ANÁLISE PREDITIVA DO MERCADO DE AÇÕES COM PROCESSAMENTO DE LINGUAGEM NATURAL

PILONETO, Julia[1]; VIEIRA, Mateus[1]; HECHT, Renan[1]
MEDEIROS, Luciano[2]

RESUMO

Este projeto tem o objetivo de prever o valor de ações diversas no mercado de ações brasileira (B3) utilizando notícias disponibilizadas em sites online sobre a empresa as quais pertencem. Para tal, foi desenvolvido uma solução utilizando a linguagem Python no ambiente do Google Colab, nele o usuário define o nome da empresa que deseja pesquisar e um *Web Crawler* automaticamente retorna os links das notícias desta empresa, estes links são então acessados e o seu conteúdo é copiado, resumido e gravado em um arquivo de texto. Em seguida, estes dados foram submetidos por um processo de análise de linguagem natural que permitiu definir o sentimento de cada frase e então classificá-las, em relação ao desempenho da empresa, em “positivo”, “negativo” e “neutro”. Esta classificação é feita com o auxílio de um banco de dados gravado em um arquivo no formato *json*, que contém frases já atribuídas à suas classificações. Este arquivo é utilizado no treinamento da rede neural e diretamente impacta no desempenho da avaliação. De forma geral, este projeto alcançou seu objetivo, o código funcionou corretamente em todas as etapas e os resultados encontrados foram condizentes com as expectativas: notícias sobre quedas na bolsa retornam sentimentos negativos, notícias sobre valorização do ativo comercializado pela empresa são compreendidos como sentimentos positivos.

Palavras-chave: inteligência artificial, big data, linguagem natural, python, mercado de ações.

[1] Estudante de Engenharia da Computação modalidade presencial da UNINTER.
[2] Professor orientador.
Artigo de Trabalho de Conclusão de Curso entregue como requisito parcial para obtenção do grau de bacharel em Engenharia da Computação ao Centro Universitário Internacional UNINTER, Curitiba – PR, 2020.

ABSTRACT

This project aims to predict various stock values in the Brazilian stock market (B3) using news available on online websites about the company that they belong to. To this end, a solution was developed using the Python language in the Google Colab environment, in which the user defines the name of the company to be searched and a web crawler automatically returns links to news of this company, these links are then accessed and their content is copied, summarized and saved to a text file. Then, this data is submitted to a natural language analysis process that defines the sentiment of each sentence and then classifies it, in relation to the company's performance, in “positive”, “negative” and “neutral”. This classification is made with the aid of a database recorded in a json file format, which contains phrases already assigned to their respective classifications. This file is used in the neural network training and directly impacts the performance of the evaluation. In general, this project achieved its objective, the code worked correctly in all stages and the results found were in line with the expectations: news about falls in the stock market returned negative sentiments, news about the value of the asset sold by the company are correctly understood as positive sentiments.

Keywords: artificial intelligence, big data, natural language, python, stock market.

1 INTRODUÇÃO

A inteligência artificial é um campo de estudo que busca criar algoritmos similar à inteligência humana, isto é, que é capaz de interpretar dados e tomar decisões rumo a um objetivo específico (RUSSEL, S.; NORVIG, P. 2004). De maneira simplificada, uma inteligência artificial estuda padrões de comportamento para depois imitá-los, as inteligências artificiais podem ser adaptadas para diversas situações, como criar imagens, criar textos e até mesmo tomar decisões financeiras.

No mercado financeiro, investidores que operam na bolsa de valores deve primeiramente pesquisar sobre as empresas em que tem intenção de se associar, para depois escolher quais são propensas a ganhar ou perder valor. Este tipo de

análise, denominada de “fundamentalista”, deve ser repetida para cada empresa, e mesmo assim, eventos especulativos frequentemente impactam os valores de diversas empresas simultaneamente, aumentando o risco deste tipo de investimento (FORTUNA, 2020).

Como esses eventos são reportados por jornais e revistas, este projeto irá validar a utilização de processamento de linguagem natural na identificação de padrões linguísticos utilizados para expressar emoções positivas ou negativas, e como afetam o mercado financeiro.

Para tal, serão utilizados como base dois projetos: o primeiro é um projeto desenvolvido por que criou uma inteligência artificial preditiva para o desempenho de empresas na bolsa de valores com base na associação de mensagens positivas ou negativas sobre a empresa na rede social *Twitter* (MITTAL, e GOEL, 2011). O segundo projeto realiza a análise de desempenho da criptomoeda com as *Tags* do *Twitter* para prever o seu comportamento (GALESHCHUK *et al*, 2018) (BOLLEN, J; MAO, H; ZENG, 2010).

2 FUNDAMENTAÇÃO TEÓRICA

2.1 A INTELIGÊNCIA ARTIFICIAL E A ANÁLISE DE DADOS

2.1.1 O sistema financeiro

O mercado financeiro é um mecanismo que permite a troca de recursos financeiros, desta maneira, unidades que tenham excesso de fundos podem aplicar naquelas que têm necessidades de fundos. Para que o retorno desse empreendimento seja maximizado, o mercado financeiro também deve fornecer informações de forma fácil, barata e rápida para o investidor, sem limitar de entrada ou saída de compradores e vendedores ou impor prazos ou limites injustos em seus participantes (LIMA, 2020).

O mercado de capitais é um conjunto de instituições e acionistas que negociam títulos e valores mobiliários. Ou seja, o mercado de capitais é a parte do mercado

financeiro que distribui valores mobiliários com o propósito de viabilizar a capitalização das empresas e dar liquidez aos títulos emitidos por elas (FORTUNA, 2020).

