UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

JOÃO FERNANDO ALMEIDA CAEMERER

Previsões de Mercado Através de Notícias

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Engenharia da Computação

Orientador: Prof. Dr. Claudio F. R. Geyer

Porto Alegre
2023

“The decision to believe
is the most important choice we ever make.”
— L. Whitney Clayton

AGRADECIMENTOS

Agradeço aos meus pais pelo apoio à minha jornada acadêmica. Ao meu filho Augusto, que é combustível para meus dias. À Camila, minha esposa paciente e inspiradora. Ao Profesor Doutor Cláudio Geyer pela orientação. Sou também grato pelos vários professores que me inspiraram a ter grande apreço pela computação.

RESUMO

O mercado de ações influencia economias em todo o globo. As variações dele podem impactar direta ou indiretamente áreas de negócios. A previsibilidade nesse ramo pode ser a diferença do lucro e prejuízo, seja para um pessoa física ou uma nação. Entretanto, há inúmeras variáveis e relações entre elas que dificultam em muito a construção de um modelo preditivo. Este trabalho apresenta um protótipo capaz de receber como entrada notícias divulgadas na Internet e dados históricos de ativos do mercado de ações para assim, treinar modelos preditivos e retornar predições sobre a variação do ativo de interesse. Os algoritmos utilizados para os modelos preditivos foram: *Random Forest*, *K-Neighbors* e *SVM*. As métricas utilizadas para verificar qual destes teve o melhor desempenho foram: MAPE (*Mean Absolute Percentage Error*) e MSE (*Mean Square Error*). O primeiro, mais voltado para o quanto, proporcionalmente, os valores de predição diferem do esperado. Já o segundo dá mais relevância aos pontos de divergência ou chamados pontos fora da curva do que o primeiro. Nos resultados obtidos é visto que o algoritmo *Random Forest* é em média 30% mais preciso que os demais. Entretanto, é necessário mencionar que há certa variação nas métricas dependendo da ação alvo de predição, Nos ativos que tiveram mais variações bruscas nos históricos, foram encontrados maiores erros de predição. Essa dificuldade na predição ocorre mesmo que um ativo tenha mais notícias ao seu respeito alimentando o modelo preditivo. Ou seja, vemos que as variáveis externas às notícias como, por exemplo, ações relacionadas podem impactar as ações. Um provável exemplo disso é visto nos dados da empresa Tesla, onde temos muitas notícias do respetivo acionista majoritário e de empresas relacionadas a esta pessoa.

Palavras-chave: Mercado de ações. Aprendizado de máquina. Notícias. Previsão.

Stock Market Prediction Through News

ABSTRACT

The stock market influences economies across the globe. Its variations can directly or indirectly impact business areas. Predictability in this segment can be the difference between profit and loss, whether for an individual or a nation. However, there are numerous variables and relationships between them that make building a predictive model very difficult. This work aims to build a predictive model based on news published on the Internet. This work presents a prototype capable of taking as input news published on the Internet and historical data of stock market assets to train predictive models and return predictions on the variation of the asset of interest. The algorithms used for the predictive models were Random Forest, K-Neighbors, and SVM. The metrics used to determine which of these performed best were MAPE (*Mean Absolute Percentage Error*) and MSE (*Mean Square Error*). The former is more focused on how much the prediction values differ proportionally from the expected ones. The latter gives more weight to divergence points or so-called outliers than the former. The results obtained show that the Random Forest algorithm is on average 30% more accurate than the others. However, it is necessary to mention that there is some variation in the metrics depending on the target prediction asset. For assets that had more drastic fluctuations in their histories, larger prediction errors were found. This prediction difficulty occurs even when an asset has more news about it feeding the predictive model. In other words, we see that variables external to the news, such as related stocks, can impact stocks. A probable example of this is seen in the data of the Tesla company, where we have much news about the respective majority shareholder and companies related to this person.

Keywords: Stock, Market, Prediction, News, Machine Learning.

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface* ou Interface de Programação de Aplicação
HTML	*Hypertext Markup Language* ou Linguagem de Marcação de HiperTexto
IA	Inteligência Artificial
JSON	*JavaScript Object Notation* ou Notação de Objetos JavaScript
MAPE	*Mean Absolute Percentage Error* ou Erro Médio Percentual Absoluto
ML	Machine Learning
MM	Milhões
MSE	*Mean Square Error* ou Erro Quadrático Médio
PLN	*Natural Language Processing* ou Processamento de Linguagem Natural
SVM	*Support Vector Machine* ou Máquina de Vetor de Suporte
URL	*Uniform Resource Locator* ou Localizador Uniforme de Recursos
WWW	*World Wide Web* ou Rede Mundial de Computadores

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Este capítulo contém quatro seções. A primeira delineia a motivação por trás deste estudo. Na segunda, é descrita a proposta deste. Os objetivos do trabalho são apresentados na terceira seção. Por fim, a última seção descreve a estrutura do presente documento.

1.1 Motivação

Nas últimas décadas os interessados pelo mercado de ações experienciaram um crescimento exponencial (BADOLIA, 2016). Também vemos um aumento recente no número de interessados por investimentos como ações no Brasil, como é ilustrado na figura 1.1. Atualmente, bilhões de dólares são negociados todos os dias por meio de ações (HOSEINZADE; HARATIZADEH, 2019). Caso um participante do mercado pudesse prever totalmente o comportamento do mercado, isso o permitiria a investir em produtos arriscados mas com grandes retornos. Essa possibilidade motiva o uso de *Machine Learning(ML)* ou aprendizado de máquina para criar modelos preditivos que possam antever as alterações da bolsa de valores (KUMBURE et al., 2022). De fato, existe um grande número de estudos publicados que tentam prever o mercado de ações com acurácia, buscando atingir isto ao desenvolver sofisticados sistemas de previsão ((SONG; LEE; LEE, 2019), (SEDIGHI et al., 2019)). E também há casos em que estudos reportaram lucros com seus modelos ((ATSALAKIS; VALAVANIS, 2009), (ARMANO; MARCHESI; MURRU, 2005), (WENG; AHMED; MEGAHED, 2017)). Em geral, a predição do mercado de ações é um dos desafios mais difíceis e relevantes (CHEN; HAO, 2017).

1.2 Proposta

Um exemplo de ferramenta atual que relaciona mudanças em ações com notícias é o *Google Finance*, para certas ações há possibilidade de ver, junto ao gráfico de variação de valor de uma ação, notícias relacionadas ao que a afetou (FINANCE, 2023a). Com a facilidade de uso e popularidade da inteligência artificial (IA) crescente, a utilização dessas técnicas em estudo de predição de mercado se tornaram cada vez mais populares. Em contrapartida aos métodos tradicionais, essas técnicas podem lidar com não-linearidade,

Figura 1.1: Relatório Trimestral de Investimentos no Brasil

Fonte: (B3, 2022)

ruído, caos e complexidade da bolsa de valores. Assim, o uso de IA pode levar a predições mais eficientes (CHEN; HAO, 2017).

Na literatura, há muitas variantes de técnicas de IA para o desenvolvimento de aplicações de previsão do mercado de ações. Entre elas, redes neurais ((RAJIHY; NERMEND; ALSAKAA, 2017), (O'CONNOR; MADDEN, 2006)), máquinas de vetores de suporte ((HUANG; NAKAMORI; WANG, 2005)) e suas variantes ((EBRAHIMPOUR et al., 2011), (ENKE; THAWORNWONG, 2005)) são as que mais frequentemente são aplicadas devido aos resultados promissores mostrados. Além disso, os insumos para a predição podem ser dados históricos como também notícias (KUMBURE et al., 2022).

Apesar dos avanços no uso de IA, há desafios na previsão de mercado, um deles é a grande quantidade de variáveis que são intrínsecas ao ambiente mundial, tais como: índices de preço, que são indicadores para medir o preço de um conjunto de ativos como o índice Ibovespa no Brasil, taxas de juros, que representam o custo do dinheiro, ou seja, a porcentagem de retorno ao emprestar recursos a outrem como um país, volatilidade, representando o quanto o preço de um ativo flutua ao longo do tempo o que pode ser usado como um indicador de risco e liquidez que é o grau de facilidade para que um ativo possa ser comprado ou vendido no mercado (ENKE; THAWORNWONG, 2005).

A proposta do trabalho, ao verificar as ferramentas, técnicas e contexto apresentados acima, é criar um protótipo que possa receber dados de notícias e histórico de valores da bolsa de valores e, ao aplicar aprendizagem de máquina, prever qual deve ser a variação do ativo.

