitetura de microserviços a Subseção 2.1.4 aborda a rede social *Twitter* e a Subseção 2.1.5 apresenta a área de análise de sentimentos.

2.1.1 Mercado financeiro

O mercado financeiro consiste em um ambiente para a negociação de instrumentos financeiros, moedas, ações, títulos ou commodities (LTD, 2020). Segundo Kenton (2020), o que faz o mercado financeiro ser vital para o bom funcionamento das economias capitalistas é a alocação de recursos na economia do país, além da criação de liquidez para as empresas.

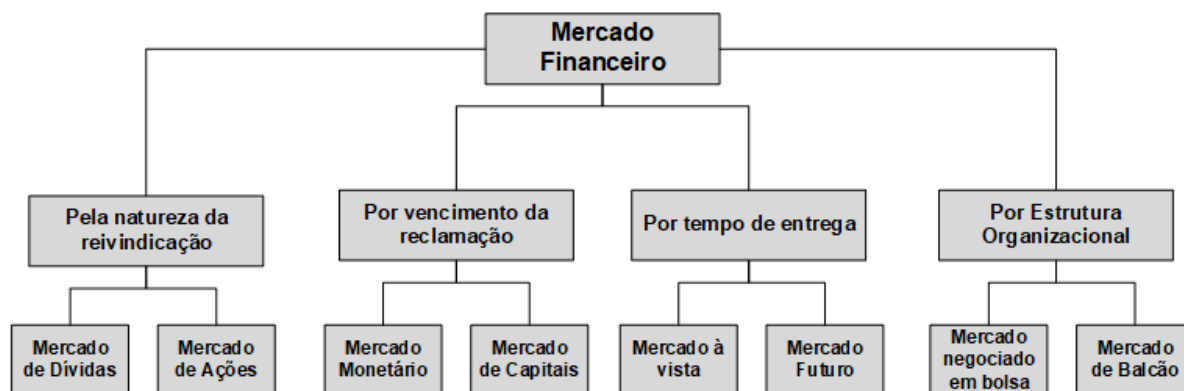
Existem vários tipos de mercados financeiros e cada país possui pelo menos um, conforme descrito em Inc (2020). Para o autor, todos os mercados financeiros são caracterizados pela transação de compra e venda de ações, títulos, moedas e derivados. Conforme apresentado por Kenton (2020), o objetivo principal dos mercados financeiros é garantir a transparência das informações para que haja uma definição de preços eficiente e apropriada.

A Figura 2.1, segundo BYJU'S (2020), apresenta a classificação do mercado financeiro e suas subdivisões.

De acordo com a Figura 2.1, o mercado financeiro é classificado de quatro formas diferentes: pela natureza da reivindicação, por vencimento da reclamação, por tempo de entrega e por estrutura organizacional. BYJU'S (2020) descreve os mercados da seguinte forma: o mercado de dívidas como sendo o local onde os investidores trocam seus títulos prefixados e debêntures¹; o mercado de ações como sendo o local onde os investidores lidam com ações; o mercado monetário, onde os investidores lidam com ativos monetários e fundos de curto prazo como

¹ Debênture é um título de crédito representativo de um empréstimo que uma companhia realiza junto a terceiros e que assegura a seus detentores direito contra a emissora, estabelecidos na escritura de emissão.

Figura 2.1 – Classificação do mercado financeiro.



Fonte: Elaborada pelo autor.

certificados de depósito, títulos do tesouro, dentre outros; o mercado de capitais como sendo o local onde se negocia ativos financeiros de médio e longo prazo; o mercado à vista, onde a negociação é liquidada em tempo real; o mercado de futuro, onde a compensação das negociações é realizada em uma data futura estipulada; o mercado negociado em bolsa, onde as transações são encaminhadas por meio de uma fonte central, ou seja, uma das partes fica responsável por ser o intermediário que conecta compradores e vendedores; e, por último, o mercado de balcão, que, em oposição ao mercado negociado em bolsa, é amplamente descentralizado, havendo uma competição entre vários intermediários para conectar compradores e vendedores.

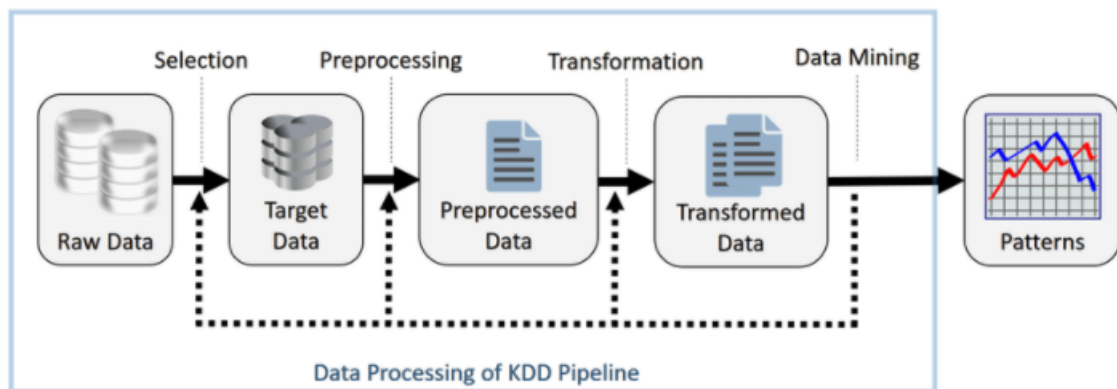
Diante das mais variadas classificações do mercado financeiro, a sua função é única: possibilitar o encontro entre vendedores e compradores (REIS, 2018). Segundo o autor, o mercado divide-se entre os credores, que fornecem capital, e os captadores, que por sua vez captam os recursos em troca de parte dos lucros. Esta interação é viabilizada por instituições responsáveis pois, mesmo ocorrendo de forma livre sem muitas interferências externas, são necessários órgãos para regular e fiscalizar a conformidade de todo o processo.

O mercado de ações, conforme visto na Figura 2.1, é uma subdivisão do mercado financeiro. Ele representa a menor parcela do capital de uma empresa, quem negocia ações de uma companhia, adquire também direitos e deveres de um sócio (INFOMONEY, 2020). Segundo o autor, o acionista ganha dinheiro com ações na valorização da empresa e consequentemente da ação. Isso vai depender do desempenho financeiro da empresa e das perspectivas para o setor em que ela atua e para a economia em geral. No Brasil, a única bolsa de valores existente é a B3 e é por ela que são feitas todas as intermediações; então, um investidor, por meio de uma corretora de valores, realiza a compra e venda de ações negociadas na B3 (TORO, 2018). A bolsa de valores brasileira, assim como as outras bolsas ao redor do mundo, só opera durante a semana, ou seja, entre segunda e sexta-feira, segundo Riconnect (2020), e não funciona durante feriados.

2.1.2 Gerência de dados na Web

O termo gerência de dados da *Web* refere-se ao estudo de técnicas para a solução de problemas relacionados à coleta, à extração, à modelagem, ao armazenamento, à transformação e à integração dos dados disponíveis na *Web*, conforme explica Abiteboul et al. (2011). O processo de gerenciar dados a partir da *Web* está muito associado à descoberta de conhecimento, exemplificada na Figura 2.2, segundo Miller (2018).

Figura 2.2 – Descoberta de conhecimento em base de dados.



Fonte: Miller (2018).

De acordo com a Figura 2.2, são apresentadas as etapas da Descoberta de Conhecimento a partir da definição sugerida por Fayyad et al. (1996). Para os autores, estas etapas definem um processo complexo que objetiva a extração de informações em grandes volumes de dados, onde cada uma das etapas desempenha uma função específica. A Figura 2.2 destaca os seguintes processos: (a) seleção dos dados, etapa que visa selecionar na base ou no repositório alvo o conjunto de dados contendo todas as possíveis variáveis envolvidas; (b) pré-processamento dos dados, onde os dados coletados passam por um filtro onde são removidos características indesejadas, alguns ruídos ou até mesmo informações incompletas; (c) transformação dos dados, etapa que localiza características úteis que representem os dados de acordo com a tarefa escolhida, além de formatá-los e armazená-los; e (d) mineração dos dados, onde são aplicados algoritmos para extração de conhecimento em cima dos dados transformados anteriormente. Regras úteis e novos padrões podem ser encontrados com base nos resultados obtidos ao longo do fluxo destas etapas.

Nas próximas subseções, são apresentados com mais detalhes conceitos importantes do processo de gerência de dados na *Web* que estão diretamente envolvidos no presente trabalho: a Subseção 2.1.2.1 discorre sobre a coleta dos dados, que é uma forma de seleção dos dados, a Subseção 2.1.2.2 descreve a extração dos dados, que é um processo intrínseco da etapa de pré-processamento e transformação dos dados, e a Subseção 2.1.2.3 aborda a mineração de dados.

2.1.2.1 Coleta de dados

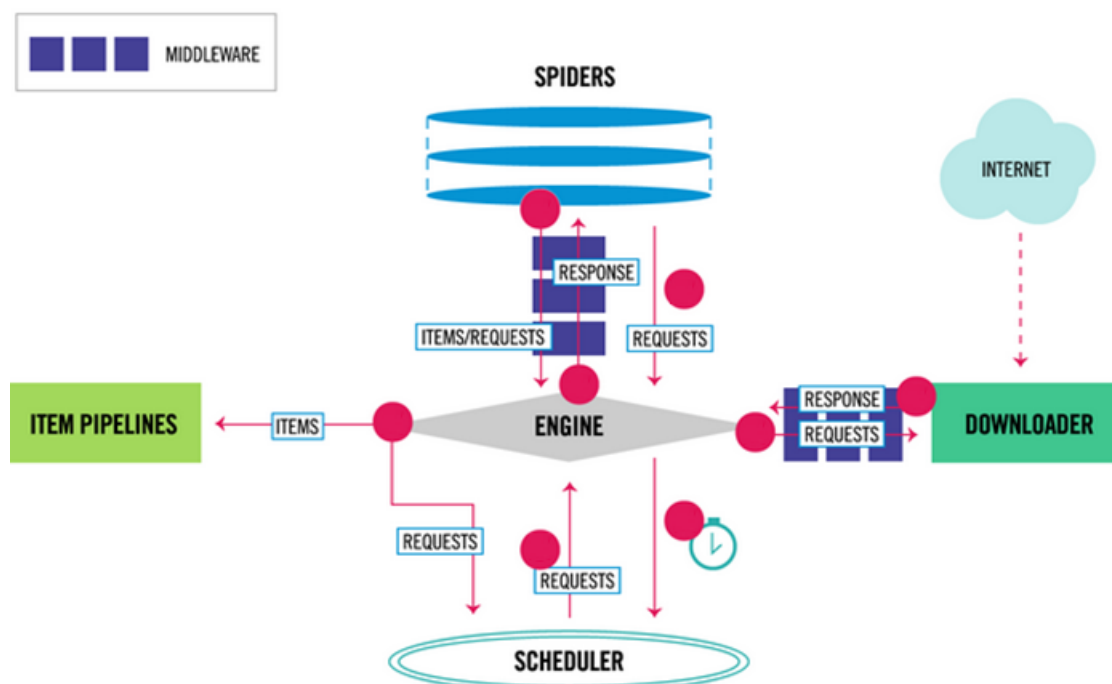
A coleta de dados na *Web* é um processo orientado a captação de qualquer tipo de informação exposta nos meios digitais (FERREIRA, 2020). O autor descreve que essa é uma prática fundamental para a realização de análises de comportamento e preferências, visto que empresas podem obter informações valiosas que evidenciam seu público de alguma forma.

Os fatores que motivam os estudos e a prática da coleta de dados são diversos (EMERSON, 2019). O autor descreve que a maior parte desses estudos são feitos para empresas que desejam acompanhar os preços e tendências do mercado, campanhas políticas, análises de sentimento do público *online* e estudos de pesquisa e desenvolvimento.

A coleta de dados em sites, ou *Web Scraping*, como também é conhecida, dá-se por meio de processos automatizados (ROCHA, 2018). Estes processos utilizam um rastreador que aplica um tipo de "raspagem" em todas as informações da página selecionada que, por meio de parâmetros pré-definidos, armazenam os dados de forma estruturada em uma planilha local ou até mesmo um banco de dados específico. Existem distintos *frameworks* ou ferramentas que realizam processos de *Web Scraping*; dentre eles, particularmente, o *framework Scrapy* que é utilizado pela plataforma proposta neste trabalho.

O *Scrapy* é um *framework* que, segundo Fan (2018), consiste em uma estrutura popular de rede baseada em eventos. Esta estrutura pode ser visualizada a partir da Figura 2.3.

Figura 2.3 – Arquitetura do *framework Scrapy*.



Fonte: Adaptado de Fan (2018)

De acordo com a Figura 2.3, a arquitetura do *framework* divide-se em cinco mecanismos principais: (a) *Engine*, mecanismo principal, responsável por controlar o fluxo de todos os

outros componentes disparando os eventos necessários; (b) *Scheduler*, mecanismo responsável por receber as solicitações diversas do mecanismo principal e as enfileirar para que sejam processadas posteriormente; (c) *Downloader*, mecanismo responsável por buscar na *Web* páginas com o conteúdo relacionado aos parâmetros pré-estabelecidos; (d) *Spiders*, classes criadas para processar as páginas coletadas pelo *Downloader* e realizar a extração dos dados solicitados; e (e) *Item pipeline*, mecanismo responsável por processar os dados após terem sido extraídos pelos *Spiders*. Esta última etapa também inclui o pré-processamento dos dados.

2.1.2.2 Extração de dados

A extração de dados ou extração de informações, como também é chamada, é o processo de encontrar, em um grande volume de dados, informações com características específicas, conforme descrito por Álvarez (2007). O autor explica que os documentos, onde se aplica esse procedimento, podem apresentar uma estruturação na organização dos dados como podem também ser totalmente desestruturados, o que dificulta o processo.

Para Silva, Barros e Prudêncio (2005), os textos podem ser classificados como estruturados, semiestruturados e não estruturados. Para o autor, os textos estruturados seguem um formato inflexível como, por exemplo, uma página *HyperText Markup Language* (HTML)² gerada de um banco de dados. Por terem uma estrutura familiar e imutável, a extração pode ocorrer com facilidade a partir de regras baseadas em delimitadores da linguagem ou da ocorrência de termos. Em contrapartida, os textos não estruturados não seguem um padrão na sua formação já que as sentenças podem ser descritas em alguma linguagem natural, inviabilizando a extração com base apenas na formatação. E quanto a textos semiestruturados, como o nome já diz, possuem algum grau de estruturação atrelado a irregularidades, como dados em ordens variadas, campos com valores nulos e ausências de delimitadores entre as informações, significando que uma análise deve ser feita para que a sua estrutura possa ser identificada e extraída (BUNEMAN, 1997).

A informação extraída é determinada por um conjunto de padrões e regras específicas de seu domínio e estes padrões podem ser definidos manual ou internamente por algum especialista conforme relata Riloff e Lehnert (1994). O objetivo das técnicas de extração de dados é a construção de sistemas que consigam encontrar, a partir da combinação de padrões selecionados, informações relevantes enquanto ignoram informações irrelevantes (COWIE; LEHNERT, 1996). Os autores explicam que, pelo fato da extração de informação ser um processo de selecionar estruturas e combinar dados encontrados em textos, a sua produção final, uma vez estruturada, é um procedimento com o claro objetivo de criação de um repositório ou de um banco de dados.

2.1.2.3 Mineração de dados

Mineração de dados ou *Data Mining* é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto

² HTML é uma linguagem de marcação utilizada na construção de páginas na *Web*.

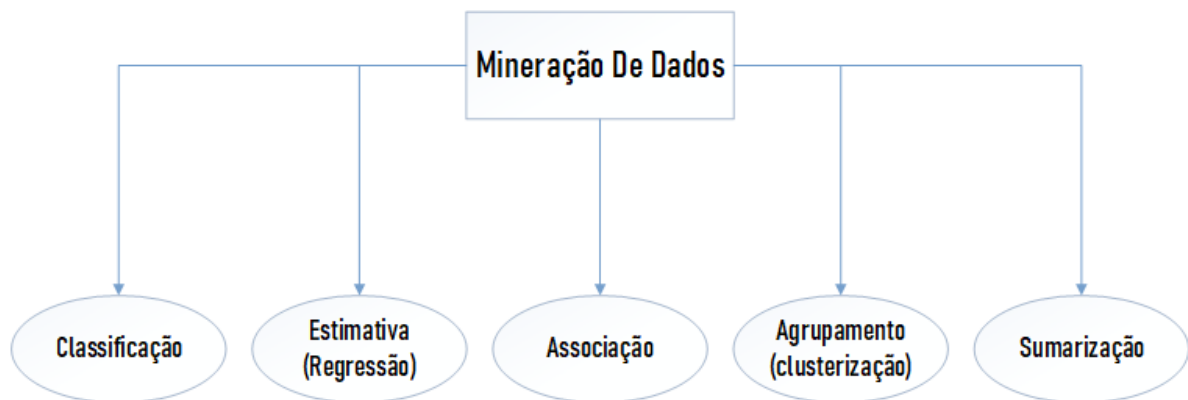
úteis quanto compreensíveis (HAND; MANNILA; SMYTH, 2001). Fayyad, Piatetsky-Shapiro e Smyth (1996) definem a mineração de dados como sendo um passo do processo de Descoberta de Conhecimento, onde a análise dos dados e a aplicação de algoritmos de descoberta são realizados para que seja possível produzir, como resultado, um conjunto de padrões para quem a utiliza.