As ações são títulos de participação negociáveis, que representam parte do capital social de uma sociedade econômica, que confere ao seu possuidor o direito de participação de sua vida social. Podem ser consideradas como um certificado ou título de propriedade, representativo das partes do capital social de uma sociedade econômica. Portanto, quem tem ações, pode se considerar sócio da empresa emissora (FORTUNA, 2020).

2.1.2 Valor de uma ação

Dependendo da situação as ações apresentam valores monetários diferentes, conforme abaixo:

- **Valor nominal**: corresponde ao capital dividido pelo número total de ações emitidas.
- **Valor patrimonial**: corresponde ao patrimônio líquido da empresa.
- **Valor contábi**l: pode ser explícito (calculado pelo valor nominal e preço de emissão) ou indiscriminado (sem valor).
- **Valor de liquidação**: corresponde ao valor avaliado em caso de encerramento da empresa.
- **Valor intrínseco**: o valor real avaliado no processo de análise fundamentalista.
- **Valor de mercado**: o valor que os compradores aceitam pagar e os vendedores recebem para fazê-lo em mercados organizados.

2.1.3 Estratégias na comercialização de ações

A análise fundamentalista é o estudo de toda informação disponível no mercado sobre determinada empresa, com a finalidade de obter seu verdadeiro valor e formular uma recomendação sobre sua compra ou venda. O analista resume e analisa a informação, parte do passado e trata de predizer o futuro, para dar sua opinião (YOSHIKAWA et al, 2019).

Esse tipo de análise busca antecipar o comportamento futuro de determinada empresa no mercado. Para tal hipotetiza-se que o preço de uma ação não reflete o verdadeiro valor da empresa, e assim seria possível descobrir os seus picos de valorização com base informações ainda não descobertas pelo mercado (YOSHIKAWA et al, 2019).

A análise fundamentalista tradicional utilizará como base o valor financeiro passado da empresa, a situação da economia atual no seu país de origem e especulação gerada por terceiros, assim como a intervenção governamental (LIMA, 2020).

Esta análise pode ser complementada com a especulação gerada por eventos e notícias, por exemplo, em novembro de 2019 a empresa fabricante de armas brasileira Taurus valorizou mais de 400% com a notícia de que as leis de portes de arma poderiam ser alteradas (BASSOTO, 202).

2.1.4 Big Data

A expressão Big Data se refere ao conjunto de dados produzidos cujo volume está além dos padrões e da capacidade das ferramentas utilizadas por modelos de bancos de dados tradicionais para capturá-los, analisá-los e gerenciá-los. Esses dados são produtos de processos tecnológicos atuais, como as mídias sociais, que geram muitos dados a todo instante no mundo inteiro (BERMAN, 2013).

O tratamento dos dados é realizado com o apoio de algoritmos que permitem que se chegue a uma conclusão sobre que tipo de ação tomar. Esses algoritmos, são a “rede neural” do sistema e podem servir para fins diversos dependendo do propósito buscado pela corporação. Os dados podem ser classificados em estruturados (que possuem formato e comprimento definido) ou, não estruturados (que não seguem um formato específico) e semiestruturados (que possui algum tipo de estrutura marcada) com base no seu gerenciamento e armazenamento (MORAIS et al, 2018).

A coleta e o armazenamento de dados têm como finalidade a extração de informações que possam gerar vantagens competitivas para as organizações, bem

como auxiliar nas tomadas de decisões. Porém, os dados ainda necessitam ser analisados. Os sistemas gerenciais têm como característica o fato de apresentarem as informações, mas a inteligência nos negócios converge para a análise detalhada dos dados, a procura de padrões, modelos ou repetições (BERMAN, 2013).

2.1.5 Mineração de dados (Data Mining)

A mineração de dados está relacionada às áreas da inteligência artificial, estatística clássica, e aprendizado de máquina. A mineração de dados faz parte do processo de KDD (*Knowledge Discovery in Databases*) e geralmente lida com grande volume de dados, mas tem a capacidade de mudar de escala com relação ao tamanho dos dados (RATNER, 2011). As principais etapas do processo de mineração são:

1. Selecionar os algoritmos que serão necessários;
2. Aplicar os algoritmos em amostras dos dados;
3. Síntese dos resultados;
4. Aplicar outras ferramentas, conforme necessário;
5. Repetir o processo.

A mineração de dados é tipicamente utilizada em uma das quatro tarefas a seguir (RATNER, 2011):

- ***Clustering***: é a descoberta e agrupamento de dados parecidos ou semelhantes, este tipo de técnica é feita sem utilizar estruturas de dados já conhecidos.
- **Classificação**: é a tentativa de aplicar uma estrutura de dados já conhecida em dados novos, a fim de classificá-los.
- **Regressão**: procura uma função matemática a qual os dados se encaixam com o menor erro possível.
- **Regra de associação**: procura encontrar padrões entre como diversos dados são adquiridos e se existe algum tipo de influência de um dado sobre o outro.

2.1.7 Inteligência Artificial

O campo de inteligência artificial tem como objetivo a criação de máquinas inteligentes, ou seja, softwares que consigam extrair, criar, e aprimorar dados. Atualmente, as inteligências artificiais são aplicadas em diversos campos (médico, mecânico, agronegócio) com o objetivo de facilitar tarefas repetitivas, ou aprimorar processos industriais (RUSSEL, S.; NORVIG, P. 2004).

O funcionamento técnico de uma inteligência artificial é feito por funções matemáticas. Em um comportamento padronizado, uma função pode ser utilizada para explicar como um resultado y surge a partir de x. Ao aplicar dados à esta função, uma inteligência artificial pode deduzir resultados plausíveis, mesmo que com uma taxa de erros.