1.3 Objetivos

O projeto visa criar um protótipo baseado em dados e aprendizado de máquina para indicar se um ativo na bolsa de valores vai se valorizar ou perder valor. Os resultados do trabalho incluem:

- Base de dados de notícias contendo título, conteúdo, data de publicação, ativos relacionados e órgão emissor. Esta sendo estruturada e capaz de alimentar modelos preditivos.
- Base de dados históricos de ativos da bolsa de valores. Esta também capaz de alimentar modelos preditivos e ser possível unificar com a base descrita no item anterior.
- Modelos de predição baseados em notícias e dados históricos do mercado de ações e suas respectivas métricas para ser averiguado qual destes teve melhor desempenho nas predições.

1.4 Estrutura

Este trabalho está organizado da seguinte forma. O capítulo 2 contém o fundamental teórico que foi base para a realização do trabalho. Já o capítulo 3 comenta sobre trabalhos existentes que abordam e propõem arquiteturas para o tema e que inspiraram partes do trabalho. No próximo capítulo, será descrita a proposta de solução para atingir o objetivo de predição de variação de mercado através de notícias. Por fim, os últimos capítulos transcrevem sobre a metodologia usada, experimentos realizados e então, a conclusão.

2 FUNDAMENTAÇÃO TEÓRICA

No transcorrer desta seção serão descritos fundamentos teóricos que foram a base para o desenvolvimento do trabalho. Os itens abordados são relacionados aos desafios e motivação relativos ao campo de pesquisa escolhido.

2.1 O Mercado

Resumidamente, o mercado de ações pode ser descrito como um local comum de compra e venda onde instrumentos financeiros são negociados. Estes produtos transacionados podem ser ações, títulos, commodities e outros. Pode-se fazer um paralelo com as praças públicas onde interessados em comprar e vender fazem suas trocas.

Segundo estudo (B3, 2022), houve um crescimento de 35% no número de brasileiros que investem em ativos em renda variável no período de um ano. Além desta alta nos investimentos mais voláteis, como vemos na tabela 2.1, o estudo também verificou um aumento de 25% no uso dos produtos do *Tesouro Direto*, este sendo um sistema do Tesouro Nacional e através dele, investidores pessoas físicas podem comprar títulos públicos, via internet. Estes dados corroboram a importância e relevância crescente dos investimentos no Brasil. Logo, pode-se notar o interesse crescente pelo mercado de valores no país.

Tabela 2.1: Variação do número de Investidores(CPFs) em milhões(MM)

Produto / Trimestre	**3º TRI 2021**	**3º TRI 2022**	**Variação**
Renda Variável	3,97 MM	5,35 MM	35%
Tesouro Direto	1,7 MM	2,1 MM	25%
Renda Fixa	9,6 MM	12,6 MM	31%

Fonte: O autor

A movimentação no mercado brasileiro teve alta nas últimas décadas, segundo a (ANBIMA, 2022) o valor em transações aumentou quatro vezes mais, isso já descontando a inflação. Em 1997, o valor foi de R$ 19,3 bilhões e 25 anos depois, R$ 365,2 bilhões foram movimentados no mercado nacional. Esses valores reforçam a tendência de aumento de investimentos no mercado ao longo do tempo.

Apesar de ser um ambiente com objetivo de crescimento financeiro, o mercado de ações tem certa volatilidade que pode ser desconfortável para boa parte das pessoas. Isto pode ser visto na diferença de público entre os usuários de *Renda Variável* e *Renda Fixa*. Certo grau do comportamento das ações se deve à percepção de quem participa do

mercado (SCHWERT, 1990).

Outro fator influenciador do mercado é a impressão dos usuários. De acordo com (KAHNEMAN; TVERSKY, 1977), há relação entre os resultados e as expectativas nos investimentos variáveis. Assim, é possível afirmar que notícias podem ter o poder de impactar as variações da bolsa. Esses indicativos sustentam, também, a relevância do trabalho proposto.

2.2 Processamento de Linguagem Natural

A linguagem dita natural é usada para a comunicação entre os seres humanos e nela há muitas variáveis, como: contexto, sarcasmo e ambiguidade. A área de processamento de linguagem natural (PLN) consiste em pesquisa e aplicação do entendimento e manipulação de língua falada ou escrita para algum propósito útil. Entretanto, a construção de programas de computador que possam entender a linguagem natural envolve três problemas: processo de pensamento, representação e significado da linguagem de entrada e o conhecimento de mundo (CHOWDHARY, 2020).

A descoberta de regras implícitas em grandes volumes de dados é o principal objetivo das técnicas de *Machine Learning* ou aprendizado de máquina (SCHÜNKE, 2015). E dados oriundos de processamento de linguagem natural podem ser usados para treinar modelos que usam essas técnicas, como visto em outros trabalhos: (SEHGAL; SONG, 2007) e (KIM; JEONG; GHANI, 2014).

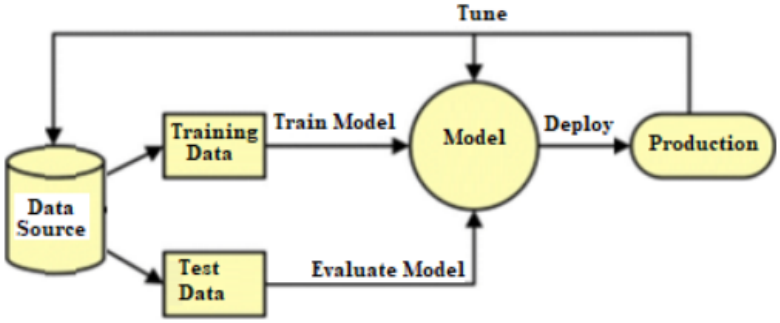
A análise de sentimentos, ou mineração de opinião é uma área ativa dentro do campo de PNL que analisa a opinião das pessoas, sentimentos, avaliações, atitudes e emoções pelo tratamento computacional em texto. Uma das abordagens da análise de sentimentos ou de opinião é o sentimento léxico, este é uma lista de palavras e expressões que uma língua contém e são classificados como positivo ou negativo. A construção de ferramentas de sentimento léxico levam muito tempo quando criadas e validadas manualmente, entretanto, está entre os métodos mais robustos para gerar uma base confiável. Uma ferramenta com bons resultados é VADER ou *Valence Aware Dictionary for sEntiment Reasoner*, ela utiliza a abordagem citada e contém uma acurácia em suas classificações de 96% em certos casos, ou seja, melhor que humanos em média (HUTTO; GILBERT, 2015).

2.3 Aprendizado de Máquina

Os humanos têm usado muitos tipos de ferramentas para realizar várias tarefas de uma maneira mais simples. A criatividade impulsionou invenções de diferentes máquinas que facilitaram a vida humana. Transporte, indústrias e computação são algumas das necessidades que estas criações auxiliam. E aprendizado de máquina está entre essas invenções.

Machine Learning (ML) é definido como um campo de estudo que dá a computadores a habilidade de aprender sem serem explicitamente programados. Com a abundância de dados a demanda por ML cresce, assim vemos muitas indústrias fazendo uso dele para extrair dados relevantes. O propósito do aprendizado de máquina é aprender através de dados. Um dos tipos de aprendizado de máquina é o supervisionado, ele acontece ao buscar uma função que mapeia entradas a saídas com base em exemplos de pares de entrada-saída. Os algoritmos de aprendizado de máquina supervisionados precisam de assistência externa. Os dados de entrada são divididos entre conjunto de treino e de teste. O conjunto de treino contém variáveis de saída que precisam ser previstas ou classificadas. Os algoritmos aprendem padrões do conjunto de treinamento e os aplicam no conjunto de teste. Ao tentar prever ou classificar quem supervisiona pode fazer ajustes nos conjuntos de dados ou no modelo, isso com intuito de melhorar o modelo construído. Vemos na imagem 2.1 a realimentação dos ajustes e testes dos algoritmos de aprendizado supervisionado (MAHESH, 2020).