Pelo fato da quantidade de dados gerados nas mais diferentes áreas do conhecimento ter crescido de maneira expressiva, segundo o autor Bakshi (2012), o desafio não está em apenas armazenar e gerenciar esse vasto volume de dados, mas também em analisar e extrair informação significativa. Neste sentido, modelos computacionais vêm sendo desenvolvidos com o intuito de simplificar o entendimento da relação entre as variáveis em grandes conjuntos de dados brutos (LIMA et al., 2016).

Neste contexto, o principal objetivo da mineração de dados, segundo Dias et al. (2001), é: fornecer subsídios para que, a partir de um histórico, sejam feitas previsões de tendências futuras e seja descoberta qual é a relação entre os dados. O autor discorre que os resultados da mineração de dados podem ser utilizados no gerenciamento de informação, processamento de pedidos de informação, tomada de decisão, controle de processos, dentre muitas outras aplicações.

As técnicas de mineração de dados podem ser aplicadas a tarefas, ou seja, para algum tipo de regularidade ou categoria de padrões desejada. A Figura 2.4 apresenta alguns exemplos de tarefas, como a de classificação, de estimativa, de associação, de agrupamento e de sumarização.

Figura 2.4 – Tarefas desempenhadas por técnicas de mineração de dados.



Fonte: Adaptado de Dias et al. (2001).

De acordo com a Figura 2.4, algumas tarefas realizadas por técnicas de mineração de dados, segundo Dias et al. (2001), podem ser descritas da seguinte maneira: (a) classificação, onde há a construção de um modelo que possa ser aplicado a dados não classificados a fim de categorizá-los em classes; (b) estimativa, onde se define um valor para alguma variável contínua desconhecida; (c) associação, onde se determina quais itens tendem a ser adquiridos juntos em uma mesma transação; (d) agrupamento, onde há o processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos; e (e) sumarização, onde são aplicadas técnicas para se encontrar uma descrição compacta para um subconjunto de dados.

2.1.3 Arquitetura de microserviços

A arquitetura de microserviços é um modelo que tem, por objetivo, estruturar uma aplicação em uma coleção de serviços e, com isso, oferecer facilidade de manutenibilidade e testabilidade, implementação independente, baixo acoplamento entre os serviços e uma melhor organização dos recursos (RICHARDSON, 2015). Para Papazoglou (2003), a arquitetura de microserviços é uma abordagem ao desenvolvimento de *software* e infraestrutura de suporte como um conjunto interconectado de serviços, acessíveis por meio de interfaces e protocolos padronizados de mensagem. Esta arquitetura, segundo o autor, é particularmente aplicável quando múltiplas aplicações e serviços, que rodam em diferentes plataformas e tecnologias, precisam comunicar-se.

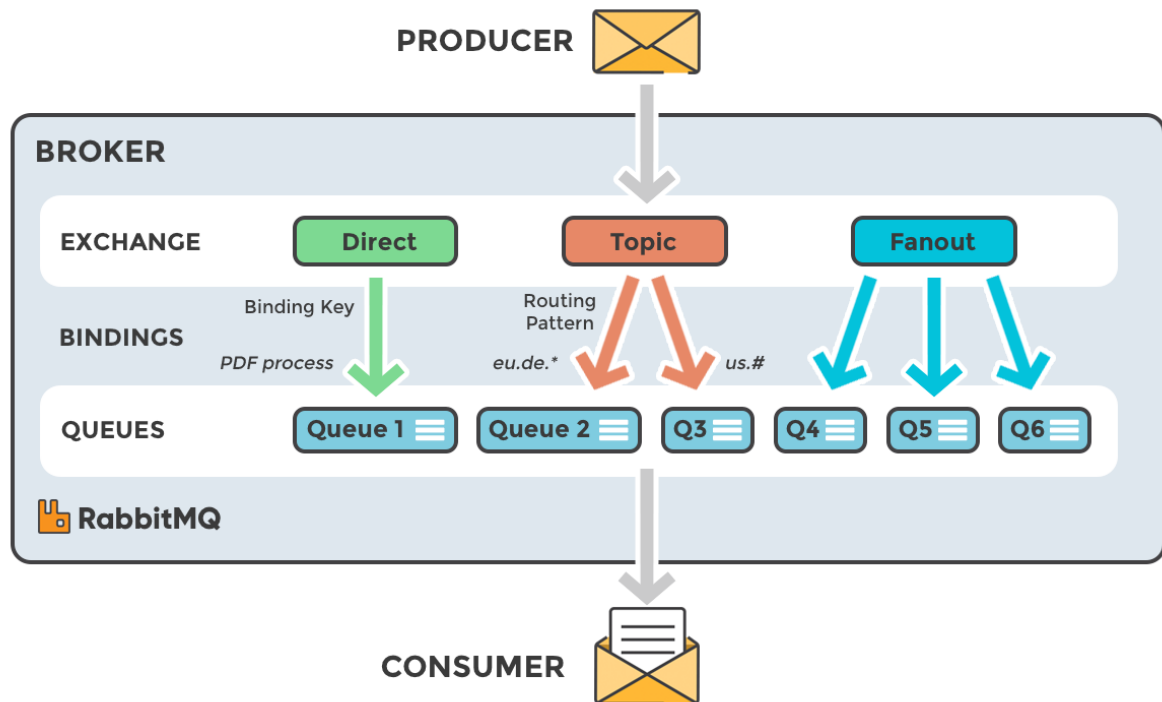
Normalmente, a comunicação entre os serviços internos desta arquitetura é realizada pelo protocolo *Hyper Text Transference Protocol* (HTTP) ou via serviços de mensagens, utilizando, por exemplo, o protocolo *Advanced Message Queuing Protocol* (AMQP) (SORENSEN, 2017). Segundo o autor, esta segunda opção é mais adotada, tendo em vista que ela provê mais confiabilidade, disponibilidade, escalabilidade, rapidez e resiliência quando comparados com as chamadas HTTP.

O *RabbitMQ* é um *framework* que, segundo Weber (2016), consiste em um *message broker* e gerenciador de filas que permite o envio e recebimento de mensagens entre aplicações, utilizando o protocolo AMQP. As vantagens de utilizar um *message broker* como o *RabbitMQ*, para Silva (2017), é o suporte que ele tem para várias linguagens de programação. Os serviços podem ter seu contexto isolado e altamente plugável no ecossistema. A sua estrutura pode ser visualizada na Figura 2.5

De acordo com a Figura 2.5, a arquitetura divide-se em cinco partes principais que compõem o fluxo de mensagens no *RabbitMQ*: (a) *Producer* é o serviço ou aplicação que publicará as mensagens para uma *exchange*; (b) *Exchange* é a entidade para onde o *producer* enviará as mensagens, sendo responsável por enviar as mensagens para a fila³ correta, de acordo com as regras definidas pelo seu tipo que pode ser: *Direct*, *Fanout*, *Topic*, *Headers*; segundo Johansson (2019), cada tipo diz respeito à forma de roteamento das mensagens para as filas de destino; (c) *Binding* constituem as regras que especificam como as *exchanges* devem rotear as mensagens para as filas; (d) *Queue* é a entidade que armazena as mensagens provenientes de uma ou mais *exchanges*; e (e) *Consumer* é o serviço ou aplicação que consumirá as mensagens da *queue*.

Neste trabalho, a arquitetura de microserviços foi utilizada como modelo para a implementação da plataforma proposta. Assim, todos os serviços necessários para o seu funcionamento foram isolados em módulos e a comunicação entre esses módulos foi realizada por mensageria, utilizando o protocolo AMQP.

³ Fila é uma estrutura de dados onde o primeiro elemento a ser inserido será o primeiro a ser retirado, ou seja, adiciona-se itens no fim e remove-se do início.

Figura 2.5 – Arquitetura do *framework RabbitMQ*

Fonte: Johansson (2019).

2.1.4 *Twitter*

Tendo sido fundado em 2006 por Jack Dorsey, Evan Williams, Biz Stone e Noah Glass, com a ideia de ser uma rede social e um servidor para *microblogging* que permitisse criar conexões entre seus usuários, o *Twitter* hoje agrega mais de 300 milhões de usuários ativos conforme aponta Statusbrew (2018). Limitado a 280 caracteres, os *tweets*, como são chamadas as mensagens publicadas na rede social, podem ser lidas por todos os seus seguidores e abordam diversos temas.

As particularidades do *Twitter*, no âmbito dos elementos hipertextuais, moldam sua comunicação. Lé (2012) menciona algumas funções hipertextuais de destaque, como: @, que identifica os usuários e suas contas, as *hashtags*, simbolizadas pelo #, e a menção a textos de outros usuários, usualmente chamado de *Retweets* (RT). Outra particularidade do *Twitter* são os chamados *Trending Topics*, que funcionam como uma espécie de *ranking* durante um determinado período de tempo para os assuntos mais comentados na rede (JUNQUEIRA, 2020). No *Twitter*, também podem ser usados os chamados *emojis*, que são uma nova geração de *emoticons*, uma representação pictorial de uma expressão facial, que tenta expressar sentimentos em suas mensagens. Segundo Smailović et al. (2015), em um espaço de 2 anos, mais de dez bilhões de *emojis* foram usados no *Twitter*.

O *Twitter*, segundo Santos, Laender e Pereira (2015), vem sendo utilizado mais expressivamente em pesquisas que objetivam extrair padrões de comportamento, influência e tendências

de mercado, pois se tornou uma importante fonte de informação. O fato é que os usuários, na maioria dos casos, afirmam que veem uma marca de forma mais positiva quando há resposta ao seu *tweet* por ela (PATEL, 2019). O autor ainda diz que a rede social é utilizada em média por 65% das empresas americanas em estratégias de *marketing*.

A rede social possui também uma *Application Programming Interface* (API) própria, que auxilia o acesso a sua base de dados e postagens. A partir de alguns argumentos necessários para a busca, a API retorna ao requisitante uma massa de dados que pode ser utilizada para várias finalidades (XAVIER, 2020). A única restrição é que esta API é gratuita até certo ponto: há um limite de dados que podem ser trafegados em sua versão gratuita; para uma busca mais elaborada e com maior volume de resultados, deve-se adquirir a sua versão paga (JOHNSON; ROTH, 2019).

Neste trabalho, o *Twitter* é utilizado como base de dados da plataforma proposta. Os *tweets* coletados desta rede social, são a fonte de dados responsável para a definição dos resultados apresentados ao usuário final.

2.1.5 Análise de sentimentos

Conhecida também como Mineração de opinião, a análise de sentimentos é uma área dentro do Processamento de Linguagem Natural (PLN) que analisa opiniões, sentimentos, avaliações, emoções e atitudes das pessoas em relação a alguma coisa, seja um produto, um serviço, organizações ou até mesmo outras pessoas (LIU, 2012). De acordo com o autor, a análise de sentimentos pode ser feita sobre qualquer fração de texto e de qualquer tamanho, tendo como exemplo comentários em redes sociais, *tweets*, notícias ou até mesmo em páginas completas da *Web*.

Além de identificar a opinião, a análise de sentimentos, segundo MTK (2018), extrai outros atributos do texto, como: o que está sendo falado, quem é a pessoa que expressa tal opinião e qual a polaridade do que foi dito. O autor cita a relevância de uma análise de sentimento bem feita pois, com o crescimento da Internet e a popularização das redes sociais, todo o conteúdo pode se tornar muito útil para empresas em análises de marketing, relações públicas, *feedbacks* de produtos e atendimento ao cliente. Chen e Zimbra (2010) apontam que grandes empresas, após perceberem o valor das opiniões expressas na *Web* e como estas opiniões as afetam de forma negativa e positiva, começaram a investir em pesquisas de mineração de opinião.

O objetivo principal da análise de sentimentos, segundo Gomes (2019), é a polarização das sentenças extraídas. Para o autor, esta polaridade pode ser uma classificação do conteúdo do texto em frases positivas, negativas ou neutras. A vantagem desta análise polar é a objetividade. A maior parte dos métodos de análise de sentimentos, que podem ser aplicados a qualquer tipo de fonte textual como, por exemplo, resenhas de livros, filmes, comentários e *tweets*, são categorizados, segundo Silva (2016), por paradigmas de classificação: (a) supervisionada, onde

já se tem uma base previamente rotulada e faz-se uso de algoritmos de aprendizado de máquina; (b) orientada a dicionários léxicos, os quais são uma lista de termos já classificados em positivos e negativos, criados manualmente ou de forma automatizada e que auxiliam no cálculo das polaridades das frases analisadas; e (c) supervisionada híbrida, quando há a união de algum algoritmo de aprendizado de máquina em conjunto com um dicionário léxico.

Analisar sentimentos na *Web* não é uma tarefa fácil, principalmente em redes sociais, aponta Santos (2010), onde os textos podem estar dispostos de qualquer forma. Para o autor, o empecilho encontra-se desde a necessidade de filtrar os textos opinativos dos objetivos até a forma como são descritos. Textos mal configurados, como os de redes sociais, carecem de um processo bem cuidadoso de tratamento e limpeza anterior à análise para que os resultados sejam satisfatórios.

Neste trabalho, a técnica de análise de sentimentos foi aplicada nos *tweets* coletados como forma de identificar a polaridade presente no texto e nos *emoticons*. De uma forma geral, o objetivo foi retornar ao usuário final o quão positivo ou negativo foi aquela publicação.

2.2 Trabalhos Relacionados

Na literatura, há diversos trabalhos que relacionam o comportamento dos usuários nas redes sociais conhecidas com a movimentação do mercado financeiro, a fim de encontrar um fator influenciador no sentimento expresso pelos usuários que justifique a geografia do mercado e os impactos observáveis, como a taxa de retorno das ações. A partir de estudos acerca de pesquisas que englobam as áreas de classificação, análise de sentimento, mercado financeiro e mineração de dados em mídias sociais, relevantes trabalhos relacionados foram encontrados. A seguir são apresentados e discutidos alguns desses.

Lima et al. (2016) desenvolveram um modelo que auxilia no processo de tomada de decisão na compra e venda de ações, baseando-se na opinião gerada a partir do sentimento humano coletivo expresso no *Twitter*. Os autores efetivaram a construção do *corpus* com a extração de 55.000 *tweets* por meio da API disponibilizada pela própria rede social; para tal, restringiram a busca para as publicações do idioma português originários do Brasil e que incluíssem a palavra-chave “Petrobras”. A partir deste *corpus* e do pré-processamento dos dados, etapa que busca suprimir caracteres desnecessários, foi selecionada a ferramenta utilizada para mineração de opinião, o léxico *Sentiment140*. A análise de sentimentos deu-se em duas etapas: inicialmente, o autor agrupou os *tweets* por dia e os analisou separadamente, objetivando o seu sentimento individual; em seguida, encontrou o sentimento coletivo diário. O autor também coletou dados do Mercado referente à ação “PETR4”⁴ do “IBOVESPA”⁵, dentro do período estudado, para que servissem de indicadores assim como os dados contendo o sentimento coletivo. Fez uso de

⁴ Código da ação da empresa Petrobras.

⁵ O principal indicador de desempenho das ações negociadas na bolsa brasileira.

técnicas de aprendizado de máquina, como o classificador *Support Vector Machine* (SVM) e aprendizado supervisionado⁶, obtendo resultados de até 70% em acurácia⁷. Por fim, os resultados validaram o *Twitter* como um termômetro social e uma fonte rica em informação, embora seja necessária uma atenção maior no pré-processamento dos dados devido à informalidade inerente aos usuários; ademais, verificou-se que eventualmente o humor coletivo, indicador coletado em etapas anteriores, não reflete o sentimento do mercado especificamente para aquele ativo.

