

**UNIVERSIDADE FEDERAL DE MINAS GERAIS  
INSTITUTO DE CIÊNCIAS EXATAS  
CIÊNCIAS ATUARIAIS**

**MARCO ANTÔNIO ARAGÃO ROCHA**

**PROCESSAMENTO DE LINGUAGEM NATURAL E MACHINE LEARNING PARA  
A CONSTRUÇÃO DE ESTRATÉGIA QUANTITATIVA DE INVESTIMENTOS**

**BELO HORIZONTE**

**2024**

MARCO ANTÔNIO ARAGÃO ROCHA

**PROCESSAMENTO DE LINGUAGEM NATURAL E MACHINE LEARNING PARA  
A CONSTRUÇÃO DE ESTRATÉGIA QUANTITATIVA DE INVESTIMENTOS**

Monografia apresentada por Marco Antônio Aragão Rocha ao Departamento de Estatística da Universidade Federal de Minas Gerais (UFMG), como parte dos requisitos necessários à obtenção do título de Bacharel em Ciências Atuariais.

Orientador: Prof. Marcos Oliveira Prates

BELO HORIZONTE

2024

## RESUMO

No mercado de investimentos a capacidade de entender cenários macroeconômicos e ponderar a alocação dos ativos no portfólio com base nisso é uma tarefa imprescindível. Atualmente, grandes empresas e consultorias de investimento possuem diversos funcionários com esse papel, os chamados especialistas de investimentos. Contudo, assim como todas as tarefas humanas, essa abordagem está fortemente susceptível a emoções e vieses, conforme apontado por Daniel Kahneman, psicólogo, economista e ganhador do prêmio Nobel em 2002. Kahneman demonstrou como as decisões humanas frequentemente se desviam da racionalidade prevista por outras teorias econômicas. Sob essa ótica, o presente trabalho foi desenvolvido com o intuito de desenvolver uma estratégia quantitativa de investimentos, que opere de forma autônoma, evitando os vieses humanos. Partindo desde o Web Scraping de notícias até a utilização de técnicas de processamento de linguagem natural como Word2Vec para auxiliar na representação do cenário macroeconômico e a partir disso, construir um portfólio de investimentos utilizando técnicas de aprendizado de máquina que busquem encontrar padrões entre as variáveis geradas a partir do processamento textual e o histórico de retorno do índice de ações brasileiro Ibovespa e o CDI. A relevância deste trabalho reside na integração de técnicas de PLN e aprendizado de máquina na análise de dados financeiros, uma área de crescente interesse e importância. A previsão de retornos de mercado com base em análises de sentimento oferece uma perspectiva inovadora para a tomada de decisões de investimento, potencialmente aumentando a eficiência e a rentabilidade das estratégias de alocação de ativos. Além disso, a capacidade de reagir de forma dinâmica às mudanças no ambiente macroeconômico e geopolítico é algo de grande relevância no contexto do mercado de capitais.

**Palavras-chave:** Aprendizado de máquina, investimentos, processamento de linguagem natural, previsão, classificação.

## ABSTRACT

In the investment market, the ability to understand macroeconomic scenarios and consider the allocation of assets in the portfolio based on this is an essential task. Currently, large companies and investment consultancies have several employees with this role, so-called investment specialists. However, like all human tasks, this approach is strongly susceptible to emotions and biases, as pointed out by Daniel Kahneman, psychologist, economist and winner of the Nobel Prize in 2002. Kahneman demonstrated how human decisions often deviate from the rationality predicted by other economic theories. From this perspective, this work was developed with the aim of developing a quantitative investment strategy, which operates autonomously, avoiding human biases. Starting from Web Scraping of news to the use of natural language processing techniques such as Word2Vec to assist in representing the macroeconomic scenario and from there, building an investment portfolio using machine learning techniques that seek to find patterns between the variables generated from textual processing and the return history of the Brazilian stock index Ibovespa and the CDI. The relevance of this work lies in the integration of NLP and machine learning techniques in the analysis of financial data, an area of growing interest and importance. Forecasting market returns based on sentiment analysis offers an innovative perspective for making investment decisions, potentially increasing the efficiency and profitability of asset allocation strategies. Furthermore, the ability to react dynamically to changes in the macroeconomic and geopolitical environment is highly relevant in the context of the capital market.