Figura 2.1: Fluxo de Trabalho de Aprendizado Supervisionado



Fonte: (MAHESH, 2020)

Alguns exemplos de algoritmos de aprendizado supervisionado, que também são apresentados pelo autor citado acima, são: *árvore de decisão*, que é representado por

um grafo, onde seus nós representam questões, as arestas respostas e as folhas a decisão. Já o *Navie Bayes*, uma técnica de classificação baseada no teorema de Bayes com a suposição de independência entre as entradas. Na equação 2.1 vemos a expressão que compõe, matematicamente, o algoritmo. Por fim, também é apresentado pelo autor a técnica de *máquinas de vetores de suporte* ou SVM. Esta abordagem consegue performar eficientemente a classificação não linear usando de um mapeamento das entradas a espaços multidimensionais. Assim, é possível encontrar margens entre grupos de dados de múltiplas entradas, podem elas serem dependentes entre si ou não.

$$P(C_k|x_1, x_2, \ldots, x_n) = \frac{P(C_k) \cdot P(x_1|C_k) \cdot P(x_2|C_k) \cdot \ldots \cdot P(x_n|C_k)}{P(x_1) \cdot P(x_2) \cdot \ldots \cdot P(x_n)} \tag{2.1}$$

Onde:

$P(C_k)$ é a probabilidade anterior da classe C_k

$P(x_i|C_k)$ é a probabilidade da característica x_i dado a classe C_k

$P(x_i)$ é a probabilidade marginal da característica x_i

2.4 Web Crawler

Um rastreador Web ou *Web Crawler* é um programa de computador que tem por objetivo analisar páginas na internet de maneira sistemática e automática. As páginas na Web muitas vezes contêm links (ligações) para outras, o que permite que o robô, dito crawler, vá de página em página buscando os dados de interesse (KAUSAR; DHAKA; SINGH, 2013).

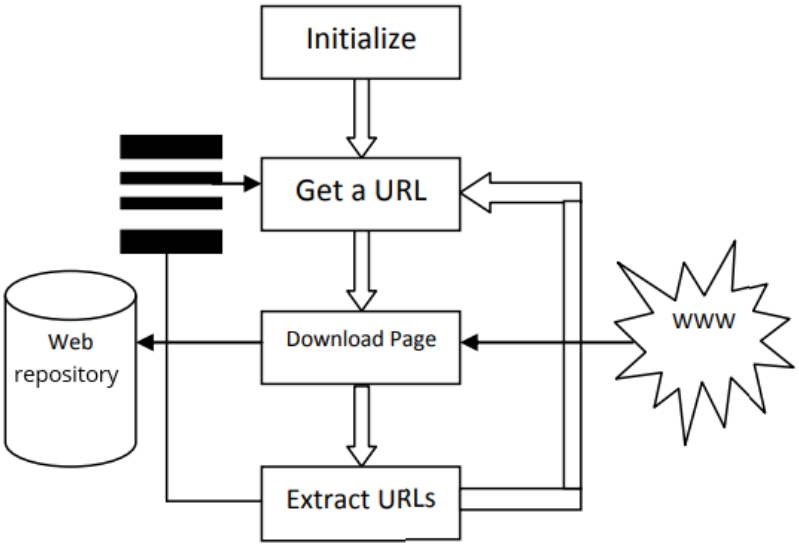
O funcionamento de um web crawler pode ser descrito nos seguintes passos:

1. Selecionar o endereço (URL) de um ou mais sites iniciais;
2. Adicionar endereços em um conjunto fronteira;
3. Escolher uma URL da fronteira e a retirar desse conjunto;
4. Buscar a página web da URL correspondente;
5. Analisar a página em busca dos dados de interesse e os salva no repositório, como outros links para URLs;

6. Adicionar novos links na fronteira;
7. Voltar ao passo 3 até a fronteira estar vazia ou até outra limitação imposta ser atingida como tempo de processamento.

Na figura 2.2 vemos o fluxo descrito acima com o a representação do repositório onde os dados extraídos serão guardados. Este repositório pode manter os dados capturados da web, e permitirá usá-lo para guardar os dados de entrada para o módulo de processamento de linguagem natural já mencionado na seção 2.2.

Figura 2.2: Processo de fluxo de um crawler.



Fonte: (KAUSAR; DHAKA; SINGH, 2013)

3 TRABALHOS RELACIONADOS

Há vários trabalhos na área da mineração de dados e previsão de mercado como: (KIM; JEONG; GHANI, 2014), (SEHGAL; SONG, 2007) e (SCHUMAKER; CHEN, 2009). Estes coletam dados e os tratam para alimentar o treino dos modelos preditivos. Na seção 3.1 é apresentado um modelo geral que engloba grande parte das ideias descritas nos trabalhos referenciados.

3.1 Mineração de Opinião em Texto

No trabalho de coleta, tratamento de sentimentos e predição de bolsa de valores descrito por (KIM; JEONG; GHANI, 2014), é visto um fluxo de tratamento de dados entre a mineração até a predição de mercado.

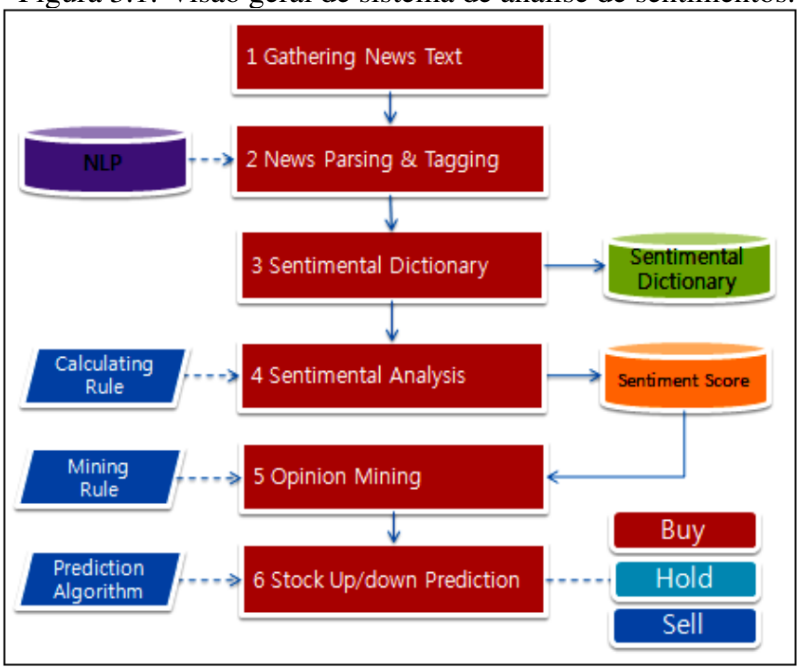
O primeiro passo, descrito na imagem 3.1, é o de extração dos dados, o que foi realizado por uma tecnologia de *scrapping*, ou seja, de raspagem de dados, esta sendo apresentada na seção 2.4 através de uma ferramenta de Web Crawler. Já no segundo passo, temos a aplicação de uma ferramenta de processamento de linguagem natural (seção 2.2), que efetua a análise e marcação do texto, preparando os dados para a etapa seguinte. A análise e o dicionário de sentimentos, vistos no ponto três e quatro, internamente ignoram os termos desnecessários, extraem os de maior importância e retiram termos duplicados, após isso são pontuados os valores para cada termo usando os dados de variação de mercado. Assim, o dicionário de sentimentos vai se formando e a análise de sentimentos também. Os últimos passos são baseados em treinamento de modelos preditivos com base nos dados de entrada e saídas esperadas, portanto, buscando predizer as próximas alterações na bolsa.

No trabalho de (SEHGAL; SONG, 2007), vemos muitos pontos parecidos com o que foi citado acima. Mas também vemos, por exemplo, outras escolhas na solução proposta como mais capilaridade na classificação de ações. Ao invés de apenas: *Comprar*, *Manter* e *Vender*; Foram adicionados: *Comprar Enfaticamente* e *Vender Enfaticamente*. Logo, essas adições permitem uma maior margem de flexibilidade para o modelo, espelhando a potência variada de frases na linguagem natural.

Já (SCHUMAKER; CHEN, 2009), difere dos dois anteriores na análise de textos. No trabalho desenvolvido por ele vemos a abordagem da *Bolsa de Palavras*, sendo ela uma maneira de determinar palavras chaves de um texto. Deste modo, termos que são

semanticamente vazios são removidos do texto analisado. Logo, as notícias treinam o modelo preditivo pelo grupo de palavras que são capturadas delas. Então, não há análise de sentimentos, mas de quais palavras nas notícias ativam as variações. Há previamente palavras registradas e a sua presença no texto avaliado liga a presença e a combinação desses pesos positivos e negativos levam a previsão da variação da ação.

Figura 3.1: Visão geral de sistema de análise de sentimentos.