Aguiar (2012) aplicou análise de sentimentos em relatórios da administração de firmas brasileiras, com o intuito de extrair a opinião presente nos textos e encontrar evidências que explicassem as variáveis do mercado de ações: volume de negócios, retorno das ações e a volatilidade das negociações. O autor utilizou uma base de dados com 1902 relatórios coletados manualmente no site da Bovespa, em um período entre 1995 e 2009; após reunidos os relatórios, foram coletados, para o mesmo período, dados com o preço de fechamento dos ativos, volume de negócios, tamanho da firma e *book-to-market*⁸ das firmas selecionadas. A técnica utilizada para mensurar a opinião dos relatórios foi a *Vector Space Model* (VSM) que, de uma forma geral, recupera a frequência das palavras nos textos, transformando-os em dados quantitativos. Posteriormente, os relatórios deram origem a um dicionário de termos, com polaridade positiva, negativa, litigiosa, incerteza e verbos modais. A polarização e criação das listas de palavras foi feita manualmente por dois pesquisadores, ou seja, de forma indutiva. O autor destaca a necessidade de um esquema de *term weighting* na etapa de classificação, método que determina o peso do termo em um *corpus*; para tal, utiliza o modelo proposto por Loughran e McDonald (2011) que, além de manifestar bons resultados nos testes realizados, atenua a desproporção no tamanho dos documentos. Como resultado da pesquisa, Aguiar (2012) diz que há uma relação entre as negociações no mercado de ações e nos relatórios divulgados pelas firmas brasileiras, o que permitiu relacionar as variáveis desejadas. Ele concluiu que as notícias positivas relacionam-se com a diminuição da volatilidade no mercado e uma estabilização dos negócios de compra e venda de ações. Para as notícias negativas, não foi encontrada evidência de uma relação direta. Como contribuição, o autor compartilhou as listas de palavras criadas ao decorrer da pesquisa, para que possam ser utilizadas em trabalhos futuros.

Peres, Vieira e Bordini (2019) realizaram um estudo para determinar se diferentes dicionários léxicos de domínio geral conseguem interpretar textos e notícias de domínio específico como o financeiro, com base em um modelo próprio desenvolvido. Os dicionários escolhidos foram o *OpLexicon*, *SentiLex* e *UniLex*. Além disso, propuseram construir um dicionário léxico utilizando técnicas comuns de PLN como *Bag of Words*⁹ e *Term Frequency – Inverse Document*

⁶ Tarefa de aprendizado de máquina que consiste em aprender uma função que mapeia uma entrada para uma saída com base em pares de entrada-saída de exemplo.

⁷ Proximidade entre o valor obtido experimentalmente e o valor verdadeiro na medição de uma grandeza física.

⁸ Métrica que mede a razão entre o valor de mercado de uma empresa e o seu valor patrimonial.

⁹ Modelo onde o texto é representado como o conjunto de suas palavras, sem considerar a ordem e sua gramática, mantendo apenas o número de vezes que determinada palavra aparece no conjunto.

Frequency (TF-IDF)¹⁰; entretanto, todo o processo foi analisado em cima de uma base de dados já rotulada e um *corpus* de opiniões de investidores disponível na ferramenta *TradingView*¹¹. Como ponto negativo, não se utilizou o processo de normalização de textos na etapa de pré-processamento dos dados, técnica de caráter relevante e que influenciaria em melhores resultados. A análise dos dicionários propostos foi feita com base nas métricas precisão¹², recall¹³ e acurácia. Concluiu-se que o *SentiLex* obteve destaque com melhor resultado, comparado aos demais, com aproximadamente 70% de precisão.

Neto e Arruda (2018) implementaram uma aplicação *Web* que analisa e exibe o sentimento das publicações feitas na rede social *Twitter* em tempo real. É uma ferramenta voltada ao mercado de *Bitcoin*, proposta para auxiliar investidores a investigarem a relação entre as notícias, publicações e a variação do preço da moeda. A mineração dos dados foi feita baseada em um extrator a partir da API disponibilizada pelo *Twitter* no formato *Streaming*¹⁴, onde os dados são enviados do servidor do *Twitter* pra o servidor dedicado sem interrupção. A etapa de análise de sentimentos deu-se pelo léxico *Vader* da biblioteca *vaderSentiment* que, aliado às técnicas de PLN, retorna os *scores* positivos, negativos e neutros do texto analisado. Como resultado, os autores não conseguiram prever, com bastante antecedência, a queda ou ascensão da moeda e nem que, de fato, a variação do preço do *Bitcoin* é causada pelo sentimento das publicações; porém, foi possível prever a existência de uma relação entre os dois e que o sentimento acompanhou os movimentos de alta e baixa da moeda, indicando anseios, medos e desejos de seus investidores.

No trabalho de Machado, Silva et al. (2017), investigou-se a dinâmica do mercado financeiro por meio de análise de 54 indústrias Brasileiras com ações negociadas na “BM&FBOVESPA”¹⁵ e do sentimento textual dos relatórios de desempenho trimestral das mesmas. Para tal, foram utilizadas métricas sugeridas por Loughran e McDonald (2011) como a análise de sentimento das palavras. O comportamento do mercado foi avaliado por meio de uma *proxy*¹⁶ para retorno anormal com cálculos baseados no modelo de mercado. Para o retorno anormal do mercado, que basicamente é a taxa de retorno da ação menos o retorno esperado, os autores utilizaram procedimentos de estudo de evento, embasados por aplicações em finanças e contabilidade, que podem ser utilizados para investigar o comportamento do mercado. Para o sentimento textual dos relatórios, foi utilizada a análise de sentimentos que, por sua vez, captura a opinião expressa em um texto com base na sua polaridade apoiada pela técnica VSM e pelo modelo proposto por Loughran e McDonald (2011) que, com base na frequência de cada palavra presente em um

¹⁰ Medida estatística que indica a importância de uma palavra em um documento em relação ao conjunto de documentos.

¹¹ <https://br.tradingview.com/>

¹² Métrica que expressa o percentual de resultados corretos dentre todos os resultados apresentados.

¹³ Métrica que indica a relação entre as previsões positivas realizadas corretamente e todas as previsões que realmente são positivas.

¹⁴ Serviço de transmissão instantânea de dados por meio da rede.

¹⁵ Bolsa de Valores, Mercadorias e Futuros.

¹⁶ *Proxy* em finanças é a hipótese utilizada como referência para se estimar o valor de uma variável, antes de conhecê-la.

determinado texto, calcula-se o seu peso em relação ao mesmo. Por fim, os autores concluíram que o tom dos relatórios, quando o desempenho acionário é inferior ao previsto, impacta negativamente no mercado, porém nada é garantido quanto ao lado positivo. Analisando o sentimento textual, viu-se uma limitação quando aplicado, pois além da amostra e do período de análise ser pequeno, a métrica utilizada não replicou corretamente o processo de leitura humana.

Terra (2016) analisou a rede social *Twitter* por meio do desenvolvimento de uma ferramenta de extração, de modo a encontrar e traçar perfis de trânsito da Serra Gaúcha que auxiliassem na detecção de adversidades. A extração de *tweets* foi realizada via API e, para mineração *Web*, foram utilizados os algoritmos *Apriori* e *Tertius*; estes algoritmos objetivam garantir, em uma busca, os itens de maior frequência e uma maior confiança avaliativa nos resultados a partir de palavras-chave pré-selecionadas. O autor não focou apenas em notícias oriundas de jornais e revistas que atuam na rede social, o que fez com que a confiabilidade do conteúdo publicado diminuísse. Ademais, a extração gerou uma base pouco volumosa, cerca de 1500 mensagens, que impactou negativamente no processamento do algoritmo de associação *Apriori* apresentando resultados distorcidos; já o *Tertius* atingiu o objetivo traçando alguns perfis do trânsito conforme esperado. Como resultado dessa pesquisa, Terra (2016) observou que o volume de dados impacta na avaliação e que a rede social escolhida em conjunto com o tema abordado não trouxe resultados relevantes conforme era esperado no início da concepção do trabalho.

Sousa (2012) implementou uma aplicação que polariza opiniões contidas em textos extraídos do *Twitter* utilizando duas abordagens para classificação dos mesmos. A extração dos *tweets* foi feita com apoio da API disponibilizada pela própria rede social e de duas formas diferentes: por *tweets* recentes, onde é possível recuperar postagens feitas nos últimos dias, ou a extração em tempo real, onde a recuperação das postagens é feita na medida em que são lançadas na rede social. Inicialmente, os dados foram classificados entre neutros e opinativos com o intuito de filtrar textos mais bem elaborados e que expressavam a opinião do usuário. Na sequência, os considerados opinativos deram origem à base de dados populada por textos positivos e negativos; para tal separação, foram utilizadas técnicas de aprendizado de máquina como o SVM, onde a minimização do erro com relação ao conjunto de dados de treinamento e de teste é objetivada. A aplicação obteve bons resultados chegando a atingir aproximadamente 70% de precisão, mesmo com um tempo de execução alto para o algoritmo de treinamento; segundo a autora, o tempo elevado é justificado pelo alto número de instâncias analisadas: aproximadamente 13.000. A má formulação dos *tweets* extraídos foi apontada como um dos principais problemas, visto que o usuário tem a liberdade de escrever da forma que desejar. O armazenamento de dados do sistema foi feito em arquivos de texto, o que inviabiliza a escalabilidade da aplicação, pois o volume de dados na *Web*, principalmente em redes sociais, é imensurável e, para uma análise mais robusta e contínua dos resultados, um Sistema de Gerenciamento de Banco de Dados (SGBD) seria mais eficiente.

Por fim, Felipe (2017) investigou o processo de alocação de capital no mercado de

*crowdfunding*¹⁷. Para tal, o autor utilizou duas abordagens, as quais tinham como objetivo a busca por referências que auxiliassem na redução do choque de informações no processo de *funding*¹⁸ coletivo. A primeira abordagem envolveu a análise semântica das notícias, de modo a entender qual o comportamento dos investidores e parametrizar informações que os levam ao investimento, bem como estratégias de redução de risco e exposição do negócio a fatores externos. A segunda abordagem, consiste na observação da geografia dos investimentos; foram observadas características geográficas das contribuições realizadas que, quando analisadas, revelariam informações econômicas e demográficas como, por exemplo, a melhor escolha de localização do empreendimento. O autor dividiu a pesquisa em quatro ensaios. O primeiro ensaio diz respeito a pesquisas e discussões teóricas sobre o tema *crowdfunding*. O segundo ensaio expõe as causas do sucesso dos projetos de *crowdfunding*: o autor utilizou regressão *logit*¹⁹ e *survival analysis*²⁰ em 4.262 projetos coletados da plataforma brasileira de *crowdfunding*, *Catarse*; os resultados indicaram que projetos de arte tendem menos a prosperar e que os que tendem a obter sucesso são aqueles desenvolvidos em regiões com maior concentração de renda *per capita*. O terceiro ensaio investiga o efeito das características geográficas das contribuições e o sentimento textual de notícias sobre *reward crowdfunding*²¹. Neste ensaio, a base de dados foi formada por dados relativos aos projetos da plataforma *Catarse*, dados socioeconômicos obtidos junto ao Instituto Brasileiro de Geografia (IBGE) das cidades nas quais as operações de investimento foram realizadas, e por dados relativos às notícias do jornal "O Estado de São Paulo". Cerca de 1.266 notícias foram coletadas manualmente e submetidas a uma análise de sentimento baseados no algoritmo de Loughran e McDonald (2011). Os resultados indicaram que uma maior distância entre o empreendedor e o investidor faz com que um maior aporte seja desestimulado. A análise de sentimentos aplicada conclui que, em um cenário negativo, os investimentos tendem a ser menores do que em um cenário positivo. Por último, o quarto ensaio verifica, por meio de análise semântica e da geografia dos aportes, o valor dos investimentos alocados em empreendimentos de *equity crowdfunding*²². Os resultados indicaram que os atributos dos empreendimentos possuem impacto no valor dos aportes e que o investidor, diante de notícias com uma maior quantidade de palavras positivas, é incentivado a aplicar maiores quantias financeiras.

Relacionando os trabalhos citados, foi produzido um comparativo ilustrado na Tabela 2.1: avaliou-se os pontos em comum de cada trabalho, qual a relação entre eles e como eles contribuem

¹⁷ *Crowdfunding* ou financiamento coletivo, como também é conhecido, consiste na obtenção de capital para iniciativas de interesse coletivo por meio da agregação de múltiplas fontes de financiamento, em geral pessoas físicas interessadas na iniciativa.

¹⁸ *Funding* é a captação de recursos financeiros para o investimento específico pré-acordado de uma empresa.

¹⁹ A regressão logística é um recurso que permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis explanatórias.

²⁰ A análise de sobrevivência corresponde a um conjunto de abordagens estatísticas usadas para investigar o tempo que um evento de interesse leva para ocorrer.

²¹ O *reward crowdfunding* consiste em indivíduos que fazem doações para um projeto ou empresa com a expectativa de receber uma recompensa não financeira em troca.

²² O *equity crowdfunding* consiste na oferta *online* de valores mobiliários de empresas privadas a um grupo de pessoas para investimento e, portanto, faz parte do mercado de capitais.

para o presente trabalho. A Tabela 2.1 apresenta os seguintes critérios: (a) extração ilimitada, que indica que o processo de coleta e extração de dados faz uso de técnicas e procedimentos ilimitados, independente da quantidade de dados; (b) Idioma pt-br, que considera que a proposta, estudo ou ferramenta é voltada para o idioma português; (c) *Twitter*, que indica que a rede social, a qual os experimentos foram realizados, é o *Twitter*; (d) Finanças, que indica que o conteúdo abordado é relacionado ao tema finanças; e (e) automatizado, que indica que todos os processos executam de forma independente uns dos outros, ou seja, não necessitam da ação humana a cada etapa.

Tabela 2.1 – Comparativo de Trabalhos Relacionados

Autores	Características				
	Extração ilimitada	Idioma pt-br	<i>Twitter</i>	Finanças	Automatizado
Lima et al. (2016)		x	x	x	
Aguiar (2012)		x		x	
Peres, Vieira e Bordini (2019)				x	
Neto e Arruda (2018)			x		x
Machado, Silva et al. (2017)		x		x	
Terra (2016)		x	x		
Sousa (2012)			x		
Felipe (2017)		x		x	
Plataforma proposta	x	x	x	x	x

Fonte: Elaborado pelo autor.

De acordo com a Tabela 2.1, observa-se, de uma forma geral, que os critérios idioma pt-br, *Twitter* e finanças foram abordados pela maior parte dos trabalhos relacionados descritos neste documento. A ideia de realizar um estudo no idioma português é de extrema importância, visto que o idioma não é tão explorado em pesquisas desta área. A rede social *Twitter* tem se mostrado uma grande aliada quando o assunto é análise de dados, pelo fato do grande volume de informações dispostas diariamente. O tema finanças, mesmo possuindo várias vertentes, relaciona-se no vocabulário, em parte dos problemas encontrados e consequentemente nas soluções propostas. Para a construção da plataforma proposta, todos estes critérios também são atendidos.

Os trabalhos citados utilizaram um meio limitado para a extração dos dados, sendo que alguns recorreram a API's disponibilizadas pelos sites, onde o acesso é restringido pela própria API, e outros utilizaram a extração manual que é uma prática inviável quando se almeja uma base de dados volumosa. A mineração dos dados é a etapa do processo caracterizada como sendo a mais importante e um volume sucinto de dados pode influenciar nos resultados esperados, conforme visto anteriormente em alguns trabalhos. Desta forma, neste trabalho, diferentemente dos demais, foi criado um módulo de *scraping* que, por meio da estrutura da página *Web* e como ela se comporta, consegue extrair todos os dados sem nenhum tipo de restrição ou impedimento.

Outro ponto importante observado foi a automatização de todo o processo. Alguns trabalhos executaram de forma separada suas etapas, criando uma independência entre os processos.

Esse tipo de modelo adotado geralmente é utilizado quando o objetivo é a pesquisa, onde se testa vários algoritmos separadamente e não carece de uma utilização completa por um usuário final. Para o presente trabalho, a automatização fez-se necessária, pois o objetivo é empregar e disponibilizar, de forma simples, todos os processos utilizados em uma única plataforma.

3 Desenvolvimento

Neste capítulo, é abordado o desenvolvimento da plataforma de auxílio ao investidor no processo de compra e venda de ativos do mercado financeiro brasileiro. Para tanto, a Seção 3.1 descreve a sua arquitetura completa de funcionamento e a Seção 3.2 apresenta a interface da plataforma, bem como a parametrização necessária para executá-la.