Para diminuir a taxa de erros, é necessária uma grande quantidade de dados verídicos, ou seja, em que temos certeza de que a fórmula de o resultado esperado. A IA irá adaptar os parâmetros de sua função para tentar encaixar estes dados. Após este processo, utilizam-se outros dados, também verídicos, para testar a função gerada. Se os resultados estiverem suficientemente próximos, a IA estará pronta para uso, senão o processo é refeito.

2.1.8 Aprendizado de Máquina

Aprendizado de máquina (*Machine learning*) pode ser definido como um conjunto métodos estatísticos computacionais que consistem no uso de uma base de dados que descrevem determinado fenômeno. Estes algoritmos utilizam de experiência e treinamento para melhorar sua performance e realizar suas previsões bem definidas com o foco em solucionar problemas práticos (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018) (BURKOV, 2019) (KREUTZ, 2015). Existem diversos tipos de metodologias de aprendizado de máquina, destacaremos as duas diferentes técnicas que lidam com os principais desafios da análise de sentimentos textuais, sendo essas o aprendizado supervisionado (*Supervised Learning*, do Inglês) e o aprendizado não supervisionado (*Unsupervised Learning*, do Inglês). A primeira abordagem composta por técnicas supervisionadas emprega o termo supervisionado

justamente pelo fato de exigir uma etapa de treinamento de um modelo com amostras previamente classificadas. O procedimento para realizar a aprendizagem de máquina compreende quatro etapas principais:

1. Obtenção de dados rotulados que serão utilizados para treino e para teste;
2. Definição das *Features* ou características que permitam a distinção entre os dados;
3. Treinamento de um modelo computacional com um algoritmo de aprendizagem;
4. Aplicação do modelo.

Esta não carece de sentenças previamente rotuladas e treinos para a criação de um modelo, sendo uma das suas principais vantagens uma vez que desta forma não mantém aplicação restrita ao contexto para o qual foram treinados. Dentre as técnicas não supervisionadas destacam-se aquelas com abordagens léxicas, estas assumem que palavras individuais possuem o que é chamado de polaridade prévia, que é, uma orientação semântica independente de contexto e que pode ser expressa com um valor numérico ou classe (TABOADA, M; ANTHONY, C e VOLL, K, 2006).

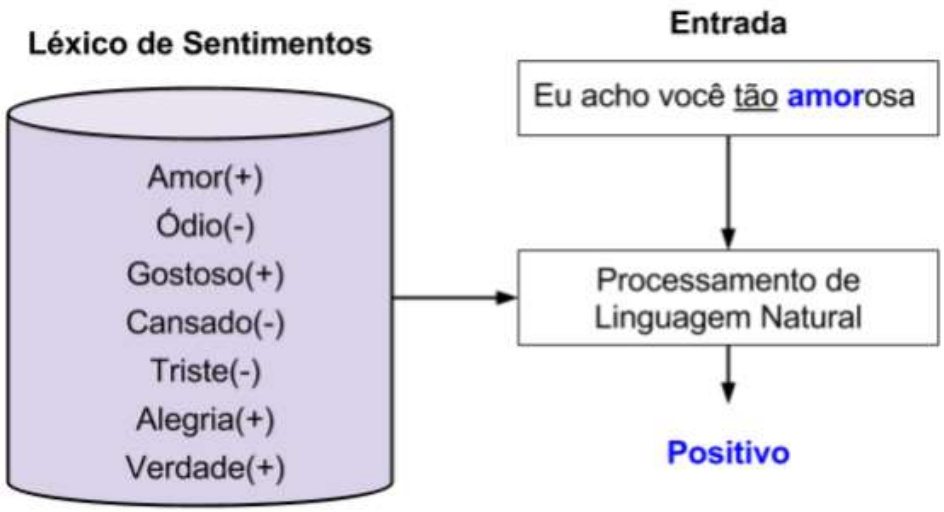

Figura 1: Léxico de sentimentos

2.1.9 Aprendizado Profundo

O aprendizado profundo ou, do inglês, *Deep Learning,* é uma subclasse da inteligência artificial que está em constante ascensão e é capaz de ser aplicada em vários tipos de tecnologias que requerem grande volume de dados. Uma arquitetura de *Deep Learning* é caracterizada por utilizar uma grande camada interna de módulos mais simples chamados de *hidden layer* em que todos, ou grande parte destes módulos, estão sujeitos a aprendizado, e no final desta sucessão produzem dados de saída. Cada módulo desta cadeia transforma seus dados de entrada (*inputs*) para aumentar tanto a seletividade quanto a invariância da representação. Com múltiplas camadas, um sistema é capaz de implementar funções extremamente complexas que sejam sensíveis aos mínimos detalhes e não sensíveis a grandes variações irrelevantes nos dados de entrada (LECUN; BENGIO; HINTON, 2015). A tendência é que a área continue evoluindo, já que *Deep Learning* é um mecanismo que aproveita o crescente número de dados computacionais disponíveis, por efeito de ser uma metodologia que necessita de pouco trabalho manual para ser implementado, ou seja, requer uma menor interferência do ser humano no processo. Essa é uma das principais vantagem do *Deep Learning* em contraste com técnicas tradicionais de *Machine Learning* (LECUN; BENGIO; HINTON, 2015) (GAO; ZHANG; WEI, 2018).

2.1.10 Frentes de Pesquisa

As frentes de pesquisa nessa área são divididas em diferentes níveis de granularidade conforme a tarefa de detecção de sentimentos nos textos. Quanto menor a granularidade, mais específica é a classificação.

Estado emocional: Esse é um recurso que permite as empresas acompanhem a satisfação pós-venda de seus produtos. Outro grande exemplo da importância da análise do estado emocional e do como ela pode ser essencial, são pesquisas capazes de caracterizar e prever experiências de depressão pós-parto em mães de recém-nascidos (DE CHOUDHURY, 2016) através de dados compartilhados no Facebook (BENEVENUTO et al., 2020).