Fonte: (KIM; JEONG; GHANI, 2014)

3.2 Aprendizado de Máquina

No trabalho de (SCHUMAKER; CHEN, 2009) são apresentados três algoritmos que podem ser aplicados para o aprendizado de máquina no assunto de interesse. Entretanto, o uso dos dessas ferramentas teve como objetivo a classificação e não a previsão de valores da bolsa de ações. O primeiro deles é o *Algoritmo Genético*, que consiste na competição entre uma 'população' de soluções candidatas para conseguir prever uma saída com base em entradas. Entretanto, essa população é avaliada e parte dela é selecionada, combinada e possivelmente aplicada mutação para formar uma nova população. O ciclo repete até que algum critério escolhido seja atingido. O segundo algoritmo citado foi *Naïve Bayesian*, este sendo ingênuo (naïve) porque supõe que as variáveis de entrada são independentes, ou seja, não influenciam uma a outra para chegar ao resultado. Este algo-

ritmo utiliza o Teorema de Bayes para calcular a probabilidade de uma entrada pertencer a uma classe. O último apresentado é o *SVM* ou Máquinas de Vetores de Suporte, este termo se refere a um conjunto de algoritmos de ML que são particularmente eficazes em problemas com muitas variáveis. A busca por um hiperplano, ou seja, função de *n* variáveis, que minimize a distância do hiperplano dos pontos de treino. É citado, no trabalho de (SEHGAL; SONG, 2007), o algoritmo de Árvores de Decisão além do Naive Bayes. Esse, capaz de lidar com relacionamentos complexos entre as variáveis de entrada, mas tende a ter *over-fitting*, ou seja, o modelo se ajusta muito aos dados de treinamento e tem bons resultados, mas não se sai tão bem quando aplicado em dados de teste. Tanto este autor quanto o anterior buscam classificação nos seus experimentos, como "Comprar", "Vender", "Manter".

3.3 Análise de Sentimentos

A natureza inerente do conteúdo das mídias sociais coloca sérios desafios para aplicações práticas de análise de sentimentos. De acordo com (HUTTO; GILBERT, 2015), a ferramenta VADER, um modelo baseado em regras para análise de sentimentos, tem bom desempenho na representação de sentimentos relacionados com textos de mídias sociais, avaliações de filmes, editoriais do *New York Times* e avaliações de produtos na *Amazon.com*. Os resultados das avaliações de acurácia da ferramenta podem ser vistos na imagem ... que utiliza a métrica F1. Esta métrica é dependente de duas outras métricas: *Precisão* que consiste na razão entre classificações verdadeiramente positivas pela soma das verdadeiramente positivas somadas às falsas positivas como apresentado na equação 3.1. Temos também a Revocação, esta é a razão entre classificações verdadeiramente positivas pela soma das verdadeiramente positivas somadas às falsas negativas como apresentado na equação 3.2. Assim podemos descrever a equação da métrica F1 como mostrado na equação 3.3.

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Positivos}} \tag{3.1}$$

$$\text{Revocação} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}} \tag{3.2}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{3.3}$$

Figura 3.2: Resultados da Métrica F1 na classificação de sentimento de ferramentas de análise de sentimentos.

	3-Class Classification Accuracy (F1 scores) Test Sets			
	Tweets	Movie	Amazon	NYT
VADER	**0.96**	0.61	**0.63**	**0.55**
NB (tweets)	0.84	0.53	0.53	0.42
ME (tweets)	0.83	0.56	0.58	0.45
SVM-C (tweets)	0.83	0.56	0.55	0.46
SVM-R (tweets)	0.65	0.49	0.51	0.46
NB (movie)	0.56	**0.75**	0.49	0.44
ME (movie)	0.56	**0.75**	0.51	0.45
NB (amazon)	0.69	0.55	0.61	0.48
ME (amazon)	0.67	0.55	0.60	0.43
SVM-C (amazon)	0.64	0.55	0.58	0.42
SVM-R (amazon)	0.54	0.49	0.48	0.44
NB (nyt)	0.59	0.56	0.51	0.49
ME (nyt)	0.58	0.55	0.51	0.50

Fonte: (HUTTO; GILBERT, 2015)

3.4 Metodologia de Avaliação

Uma medida popular e que é capaz de descrever a diferença de magnitude entre o predito e o observado é o *MAPE* ou *Erro Médio Percentual Absoluto*. Esta métrica descrita por (MOHAN et al., 2019) é feita verificando a média da porcentagem de erro de cada predição. Ela é descrita na equação 3.4 e pode indicar o quão, proporcionalmente, a predição está errando. Esta medida é muito interessante para o assunto do trabalho pois ações tem valores variados, logo o valor absoluto não é tão relevante quanto a variação proporcional dos ativos.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \times 100 \tag{3.4}$$

Onde:

n é o número total de observações ou pontos de dados.

Y_i representa o valor real (observado) para o ponto de dados i.

$\hat{Y}_i$ representa a previsão ou estimativa do modelo para o ponto de dados i.

Outra métrica interessante apresentada por (DZIKEVIČIUS; ŠARANDA, 2010), é o *Erro Quadrático Médio* ou MSE, este é definido como o calculo da média dos erros entre o valor previsto e o esperado elevado ao quadrado. A expressão matemática correspondente está descrita na equação 3.5. Ou seja, caso tenhamos pontos atípicos de predição eles farão o valor do erro se sobressair. Este tipo de calculo pode ser importante pois caso um erro grande, proporcionalmente ao valor desejado, ocorra há possibilidade de considerável perda de recursos. Logo, a métrica MSE é relevante para verificar a qualidade de previsão do protótipo construído.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2 \tag{3.5}$$

Onde:

n é o número total de observações ou pontos de dados.

Y_i representa o valor real (observado) para o ponto de dados i.

$\hat{Y_i}$ representa a previsão ou estimativa do modelo para o ponto de dados i.

4 PROJETO PROPOSTO

Este capítulo apresenta o trabalho proposto, como seu escopo e anseios. Serão detalhados: a construção de um fluxo de extração de notícias e variação da bolsa de valores, tratamento de dados e por fim, a previsão de alterações no mercado com a classificação das ações.

O módulo da ferramenta que será responsável pela parte de extração de dados será um *Web Crawler* que percorrerá todas as páginas de um *website* (descrito na seção 4.2), filtrar as informações de interesse e retornar um conjunto de dados relacionados à busca. Um ponto de relevância é o cuidado para deduplicação de dados ao passar para o próximo nível, evitando que uma mesma notícia tenha, possivelmente, o peso positivo ou negativo dobrado.

A segunda parte do trabalho proposto é o processamento de linguagem natural dos dados colhidos anteriormente. Será realizada utilizando uma biblioteca de código (ALLENNLP, 2023) e terá como objetivo a classificação de sentimento das notícias e identificação de ações relacionadas. Será então, retornado um cenário positivo ou não para o mercado. Então, com os dados tratados e prontos, os modelos preditivos (seção 4.2), último módulo, é acionados e executado algoritmos preditivos a fim de dizer quais serão as variações de mercado.

4.1 Arquitetura Escolhida

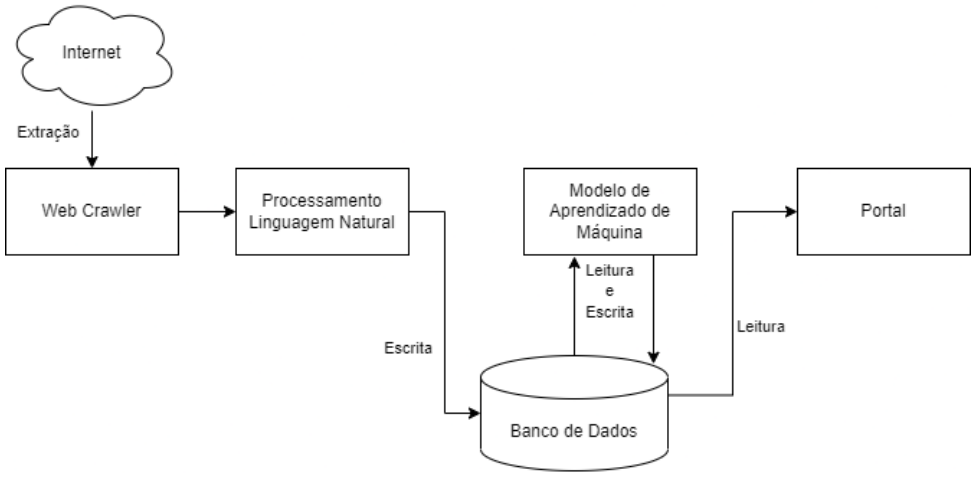
A arquitetura escolhida é baseada em trabalhos usados como referência ((KIM; JEONG; GHANI, 2014; SCHUMAKER; CHEN, 2009; SEHGAL; SONG, 2007)) que demonstraram resultados favoráveis e isolamento de funcionalidade, permitindo uso de soluções específicas em módulos. As soluções implementadas para atingir as funcionalidades dos módulos descritos acima são ilustradas na figura 4.1 e citadas abaixo:

1. **Web Crawler** - Para a raspagem de dados na *web*, o *framework Scrapy*, também contruído na linguagem de programação Python, foi implantado. Este ferramental sendo *opensource* foi a base para o módulo de extração de notícias da *web* do projeto (SCRAPY, 2023);
2. **Processamento de Linguagem Natural** - Recebendo os dados extraídos da *Internet*, eles serão consumidos por um módulo de Processamento de Linguagem Natu-

ral (PLN). Este foi desenvolvido usando a biblioteca *AllenNLP* que recebe textos e analisa o grau de positividade ou negatividade de um texto e também palavras chave. Assim, será possível dar pontuações para as notícias e representar os diferentes níveis de impacto no mercado (ALLENNLP, 2023);

3. **Aprendizado de Máquina** - A biblioteca *Scikit-learn* será usada para o desenvolvimento dos modelos de previsão que serão treinados com os dados resultantes das etapas anteriores. Os algoritmos: *Nearest Neighbors*, *Support Vector Machines* e *Random Forest* são usados para verificar qual deles terá o melhor desempenho preditivo (SCIKIT-LEARN, 2023). O primeiro algoritmo foi escolhido para testar a relação entre variações de uma ação em outras. Já os últimos dois foram inspirados no trabalho (SEHGAL; SONG, 2007);
4. **Portal de Resultados** - Um portal *web* terá os resultados, ou seja, descrição de quão próximo as predições foram da realidade. A disponibilidade visa facilitar a pesquisa em projetos futuros.