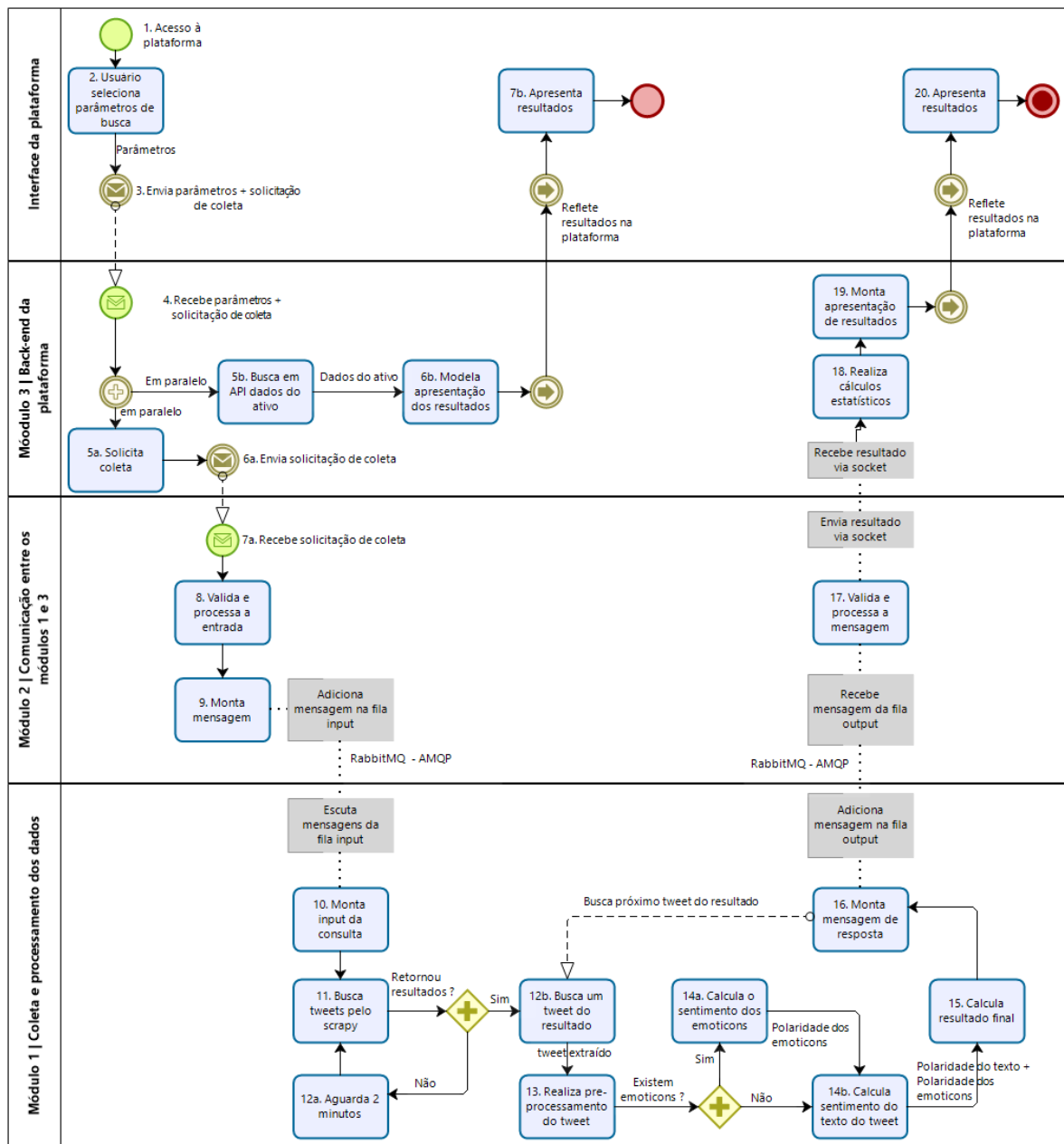
3.1 Arquitetura de funcionamento

A arquitetura da plataforma foi desenvolvida baseada em três módulos principais: (a) coleta e processamento dos dados, módulo que visa coletar as páginas da rede social *Twitter*, realizar as etapas de extração e pré-processamento dos *tweets* das páginas coletadas e classificar o sentimento de cada um deles; (b) *back-end* da plataforma, módulo que realiza o recebimento, organização e apresentação de dados; e (c) comunicação entre módulos de coleta e processamento dados e o *back-end* da plataforma, responsável por garantir a segurança e a padronização das mensagens enviadas e recebidas utilizando um sistema de enfileiramento de mensagens. O macrofluxo da arquitetura de funcionamento da plataforma encontra-se na Figura 3.1.

Conforme apresentado na Figura 3.1, a plataforma, de modo geral, comporta-se da seguinte maneira: a partir de parâmetros de entrada definidos pelo usuário final, como por exemplo o código de um ativo negociado na bolsa e um intervalo de datas, é realizada uma busca no *Twitter* por publicações relacionadas a este ativo. Destas publicações, são extraídos os *tweets* que, em seguida são pré-processados, resultando em *tweets* melhor estruturados. Após o pré-processamento, os *tweets* são analisados para verificar a existência de *emoticons*. Caso exista, os *emoticons* são retirados do texto e classificados isoladamente, para que seja possível saber qual o sentimento expresso por cada um deles. Após a classificação dos *emoticons*, o texto do *tweet* passa pelo processo de classificação retornando seu teor positivo, negativo ou neutro. Ao final, os valores da polaridade dos *emoticons* e do texto do *tweet* são unidos, afim de obter-se um resultado final. Por fim, aplica-se um algoritmo estatístico e de análise que devolve ao usuário final na plataforma, indicadores de tendência para o ativo escolhido, baseados no sentimento expresso nos textos para auxílio na tomada de decisão. Este processo tende a ser assíncrono e em tempo real. A sequência dos passos, apresentados na Figura 3.1, são explicados na Tabela 3.1.

Nas próximas subseções, são apresentados com mais detalhes os três módulos desenvolvidos para o funcionamento da plataforma: a Subseção 3.1.1 descreve o primeiro módulo, onde é feita a coleta das páginas do *Twitter*, a extração dos *tweets*, seu pré-processamento e a análise de sentimentos; a Subseção 3.1.2 aborda o terceiro módulo, onde é realizado o recebimento, organização e apresentação dos dados; e a Subseção 3.1.3 apresenta o segundo módulo, onde é feita a comunicação entre o primeiro e o terceiro módulos.

Figura 3.1 – Arquitetura de funcionamento da plataforma proposta.



Fonte: Elaborado pelo autor.

3.1.1 Coleta e processamento dos dados

A coleta e o processamento dos dados do *Twitter* compõem o primeiro módulo da plataforma. Para esta etapa, foi criado um *script* escrito na linguagem *Python* com o auxílio do *framework Scrapy* (vide Figura 2.3).

O *script* recebe uma solicitação de coleta de acordo com os parâmetros definidos pelo usuário no painel de busca da plataforma (passo 2 da Figura 3.1); estes parâmetros formam uma *query* que dá início a coleta de páginas no *Twitter* a partir do mecanismo *Downloader* (passo 11 da Figura 3.1). No momento em que as páginas retornam ao módulo, o mecanismo *Spider* as recebe e realiza a extração dos dados mais relevantes, como a data e hora da postagem, o texto

Tabela 3.1 – Manual de funcionamento da plataforma proposta

Passo	Camada	Tipo	Descrição
1	Interface da plataforma	Ação	O usuário final acessa a plataforma.
2	Interface da plataforma	Ação	O usuário final digita os parâmetros desejados para a coleta e análise.
3	Interface da plataforma	Comunicação	A plataforma recebe os parâmetros de entrada do usuário final e os envia em uma solicitação de coleta para o Módulo 3.
4	Módulo 3 / Back-end	Comunicação	O Módulo 3 recebe da plataforma a solicitação de coleta com os parâmetros definidos pelo usuário final.
5b	Módulo 3 / Back-end	Ação	O Módulo 3, com o código do ativo, faz uma requisição à API de ações negociadas na bolsa e retorna dados relativos aquele ativo (assíncrono).
6b	Módulo 3 / Back-end	Ação	O Módulo 3, com a resposta dos dados retornados pela API, modela um gráfico inicial com indicadores que são apresentados para o usuário final.
7b	Interface da plataforma	Ação	A plataforma reflete para o usuário final os resultados iniciais montados pelo Módulo 3.
5a	Módulo 3 / Back-end	Ação	O Módulo 3 abre uma comunicação HTTP com o módulo Módulo 2.
6a	Módulo 3 / Back-end	Comunicação	O Módulo 3 envia uma requisição de coleta, a partir dos parâmetros definidos pelo usuário final, para o Módulo 2.
7a	Módulo 2 / Comunicação entre módulos	Comunicação	O Módulo 2 recebe a requisição HTTP de coleta.
8	Módulo 2 / Comunicação entre módulos	Ação	O Módulo 2 valida os dados de entrada e os processa.
9	Módulo 2 / Comunicação entre módulos	Ação	O Módulo 2 monta a mensagem a ser enviada para o Módulo 1 e a coloca na fila de <i>input</i> criada pelo <i>RabbitMQ</i> .
10	Módulo 1 / Coleta e processamento	Ação	O Módulo 1, a partir do <i>RabbitMQ</i> , é informado de que há uma mensagem na fila de <i>input</i> para ser processada, buscando pela mensagem e montando o <i>input</i> da consulta para a coleta dos dados no <i>Twitter</i> .
11	Módulo 1 / Coleta e processamento	Ação	O Módulo 1 busca os dados no <i>Twitter</i> pelo <i>framework Scrapy</i> (assíncrono).
12a	Módulo 1 / Coleta e processamento	Ação	Se a busca não retornar resultados, o módulo 1 aguarda um tempo estimado de 2 minutos e realiza a busca novamente. (Processo retorna ao Passo 11).
12b	Módulo 1 / Coleta e processamento	Ação	Se a busca retornar resultados, o Módulo 1 enfileira os <i>tweets</i> resultantes (assíncrono) e os processa um a um.
13	Módulo 1 / Coleta e processamento	Ação	O Módulo 1 realiza o pré-processamento do <i>tweet</i> .
14a	Módulo 1 / Coleta e processamento	Ação	Se no <i>tweet</i> pré-processado houver <i>emoticons</i> , o Módulo 1 separa-os do texto original e calcula a polaridade individual dos <i>emoticons</i> . (Processo segue para o Passo 14b).
14b	Módulo 1 / Coleta e processamento	Ação	Se no <i>tweet</i> não houver <i>emoticons</i> , a polaridade do texto é calculada. Caso o fluxo passe pelo passo 14a, esta etapa recebe o texto do <i>tweet</i> e o valor da polaridade dos <i>emoticons</i> .
15	Módulo 1 / Coleta e processamento	Ação	O Módulo 1, uma vez calculado a polaridade dos <i>emoticons</i> , caso existam, e do texto do <i>tweet</i> , realiza um cálculo final para conhecer a polaridade final do <i>tweet</i> .
16	Módulo 1 / Coleta e processamento	Ação	O Módulo 1 monta a mensagem de resposta e a coloca na fila de <i>output</i> criada pelo <i>RabbitMQ</i> . Após este processo, o Módulo 1 busca o próximo <i>tweet</i> a ser processado. (Passo 12b).
17	Módulo 2 / Comunicação entre módulos	Ação	O Módulo 2, a partir do <i>RabbitMQ</i> , é informado de que há uma mensagem na fila de <i>output</i> para ser processada. Buscando pela mensagem, validando os dados de entrada, processando-os e enviando o objeto final via <i>websocket</i> ao módulo 3.
18	Módulo 3 / Back-end	Ação	O Módulo 3 recebe via <i>websocket</i> o objeto final e realiza os cálculos estatísticos esperados, alinhando-os com os resultados iniciais processados no Passo 6b.
19	Módulo 3 / Back-end	Ação	O Módulo 3 modela os gráficos finais com os resultados, para apresentá-los ao usuário final.
20	Interface da plataforma	Ação	A plataforma reflete para o usuário final os resultados finais gerados pelo Módulo 3.

Fonte: Elaborado pelo autor.

do *tweet* e o usuário que o postou (passo 12b da Figura 3.1). Para realizar essa extração, um conjunto de regras XPath¹ foram criadas e podem ser visualizadas na Figura 3.2.

De acordo com a Figura 3.2, observa-se que cada regra XPath esta sendo usada para extrair um conteúdo específico de uma página HTML do *Twitter*. Nas linhas de 1 a 4, é executada a regra que extrai o nome do usuário que publicou o *tweet*. Nas linhas de 6 a 9, o texto completo do *tweet* é extraído; e, nas linhas de 11 a 14, a regra XPath extrai a data e hora em que o *tweet* foi publicado. Estas três regras são aplicadas para cada *tweet* encontrado na página.

Após a extração dos *tweets*, conforme apresentado na arquitetura de funcionamento da

¹ O XPath é uma linguagem de consulta que ajuda a navegar por documentos que usam marcadores, como os arquivos XML e HTML.

Figura 3.2 – Fragmento do *script* que realiza a extração dos *tweets* a partir de regras Xpath

```

1  √ tweet['usernameTweet'] =
2  √      item.xpath(
3      |      './span[@class="username u-dir u-textTruncate"]/b/text()'
4      |      ).extract()[0]
5
6  √ tweet['text'] =
7  √      '.join(item.xpath(
8      |      './div[@class="js-tweet-text-container"]/p//text()'
9      |      ).extract())
10
11 √ tweet['datetime'] =
12 √      datetime.fromtimestamp(int(item.xpath(
13 |      './div[@class="stream-item-header"]/small[@class="time"]/a/span/@data-time'
14 |      ).extract()[0])).strftime('%Y-%m-%d %H:%M:%S')
15

```

Fonte: Elaborado pelo autor.

Figura 3.1, estes passam para a etapa de pré-processamento (passo 13 da Figura 3.1). Esta etapa consiste em executar uma série de tratamentos no texto e na data, com o propósito de padronizar todos os *tweets* extraídos. Estes tratamentos buscam estruturar os *tweets* puros, melhorando a sua qualidade e preparando-os para serem submetidos ao processo de classificação. Neste contexto, algumas técnicas e algoritmos foram utilizados. A Figura 3.3 apresenta um fragmento do código criado para o pré-processamento dos *tweets*.

A Figura 3.3 apresenta um fragmento do código do primeiro módulo que, após coletar as páginas do *Twitter* e extrair os *tweets*, exemplifica a etapa de pré-processamento dos *tweets* extraídos por meio dos seguintes métodos:

- *duplicatesRemover*. O *Twitter* possui a funcionalidade de *retweet*, ou seja, a mesma informação uma vez publicada pode ser “*retweetada*” por outros perfis da rede e coletadas pelo coletor. Consequentemente, isso gera uma grande quantidade de *tweets* desnecessários. Este método realiza um filtro destes *tweets* idênticos, identificando-os e mantendo apenas uma versão de cada um.
- *stopwordsRemover*. As *stopwords* são palavras sem significado relevante como algumas preposições e artigos. A remoção delas diminui o tempo de processamento da plataforma e o espaço em disco, aumentando a precisão da classificação e análise. Este método faz uso da biblioteca *Natural Language Toolkit* (NLTK) no intuito de identificar e remover tais palavras do texto.
- *dateFormatter*. Os *tweets* são recebidos com um formato de data que precisa ser alterado; logo, este método realiza a transformação para o formato “YYYYMMDD”.
- *normalizer*. Como os *tweets* são textos livres escritos pelo usuário, eles podem conter alguns ruídos que também devem ser removidos, como caracteres especiais, pontuações e

Figura 3.3 – Fragmento do *script* que realiza o pré-processamento dos *tweets*

```

12 # remove itens duplicados do array
13 def duplicatesRemover(tweets):
14     return list(set(tweets))
15
16 # remove palavras consideradas stopwords do idioma portugues
17 def stopwordsRemover(tweet):
18     sentences = nltk.sent_tokenize(tweet)
19
20     for i in range(len(sentences)):
21         words = nltk.word_tokenize(sentences[i])
22         newwords = [word for word in words if word not in stopwords.words('portuguese')]
23         sentences[i] = ' '.join(newwords)
24
25     return sentences
26
27 # formata a data para YYYYMMDD
28 def dateFormatter(date):
29     months = ["Jan", "Fev", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"]
30     day = date.split(' ')[2] # example: "10"
31     month = date.split(' ')[1] # example: "Dec"
32     year = date.split(' ')[-1] # example: "2020"
33
34     try:
35         month = str(months.index(month)+1)
36     except:
37         print("Month not found")
38
39     if(len(month) < 2): month = "0" + month
40     if(len(day) < 2): day = "0" + day
41     return year + month + day
42
43 # transforma o texto bruto em um texto limpo
44 def normalizer(tweet):
45     # remove pontuação
46     text = "".join([char for char in tweet if char not in string.punctuation])
47     nfkd = unicodedata.normalize('NFKD', text)
48     text = u"".join([c for c in nfkd if not unicodedata.combining(c)])
49
50     # remove urls e links
51     text = re.sub("https?:\V\)?([\da-z\.-]+)\.([a-z\.\]{2,6})([\V\w \.-]*)", "", text)
52     #remove caracteres especiais
53     text = re.sub("@ [A-Za-z0-9+]|(^0-9A-Za-z $)|(\w+:\V\V\S+)", "", text)
54     text = text.split('http')[0]
55     return text.lower()
56
57 # Retira todos os emojis encontrados no texto
58 def getEmojis(text):
59     emoji_list = []
60     data = regex.findall(r'\X', text)
61
62     for word in data:
63         if any(char in emoji.UNICODE_EMOJI['en'] for char in word):
64             emoji_list.append(word)
65
66     emoji_list = sorted(set(emoji_list))
67     return emoji_list

```

Fonte: Elaborado pelo autor.

links. Além disso, todos os *tweets* foram convertidos para minúsculo. Este método faz uso de expressões regulares para identificar estes padrões e removê-los do texto.