**Keywords:** Machine learning, investments, natural language processing, classification.

## LISTA DE FIGURAS

Figura 1 : Representação da arquitetura CBOW .....	14
Figura 2 : Representação da arquitetura Skip-gram.....	14
Figura 3 : Exemplo de árvore de decisão .....	17
Figura 4 : Estruturas de um neurônio de uma rede neural recorrente LSTM. ....	18
Figura 5 : Arquitetura de uma rede neural recorrente LSTM. ....	20
Figura 6 : Índice de sentimento de 2014 até 2017 .....	23
Figura 7 : Índice de sentimento de 2020 até 2023 .....	24
Figura 8 : Aplicação do modelo De Word2Vec.....	25
Figura 9 : Retorno Acumulado IBOV x CDI .....	26
Figura 10 : Semanas com retorno acumulado IBOV > CDI.....	27
Figura 11 : Comportamento histórico do Target.....	28
Figura 12 : Métricas modelos sem features .....	29
Figura 13 : Métricas de classificação modelos sem features .....	30
Figura 14 : Métricas de classificação modelos com features .....	31
Figura 15 : Aplicação portfólio modelo com features.....	32
Figura 16 : Métricas de classificação modelos com features e sem leakage .....	33
Figura 17 : Aplicação portfólio modelo com features e sem Leakage .....	34

## LISTA DE TABELAS

Tabela 1 : Descrição e tipo das variáveis 1° base de dados .....	10
Tabela 2 : Descrição e tipo das variáveis 2° base de dados .....	11
Tabela 3 : Exemplificação Data Leakage .....	33

## SUMÁRIO

<b>1. INTRODUÇÃO</b>	<b>8</b>
1.1. APRESENTAÇÃO	8
1.2. OBJETIVOS	9
1.2.1. <i>Objetivos Gerais</i>	9
1.2.2. <i>Objetivos Específicos</i>	9
<b>2. BANCO DE DADOS</b>	<b>10</b>
2.1. BANCO DE NOTÍCIAS	10
2.2. BANCO DE ATIVOS	10
<b>3. METODOLOGIA</b>	<b>12</b>
3.1. PROCESSAMENTO TEXTUAL	12
3.1.1. <i>Bag Of Words</i>	12
3.1.2. <i>Word2Vec</i>	13
3.2. MÉTODOS PARA PREVISÃO	15
3.2.1. <i>ARIMA</i>	15
3.2.2. <i>SARIMA</i>	15
3.2.3. <i>SARIMAX</i>	16
3.2.4. <i>DTSF</i>	16
3.2.5. <i>CATBOOST</i>	17
3.2.6. <i>LSTM</i>	18
3.3. ANÁLISE E MÉTRICAS DE AVALIAÇÃO	20
<b>4. APLICAÇÃO</b>	<b>22</b>
4.1. PRÉ-PROCESSAMENTO	22
4.1.1. <i>Remoção Caracteres Especiais e Normalização</i>	22
4.1.2. <i>Tokenizer</i>	22
4.1.3. <i>Remoção de Stop words</i>	22
4.2. ANÁLISE DESCRITIVA	22
4.2.1. <i>Aplicação dos Métodos de Processamento Textual</i>	22
4.2.2. <i>Aplicação Word2Vec</i>	25
4.2.3. <i>Análise Retornos Dos Investimentos</i>	26
<b>5. RESULTADOS</b>	<b>28</b>

**6. CONCLUSÃO E AVANÇOS FUTUROS**

**35**

**7. REFERÊNCIAS**

**36**

## 1. INTRODUÇÃO

### 1.1. APRESENTAÇÃO

A era digital trouxe consigo uma avalanche de dados e avanços tecnológicos, com isso, cada vez mais informações de notícias do mundo inteiro estão disponíveis em tempo recorde, transformando a maneira como os mercados financeiros operam e como os investidores tomam decisões.

A capacidade de análise constante das atualidades e extração de insights tornou-se um diferencial competitivo. Contudo, do ponto de vista humano, pode-se dizer que é inviável estar antenado em todas as notícias e avaliar como cada mudança macroeconômica pode impactar os investimentos, em adição a isso, essa abordagem de análise humanizada está muito susceptível a vieses de acordo com Tversky e Kahneman (1974).