Figura 4.1: Visão geral da arquitetura escolhida.



Fonte: O autor

4.2 Metodologia

A aplicação do trabalho foi realizada em portais de notícias mais conhecidos como: *Yahoo Finance*, *Google Finance* e outros. Os portais reconhecidos tiveram preferência pois, além de terem volume de dados relevante, também continham bibliotecas para exportação de dados, facilitando a entrega para o módulo de PLN. Com os dados

concisos das notícias, há também a necessidade de ter o histórico da cotação das ações de interesse. Estas, serão capturadas no portal da (NASDAQ, 2023) e (FINANCE, 2023b) dos últimos cinco anos (limitação dos históricos nos portais) e verificar o que é mais completo para o treinamento dos modelos preditivos, ou seja, o que trará melhor precisão nas previsões. Por último, serão desenvolvidos modelos de predição com base em três algoritmos: *Nearest Neighbors*, *Naive Bayes* e *Random Forest*. Foi verificado qual destes se demonstrou ser a melhor técnica para a construção de um modelo preditivo para os casos considerando a métrica MAPE e também MSE. Os resultados dos indicadores para cada algoritmo aplicado serão descritos no capítulo 6.

5 EXPERIMENTOS

Esta capítulo discorrerá sobre como foi implementado o modelo escolhido. Além disso também descreverá as mudanças realizadas durante o desenvolvimento e suas motivações. Este foi seccionado em subseções para que cada uma aborde uma etapa do projeto construído.

5.1 Implementação

O objetivo do protótipo construído foi testar quais algoritmos de aprendizado de máquina desempenham melhor predição com base em notícias divulgadas na Internet. O protótipo implementado foi desenvolvido através de módulos ou etapas, e estes foram: a extração de dados históricos, extração de dados de notícias, análise de sentimentos das notícias, aprendizado de máquina e por fim a avaliação de acurácia dos modelos treinados. A primeira etapa consiste na captura de dados históricos de ações através de uma API, ela recebe uma requisição contendo a ação e período desejado e assim, retorna os dados em formato JSON. Na próxima etapa temos a extração de notícias, muito similarmente com a etapa anterior, ela foi desenvolvida em uma API que recebe variáveis de filtros e outras configurações para, então, retornar os dados em formato JSON. Já na etapa de análise de sentimentos, pressupõe-se que temos os dados necessários, e a ferramenta de análise escolhida dá uma nota para os sentimentos de cada notícia. Com a nota do sentimento de cada notícia, esses dados são inseridos na etapa de aprendizado de máquina. E então, são realizados os treinamentos em cada algoritmo escolhido e depois aplicada a previsão no conjunto de dados. Com as previsões feitas a última etapa pode iniciar para verificar a diferença entre o que foi previsto e os valores reais. As métricas escolhidas são usadas para avaliar a qualidade das predições entre as técnicas de predição realizadas.

5.2 Experimentos Realizados

O protótipo descrito nas subseções a seguir foi desenvolvido na linguagem Python rodando em um Sistema Operacional Linux, distribuição Ubuntu, versão 22.04. Logo, quando bibliotecas forem citadas, serão relacionadas com a linguagem de programação mencionada.

5.2.1 Web Scraping

Inicialmente, foi realizada a raspagem de dados do portal *Yahoo Finance* usando a biblioteca **Scrapy** (SCRAPY, 2023), versão 2.10, para buscar as páginas HTML do portal e buscando outros links iterativamente. Alguns problemas nessa abordagem foram, por exemplo, a demora, que estava relacionada ao *overhead* (sobrecarga) das requisições no portal e também ao limite de recursos da máquina executando a extração. Outro contratempo foi a duplicação de dados que ocorria ao existirem links nas páginas extraídas que formavam *loops*, assim um filtro foi construído para manter as notícias salvas apenas uma vez. A biblioteca **yfinance** (YFINANCE, 2023), versão 0.2.28, também foi testada na etapa de extração de dados de notícias na web, entretanto, mesmo facilitando muito o desenvolvimento por retornar apenas dados específicos das notícias e não ter o *overhead* da biblioteca citada anteriormente, não há histórico suficiente para os propósitos do trabalho. Havia cerca de cinco últimas notícias sendo retornadas por cada ação testada e não havendo possibilidade de percorrer mais dados historicamente. Inspirado nos impactos positivos relativos ao uso da biblioteca acima, foi também testada a **NewsAPI** (NEWSAPI, 2023). Esta sendo mais robusta e conhecida, entretanto, com um plano grátis muito limitado. Nos testes foi possível testar filtros na API, consultar dados de até 30 dias antes da data da consulta, mas com limite diário de apenas algumas centenas de resultados. Logo, essas limitações impediram a continuação desta biblioteca no decorrer do desenvolvimento do trabalho. Por fim, a API do portal **EODHD** (EODHD, 2023) foi testada. Este, teve o melhor balanço entre as opções já citadas pois foi possível consultar mais de um mês de dados de um ativo da bolsa de valores por dia e a limitação de dados históricos foi de um ano, apesar da versão ser gratuita. Então os dados de notícias já filtrados e semi estruturados foram extraídos por meio dessa ferramenta, assim, diminuindo a complexidade do projeto, tanto no processo de extração quanto no de estruturação de dados.

5.2.2 Dados Históricos de Ações

A extração dos dados históricos de ações ocorreu sem contratempos, em comparação à etapa descrita na subseção anterior. Foi utilizada a biblioteca *yfinance*, versão 0.2.28, para capturar as informações desejadas de ativos na bolsa de valores americana. As informações, agrupadas por dia, de interesse das ações foram: valor de abertura, valor

de fechamento, valor de pico, valor de baixa e data. Assim, é possível relacionar a variação do valor das ações com os sentimentos das notícias nas próximas etapas. Na tabela 5.1 vemos um exemplo de conjunto de dados relativos à ação da empresa Tesla (TSLA) que foram extraídos.

Tabela 5.1: Exemplo de Valores de Ações Extraídas

Data(yyyy-mm-dd)	**Ação**	**Abertura**	**Alta**	**Baixa**	**Fechamento**
2023-08-03	TSLA	199,91	202,69	192,20	194,77
2023-08-04	TSLA	197,32	198,74	190,32	192,58
2023-08-05	TSLA	190,52	190,67	183,76	185,52
...	...	...	...	...	...

Fonte: Retorno da biblioteca **yfinance**

5.2.3 Manipulação de Dados

Ao realizar as requisições para as bibliotecas e APIs descritas nas subseções da seção 5.2, estas são respondidas em formato JSON. Após a seleção das informações de interesse, os dados são aglutinados e salvos em arquivos do mesmo formato de origem. Este caminho foi escolhido para facilitar a etapa de estruturação em DataFrame dos dados. Para trabalhar os dados a biblioteca *pandas*, versão 2.1, foi utilizada. Efetuando a leitura dos arquivos salvos, aplicando funções aritméticas ou de PLN, foi possível construir resultados estruturados, em tabela, das predições de mercado.

Ao ter os dados estruturados, foi possível realizar o tratamento dos mesmos. A deduplicação foi realizada verificando se uma mesma notícia foi recebida mais de uma vez de uma mesma fonte, se sim, exclui-se as duplicatas e é mantido um exemplar dos registros. Outra processo realizado nos dados foi a verificação de caracteres estranhos, ou seja, caso uma notícia tenha sido recebida com caracteres não pertencentes ao conjunto padrão da língua inglesa o registro é descartado para não gerar ruído na etapa de análise de sentimentos. Foram mantidos os registros de notícias que são parecidas ou até iguais, no quesito de título, que foram publicadas em veículos diferentes, assim multiplicando a influência da notícia por ter sido veiculada em múltiplos portais.