- *getEmojis*. O texto dos *tweets* podem conter *emoticons* que devem ser classificados de forma isolada. Este método faz uso da biblioteca *emoji* no intuito de identificar os *emoticons* presentes no texto e armazená-los em uma lista separadamente.

Na medida em que as páginas são coletadas, os *tweets* são extraídos e pré-processados: um objeto contendo o nome do usuário que efetivou a publicação, a lista de *emoticons*, o texto e a data de cada *tweet* são armazenados em uma estrutura de dados em memória. A próxima etapa do primeiro módulo é a análise de sentimentos, onde os objetos inseridos na fila anteriormente, são retirados e processados pelos métodos responsáveis pela classificação.

A análise de sentimentos compõe a próxima etapa do primeiro módulo da plataforma. Esta etapa inicia-se com a chegada de um objeto proveniente de uma fila criada na etapa anterior contendo seus próprios campos. Assim que chegam, os *tweets* e a lista de *emoticons* passam por um processo de classificação e análise de sentimentos que objetivam identificar, quantificar e extrair qualquer sentimento presente neles. Primeiramente, a lista contendo os *emoticons* é classificada; para isso, foi utilizada a biblioteca em *Python*, *emosent*. Esta biblioteca, consegue interpretar os caracteres *unicode* que formam os *emoticons* e compará-los com uma base de dados já classificada (SMAILOVIĆ et al., 2015). Ao final deste processo, para cada *emoticon* presente na lista de entrada, é retornada a sua polaridade que pode variar de -1.0 a 1.0. O algoritmo 1 apresenta o pseudocódigo utilizado para classificar os *emoticons*.

Algorithm 1 Classificador de *emoticons*

```

1: Input: arrayEmoticons
2: Output: emoticonsPolarity
3: if not arrayEmoticons then
4:   return 0
5: end if
6: polarity = 0.0
7: sentiment = { }
8: for emoticon ∈ arrayEmoticons do
9:   sentiment = getEmojiSentimentRank(emoticon)
10:  polarity += sentiment.score
11: end for
12: emoticonsAmount = len(arrayEmoticons)
13: emoticonsPolarity = polarity / emoticonsAmount
14: return emoticonsPolarity

```

De acordo com o Algoritmo 1, a linha 1 indica o dado de entrada, sendo ele a lista de *emoticons* extraída do texto a partir do método “getEmojis” (Vide Figura 3.3) e a linha 2 indica o dado de saída, a polaridade final dos *emoticons* presentes na lista. Em seguida é verificado se a lista de *emoticons* está vazia; se estiver, significa que nenhum *emoticon* foi encontrado no texto daquele *tweet* extraído e a polaridade retornada será zero (linhas 3 e 4). Inicialmente, a variável “polarity” recebe o valor zero (linha 6), para que acumule a polaridade dos *emoticons* presentes na lista e a variável “sentiment” é inicializada como um objeto vazio (linha 7), para que armazene o resultado do cálculo de polaridade de cada *emoticon*. Quanto ao cálculo da polaridade dos *emoticons*, nas linhas 8 a 11, um processo de repetição ocorre: para cada *emoticon* presente na lista, é realizada a sua análise de sentimento a partir do método “getEmojiSentimentRank” da biblioteca *emosent*

(linha 9). O objeto resultante é atribuído à variável “*sentiment*” para que posteriormente possa ser extraído e acumulado na variável “*polarity*” o valor do atributo “*score*”, que representa a polaridade daquele *emoticon* (linha 10). Na linha 12, a variável “*emoticonsAmount*” é inicializada com a quantidade de *emoticons* presentes na lista de entrada. A linha 13 realiza a média do valor total das polaridades encontradas pela quantidade de *emoticons* existentes, para que, caso haja duplicidade de *emoticons*, isto não impacte no resultado final. Por fim, na linha 14, o resultado final é retornado pelo algoritmo.

Após a classificação dos *emoticons*, a próxima etapa consiste em encontrar o sentimento expresso no texto do *tweet* e conhecer o seu teor de positividade e de negatividade. A biblioteca em *Python* para processamento de dados textuais, *TextBlob*, foi utilizada. Esta biblioteca faz uso da técnica de classificação *Naive Bayes* que, segundo Rish et al. (2001), é um algoritmo rápido, preciso e confiável. O *TextBlob* realiza uma análise de sentimentos mais quantitativa, retornando para cada *tweet* processado a sua pontuação de polaridade em um intervalo de -1.0 a 1.0 e de subjetividade dentro de um intervalo de 0.0 a 1.0. O Algoritmo 2 apresenta o pseudocódigo utilizado.

Algorithm 2 Classificador de *tweets*

```
1: Input: tweet
2: Output: tweetPolarity
3: translatedTweet = GoogleTranslator.translate(tweet)
4: analysis = TextBlob(translatedTweet)
5: polarity = analysis.sentiment.polarity
6: return polarity
```

De acordo com o Algoritmo 2, a linha 1 indica o dado de entrada, sendo ele o texto do *tweet* extraído e limpo, a partir dos métodos de pré-processamento realizados anteriormente (vide Figura 3.3) e a linha 2 indica o dado de saída, a polaridade final do texto do *tweet*. Como a extração dos *tweets* está configurada para buscá-los independente do idioma e a biblioteca que realiza a análise de sentimentos aceita apenas *tweets* na língua inglesa, foi necessário realizar uma tradução por meio do *google translate* a partir da biblioteca *deep translator* (linha 3). Após a tradução, o *tweet* foi passado via parâmetro para a biblioteca *Textblob* que realiza a análise de sentimentos (linha 4). O objeto resultante dessa classificação foi atribuído à variável “*analysis*”. Na linha 5, a variável “*polarity*” recebe o valor do atributo “*sentiment.polarity*” presente no objeto resultante da classificação, que contém a polaridade final encontrada para o texto do *tweet* passado via parâmetro. Por fim, na linha 6, o resultado final é retornado pelo algoritmo. Em casos de falha nos processos das linhas 3 e 4, o *tweet* analisado é descartado; isso também acontece para os *tweets* em que a polaridade calculada é zero, já que não revelam nenhum posicionamento do autor ou são considerados isentos de opinião.

Ao final da execução do Algoritmo 2, é aplicado um cálculo para conhecer a polaridade final do *tweet*, ou seja, o sentimento expresso pelos *emoticons* junto ao sentimento expresso

no texto. Para este cálculo, foi utilizada uma média ponderada; neste caso, um peso com valor 3 foi atribuído à polaridade dos *emoticons* (*output* do Algoritmo 1) e o peso 7 foi atribuído à polaridade do texto do *tweet* (*output* do Algoritmo 2). Os valores para os pesos foram medidos empiricamente.

Uma vez calculada a polaridade final para um *tweet*, tal valor é incluído no objeto recebido na etapa inicial da análise de sentimentos. Este objeto, agora contendo o nome do usuário que efetivou a publicação, a lista de *emoticons*, o texto e a data do *tweet* e a sua polaridade final, deve ser retornado para o primeiro módulo da plataforma. O primeiro módulo, por fim, abre uma conexão com *RabbitMQ*, monta a mensagem de resposta e a insere na fila de *output* criada pelo segundo módulo (passo 16 da Figura 3.1).

Pelo fato do processo de extração acontecer de forma assíncrona (passo 11 da Figura 3.1), os *tweets* resultantes chegam separadamente no primeiro módulo. Portanto, o fluxo, entre as etapas 12b e 16 da Figura 3.1, acontece repetidamente para cada *tweet* obtido como resultado.

3.1.2 Recebimento, organização e apresentação de dados

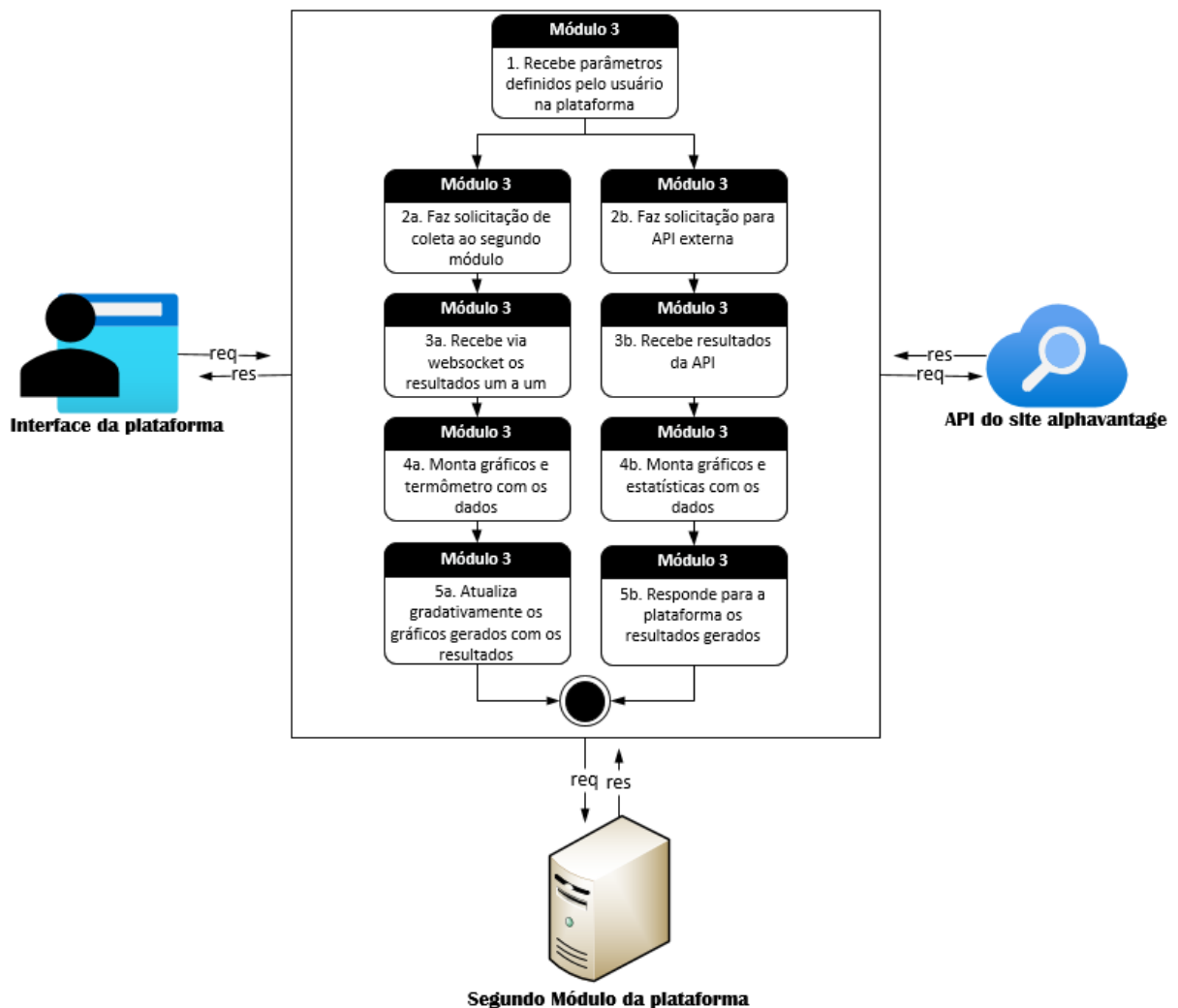
O processo de recebimento, organização e apresentação de dados compõem o terceiro módulo da plataforma, sendo responsável, então, por receber solicitações e promover requisições a outros componentes da plataforma, além de receber, organizar e apresentar os resultados para o usuário final. Para este módulo, foi utilizada a linguagem *Javascript* com o auxílio da biblioteca *React*.

De uma forma geral, este módulo está diretamente ligado à forma como a plataforma é visualizada pelo usuário final, recebendo, organizando e apresentando dados de forma amigável na interface da plataforma. Para isso, realiza alguns processos internos que são divididos em duas etapas. A primeira etapa diz respeito à forma como são recebidos os valores que foram preenchidos pelo usuário final na plataforma e como é iniciada as requisições a outros componentes. A segunda etapa envolve o recebimento dos resultados das solicitações feitas, advindas do segundo módulo, a organização e a apresentação dos mesmos na plataforma. A Figura 3.4 apresenta o funcionamento do terceiro módulo.

Conforme apresentado pela Figura 3.4, o terceiro módulo centraliza a comunicação entre uma API externa, o segundo módulo e a plataforma. O seu funcionamento ocorre da seguinte maneira: o usuário final preenche na plataforma o nome de um ativo do mercado financeiro, uma data inicial e uma data final. Após preenchidos, esses dados são enviados ao terceiro módulo, que, de forma paralela, realiza duas requisições: uma para o segundo módulo, onde se inicia a solicitação de coleta (passo 5a Figura 3.1) e outra para a API do site *alphavantage*² (passo 5b Figura 3.1), onde se obtém informações sobre o ativo solicitado. Após disparar essas requisições, o terceiro módulo aguarda os resultados, de forma assíncrona. O objeto retornado pela API contém:

² <https://www.alphavantage.co/>

Figura 3.4 – Funcionamento do terceiro módulo



Fonte: Elaborado pelo autor.

o volume de negociações e os preços de alta, baixa, abertura e fechamento daquele ativo para o intervalo de datas estipulado. Quando estes valores são retornados, o terceiro módulo organiza-os por datas e gera um gráfico para o volume de negociações e para o preço de fechamento do ativo, além de um relatório de estatísticas referentes à data mais atual retornada pela API. Já em relação ao segundo módulo, os objetos retornados via *websocket* (vide Figura 3.7) contém, em cada um deles, um *tweet* referente ao ativo pesquisado, a sua data de postagem e o valor calculado da polaridade desse *tweet* (vide Capítulo 3.1.1). Conforme estes objetos vão sendo recebidos pelo terceiro módulo e a polaridade dos *tweets* vão sendo somadas em uma variável, um termômetro inicia-se a fim de indicar o pessimismo, neutralidade ou otimismo dos usuários do *Twitter* para aquele determinado ativo. Por fim, um último gráfico diário é gerado, traçando um paralelo entre o retorno linear do ativo e o seu valor final de sentimento para todas as datas dentro do intervalo estipulado. A Figura 3.5 apresenta a fórmula utilizada para calcular o retorno linear de um ativo.

De acordo com a Figura 3.5, o retorno linear, baseado no preço de fechamento de um

Figura 3.5 – Fórmula do retorno linear de um ativo

$$\text{Retorno linear} = \frac{\text{Preço}_{\text{fechamento dia}} - \text{Preço}_{\text{fechamento dia anterior}}}{\text{Preço}_{\text{fechamento dia anterior}}} * 100$$

Fonte: Elaborado pelo autor.

ativo, é dado pelo preço de fechamento do dia, menos o preço de fechamento do dia anterior dividido pelo preço de fechamento do dia anterior, vezes cem. Este cálculo foi utilizado pois, além de mostrar o quanto uma ação subiu ou caiu, o retorno diário de uma ação normalmente tem valores pequenos, variando entre -5% e 5% segundo [Investidor \(2017\)](#), e a sua relação no gráfico com o valor final da polaridade dos *tweets* fica mais visível.

Após realizar a modelagem e a estruturação dos resultados obtidos, o terceiro módulo reflete-os na plataforma.

3.1.3 Comunicação entre módulos

A comunicação entre o primeiro e o terceiro módulos da plataforma compõe o segundo módulo da mesma. Para este módulo, foi criada uma aplicação na linguagem *Javascript* que realiza a comunicação entre os módulos por meio de mensagens. Para isso, foi utilizado o *framework RabbitMQ* (vide Figura 2.5) pois, como a coleta e a extração de *tweets* podem gerar um volume muito grande de informações, um *framework* assíncrono e distribuído, que atenda aos requisitos de alta escala, alta disponibilidade e confiabilidade na entrega, é de extrema importância para a plataforma proposta.