Nesse cenário, as estratégias quantitativas de investimentos emergem como uma abordagem chamativa devido a sua capacidade de analisar grandes volumes de dados de forma objetiva e sistemática, reduzindo o impacto de vieses emocionais e subjetivos na tomada de decisões. Utilizando modelos matemáticos e estatísticos para tomar decisões de investimento, baseando-se em dados históricos e informações atuais do mercado.

Este trabalho parte do pressuposto de que as notícias divulgadas geram um impacto no comportamento do investidor e que ao se aplicar técnicas de processamento de linguagem natural para extração de features textuais como, por exemplo, a construção de um índice de sentimento das notícias pode-se obter uma vantagem na previsão do mercado.

Tem sido relativamente comum na literatura o uso de sentimento geopolítico para a construção de portfólios e análise de poder preditivo e/ou inferencial para dados de mercado, especialmente após a modernização e a popularização de técnicas de PLN e raspagem de dados. O trabalho recente de Caldara e Iacoviello (2022), tornou-se referência, nos últimos artigos para a construção do índice de sentimento geopolítico, tais quais. *“Russia–Ukraine crisis: The effects on the European stock market”* (AHMED; HASAN e KAMAL, 2023) e *“The Russia-Ukraine conflict and volatility risk of commodity markets. Finance Research Letters”* (FANG e SHAO, 2022), inspirou parcialmente o trabalho.

Este trabalho, portanto, busca contribuir para o avanço do conhecimento na área de investimentos, explorando novas fronteiras tecnológicas que aos poucos estão mudando a maneira como decisões financeiras são tomadas. Ao integrar conceitos de Processamento de Linguagem Natural (PLN) e Machine Learning (ML) em estratégias quantitativas de investimentos, como explorado em *“Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow”* (GÉRON, 2022) e em *“Advances in financial machine learning”* (DE PRADO, 2018). Dessa forma, espera-se demonstrar o potencial dessas tecnologias para aumentar a eficiência e a performance de portfólios de investimento no mercado financeiro.

## 1.2. OBJETIVOS

### 1.2.1. Objetivos Gerais

O objetivo geral deste trabalho é explorar a aplicação de Processamento de Linguagem Natural e Machine Learning na construção de estratégias quantitativas de investimentos.

### 1.2.2. Objetivos Específicos

- Aplicar o uso de PLN na análise de notícias.
- Desenvolver e implementar uma estratégia de trade quantitativo que performe acima do CDI.
- Compreender como o PLN pode ser utilizado para extrair informações relevantes de textos financeiros, auxiliando na tomada de decisões de investimento.
- Desenvolver modelos de ML para predição de tendências de mercado: criar modelos preditivos utilizando técnicas de ML para identificar tendências e padrões no mercado financeiro.
- Avaliar o desempenho das estratégias quantitativas desenvolvidas comparando seu desempenho com métodos tradicionais que não utilizam notícias para auxiliar na predição.
- Explorar a integração de dados textuais e numéricos: investigar como a combinação de dados textuais (notícias, relatórios) e numéricos (cotações, indicadores financeiros) pode aprimorar as estratégias de investimento.

## 2. BANCO DE DADOS

Para construção desse trabalho foram utilizadas duas bases de dados principais que serão apresentadas a seguir. A primeira com informações das notícias que foram usadas para construção de features e a segunda base com as variáveis resposta de interesse com informações do mercado financeiro, especificamente dados dos retornos dos ativos selecionados.

### 2.1. BANCO DE NOTÍCIAS

As notícias utilizadas para treinar os modelos de PLN foram retiradas do jornal The Diplomat por meio de Web Scraping feito em linguagem Python (PYTHON SOFTWARE FOUNDATION, 2024) através do pacote BeautifulSoup (RICHARDSON, 2021) e abrangem um período que vai de 2014 a 2023. A partir dos arquivos de artigos do jornal, foram extraídos título, texto e data de todos os artigos publicados no segmento de geopolítica do The Diplomat. A escolha desse jornal se deu pela disponibilidade de fácil recuperação de grande número de artigos de forma automatizada, juntamente com seu reconhecimento internacional em apresentar artigos especializados respeitados ao redor do mundo. Para utilizar os dados de forma adequada, foi realizado um processo de tratamento de forma a torná-los apropriados como entrada para os modelos, exemplificado na Tabela 1, sendo necessária a filtragem dos casos em que o Web Scraping não realizou toda extração do texto do artigo.