5.2.4 Análise de Sentimentos

Primeiramente, a biblioteca *AllenNLP*, versão 2.10.1, foi escolhida para realizar a análise de sentimentos dos textos das notícias. Entretanto, foram verificadas duas principais dificuldades que desencorajaram o uso dessa ferramenta. A biblioteca foi arquivada no fim de 2022, e após um período em modo de manutenção, ou seja, recebendo questões dos usuários e consertando falhas, ela parou de receber qualquer tipo de atualização. Isso dificultou o processo de instalação de dependências, pois certas bibliotecas precisavam ser usadas em versões especificas e que não eram encontradas tão facilmente, como a biblioteca *torchvision* na versão 0.8.2, que não foi encontrada tentar a instalação pelo gerenciador de pacotes Python *pip*. Encontrando portais na internet que continham arquivos das versões de interesse das bibliotecas, foi possível utilizar a ferramenta de análise de sentimentos da AllenNLP. Entretanto, muitos dos resultados, cerca de 60%, foram inconclusivos, ou seja, a ferramenta não conseguiu concluir se o texto tinha uma conotação negativa ou positiva.

Após esses contratempos, foi encontrada a biblioteca *Valence Aware Dictionary for sEntiment Reasoner* ou VADER que é um analisador léxico que foi especialmente testado para sentimentos expressados nas redes sociais, mas também usado para demais textos (VADER, 2023). Os resultados obtidos através dele foram bem mais relevantes, ocorrendo muito menos casos em que a ferramenta não soubesse classificar se um texto era positivo ou não. Na tabela 5.2 vemos que as notícias com sentimentos negativos tendem a ter uma nota baixa, isso em relação à faixa de valores que é de *-1* (nota mais negativa) até *+1* (nota mais positiva). As faixas de valores recomendadas para decidir qual conotação o resultado indica são:

- Sentimento positivo: `Composto` >= 0,05
- Sentimento neutro: `Composto` < 0,05 e `Composto` > -0,05
- Sentimento negativo: `Composto` <= -0,05

O valor composto é a soma da pontuação de cada palavra do texto analisado e ajustado de acordo com regras da ferramenta, e então o valor sendo normalizado entre -1 e 1.

Tabela 5.2: Análise de Sentimentos com VADER

Texto de Exemplo	Negativo	Neutro	Positivo	Composto
August jobs report: U.S. added 187,000 jobs, unemployment rate rises.	0,24	0,76	0,00	-0,44
Amazon is quietly building the most powerful advertising machine in the world	0,00	0,68	0,32	0,59
Tesla investigated for reportedly funding Musk's glass house	0,00	1,00	0,00	0,00
Microsoft vs Apple: Which Is the Better Dividend Stock?	0,00	0,73	0,27	0,44

Fonte: O autor.

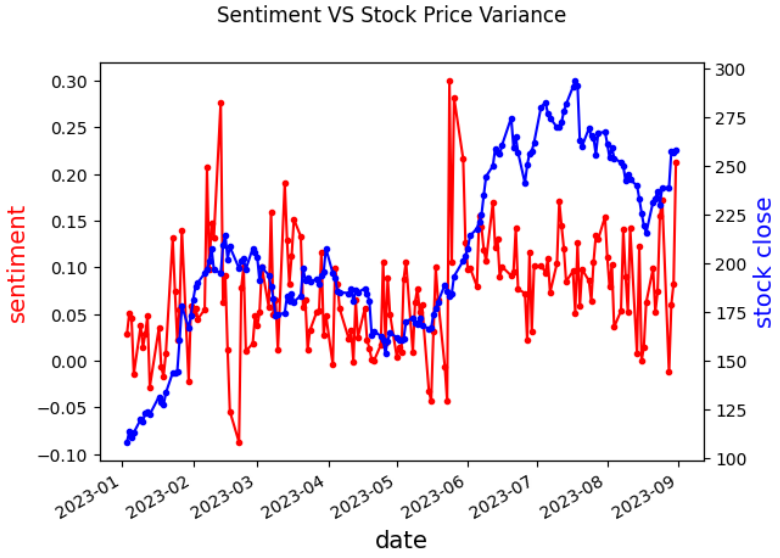
5.2.5 Aprendizado de Máquina

A etapa de aprendizado de máquina acontece após a execução dos passos anteriores, pois temos os dados necessários para treinar os modelos preditivos, ou seja, com a entrada sendo os resultados da análise de sentimentos e a saída esperada que seria a variação da ação após as notícias serem divulgadas. Para realizar esse experimento foi usada a biblioteca *scikit-learn*, versão 1.3, e usado o algoritmo de Regressão Florestal Aleatório ou *Random Forest Regressor* que tem como entrada a média aritmética do valor normalizado da análise de sentimentos das notícias de cada dia. Após o treinamento, foi feita uma verificação de quão próximo do valor esperado (real variação da ação) ficou das previsões do modelo.

Os dados de entrada para o treino dos algoritmos para cada ativo foram: a média dos sentimentos das notícias do ativo, valor de fechamento do dia anterior, valor de abertura, valor de baixa, valor de alta e variação entre o fechamento do dia anterior e o fechamento do dia. Essas entradas são insumos para a predição diária do ativo, ou seja, as entradas são incrementadas diariamente para melhorar a predição ao longo do tempo. Usando a ação TSLA, da empresa *Tesla, Inc.*, podemos ver no gráfico da figura 5.1 os dados de entrada e os alvos para cada dia. Nela, podemos verificar a oscilação na variação do fechamento da ação, entretanto também é possível observar que a média da variação de sentimento ao longo do tempo acompanha os movimentos de subida e descida do valor alvo.

O modelo treinado mostrou certa precisão na predição de fechamento de valor das ações. No teste de dados, da empresa citada, entre janeiro de 2023 até o do mês de

Figura 5.1: Variação do Sentimento Analisado nas Notícias e Valor da Ação.



Fonte: (FINANCE, 2023b) e o autor.

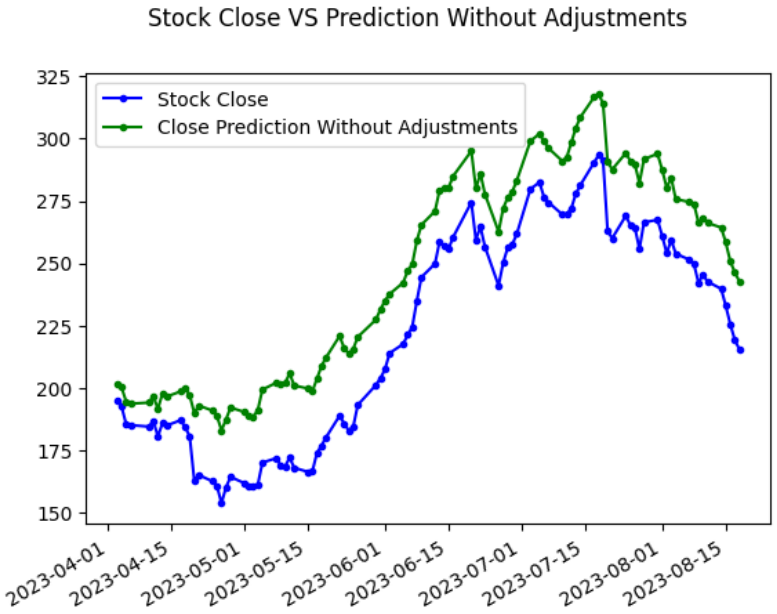
agosto de 2023 a variação de valores entre o fechamento e a previsão usando o algoritmo *Random Forest Regressor* foi de:

- Intervalo de [0 - 1] dólares - 46,70%
- Intervalo de (1, 3] - 38,92%
- Intervalo de (3, 5] - 9,58%
- Intervalo de (5, 10] - 4,19%
- Intervalo de (10, ∞) - 0,00%

Relacionando esses valores com o valor médio de fechamento das ações no período (cerca de $205,13), vemos que cerca de 50% dos dias a previsão errou por volta de 0,5% do valor alvo. Já englobando 95% das previsões, temos um erro de apenas 2,4%, isso na aplicação do modelo construído em notícias divulgadas em cada dia individualmente. Caso não fosse recalibrado o valor inicial da previsão diária os valores previstos teriam muito mais erros ao logo dos meses. Na figura 5.2, vemos que o acúmulo de erros e pontuais dias levam a predição a se afastar do valor desejado. Alguns dos dias que divergem e conduzem a falha da previsão seriam os que estão logo após a segunda quinzena de abril, entretanto, vemos que mesmo com essa falha pontual as previsões seguem o comportamento da curva de fechamento real.