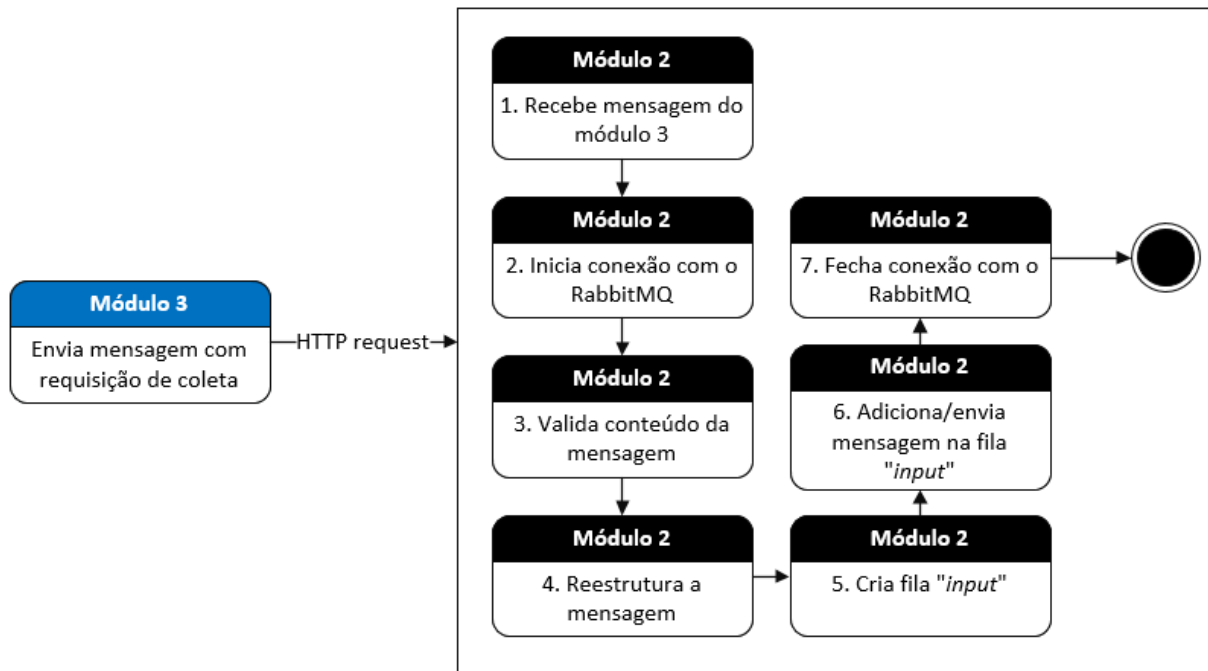
Os processos internos realizados pelo segundo módulo são divididos em duas etapas. A primeira etapa (vide Figura 3.6) consiste no envio de uma requisição de coleta, onde a comunicação é feita do terceiro para o primeiro módulo. A segunda etapa (vide Figura 3.7) é formada pelo retorno dos resultados obtidos, do primeiro para o terceiro módulo.

De acordo com a Figura 3.6, a primeira etapa de comunicação inicia-se com a chegada de uma requisição HTTP enviada pelo terceiro módulo. Esta requisição aciona um serviço que realiza uma série de tarefas antes de repassar a mensagem recebida ao seu destino final. Nesta etapa, o serviço acionado é responsável por: receber o objeto contendo a mensagem que foi enviada; abrir uma conexão AMQP para que o *rabbitMQ* possa publicar mensagens dentro da rede; validar o conteúdo da mensagem recebida e verificar se os atributos estão no formato adequado; reestruturar a mensagem para o formato estabelecido pelo *rabbitMQ* para envio; criar a fila “*input*” e uma *exchange* do tipo *fanout exchange*³ (vide Figura 2.5); adicionar a mensagem já estruturada na fila e enviá-la; e fechar a conexão AMQP temporariamente.

De acordo com a Figura 3.7, a segunda etapa de comunicação é feita por meio de Espera

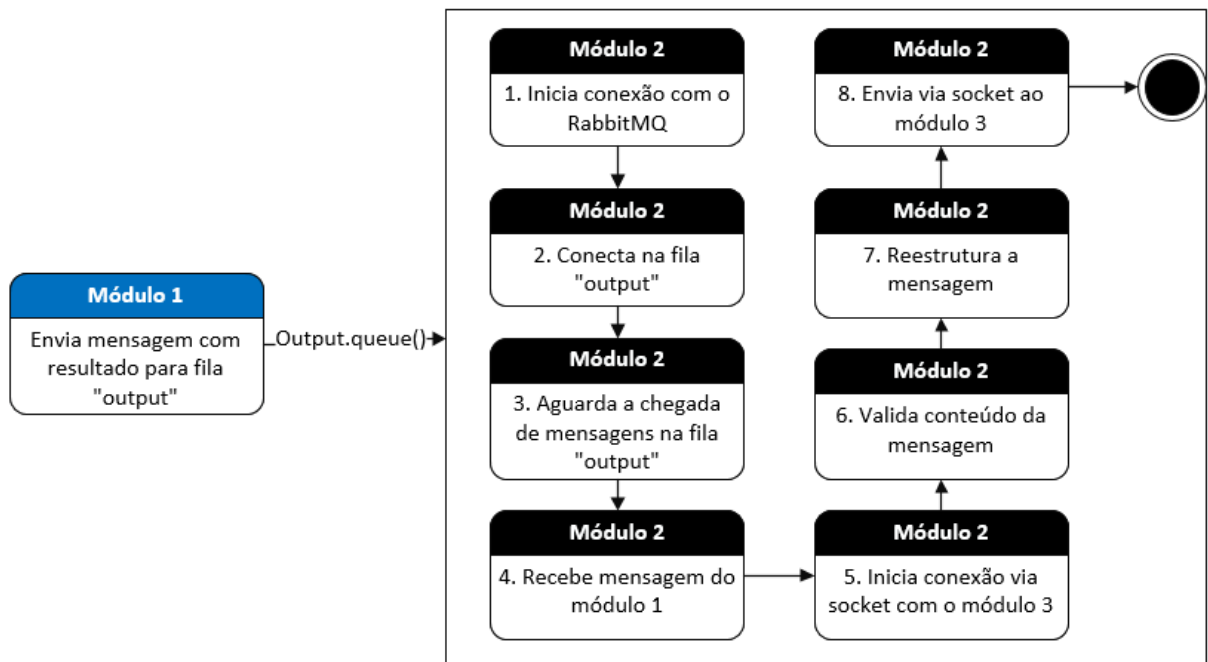
³ A *exchange* do tipo *fanout* entrega a mensagem a todas as filas que estão vinculadas a esta *exchange*.

Figura 3.6 – Primeira etapa do módulo de comunicação



Fonte: Elaborado pelo autor.

Figura 3.7 – Segunda etapa do módulo de comunicação



Fonte: Elaborado pelo autor.

ativa⁴. O serviço responsável inicia uma conexão AMQP para que o *rabbitMQ* possa consumir mensagens, conecta-se na fila “*output*” para que seja possível receber mensagens postadas nela e fica aguardando a chegada de uma nova mensagem. Quando o primeiro módulo posta uma

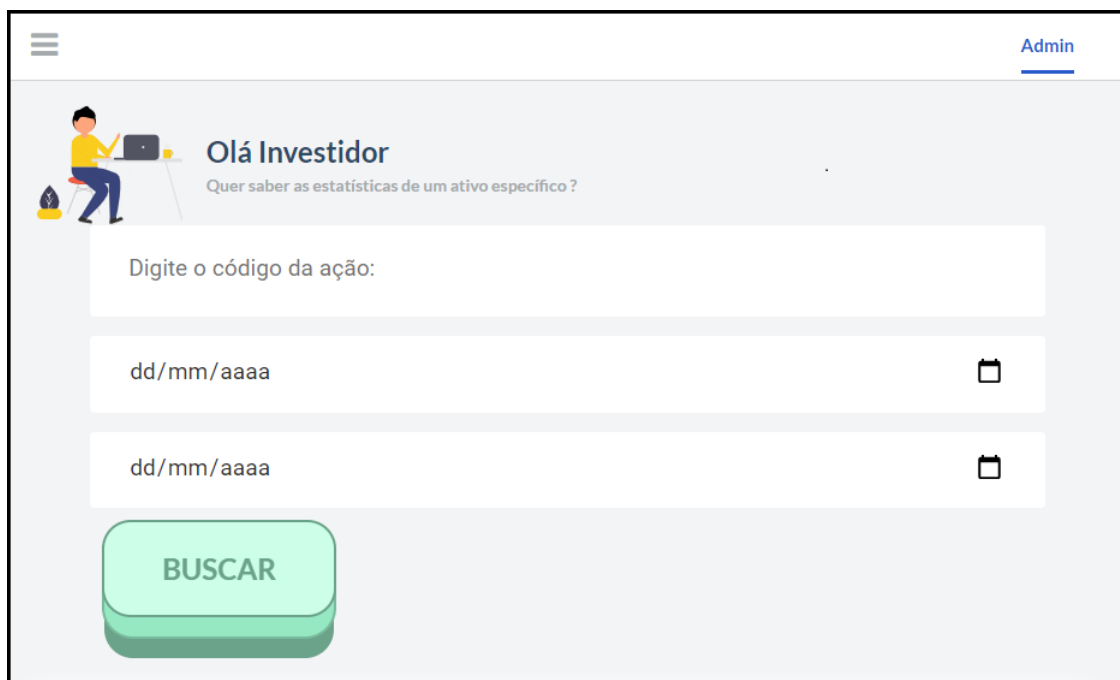
⁴ Técnica em que um processo verifica uma condição repetidamente até que ela seja verdadeira.

mensagem na fila “*output*”, imediatamente o serviço ativo no segundo módulo recebe-a e dá início ao seu processamento. Nesta etapa, o serviço é responsável por: iniciar uma nova conexão via *websocket*⁵ sendo o terceiro módulo o cliente; validar o conteúdo da mensagem recebida e verificar se os atributos estão no formato adequado; reestruturar a mensagem para o formato esperado pelo terceiro módulo; e enviar a mensagem na conexão aberta anteriormente. Este processamento ocorre para cada mensagem recebida na fila “*output*”.

3.2 Interface e parametrização da plataforma

Nesta seção, é apresentada a interface da plataforma proposta envolvendo a parametrização necessária para executá-la. A tela inicial da plataforma, apresentada na Figura 3.8, foi planejada para que a interação com o usuário aconteça de forma simples e fácil.

Figura 3.8 – Tela inicial da plataforma

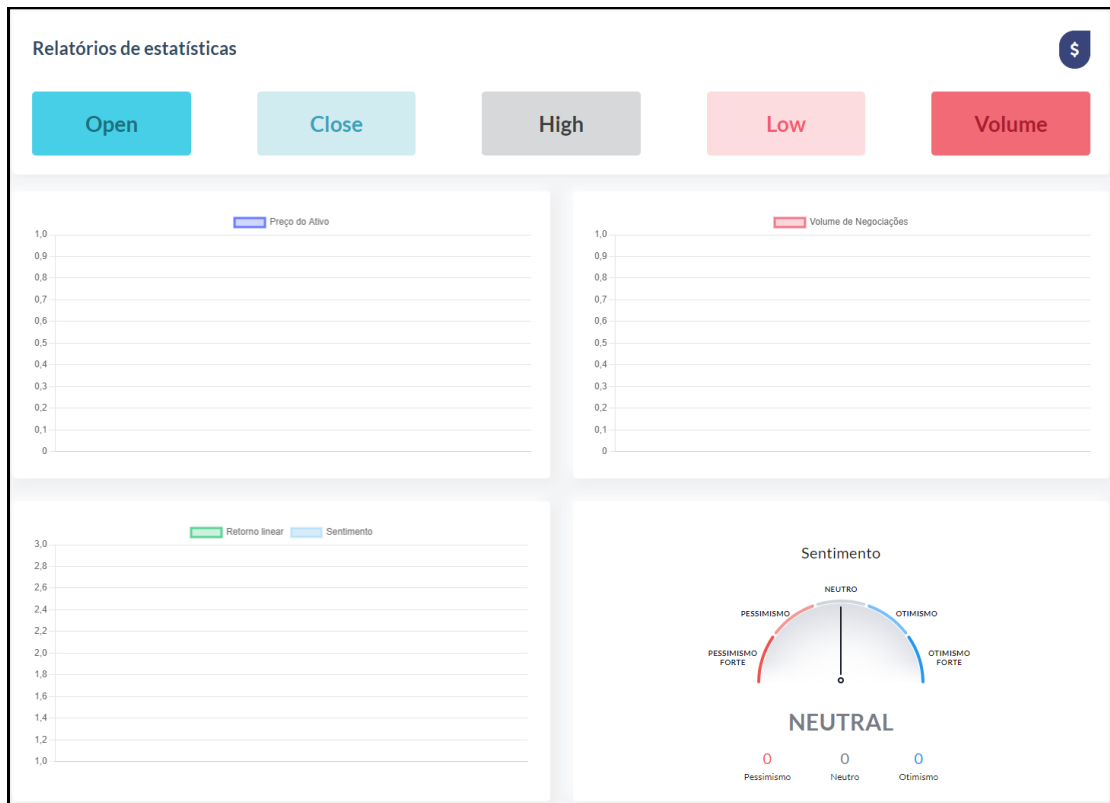


Fonte: Elaborado pelo autor.

De acordo com a Figura 3.8, para dar início ao processo de geração de resultados a partir de dados do *Twitter* e do mercado financeiro, o usuário deve pressionar o botão “BUSCAR” após informar os parâmetros: código da ação desejada, data inicial da coleta e data final da coleta. Deixar algum desses parâmetros sem preenchimento, preencher com um código de ação inválido ou selecionar uma data final da consulta menor que a data inicial são fatores que invalidam a busca. Após a realização de uma busca, os resultados obtidos, seguindo a arquitetura de funcionamento da plataforma descrita na Figura 3.1, são apresentados na tela ilustrada na Figura 3.9.

⁵ *Websocket* define uma API que estabelece conexões de “soquete” entre um navegador da *web* e um servidor. Em outras palavras, há uma conexão persistente entre o cliente e o servidor e ambas as partes podem começar a enviar dados a qualquer momento.

Figura 3.9 – Tela de apresentação dos resultados obtidos



Fonte: Elaborado pelo autor.

De acordo com a Figura 3.9, no “Relatório de estatísticas”, são apresentadas as médias de valores para a abertura, o fechamento, a alta, a baixa e o volume de negociação da ação desejada naquele determinado período. Estes valores possuem, como data de referência, a data final escolhida pelo usuário. Após o relatório de estatísticas, são apresentados os gráficos diários do preço da ação, do volume de negociações e um último que correlaciona o retorno linear normal da ação com o valor médio da polaridade dos *tweets* por dia. Por fim, é apresentado um termômetro de sentimentos, onde a cada *tweet* recebido como resultado, o seu valor de sentimento será acumulado e pontuado na tela. O objetivo deste termômetro é conhecer o grau de pessimismo, neutralidade ou otimismo para aquela ação no intervalo de datas estipulado.

4 Experimentos Computacionais

Neste capítulo, são apresentados os experimentos computacionais relativos à plataforma proposta. A Seção 4.1 descreve os experimentos realizados na plataforma, abrangendo os três módulos existentes, e a Seção 4.2 trata e avalia os resultados obtidos nos experimentos.

4.1 Descrição dos experimentos

Esta seção descreve, de forma detalhada, a configuração dos experimentos propostos para validação da plataforma analisando dois focos distintos: (a) foco na contribuição da plataforma para os usuários e (b) foco na avaliação e performance da arquitetura utilizada na construção da plataforma. Os experimentos realizados utilizaram uma máquina local com sistema operacional *Linux Ubuntu*, 16GB de memória *Random Access Memory* (RAM) e um processador 7^a geração *Intel Core i7*.

Como forma de verificar a contribuição da plataforma para os usuários, alguns casos de testes foram estabelecidos. A Tabela 4.1 ilustra a parametrização dos casos executados.

Tabela 4.1 – Parametrização dos casos de teste

Caso de teste	Ativo	Data inicial	Data final
1	BOVA11	03/05/2021	08/08/2021
2	HGLG11	01/05/2021	01/08/2021
3	PETR4	04/06/2021	06/08/2021

Fonte: Elaborado pelo autor.

A Tabela 4.1 apresenta a parametrização dos casos de testes que objetivam avaliar o funcionamento e a contribuição da plataforma para os seus usuários. Cada caso de teste possui os seguintes parâmetros: (a) o número de identificação do caso de teste; (b) o nome do ativo negociado no mercado financeiro brasileiro; (c) uma data inicial para a coleta; e (d) uma data final para a coleta. O nome do ativo, passado via parâmetro, será utilizado como índice na coleta das páginas do *Twitter* e o critério de parada deste processo de coleta, ou seja, a quantidade de *tweets* que serão extraídos dessas páginas, estará limitado no intervalo de datas definido.

Por fim, para a análise de performance da arquitetura proposta para a plataforma, foram avaliados tópicos como: confiabilidade na entrega das mensagens, velocidade de extração e processamento dos dados, vazão e tempo de resposta da plataforma.

4.2 Análise dos resultados obtidos

Esta seção apresenta e analisa os resultados obtidos pelos experimentos realizados e descritos na Seção 4.1.