Análise de Sentimentos para comparação ou *Comparative Sentiment Analysis*: Em diversos casos usuários não informam a opinião direta sobre um produto ou pessoa, no entanto, eles fornecem opiniões comparativas em sentenças

como “Este computador Apple aparenta ser bem melhor do que aquele Asus”, “Eu dirijo um carro X, mas a mudança de marcha é bem pior que a do carro Y”. O objetivo da análise de sentimentos neste caso é identificar as sentenças que contém as opiniões para ser comparadas (utilizando, por exemplo, advérbios como piores que, melhor que) e assim extrair a entidade referida daquela opinião (FELDMAN, 2013) (BENEVENUTO et al., 2020).

Nível de Documento: Nesse nível de granularidade, a classificação de sentimentos ocorre com a análise de um texto como um todo. Ou seja, nesse nível, assume-se que todo o texto está relacionado a um único assunto que possui certa polaridade. Na prática, se no documento possuir várias entidades com opiniões diferentes, então seus sentimentos podem ser diferentes. Desta forma é difícil assimilar um sentimento ao documento todo, mas um caso interessante em que a análise em nível de documento pode ser utilizada é em reviews de produtos ou filmes por exemplo (LIU, 2010).

Nível de Sentença: É nesse nível de análise que este trabalho se dedica, pois um único documento pode conter múltiplas opiniões ou mesmo entidades. Neste caso é assumido que o texto foi dividido em frases ou sentenças que possam conter uma opinião individualmente. Cabe ressaltar que, em geral, postagens e comentários em mídias sociais seguem um padrão de sentenças curtas. Quando se pode monitorar as redes sociais, abre-se uma variedade de oportunidades de estudo, um caso interessante é o monitoramento do Twitter para previsão de bolsa de valores (BOLLEN, J; MAO, H; ZENG, 2010).

Nível de Palavra ou Dicionário: Nessa frente de pesquisa os trabalhos focam em otimizar os Léxicos de sentimentos existentes na literatura. Não é claro a melhor maneira de se construir um dicionário de sentimentos. No entanto, existem diversos dicionários e suas principais diferenças são constituídas pelas palavras que os formam e às vezes na adição de gírias e acrônimos vindas das redes sociais, como “vc”, “blz”, “tb”. A inclusão de diferentes termos é importante para alcançar melhor desempenho quando se trabalha com o foco em mídias sociais. Existem outras diferenças entre tais dicionários como a forma que é avaliada a palavra, binária (positivo/negativa) ou proporcional à força do sentimento (-1 a 1) (NIELSEN, 2011).

Nível de Aspecto: Nesse nível de granularidade, uma sentença pode ser julgada por várias entidades e pode conter múltiplos sentimentos associados a ela. Por exemplo, a sentença “Esse hotel, apesar de possuir um ótimo quarto, tem um atendimento péssimo!” possui duas diferentes polaridades associadas a “quarto” e “atendimento” para o mesmo hotel. Enquanto “quarto” é considerado positivo, “atendimento” pode ser analisado de forma negativa. Esta necessidade de avaliar a opinião para cada entidade é comum em reviews de produtos ou em fóruns de discussões. Existem dois motivos para a popularidade desse nível de sentença. O primeiro deles é a aplicabilidade no contexto de redes sociais em que grande parte dos textos produzidos são sentenças ou textos curtos. Em casos como o do *Twitter*, por exemplo, existe a limitação no número de caracteres postados. Além disso, a análise de sentimentos em sentença é, geralmente, a base para os demais níveis. Os métodos atuais de detecção de sentimentos em sentenças podem ser divididos em duas classes: os baseados em aprendizado de máquina e os métodos léxicos. Métodos baseados em aprendizado de máquina geralmente dependem de bases de dados rotuladas para treinar classificadores (PANG, 2002), o que pode ser considerado uma desvantagem, devido ao alto custo na obtenção de dados rotulados. Por outro lado, métodos léxicos utilizam listas e dicionários de palavras associadas a sentimentos específicos. Apesar de não dependerem de dados rotulados para treinamento, a eficiência dos métodos léxicos está diretamente relacionada a generalização do vocabulário utilizado, para os diversos contextos existentes. A seção seguinte apresentará em detalhes as abordagens supervisionadas e léxicas que vêm sendo utilizadas nesse contexto.

2.1.11 Considerações sobre Aprendizado de Máquina

Uma importante consideração a ser destacada a respeito de técnicas de aprendizado de máquina é que neste tipo de estratégia o modelo gerado pode ir muito bem nos conjuntos de dados para o qual ele foi treinado fazendo com que resultados sejam razoáveis no treino mas no momento de testes os resultados apresentem resultados bem diferentes. Tal situação, conhecida como *Overfitting*, deve ser evitada e existem metodologias que devem ser seguidas para que isto não aconteça.

As técnicas de aprendizado apresentam algumas dificuldades que serão descritas a seguir. A primeira delas diz respeito à aplicabilidade do modelo que, em geral, são bem restritos ao contexto para o qual foram criados. Outro ponto é a necessidade de boa quantidade dados para treinamento, previamente validados, muitas vezes de difícil obtenção. A escolha dos dados para treinamento deve ainda ser cuidadosa pois caso sejam mal escolhidos podem criar um viés muito grande no modelo tornando-o tendencioso a dar como saída uma classe específica. Além disso, a abordagem supervisionada pode ser computacionalmente cara em termos de processamento da CPU e memória gerar o modelo da aprendizagem. Esta característica pode restringir a capacidade de avaliar um sentimento em dados de streaming por exemplo.

Por fim, algumas características utilizadas para alimentar a aprendizagem de máquina são derivadas de algoritmos que geram um modelo dificilmente interpretável por seres humanos. Isto torna os modelos difíceis de generalizar, modificar ou estender (para outros domínios por exemplo) (HUTTO; GILBERT, 2014).