Avaliando o desempenho dos algoritmos utilizando as métricas MAPE e MSE vemos que o *Random Forest* se sobressaiu em ambos testes. Pontuar menos na métrica

Figura 5.2: Fechamento de Ação vs Previsão de Fechamento de Ação - TSLA.



Fonte: (FINANCE, 2023b) e o autor.

Erro Médio Percentual Absoluto significa que o modelo está errando proporcionalmente menos em média, já no *Erro Médio Quadrático* significa que a diferença entre o previsto e o real não ocorreram muitos pontos discrepantes, ou seja, pontos com muita divergência. Podemos observar os resultados na tabela 5.3 que mostram como os algoritmos se saíram ao serem aplicados nos primeiros oito meses de 2023 no ativo da empresa Tesla.

Tabela 5.3: Resultados dos Algoritmos em Métricas de Erros - TSLA

Ação	**Data Inicial**	**Data Final**	**Notícias**	**Algoritmo**	**MAPE**	**MSE**
TSLA	2023-01-01	2023-08-31	7235	Random Forest	0.99	1469
TSLA	2023-01-01	2023-08-31	7235	K-Neighbors	1.29	1905
TSLA	2023-01-01	2023-08-31	7235	SVM-linear	1.50	2598
TSLA	2023-01-01	2023-08-31	7235	SVM-rbf	1.55	3320
TSLA	2023-01-01	2023-08-31	7235	SVM-poly	1.99	4092

Fonte: o autor.

Aplicando as mesmas técnicas e avaliações em outros conjuntos de dados, agora da empresa Apple e da Netflix, temos os resultados da tabela 5.4. Vemos, principalmente, a diferença positiva nos resultados das previsões para a ação da Apple, isso em relação as outras duas avaliadas. Os algoritmos que foram avaliados foram o Random Forest (familiar ao Árvore de Decisão), *Nearest Neighbors*, *Support Vector Machines* com três variações de *kernel* ou dependência entre variáveis que são: linear, polinomial e *RBF* ou Função de Base Radial que seria um tipo de expressão exponencial.

Tabela 5.4: Resultados dos Algoritmos em Métricas de Erros - APPL & NFLX

Ação	**Data Inicial**	**Data Final**	**Notícias**	**Algoritmo**	**MAPE**	**MSE**
AAPL	2023-01-01	2023-08-31	4544	Random Forest	0.39	171
AAPL	2023-01-01	2023-08-31	4544	K-Neighbors	0.51	229
AAPL	2023-01-01	2023-08-31	4544	SVM-linear	0.57	292
AAPL	2023-01-01	2023-08-31	4544	SVM-rbf	0.56	306
AAPL	2023-01-01	2023-08-31	4544	SVM-poly	0.59	309
NFLX	2023-01-01	2023-08-31	1828	Random Forest	0.76	3322
NFLX	2023-01-01	2023-08-31	1828	K-Neighbors	1.01	4210
NFLX	2023-01-01	2023-08-31	1828	SVM-linear	1.03	4870
NFLX	2023-01-01	2023-08-31	1828	SVM-rbf	1.08	6404
NFLX	2023-01-01	2023-08-31	1828	SVM-poly	1.26	7043

Fonte: o autor.

6 ANÁLISE DOS RESULTADOS

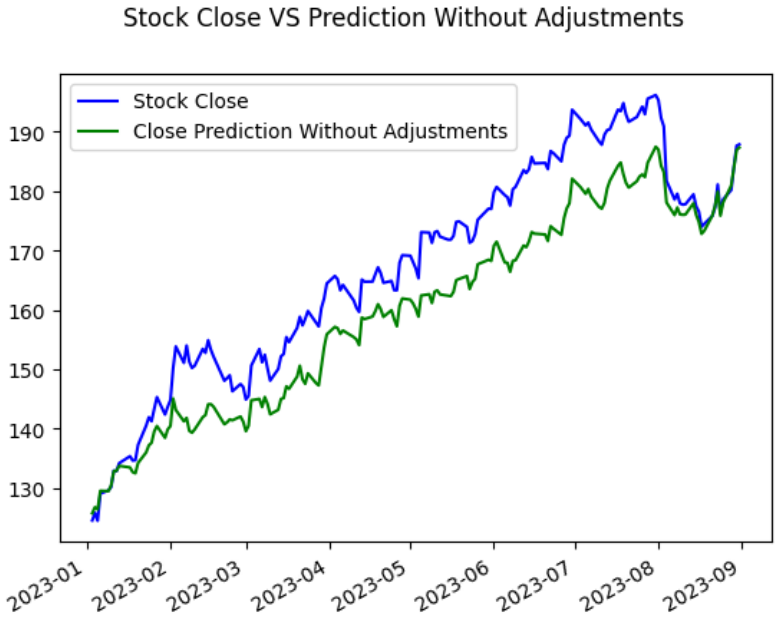
Este capítulo tem com objetivo discorrer sobre os resultados alcançados durante o desenvolvimento do protótipo e a realização do experimentos. Será apresentado uma visão crítica das escolhas feitas na construção de cada etapa do trabalho e as métricas resultantes da aplicação do protótipo nos conjuntos de dados escolhidos.

A extração dos dados históricos das ações de interesse foi a etapa de menor complexidade e pelas informações serem públicas e estruturadas foi aplicada uma solução simples, sendo ela uma biblioteca que recebe as requisições e retorna os dados solicitados. Entretanto, a etapa posterior, de extração de notícias, foi mais complicada. Esta segunda etapa necessita de dados de múltiplos portais na web, o que dificulta serem estruturados e gratuitos. Foram investidos vários dias na procura e testes da melhor ferramenta para o protótipo. Fazer uma pesquisa mais abrangente antes de efetuar vários testes aceleraria o processo de encontrar a solução mais próxima do ideal. Estes dados são limitados com o título da notícia e parte do conteúdo, mas foram suficientes para indicarem os sentimentos relativos às notícias, sendo que o título tem maior peso no texto.

Já a análise de sentimentos, próxima etapa no fluxo do protótipo, utilizando a ferramenta VADER mostrou bons resultados na classificação de sentimentos, mostrando raramente dificuldade e classificando textos como neutros, q que não é tão comum no conjunto utilizado. Uma possível melhoria seria a utilização de mais ferramentas de análise de sentimentos para verificar quais combinações de analisador de sentimentos e algoritmo de aprendizagem de máquina têm melhores resultados nos conjuntos de dados.

Por fim, os resultados da etapa de aprendizagem de máquina mostraram que o algoritmo de Random Forest tem melhor desempenho entre os testados, ou seja, Random Forest, K-Neighbors e SVM com kernel RBF, Linear ou Polinomial. Na métrica MAPE o algoritmo 'vencedor' ficou com uma margem de pelo menos **25-30% de melhor** precisão que os demais. Já no MSE, vemos os resultados muito parecidos, mesmo com a métrica realçando os pontos divergentes nos dados da saída. Logo, vemos que os resultados do **Random Forest** estão **26%-34% melhor** que o segundo colocado, que no caso foi o K-Neighbors. Estes dados podem ser verificados nas tabelas 5.3 e 5.4. A respeito das diferenças de valores entre os dados dos ativos de interesse é possível inferir que as ações com menos variações bruscas e com mais notícias podem ter melhores resultados como é o caso do ativo AAPL da Apple que podemos ver na figura 6.1.

Figura 6.1: Fechamento de Ação vs Previsão de Fechamento de Ação - AAPL



Fonte: (FINANCE, 2023b) e o autor.

7 CONCLUSÃO

Após a aplicação do protótipo construído concluímos que é possível, em certa medida, prever variações de mercado com margem de erro baixa, algo próximo de 1% nos experimentos realizados. Entretanto, também são verificados pontos de erro por variações bruscas no mercado como visto na figura 5.2 descrevendo o preço da ação da empresa Tesla e a previsão sem reajuste diário. Os experimentos se basearam em um conjunto sucinto de dados, este referente a três ativos nos primeiros oito meses de 2023. O uso de mais ativos e utilizá-los ao mesmo tempo para o treinamento dos modelos preditivos podem levar a diferentes resultados, podendo encontrar modelos mais genéricos e robustos. Outra recomendação para trabalhos futuros seriam o uso de demais ferramentas de análise de sentimentos. O recurso utilizado tem grande foco na classificação de textos das mídias sociais, o que pode não ser ideal para os dados de notícias de finanças que servem de entrada. Ao testar múltiplos classificadores junto dos algoritmos seria possível encontrar as melhores combinações, logo desenvolvendo um protótipo mais robusto e específico para o tema abordado.