Relativo ao foco associado à contribuição da plataforma para os usuários, a Figura 4.1 apresenta os valores obtidos no funcionamento da plataforma para o primeiro caso de teste escolhido.

De acordo com a Figura 4.1, é possível visualizar os resultados obtidos com a consulta na plataforma pelo ativo “BOVA11”, que se trata de um fundo de índice, o qual a sua rentabilidade assemelha-se bastante com o índice Ibovespa (IBOV), o principal índice da bolsa de valores brasileira (MODALMAIS, 2020). Foi possível identificar uma relação entre o retorno linear normal e o sentimento diário dos *tweets*, onde uma redução na positividade dos *tweets* entre os dias 25/06 e 01/07 refletiu em uma posterior queda no preço da ação. Isso mostra que o pessimismo com relação ao ativo impactou diretamente nas negociações destes dias. Com relação ao volume de negociações, houve uma queda de aproximadamente 46% entre os dias 04/08 e 06/08; geralmente, isso identifica uma tendência de queda nos preços da ação. O termômetro de sentimento apresenta um número bem maior de *tweets* negativos do que positivos ou neutros; isso se relacionado com a queda no volume de negociações nos últimos dias, pode justificar pessimismo por parte dos investidores.

Também relativo ao foco associado à contribuição da plataforma para os usuários, a Figura 4.2 apresenta os valores obtidos no funcionamento da plataforma para o segundo caso de teste escolhido.

De acordo com a Figura 4.2, observa-se os resultados da consulta na plataforma pelo ativo “HGLG11”, um fundo imobiliário que tem como base investimentos em empreendimentos de galpões logísticos (FUNDSEXPLORER, 2019). Por se tratar de um fundo imobiliário, segundo Campagnaro (2020), a sua volatilidade é menor quando comparado com as ações, o que traz mais tranquilidade para pessoas com menor tolerância ao risco. Essa baixa volatilidade pode ser vista no gráfico do retorno linear do ativo, trazendo um variação bem mais suave no seu preço para o período estipulado. Avaliando a relação entre o retorno linear e o sentimento dos usuários no *Twitter*, é observado que, em alguns momentos como entre as datas 08/06 e 16/06, os valores seguem um padrão de alta e baixa, podendo identificar uma relação otimista e pessimista sobre o ativo. O sentimento dos *tweets* para o ativo foram, em sua maioria, positivos, de acordo com o termômetro apresentado. Analisando o gráfico do volume de negociações nos últimos dias do período estipulado, houve um aumento de mais de 300%; isso pode indicar uma tendência de alta nos preços do ativo e uma boa oportunidade para compra dentro deste período.

Como último exemplo relativo ao foco associado à contribuição da plataforma para usuários, a Figura 4.3 apresenta os valores obtidos no funcionamento da plataforma para o terceiro caso de teste escolhido.

De acordo com a Figura 4.3, verifica-se os resultados para o ativo “PETR4”, código que se refere à empresa Petrobras. O retorno linear do preço desta ação, no período estipulado, não teve muitas variações e o sentimento do investidor baseado nos *tweets* mostrou-se tímido, mas mantendo-se na faixa positiva do gráfico. Os *tweets* dos investidores são, em sua maioria, otimistas com relação à empresa, como é observado no termômetro de sentimentos. Relacionando o retorno linear e o volume de negociações, é observada uma queda a partir do dia 05/08, de acordo com Deschatre e Majer (2006); isso pode identificar que uma tendência de baixa está para ser revertida, ocorrendo uma diminuição no ritmo da queda dos preços para, então, haver um aumento. Considerando esses fatores, o termômetro aponta uma oportunidade de compra para os próximos dias do período estipulado.

A Tabela 4.2 apresenta os resultados coletados referentes aos segundo foco dos experimentos realizados: a performance da arquitetura utilizada no desenvolvimento da plataforma. Para tanto, a performance foi avaliada a partir dos casos de teste apresentados na Tabela 4.1 e, para cada *tweet* referente a tais testes, foi calculada a média dos tempos descritos.

Tabela 4.2 – Resultados da performance da plataforma

Teste	Resultado
Requisição	12 ms
Resposta	11 ms
Extração	42 ms
Processamento e classificação	29 ms
Vazão	82 ms
Total de <i>tweets</i> extraídos	38210
Total de <i>tweets</i> processados	11463

Fonte: Elaborado pelo autor.

A Tabela 4.2 apresenta os seguintes resultados: (a) requisição: contabiliza o intervalo de tempo entre a solicitação de uma requisição na interface da plataforma e a chegada desta requisição no primeiro módulo; (b) resposta: contabiliza o intervalo de tempo entre a postagem de um objeto de resposta, pelo primeiro módulo, na fila de resposta e a sua chegada no terceiro módulo; (c) extração: contabiliza o tempo gasto na extração de cada *tweet*; (d) processamento e classificação: contabiliza o tempo gasto nas etapas de processamento e classificação do texto de cada *tweet*; (e) vazão: contabiliza o tempo total gasto desde a extração de um *tweet* pelo primeiro módulo, até o recebimento do objeto final no terceiro módulo; e (f) total de *tweets* extraídos: contabiliza o número total de *tweets* extraídos pelo *framework scrapy*; e (g) total de *tweets* processados: contabiliza o número total de *tweets* processados, sem falhas, pela plataforma.

O sistema de filas adotado na plataforma, por meio do *RabbitMQ*, mostrou-se rápido e confiável, com um média de 12ms para envio da requisição e 11ms para resposta, além de ter garantido a entrega para 100% das mensagens trafegadas. Em média, a extração de cada *tweet* foi rápida; isso pode ser atribuído ao isolamento do módulo de extração pelo *Scrapy* o que facilitou o processo. O processamento e classificação dos *tweets* aconteceu de forma efetiva, levando em

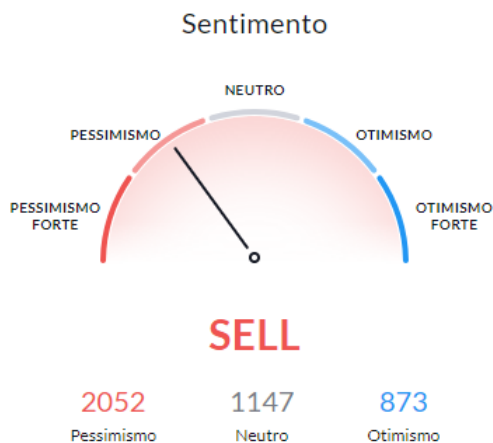
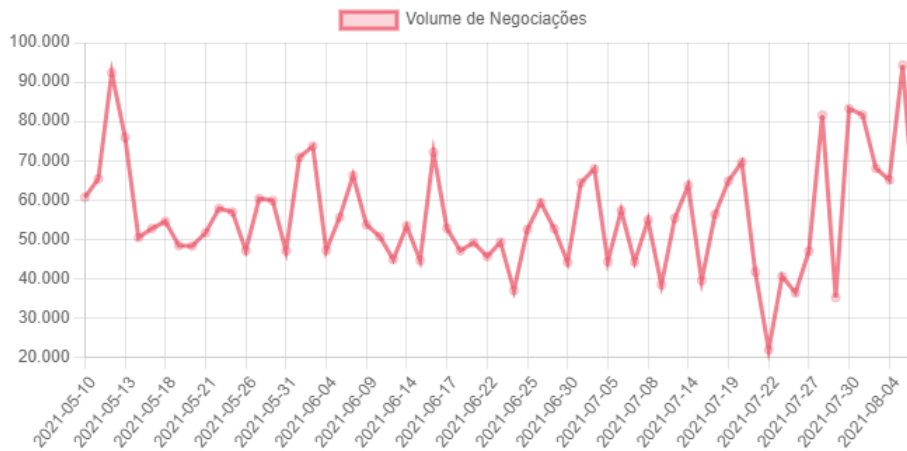
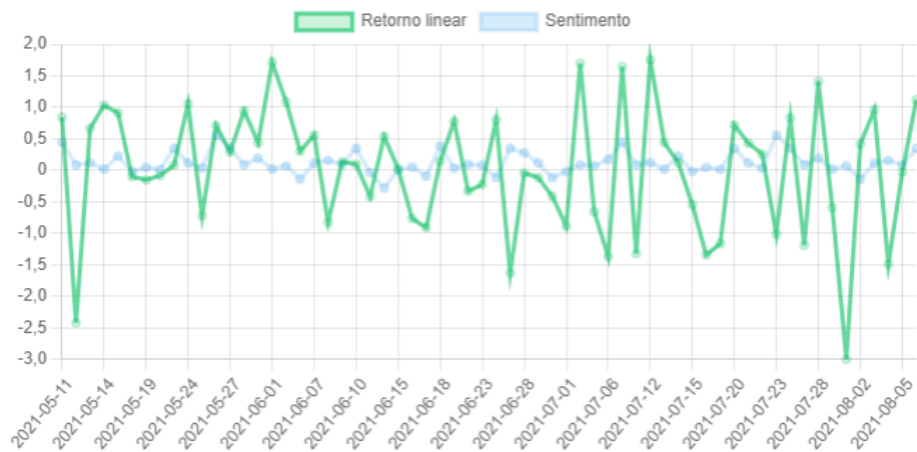
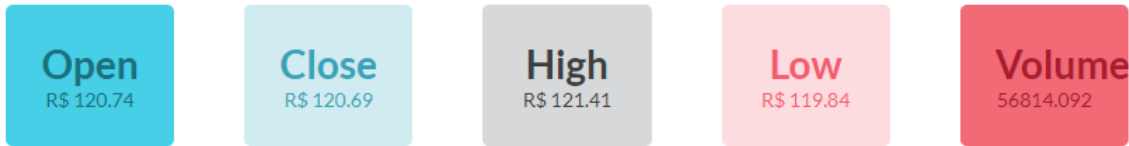
conta as diversas etapas realizadas e o curto espaço de tempo entre a chegada de um *tweet* e outro. Pelo fato do experimento ter sido realizado em uma máquina local e isso ocasionar uma baixa latência entre os módulos, a vazão também obteve uma média satisfatória. O sistema extraiu 38210 *tweets*, porém processou apenas 30% deste total; esta perda está relacionada à etapa de classificação pois, em caso de falhas no consumo da biblioteca externa (vide Algoritmo 2) ou quando a classificação retorna zero, o *tweet* é descartado por não ser relevante no processo. Para o volume de *tweets* processados, baseado no tempo médio deste processamento, a arquitetura no geral mostrou-se robusta e eficiente.

De modo geral, os resultados obtidos na plataforma foram satisfatórios. É possível observar que o objetivo de entregar análises detalhadas, baseando-se em *tweets* e suas polaridades, auxilia de alguma forma os investidores que tenham um conhecimento básico acerca de tendências de mercado. Com o uso da plataforma eles podem ponderar seu processo de compra e venda de um ativo no mercado financeiro. Os casos de teste, trouxeram momentos em que a relação entre o retorno linear e o sentimento presente nos *tweets* foi forte, bem como o termômetro de sentimentos retornou resultados relevantes, o que resultou, nas análises, possíveis oportunidades de compra e venda dos ativos avaliados.

Figura 4.1 – Caso de teste 1

Relatórios de estatísticas

Média dos valores retornados para o intervalo selecionado.
 Data de Referência: 2021-08-08
 Ativo: BOVA11



Fonte: Elaborado pelo autor.

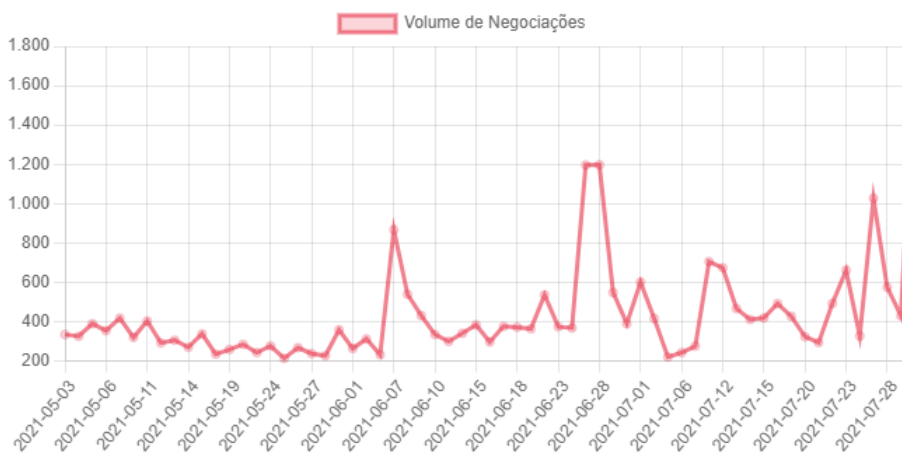
Figura 4.2 – Caso de teste 2

Relatórios de estatísticas

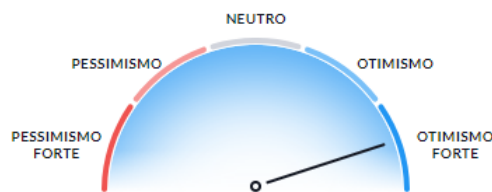
Média dos valores retornados para o intervalo selecionado.
 Data de Referência: 2021-08-01
 Ativo: HGLG11



Open R\$ 167.81	Close R\$ 167.76	High R\$ 168.61	Low R\$ 166.81	Volume 439.092
---------------------------	----------------------------	---------------------------	--------------------------	--------------------------



Sentimento

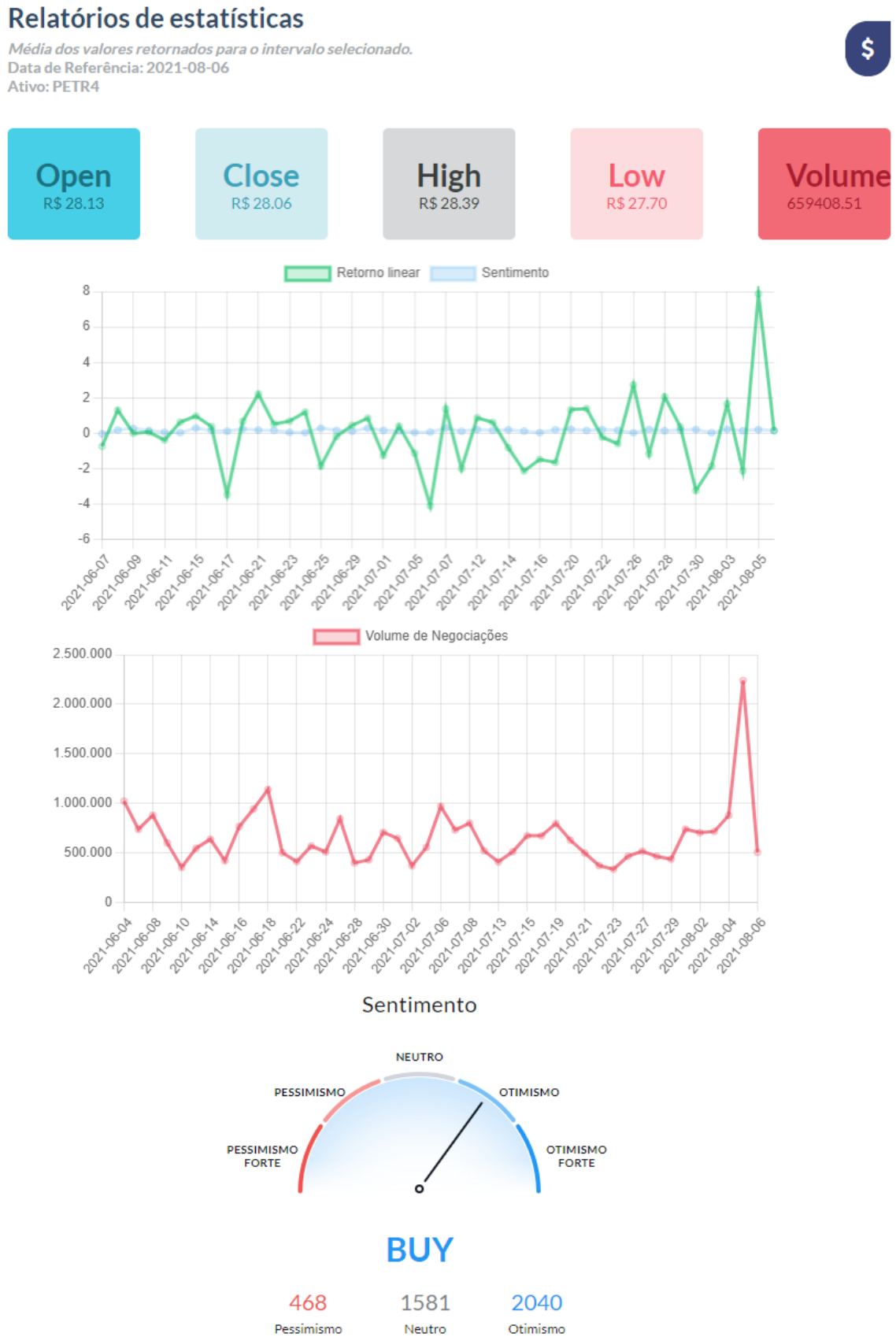


STRONGBUY

736 Pessimismo 64 Neutro 2502 Otimismo

Fonte: Elaborado pelo autor.