2.1.12 Processamento de linguagem natural

O campo de processamento de linguagem natural tem como objetivo permitir que um computador seja capaz de analisar, compreender e gerar a linguagem utilizada por seres humanos (BIRD et al, 2019). Esse tipo de linguagem, denominada de linguagem natural, é bem menos estruturada do que a linguagem utilizada por um computador, e além disso está em perpétua mudança, novas gírias e expressões são adicionadas e esquecidas o tempo todo.

Uma aplicação das metodologias recentes de processamento de linguagem natural envolveria o treinamento de um modelo para reconhecer quais palavras provavelmente tem significado afirmativo, e com estes dados seria capaz de reconhecer e inferir mais comandos em linguagem natural. Esse tipo de aplicação já é utilizado em larga escala com *chatbots,* assistentes virtuais que simulam o atendimento com um cliente por um meio escrito a fim de resolver problemas simples e comuns.

Outras aplicações incluem a coleta de dados inteligente, o computador analisa um texto para determinar algum tipo de informação, por exemplo, encontrar o maior problema em reclamações feitas pela internet.

2.1.13 Coleta de dados

A coleta de dados é a base para realizar a análise destes. Com isso são utilizados alguns métodos em conjunto para ter o dado final.

2.1.13.1 Web Scraping

O *Web Scraping* pode ser definido como: “o ato de baixar automaticamente os dados de uma página web e extrair informações muito específicas dela. As informações extraídas podem ser armazenadas praticamente em qualquer lugar (banco de dados, arquivo, etc.)” (DATA SCIENCE ACADEMY, 2018).

Podendo ser traduzido para extração de dados da *Web*, *Web Scraping* nada mais é que uma maneira automatizada de coletar informações ou conteúdos usando *Bots* (robôs), conhecidos como *Scrapers*. Após essa coleta os dados serão armazenados ou já utilizado em seguida em outro site.

Podemos resumidamente separar em 3 partes a coleta de dados:

- Definir quais os dados que vão ser coletados (assunto, sites específicos, etc...);
- Os dados são coletados direto do HTML dos sites, e transformados para um formato *raw*, como em um *json* um até mesmo um simples formato de texto.
- Após isso é enviado para os próximos passos, seja direto para o processamento ou armazenamento.

2.1.13.2 Web Crawling

Temos também uma outra técnica para a coleta de dados, a qual podemos definir como: “o ato de baixar automaticamente os dados de uma página web, extrair os

hiperlinks contidos nela e segui-los. Os dados baixados são geralmente armazenados em um índice ou banco de dados para facilitar sua busca.” (DATA SCIENCE ACADEMY, 2018). Neste método apenas é feito uma indexação de dados na internet, e não é feito nenhuma tratativa em cima deles. Um exemplo são os motores de pesquisa como Google, quando realizamos uma pesquisa nele, ele nos retorna tudo o que possui indexado sobre o que foi digitado.

Crawlers também podem ser usados para tarefas de manutenção automatizadas em um Web Site, como chegar os links ou validar o código HTML. Também podem ser usados para obter tipos específicos de informações das páginas da Web, como minerar endereços de e-mail (mais comumente para spam). Em geral, ele começa com uma lista de URLs para visitar (também chamado de *Seeds*). À medida que os *Crawlers* visitam essas URLs, eles identificam todos os links na página e os adicionam na lista de URLs para visitar. Tais URLs são visitadas recursivamente de acordo com um conjunto de regras. (Redação Global AD, 2019)

Com *Web Crawling* obtemos informações genéricas e com *Web Scraping*, obtemos informações específicas. Também é importante entender a diferença entre *Web Scraping* e a Mineração de Dados (*Data Mining*). Resumindo, enquanto a mineração de dados pode acontecer em qualquer matriz de dados e pode ser feita manualmente, o *Web Scraping* ocorre apenas nas páginas web sendo executados por robôs especiais – *Scrapers*/*Crawlers*. Atualmente grande parte dos Scrapers são escritos utilizando a linguagem Python, para facilitar as próximas etapas de processamento. Junto são utilizados alguns *Frameworks* e bibliotecas, como o *Scrapy* e *Selenium*.

2.1.13.3 Armazenamento de dados

O armazenamento de dados é vital após a sua coleta, já que nem sempre será utilizado de forma imediata. Neste projeto, os dados são armazenados em arquivos de texto (extensão “.txt”) no formato JSON, que é um formato de texto composto de duas estruturas: um conjunto de pares de nome e valores, semelhantes aos *dict* da

linguagem Python. Sua outra estrutura é uma sequência ordenada de valores, semelhantes à uma lista (ECMA, 2018).

2.2 METODOLOGIA

O projeto foi desenvolvido no Google Colab, um ambiente de desenvolvimento baseado em Jupyter Notebook, que permite a adição de células de códigos que podem ser executadas em sequência ou individualmente. As células também podem ser para apenas textos ou conter formulários para entrada de dados. O ambiente do Google Colab é utilizado com a linguagem de programação Python (versão 3.7), que é a linguagem utilizada integralmente neste projeto.

O desenvolvimento foi divido em quatro partes, a coleta de dados, o *Web Scraping*, o processo de treinamento da inteligência artificial, e por último a sumarização dos resultados.

Para coletar dados, o usuário pode selecionar em uma célula de formulário se deseja procurar os dados de uma empresa especifica ou utilizar uma empresa de uma carteira pré-selecionada. Caso, utiliza-se uma empresa da carteira, a API *Investpy* retorna o nome da empresa por extenso.