Tabela 1 : Descrição e tipo das variáveis 1º base de dados

Variável	Tipo	Descrição
Date	Date Time	Dia de publicação da notícia no formato "yyyy-mm-dd"
Title	String	Título da notícia
Description	String	Lide da notícia
Article Text	String	Todo corpo da notícia com o artigo inteiro
Topics	String	Lista de tópicos que a notícia se encaixa
Link	String	Link da notícia no site The Diplomat

Fonte: Elaborado pelo autor

### 2.2. BANCO DE ATIVOS

O banco de dados de ativos financeiros utilizado neste trabalho inclui informações sobre o índice Ibovespa (IBOV) e a taxa do Certificado de Depósito

Interbancário (CDI). Estas são duas das principais referências do mercado financeiro brasileiro.

O Ibovespa é o principal indicador de desempenho das ações negociadas na B3 (Brasil, Bolsa, Balcão). Ele é composto pelas ações das empresas mais representativas e negociadas no mercado brasileiro, refletindo o comportamento médio do mercado de ações brasileiro. Foi criado em 1968 e, ao longo desses mais de 55 anos, consolidou-se como referência para investidores no Brasil. O índice é calculado com base no desempenho das ações, considerando fatores como liquidez e volume de negociações.

O CDI a taxa de juros de referência para operações de empréstimos entre bancos. Esta taxa é amplamente utilizada como benchmark para diversas aplicações financeiras, incluindo fundos de investimento e títulos privados. O CDI é considerado um indicador importante do custo do dinheiro no mercado interbancário e, por extensão, influencia diretamente o rendimento de diversas modalidades de investimento.

Os dados de retorno diário do IBOV e do CDI foram extraídos de fontes confiáveis sendo o primeiro do site da CVM (Comissão de Valores Mobiliários) e o segundo do portal de dados abertos do Bacen (Banco Central do Brasil), exposto na Tabela 2.

Tabela 2 : Descrição e tipo das variáveis 2º base de dados

Variável	Tipo	Descrição
Date	Date Time	Data no formato "yyyy-mm-dd"
IBOV	Numérica	Retorno Diário CDI
CDI	Numérica	Retorno Diário Ibovespa

Fonte: Elaborado pelo autor

### 3. METODOLOGIA

Dentro do contexto da construção de estratégias quantitativas de investimento, o aprendizado de máquina é amplamente utilizado porque auxilia no desenvolvimento de métodos computacionais que reconhecem padrões nos dados históricos de mercado a partir de fundamentos estatísticos e matemáticos. Nesse trabalho os algoritmos de aprendizado de máquina utilizados podem ser divididos em duas categorias. Primeiro os métodos de PLN para lidar com as informações macroeconômicas trazidas pelas notícias. Aqui destacam-se os modelos de Bag Of Words (JURAFSKY e MARTIN, 2000) e Word2Vec (MIKOLOV et. al 2013). E finalmente, os métodos de aprendizado de máquina supervisionado nos quais a variável de interesse será o retorno dos investimentos e as variáveis originadas do PLN entraram como features. Aqui destacam-se os modelos ARIMA (EDIGER e AKAR, 2007), DTSF (Costa et al. 2021) e CatBoost (PROKHORENKOVA, 2018), vale ressaltar que alguns desses métodos não aceitam variáveis explicativas, mas são usados como referência para analisar se a inclusão das features provenientes das notícias ajuda na previsão de mercado.

#### 3.1. PROCESSAMENTO TEXTUAL

Essa etapa é uma etapa central do projeto que busca transformar os textos em representações numéricas permitindo assim a aplicação dos algoritmos posteriores.

##### 3.1.1. Bag Of Words

A primeira metodologia de Word Embedding aplicada foi a Bag Of Words. Goldberg (2022) define Word Embedding como sendo um conjunto de técnicas que mapeia a semântica e sintática de uma linguagem natural em um espaço real utilizando estatísticas. Dessa forma, palavras de um conjunto de texto são mapeadas para vetores reais. O espaço destes vetores é denominado "embedding space".