REFERÊNCIAS

ALLENNLP. **An Apache 2.0 NLP research library, built on PyTorch, for developing state-of-the-art deep learning models on a wide variety of linguistic tasks.** 2023. Último acesso em 18/03/2023. Available from Internet: <https://docs.allennlp.org/main/>.

ANBIMA. **Mercado de capitais brasileiro aumentou quatro vezes em 25 anos**. 2022. Último acesso em 16/03/2023. Available from Internet: <https://www.anbima.com.br/pt_br/noticias/mercado-de-capitais-brasileiro-aumentou-quatro-vezes-em-25-anos.htm>.

ARMANO, G.; MARCHESI, M.; MURRU, A. A hybrid genetic-neural architecture for stock indexes forecasting. **Information Sciences**, v. 170, n. 1, p. 3–33, 2005. ISSN 0020-0255. Computational Intelligence in Economics and Finance. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S002002550300433X>.

ATSALAKIS, G. S.; VALAVANIS, K. P. Forecasting stock market short-term trends using a neuro-fuzzy based methodology. **Expert Systems with Applications**, v. 36, n. 7, p. 10696–10707, 2009. ISSN 0957-4174. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0957417409001948>.

B3. **Número de investidores na B3 cresce mesmo em cenário de alta volatilidade**. 2022. Último acesso em 05/03/2023. Available from Internet: <https://www.b3.com.br/pt_br/noticias/numero-de-investidores-na-b3-cresce-mesmo-em-cenario-de-alta-volatilidade.htm>.

BADOLIA, L. **How can i get started investing in the stock market**. [S.l.]: Educreation Publishing, 2016.

CHEN, Y.; HAO, Y. A feature weighted support vector machine and k-nearest neighbor algorithm for stock market indices prediction. **Expert Systems with Applications**, v. 80, p. 340–355, 2017. ISSN 0957-4174. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0957417417301367>.

CHOWDHARY. Natural language processing. **Fundamentals of artificial intelligence**, Springer, p. 603–649, 2020.

DZIKEVIČIUS, A.; ŠARANDA, S. Ema versus sma usage to forecast stock markets: the case of s&p 500 and omx baltic benchmark. **Business: Theory and Practice**, v. 11, n. 3, p. 248–255, 2010.

EBRAHIMPOUR, R. et al. Mixture of mlp-experts for trend forecasting of time series: A case study of the tehran stock exchange. **International Journal of Forecasting**, v. 27, n. 3, p. 804–816, 2011. ISSN 0169-2070. Special Section 1: Forecasting with Artificial Neural Networks and Computational Intelligence Special Section 2: Tourism Forecasting. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0169207010000920>.

ENKE, D.; THAWORNWONG, S. The use of data mining and neural networks for forecasting stock market returns. **Expert Systems with Applications**, v. 29, n. 4, p. 927–940, 2005. ISSN 0957-4174. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0957417405001156>.

EODHD. **30+ years of data. Financial Data APIs.** 2023. Último acesso em 01/09/2023. Available from Internet: <https://eodhd.com/>.

FINANCE, G. **First Republic Bank (FRC) Stock Price & News - Google Finance**. 2023. Último acesso em 18/03/2023. Available from Internet: <https://www.google.com/finance/quote/FRC:NYSE?hl=en&window=1Y>.

FINANCE, Y. **An open source and collaborative framework for extracting the data you need from websites. In a fast, simple, yet extensible way.** 2023. Último acesso em 18/03/2023. Available from Internet: <https://finance.yahoo.com/quote/\%5EIXIC/history/>.

HOSEINZADE, E.; HARATIZADEH, S. Cnnpred: Cnn-based stock market prediction using a diverse set of variables. **Expert Systems with Applications**, v. 129, p. 273–285, 2019. ISSN 0957-4174. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0957417419301915>.

HUANG, W.; NAKAMORI, Y.; WANG, S.-Y. Forecasting stock market movement direction with support vector machine. **Computers & Operations Research**, v. 32, n. 10, p. 2513–2522, 2005. ISSN 0305-0548. Applications of Neural Networks. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0305054804000681>.

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

KAHNEMAN, D.; TVERSKY, A. **Intuitive prediction: Biases and corrective procedures**. [S.l.], 1977.

KAUSAR, M. A.; DHAKA, V.; SINGH, S. K. Web crawler: a review. **International Journal of Computer Applications**, Foundation of Computer Science, 244 5 th Avenue,# 1526, New York, NY 10001 . . . , v. 63, n. 2, p. 31–36, 2013.

KIM, Y.; JEONG, S. R.; GHANI, I. Text opinion mining to analyze news for stock market prediction. **Int. J. Advance. Soft Comput. Appl**, v. 6, n. 1, p. 2074–8523, 2014.

KUMBURE, M. M. et al. Machine learning techniques and data for stock market forecasting: A literature review. **Expert Systems with Applications**, v. 197, p. 116659, 2022. ISSN 0957-4174. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0957417422001452>.

MAHESH, B. Machine learning algorithms-a review. **International Journal of Science and Research (IJSR).[Internet]**, v. 9, n. 1, p. 381–386, 2020.

MOHAN, S. et al. Stock price prediction using news sentiment analysis. In: IEEE. **2019 IEEE fifth international conference on big data computing service and applications (BigDataService)**. [S.l.], 2019. p. 205–208.

NASDAQ. **Historical data**. 2023. Último acesso em 18/03/2023. Available from Internet: <https://www.nasdaq.com/market-activity/quotes/historical>.

NEWSAPI. **News API is a simple, easy-to-use REST API that returns JSON search results for current and historic news articles published by over 80,000 worldwide sources.** 2023. Último acesso em 01/09/2023. Available from Internet: <https://newsapi.org/>.

O'CONNOR, N.; MADDEN, M. G. A neural network approach to predicting stock exchange movements using external factors. **Knowledge-Based Systems**, v. 19, n. 5, p. 371–378, 2006. ISSN 0950-7051. AI 2005 SI. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0950705106000153>.

RAJIHY, Y.; NERMEND, K.; ALSAKAA, A. Back-propagation artificial neural networks in stock market forecasting. an application to the warsaw stock exchange wig20. **The IEB International Journal of Finance**, v. 15, p. 88–99, 01 2017.

SCHUMAKER, R. P.; CHEN, H. Textual analysis of stock market prediction using breaking financial news: The azfin text system. **ACM Transactions on Information Systems (TOIS)**, ACM New York, NY, USA, v. 27, n. 2, p. 1–19, 2009.

SCHÜNKE, M. A. Aplicação de algoritmos de classificação para análise dos fatores que influenciam na predição do fator de impacto nas redes sociais. **PPGC UFRGS**, 2015.

SCHWERT, G. W. Stock market volatility. **Financial analysts journal**, Taylor & Francis, v. 46, n. 3, p. 23–34, 1990.

SCIKIT-LEARN. **Simple and efficient tools for predictive data analysis.** 2023. Último acesso em 18/03/2023. Available from Internet: <https://scikit-learn.org/stable/>.

SCRAPY. **An open source and collaborative framework for extracting the data you need from websites. In a fast, simple, yet extensible way.** 2023. Último acesso em 18/03/2023. Available from Internet: <https://scrapy.org/>.

SEDIGHI, M. et al. A novel hybrid model for stock price forecasting based on metaheuristics and support vector machine. **Data**, v. 4, n. 2, 2019. ISSN 2306-5729. Available from Internet: <https://www.mdpi.com/2306-5729/4/2/75>.

SEHGAL, V.; SONG, C. Sops: stock prediction using web sentiment. In: IEEE. **Seventh IEEE international conference on data mining workshops (ICDMW 2007)**. [S.l.], 2007. p. 21–26.

SONG, Y.; LEE, J. W.; LEE, J. A study on novel filtering and relationship between input-features and target-vectors in a deep learning model for stock price prediction. **Applied Intelligence**, v. 49, n. 3, p. 897–911, Mar 2019. ISSN 1573-7497. Available from Internet: <https://doi.org/10.1007/s10489-018-1308-x>.

VADER. **VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains.** 2023. Último acesso em 01/09/2023. Available from Internet: <https://github.com/cjhutto/vaderSentiment>.

WENG, B.; AHMED, M. A.; MEGAHED, F. M. Stock market one-day ahead movement prediction using disparate data sources. **Expert Systems with Applications**, v. 79, p. 153–163, 2017. ISSN 0957-4174. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0957417417301331>.

YFINANCE. **yfinance is not affiliated, endorsed, or vetted by Yahoo, Inc. It's an open-source tool that uses Yahoo's publicly available APIs, and is intended for**

research and educational purposes. 2023. Último acesso em 01/09/2023. Available from Internet: <https://github.com/ranaroussi/yfinance>.