Utilizando a biblioteca do *Selenium*, o *Web Crawler* faz uma pesquisa no motor de pesquisas da Google por notícias relacionadas que tenham data de publicação do mesmo dia. Os links das notícias da primeira página são salvos no arquivo “links.txt”.

Figura 2: Exemplo do arquivo com os endereços das páginas dos sites.

Para o *Web Scraping* foi utilizada a biblioteca do *Selenium* para acessar cada link, com a as bibliotecas *urlopen* em conjunto a *BeautifulSoup* para capturar as informações da página, as informações que precisamos estão dentro das *tags* "*article*" ou "*main*" HTML, foi aplicada uma função do *BeautifulSoup* que extrai estes dados de forma íntegra.

Para resumir o texto das notícias, foi utilizada a biblioteca NLTK, especificamente os pacotes *sent_tokenize, word_tokenize* e *stopwords*, juntamente com o pacote de pontuação da biblioteca *String*. Para trabalhar com a análise por trechos menores, o texto foi dividido em sentenças, em seguida foram removidas as *stopwords* e a pontuação do texto, mantendo só as palavras de maior importância para a o contexto.

Com o pacote *FreqDistt* do NLTK, junto ao *nlargest* do *heapq*, foi realizada uma contagem de cada termo presente no texto, à estes termos foram atribuídos pesos, dado que as palavras que mais aparecem em notícias são as palavras com mais relevância, e o parágrafo com maior número de palavras relevantes deve ser o com mais importância para a análise do texto. Como exemplo, utilizamos o link de uma notícia do rompimento da represa do vale, através da referência (REDAÇÃO G1 et al., 2019), podemos validar na imagem abaixo, as palavras que o nosso código identificou como sendo as de maior relevância para essa notícia:

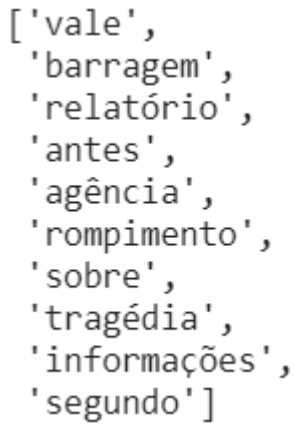

Figura 3: Retorno das palavras mais relevante ao texto.

Das palavras encontradas sete são palavras relacionadas especificamente ao assunto, dessa maneira palavras como “antes”, “sobre” e “segundo” passam a ser consideradas neutras na análise destes tópicos.

Para dar prosseguimento a análise, foi necessário armazenar o valor de significância (score) de cada palavra, foi utilizado um dicionário nas quais as chaves são as sentenças e os valores são os *scores* de significância, com isso foram selecionadas as 4 sentenças mais importantes para serem analisadas.

Para a etapa de analise foram utilizadas bibliotecas *sys, json, tflearn, random, unicodedata, numpy, tensorflow,* além das bibliotecas já importadas nas etapas anteriores. Com os textos já armazenados em arquivo, o algoritmo de análise de sentimentos tem o seguinte funcionamento:

1. É criado uma estrutura para armazenar pontuações:

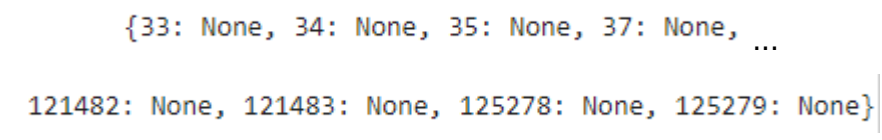

Figura 4: o dicionário contendo o caractere ascii e ao que serão traduzidos, neste caso, nada.

2. O *Stemmer* de Lancaster é inicializado e é importado um arquivo (*data.json*) que contém as informações para o treinamento da inteligência artificial:

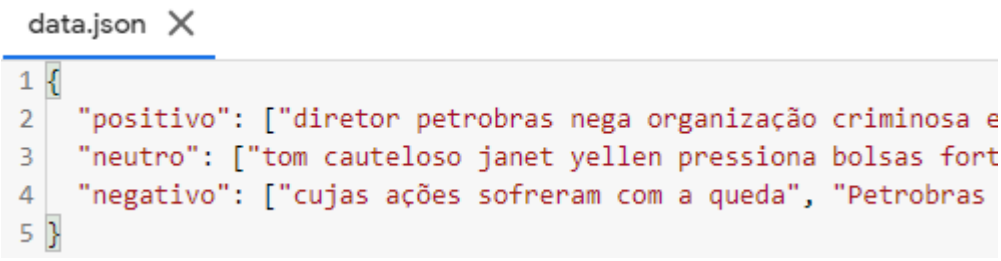

Figura 5: o conteúdo do arquivo, a estrutura contém a classificação e os textos pré-selecionados.

3. Utilizando a estrutura de pontuações, os textos para análise têm suas pontuações removidas:

```
diretor petrobras nega organização criminosa estatal notícias brasil
bovespa caminha nova máxima ano quarta alta seguida
paulo miranda reivindica iluminação telefonia zona rural
petrobras fatos dados contribuição econômica país atingiu 747 bilhões 2013
```

Figura 6: o texto do arquivo data.json, sem nenhuma pontuação.

4. Para treinar a IA, cada sentença do arquivo é reduzida em *tokens*:

```
['diretor', 'petrobras', 'nega', 'organização', 'criminosa', 'estatal',
```

Figura 7: as frases do arquivo são reduzidas a *tokens*: o seu menor significado.

5. Em seguida, os *tokens* são reduzidos em *stems* e normalizados para ficarem em minúsculo:

```
'a', 'abaf', 'abaixo', 'abandon', 'abastec', 'abastecimento',
```

Figura 8: os tokens são reduzidos para suas raízes, ou *stems*.