Figura 4.3 – Caso de teste 3



Fonte: Elaborado pelo autor.

5 Considerações Finais

Neste capítulo, são apresentados aspectos conclusivos sobre o trabalho desenvolvido (vide Seção 5.1) e também perspectivas de trabalho futuro (vide Seção 5.2) que podem ser realizados na área de análise do *Twitter* para auxílio à tomada de decisão do investidor.

5.1 Conclusão

Conforme apresentado, o objeto geral deste trabalho consistiu na proposta, no desenvolvimento e na validação de uma plataforma que auxilia o processo de compra e venda de ações mediante a divulgação de *tweets* e suas polaridades.

Por meio dos experimentos realizados, quanto ao foco associado à contribuição da plataforma para os usuários, pode-se concluir que os resultados obtidos foram bem satisfatórios o que torna a plataforma eficaz. As análises realizadas, junto a um conhecimento prévio sobre tendências e a volatilidade do mercado financeiro, podem auxiliar o investidor no estudo, baseando-se em datas passadas, sobre os padrões para compra e venda de ações que a plataforma responde, bem como no aproveitamento de oportunidades por meio dos gráficos atuais gerados. Quanto ao foco associado à performance da arquitetura da plataforma, o coletor de páginas do *Twitter* mostrou-se robusto e assertivo. Os métodos utilizados para o processamento e classificação do texto e dos *emojicons* dos *tweets* foram eficientes e seus resultados foram satisfatórios. O sistema de filas adotado provou-se eficaz e confiável, garantindo de forma rápida 100% na entrega das mensagens; isso auxiliou também na divisão do projeto em módulos, o que trouxe resultados positivos, já que todos os processos foram rápidos e sem travamentos por motivo de sobrecarga. A estratégia utilizada para buscar *tweets* foi uma das limitações do trabalho, pois, baseando apenas no código do ativo, o volume de *tweets* retornados foi muito baixo quando comparado ao número de *tweets* diários publicados na plataforma; ademais, o texto destes *tweets*, na maioria dos casos, traziam diversos outros códigos e palavras que eram removidas na etapa de processamento, por não trazerem significado relevante. Adicionar ao grupo de palavras-chave o nome da empresa a qual o código identifica e o nome de usuário de pessoas conceituadas relativas a esta empresa, pode acarretar em melhores resultados na classificação e na análise final.

5.2 Trabalho futuro

Como perspectivas de trabalho futuro, pretende-se: (1) identificar um novo grupo de palavras-chave para serem o índice da pesquisa no *Twitter*; (2) realizar novos experimentos na plataforma para analisar novos padrões entre os resultados; (3) realizar análise numérica

da eficácia dos métodos utilizados na plataforma; (4) realizar um estudo sobre experiência do usuário quanto ao uso da plataforma, a fim de melhorar a interface e a interação.

Referências

- ABITEBOUL, S.; MANOLESCU, I.; RIGAUX, P.; ROUSSET, M.-C.; SENELLART, P. *Web Data Management*. [S.l.]: Cambridge University Press, 2011.
- AGUIAR, M. *Sentiment analysis em relatórios da administração divulgados por firmas brasileiras*. Vitória, 2012. Tese (Doutorado) — Dissertação. Programa de Pós-Graduação em Administração da Fundação . . . , 2012.
- ÁLVAREZ, A. C. *Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem*. Tese (Doutorado) — Universidade de São Paulo, 2007.
- ATSALAKIS, G. S.; VALAVANIS, K. P. Surveying stock market forecasting techniques—part ii: Soft computing methods. *Expert Systems with applications*, Elsevier, v. 36, n. 3, p. 5932–5941, 2009.
- BAKSHI, K. Considerations for big data: Architecture and approach. In: IEEE. *2012 IEEE aerospace conference*. [S.l.], 2012. p. 1–7.
- BOLLEN, J.; MAO, H.; PEPE, A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: *Proceedings of the International AAAI Conference on Web and Social Media*. [S.l.: s.n.], 2011. v. 5, n. 1.
- BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. *Journal of computational science*, Elsevier, v. 2, n. 1, p. 1–8, 2011.
- BONDT, W. F. D.; MURADOGLU, Y. G.; SHEFRIN, H.; STAIKOURAS, S. K. Behavioral finance: Quo vadis? *Journal of Applied Finance (Formerly Financial Practice and Education)*, v. 18, n. 2, 2008.
- BROWN, G. W.; CLIFF, M. T. Investor sentiment and the near-term stock market. *Journal of empirical finance*, Elsevier, v. 11, n. 1, p. 1–27, 2004.
- BUNEMAN, P. Semistructured data. In: *Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*. [S.l.: s.n.], 1997. p. 117–121.
- BYJU’S. *O que é mercado financeiro*. 2020. Acesso em: 30 Set. 2020. Disponível em: <<https://byjus.com/commerce/what-is-financial-market/>>.
- CAMPAGNARO, R. *Renda variável varia!* 2020. Acesso em: 10 Ago. 2021. Disponível em: <<https://fiis.com.br/artigos/renda-variavel-varia/>>.
- CHEN, H.; ZIMBRA, D. Ai and opinion mining. *IEEE Intelligent Systems*, IEEE, v. 25, n. 3, p. 74–80, 2010.
- COMSCORE, I. *2015 Brazil Digital Future in Focus*. 2015. Acesso em: 01 jun. 2020. Disponível em: <<https://www.comscore.com/por/Insights/Apresentacoes-e-documentos/2015/2015-Brazil-Digital-Future-in-Focus>>.
- COWIE, J.; LEHNERT, W. Information extraction. *Communications of the ACM*, ACM New York, NY, USA, v. 39, n. 1, p. 80–91, 1996.

- D'AGOSTO, M. *Redes sociais no mercado financeiro*. 2012. Acesso em: 24 jun. 2020. Disponível em: <<https://www.valor.com.br/valor-investe/o-consultor-financeiro/2823600/redes-sociais-no-mercado-financeiro>>.
- DESCHATRE, G. A.; MAJER, A. *Aprenda a investir com sucesso em ações: análise técnica e fundamentalista*. [S.l.]: Ciência Moderna, 2006.
- DIAS, M. M. et al. Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados. Florianópolis, SC, 2001.
- DISTRITO. *Data mining: conceito e importância da mineração de dados para a sua empresa*. 2020. Acesso em: 13 jul. 2020. Disponível em: <<https://distrito.me/data-mining/>>.
- DONDIO, P. Predicting stock market using online communities raw web traffic. In: IEEE. *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. [S.l.], 2012. v. 1, p. 230–237.
- ELBOGHADY, D. *Market quavers after fake AP tweet says Obama was hurt in White House explosions*. 2013. Acesso em: 30 may. 2021. Disponível em: <https://www.washingtonpost.com/business/economy/market-quavers-after-fake-ap-tweet-says-obama-was-hurt-in-white-house-explosions/2013/04/23/d96d2dc6-ac4d-11e2-a8b9-2a63d75b5459_story.html>.
- EMERSON, K. *An Introduction to Web Scraping for Research*. 2019. Acesso em: 12 nov. 2020. Disponível em: <<https://researchdata.wisc.edu/news/an-introduction-to-web-scraping-for-research/>>.
- FALCÃO, M. C. dos S. *ANÁLISE DE SENTIMENTO DE NOTÍCIAS DO MERCADO FINANCEIRO*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2020.
- FAN, Y. Design and implementation of distributed crawler system based on scrapy. In: *IOP Conference Series: Earth and Environmental Science*. [S.l.: s.n.], 2018. v. 108, n. 4, p. 042086.
- FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37–37, 1996.
- FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. et al. Knowledge discovery and data mining: Towards a unifying framework. In: *KDD*. [S.l.: s.n.], 1996. v. 96, p. 82–88.
- FELIPE, I. J. d. S. *Determinantes do sucesso de campanhas de equity e de reward crowdfunding*. Tese (Doutorado), 2017.
- FERREIRA, F.; UNGARETTI, M. *Ritmo de crescimento de pessoas físicas na Bolsa desacelera no início de 2021*. 2021. Acesso em: 30 may. 2021. Disponível em: <<https://conteudos.xpi.com.br/acoes/relatorios/crescimento-de-pessoas-fisicas-na-bolsa-desacelera-no-inicio-de-2021/>>.
- FERREIRA, K. *Coleta de dados: o que é, ferramentas e como fazer no marketing?* 2020. Acesso em: 12 nov. 2020. Disponível em: <<https://rockcontent.com/br/blog/coleta-de-dados/>>.
- FUNDSEXPLORER. *HGLG11*. 2019. Acesso em: 10 Ago. 2021. Disponível em: <<https://www.fundsexplorer.com.br/funds/hglg11>>.
- GOMES, P. C. T. *Análise de Sentimentos com Machine Learning*. 2019. Acesso em: 08 Out. 2020. Disponível em: <<https://www.datageeks.com.br/analise-de-sentimentos/>>.

MILLER, M. *Visual Analytics of Spatio-Temporal Event Predictions: Investigating Causes for Urban Heat Islands*. Tese (Doutorado), 03 2018.

MODALMAIS. *BOVA11: O que é e como investir*. 2020. Acesso em: 09 Ago. 2021. Disponível em: <<https://www.modalmmais.com.br/blog/bova11>>.

MTK, C. F. *O que é análise de sentimentos?* 2018. Acesso em: 08 Out. 2020. Disponível em: <<https://controllf5mkt.com.br/blog/o-que-e-analise-de-sentimentos/>>.

MULLER, N. *O impacto da tecnologia em nossas vidas*. 2016. Acesso em: 13 jul. 2020. Disponível em: <<https://www.oficinadanet.com.br/post/16174-o-impacto-da-tecnologia-em-nossas-vidas>>.

NETO, B.; ARRUDA, N. Análise de sentimentos do twitter como suporte aditivo para a previsão da volatilidade do bitcoin. 2018.

NEWBURGER, E. *JP Morgan has created an index to track the effect of Trump's tweets on financial markets: 'Volfefe index'*. 2019. Acesso em: 24 jun. 2020. Disponível em: <<https://www.cnbc.com/2019/09/08/donald-trump-is-tweeting-more-and-its-impacting-the-bond-market.html>>.

PALMER, D. *Handbook of Natural Language Processing: Text Pre-processing*. [S.l.]: 2. ed. Boca Raton, Florida: CRC Press, Taylor and Francis Group, 2010.

PAPAZOGLU, M. P. Service-oriented computing: Concepts, characteristics and directions. In: IEEE. *Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003. WISE 2003*. [S.l.], 2003. p. 3–12.

PATEL, N. *As 10 Redes Sociais Mais Usadas no Brasil (e no Mundo) em 2018 e 2019*. 2019. Acesso em: 07 Out. 2020. Disponível em: <<https://neilpatel.com/br/blog/redes-sociais-mais-usadas/>>.

PERES, V.; VIEIRA, R.; BORDINI, R. Análises de sentimentos: abordagem lexical de classificação de opinião no contexto mercado financeiro brasileiro. 2019.

RANCO, G.; ALEKSOVSKI, D.; CALDARELLI, G.; GRČAR, M.; MOZETIČ, I. The effects of twitter sentiment on stock price returns. *PLoS one*, Public Library of Science, v. 10, n. 9, 2015.

REGAL, A. A.; AÑON, J. M.; FABBRI, C.; HERRERA, G.; YAULLI, G.; PALOMINO, A.; GIL, C. Proyección del precio de criptomonedas basado en Tweets empleando LSTM. *Ingeniare. Revista chilena de ingeniería*, scielocl, v. 27, p. 696 – 706, 12 2019. ISSN 0718-3305. Disponível em: <http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-33052019000400696&nrm=iso>.

REIS, T. *O que é mercado financeiro*. 2018. Acesso em: 06 Out. 2020. Disponível em: <<https://www.sunoresearch.com.br/artigos/mercado-financeiro/#:~:text=Em%20economia%2C%20mercado%20financeiro%20%C3%A9,ativos%20com%20algum%20valor%20financeiro./>>>.

RICHARDSON, C. Introduction to microservices. *Recuperado de https://www.nginx.com/blog/introduction-to-microservices*, 2015.

- RICONNECT. *Horários de Abertura das Bolsas de Valores (Tabela Completa)*. 2020. Acesso em: 09 Ago. 2021. Disponível em: <<https://riconnect.rico.com.vc/blog/horarios-de-abertura-das-bolsas-de-valores-tabela-completa>>.
- RILOFF, E.; LEHNERT, W. Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems (TOIS)*, ACM New York, NY, USA, v. 12, n. 3, p. 296–333, 1994.
- RISH, I. et al. An empirical study of the naive bayes classifier. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. [S.l.: s.n.], 2001. v. 3, n. 22, p. 41–46.
- ROCHA, T. *Web Scraping: Como encontrar insights valiosos analisando os números de qualquer blog*. 2018. Acesso em: 12 nov. 2020. Disponível em: <<https://resultadosdigitais.com.br/blog/web-scraping/>>.
- SANTOS, H.; LAENDER, A.; PEREIRA, A. C. Uma visão do mercado brasileiro de ações a partir de dados do twitter. In: *SBC. Anais do IV Brazilian Workshop on Social Network Analysis and Mining*. [S.l.], 2015.
- SANTOS, L. M. Protótipo para mineração de opinião em redes sociais: estudo de casos selecionados usando o twitter. *Monografia. Departamento de Ciência da Computação, Universidade Federal de Lavras*, 2010.
- SILVA, E. F.; BARROS, F. A.; PRUDÊNCIO, R. B. Uma abordagem de aprendizagem híbrida para extração de informação em textos semi-estruturados. In: *Anais do XXV Congresso da Sociedade Brasileira de Computação*. [S.l.: s.n.], 2005. v. 1, p. 504–13.
- SILVA, J. *Enterprise Messaging with RabbitMQ and AMQP*. 2017. Acesso em: 30 jun. 2021. Disponível em: <<http://throughaglass.io/technology/Enterprise-Messaging-with-RabbitMQ-and-AMQP.html>>.
- SILVA, N. F. F. d. *Análise de sentimentos em textos curtos provenientes de redes sociais*. Tese (Doutorado) — Universidade de São Paulo, 2016.
- SMAILOVIĆ, J.; SLUBAN, B.; MOZETIČ, I. et al. Sentiment of emojis. *Plos one*, v. 10, n. 12, p. e0144296–e0144296, 2015.
- SORENSEN, F. *AMQP vs HTTP*. 2017. Acesso em: 30 jun. 2021. Disponível em: <<https://dev.to/fedejsoren/amqp-vs-http>>.
- SOUSA, G. L. S. de. *Tweetmining: Análise de opinião contida em textos extraídos do twitter*. 2012.
- SOUZA, T. T. P.; KOLCHYNA, O.; TRELEAVEN, P. C.; ASTE, T. Twitter sentiment analysis applied to finance: A case study in the retail industry. *arXiv preprint arXiv:1507.00784*, 2015.
- STATUSBREW. *100 Social Media Statistics For Businesses 2019 + [Infographic PDF]*. 2018. Acesso em: 04 jun. 2020. Disponível em: <<https://statusbrew.com/insights/social-media-statistics-2019/>>.
- SULZ, P. *O guia completo de Redes Sociais: saiba tudo sobre as plataformas de mídias sociais!* 2019. Acesso em: 13 jul. 2020. Disponível em: <<https://rockcontent.com/blog/tudo-sobre-redes-sociais/>>.

TEAM, M. *Redes Sociais e Big Data: a melhor estratégia para conhecer seus clientes*. 2020. Acesso em: 04 jun. 2020. Disponível em: <<https://www.mjvinnovation.com/pt-br/blog/redes-sociais-e-big-data-a-melhor-estrategia-para-conhecer-seus-clientes/>>.

TERRA, É. D. Ferramenta para extração de dados do twitter para mineração de dados. 2016.

TORO. *Mercado de Ações - o que é e como funciona*. 2018. Acesso em: 09 Ago. 2021. Disponível em: <<https://artigos.toroinvestimentos.com.br/mercado-de-acoes-como-funciona-curso>>.

WEBER, P. M. B. Middleware com escalonamento de aplicações. 2016.

XAVIER, O. C. *Utilizando a API do Twitter no desenvolvimento de aplicações web com PHP e cURL*. 2020. Acesso em: 07 Out. 2020. Disponível em: <