6. O conjunto de treinamento recebe um modelo *Bag of Words* e a linha de saída que informa a qual sentença pertence:

```
 0, 0, 0, 0, 0, 0, 0, 0], [1, 0, 0]]
```

Figura 9: a estrutura com o *Bag of Words* e a informação de onde pertence.

7. O conjunto de treinamento passa pelo processo de *shuffle*, e passa a ser do tipo *np.array:*

```
list([0, 0, 0, 0, 0, 0,
list([1, 0, 0])]
```

Figura 10: o conjunto anterior, reordenado e em nova estrutura.

8. Esse conjunto resultante é dividido em *Labels* e *Bag Of Words*. O processo de treinamento é iniciado, com o padrão de 1000 épocas:

```
Training Step: 315  | total loss: 0.94850 | time: 0.453s
| Adam | epoch: 002 | loss: 0.94850 - acc: 0.4593 -- iter: 0392/2125
```

Figura 11: os dados do treinamento: Step se refere às etapas do processo de treinamento. O processo de treinamento será repetido por completo por um determinado número de iterações (2125 vezes). O ciclo de iterações é então repetido pela quantidade de épocas (1000 vezes).

9. Ao concluir o treinamento, as sentenças extraídas com o *Web Crawler* são avaliadas:

textos avaliados.txt

```
1 negativo = Vale omitiu problemas na barragem de Bruma
2 negativo = Vale omitiu problemas na barragem de Bruma
3 negativo = Segundo a Agência Nacional de Mineração, r
4 negativo = O radar que mede as movimentações da barra
```

Figura 12: os textos avaliados, para todas as sentenças o sentimento encontrado foi negativo.

10. No final, é retornada à totalização da avaliação e a sugestão do programa:

```
Ação em baixa. Comprar
negativo:  4  positivo:  0  neutro:  0
```

Figura 13: conforme os textos avaliados, foram encontrados mais sentimentos negativos do que positivos ou neutros, o que sugere que esta ação esteja desvalorizada e portanto deve ser comprada.

3 RESULTADOS E DISCUSSÃO

Todas as quatro etapas do código retornaram seus resultados com êxito. Para validar o treinamento e utilização do analisador de sentimentos o resultado da avaliação foi comparado com o valor da ação no momento da execução do código:

O código foi executado no dia 04 de dezembro de 2020, para a empresa Petrobrás e foram coletados os seguintes dados:

```
negativo:  6  positivo:  23  neutro:  2
```

Figura 14: o resultado final da execução do código.

Figura 15: o valor da ação PETR4 no momento da execução do código.

Figura 16: o valor da ação PETR3.

Em seguida, o código foi executado com a data de outra semana, no dia 01 de dezembro de 2020, para a empresa Petrobrás e foram coletados os seguintes dados:

```
negativo:  34  positivo:  29  neutro:  4
```

Figura 17: o resultado final da execução do código.

Figura 18: o valor da ação PETR4 no momento da execução do código.

Figura 19: o valor da ação PETR3.

Analisando os dados coletados, observa-se que na primeira execução do código tínhamos a proporção de 6 notícias negativas e 23 notícias positivas sobre a empresa Petrobras, nas quais as ações PETR4 e PETR3 estavam com o perfil de aumento por 5 dias consecutivos, já na semana seguinte a analise identificou um total de 34 notícias negativas e 29 notícias positivas, tendo um aumento de 28 notícias negativas e o aumento de apenas 6 notícias positivas, em conjunto a isto, as ações PETR4 e PETR3 obtiveram uma queda nos 5 dias anteriores à execução.

4 CONCLUSÃO

Concluímos que o objetivo deste projeto de desenvolver uma solução de avaliação de ações utilizando análise de linguagem natural obteve sucesso e que por meio dele avaliamos que a utilização de NLTK é viável para análise de sentimentos na língua portuguesa, e que sua utilização na análise de notícias retorna um valor correspondente com a queda/subida da ação.

Das dificuldades encontradas no desenvolvimento e execução do programa, destaca-se a coleta de dados, os sites encontrados pelo *Web Crawler* abordavam a configuração de HTML da página de formas distintas, ocorrendo casos em que as informações necessárias estavam nas *tags* “article” ou “main”, tornando a captura dos

dados mais complexa. Como agravante, muitos sites só possuíam a informação principal na *tag* “article” mas blogs comumente tinham propagandas ou outros assuntos nesta tag. Resolvemos este problema definindo os sites que utilizaríamos a captura do link, mantendo um formato mais uniforme e prevenindo problemas futuros.

Um segundo problema que pode ser destacado seria a complexibilidade da análise de sentimentos no quesito de preparação de uma base de dados solida para o treinamento da IA. Fatores como tempo da gramática (por mudanças de escrita ao longo dos anos) ou regionalidades dificultam a compreensão da escrita para a inteligência artificial pois ela tem que ser preparada especificamente para esse tipo de escrita. Este tipo de problema só pode ser resolvido com a reavaliação e aprimoramento constantes do banco de dados.

4 REFERÊNCIAS

RUSSEL, S.; NORVIG, P. **Inteligência artificial**. 2ª edição. Rio de Janeiro: Campus, 2004.

MITTAL, A.; GOEL, A. **Stock prediction using twitter sentiment analysis**. Stanford University, 2011.

GALESHCHUK, S., VASYLCHYSHYN, O.; KRYSOVATYY, A. **Bitcoin Response to Twitter Sentiments.** ICTERI Workshops, 2018.

FORTUNA Eduardo. **Mercado financeiro: produtos e serviços**. 22ª ed. QualityMark. Rio de Janeiro: Qualitymark Ed., 2020.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit**. Natural Language Toolkit, 2019

DEDIC, N.; STANIER, C. **Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery. Innovations in Enterprise Information Systems Management and Engineering**. Heidelberg: Springer International Publishing, 2017.

DATA SCIENCE ACADEMY et al. **Web scraping e web crawling são legais ou ilegais?**. [S. l.], 1 jul. 2018. Disponível em: http://datascienceacademy.com.br/blog/web-scraping-e-web-crawling-sao-legais-ou-ilegais/. Acesso em: 21 jun. 2020.

Vinicius et al. I**ntrodução aos bancos de dados NoSQL**. [S. l.], 2012. Disponível em: https://www.devmedia.com.br/introducao-aos-bancos-de-dados-nosql/26044. Acesso em: 21 jun. 2020.

MEDEIROS, **Higor et al. Introdução ao MongoDB**. [S. l.], 2014. Disponível em: https://www.devmedia.com.br/introducao-ao-mongodb/30792. Acesso em: 21 jun. 2020.

REDAÇÃO GLOBAL AD. **O que é Crawler?**. [S. l.], 2019. Disponível em: https://globalad.com.br/blog/o-que-e-crawler/. Acesso em: 28 jun. 2020.

LIMA, Juliano P.. **Mercado de Capitais**; Grupo GEN, 07/2019. 9788597021752. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788597021752/. Acesso em: 14 Jun 2020

YOSHIKAWA, Kleber.K.C; MALAIA, Maria.C.B.T; MATTEI, Cesar. **Mercado de Capitais: técnicas para avaliação de carteira de ações para pessoa física**. 2019. Disponivel em: https://www.aedb.br/seget/arquivos/artigos09/401_401%20Mercado%20de%20capitais%20-%20tecnicas%20para%20avaliacao%20de%20carteira%20de%20acoes%20para%20pessoa%20fisica.pdf. Acesso em 15 jun 2020.

FORTUNA Eduardo. **Mercado financeiro: produtos e serviços**. 22ª ed. QualityMark. Rio de Janeiro: Qualitymark Ed., 2020.

BASSOTO, Lucas. **Ações da Taurus valorizam mais de 400%: euforia ou realidade?**. Cointimes, nov 2019. Disponível em: https://cointimes.com.br /acoes-da-taurus-euforia-ou-realidade/. Acesso em: 01 jun 2020

MORAIS, Izabelly S. et al. **Introdução a Big Data e Internet das Coisas (IoT)**; Grupo A, 2018. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788595027640/. Acesso em: 15 Jun 2020

BERMAN, Jules J. **Principles of Big Data**. 1ª Ed. Massachusetts, EUA. Elsevier, 2013.

RATNER, Bruce. **Statistical and Machine-Learning Data Mining**. 2ª Ed. Florida, EUA. CRC Press, 2011

TABOADA, M; ANTHONY, C e VOLL, K. **Methods for creating semantic orientation dictionaries**. Conference on Language Resources and Evaluation (LREC), pg. 427–432, 2006.

TABOADA et al. **Lexicon-based methods for sentiment analysis**. Comput. Linguist., 37, pg. 267– 307, 2011.

LIU, B.. **Sentiment analysis and subjectivity.** 2010.

BOLLEN, J., MAO, H., ZENG, X. **Twitter mood predicts the stock market**. CoRR, abs/1010.3003, 2010.

NIELSEN, F. Å. **A new anew: Evaluation of a word list for sentiment analysis in microblogs**. 2011.

PANG et al. **Thumbs up? sentiment classification using machine learning techniques**. In Proceedings of EMNLP, pg. 79–86, 2002.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of Machine Learning**. 2. ed. Mit Press, 2018.

BURKOV, A. **The Hundred-Page Machine Learning Book**. 1. ed. [S.l.]: Kindle Direct Publishing, 2019.

KREUTZ, D. et al. **Software-defined networking: A comprehensive survey**. Proceedings of the IEEE, v. 103, n. 1, p. 14–76, Jan 2015.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016

LECUN, Y.; BENGIO, Y.; HINTON, G. **Deep learning**. Nature, Nature Publishing Group. v. 521, p. 436, Maio 2015.

GAO, X.; ZHANG, J.; WEI, Z. **Deep learning for sequence pattern recognition**. In: 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). pg. 1–6, 2018

DE CHOUDHURY et al. **Characterizing and predicting postpartum depression from shared facebook data**. In Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '14, pages 626–638, New York, NY, USA. ACM, 2014.

FELDMAN, R. **Techniques and applications for sentiment analysis**. Communications of the ACM, 56(4):82–89, 2013.

HUTTO, C; GILBERT, E. **Vader: A parsimonious rule-based model for sentiment analysis of social media text**. 2014.

BENEVENUTO, Fabrício *et al*. **Métodos para Análise de Sentimentos em mídias sociais**. [*S. l.*], 19 jul. 2020. Disponível em: https://homepages.dcc.ufmg.br/~fabricio/download/. Acesso em: 30 jul. 2020.

REDAÇÃO G1 *et al*. **Vale omitiu problemas na barragem de Brumadinho antes do rompimento, diz relatório da ANM**. [*S. l.*], 5 nov. 2019. Disponível em: https://g1.globo.com/jornal-nacional/noticia/2019/11/05/vale-omitiu-problemas-na-barragem-de-brumadinho-antes-do-rompimento-diz-relatorio-da-anm.ghtml. Acesso em: 30 jul. 2020.

OLIVEIRA, Lucas. **Um guia abrangente sobre NLP - Processamento de texto**. [*S. l.*], 20 jan. 2020. Disponível em: https://medium.com/@lucasoliveiras/um-guia-abrangente-sobre-nlp-processamento-de-texto-60b852125202. Acesso em: 30 jul. 2020.

ECMA International. **The JSON Data Interchange Syntax**. Standard ECMA-404. 2ª Ed. Dezembro de 2017. Disponível em: http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf. Acesso em: 30 jul. 2020.