O modelo Bag of Words, também foi descrito por Jurafsky (2000), é uma técnica simples e amplamente utilizada no processamento de linguagem natural para representar texto. Ele funciona ao criar uma matriz que lista os termos presentes em um documento e quantifica quantas vezes cada termo aparece, sem considerar a sequência das palavras. Essa abordagem simplifica a análise textual, focando apenas na frequência das palavras. A presença ou ausência de uma palavra pode ser representada de forma binária, com 1 indicando presença e 0 ausência, uma técnica conhecida como one-hot encoding.

No entanto, essa simplicidade pode ser aprimorada para capturar mais nuances no texto. Silge (2017) explica a técnica de term frequency (tf), onde os valores binários são substituídos por pesos que refletem a importância de cada palavra dentro de um documento específico. Para ajustar ainda mais essa importância em um contexto mais amplo, a técnica inverse document frequency (idf) pondera os pesos de acordo com a raridade de uma palavra em um conjunto de documentos. A combinação dessas abordagens resulta no método tf-idf, que equilibra a frequência de termos em um único documento com sua distribuição geral, permitindo que o modelo valorize palavras que são informativas, mas não excessivamente comuns.

Embora o Bag of Words ignore a semântica e a ordem das palavras, o que pode ser uma limitação em alguns contextos, ele ainda é uma ferramenta poderosa para tarefas que exigem uma representação rápida e eficiente do texto. Quando combinado com técnicas menos simplificadas, como tf-idf, ele se torna mais eficaz para capturar o conteúdo informativo dos documentos.

### 3.1.2. Word2Vec

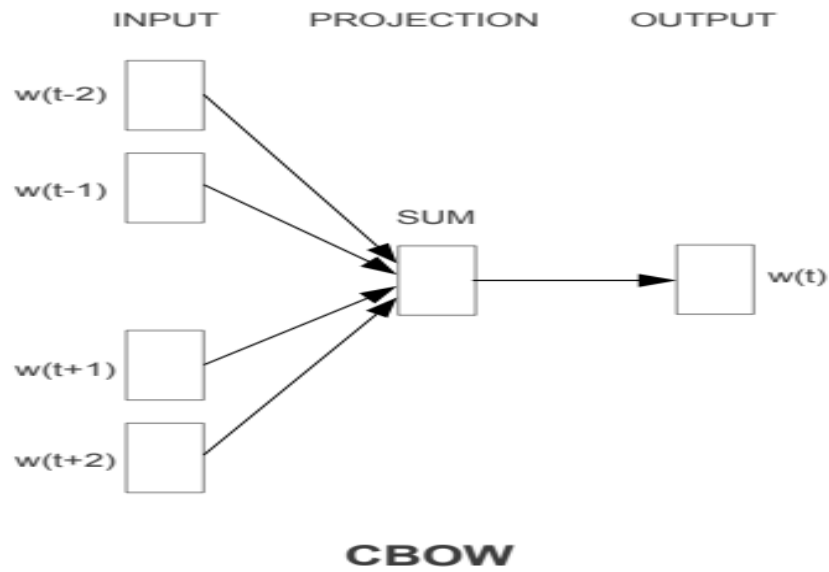
Mikolov et. al (2013) propuseram em seu trabalho duas diferentes arquiteturas para computar palavras em representação vetorial. O objetivo era criar modelos com redes neurais que fossem treinados mais rapidamente e tivessem melhor acurácia em tarefas de processamento da linguagem natural. Em sua pesquisa enfatizou que word embeddings são capazes de extrair conhecimento semântico indo além da ideia de alguns outros pesquisadores que limitavam a análise na extração sintática. É possível justificar o poder de embutir conhecimento semântico sobre vetores com a seguinte operação mostrada:

$$V(\textit{“rainha”}) \approx V(\textit{“rei”}) - V(\textit{“homem”}) + V(\textit{“mulher”})$$

Essa fórmula apresenta uma ideia em que o Embedding Space pode assimilar o contexto semântico, sendo  $V$  o vetor da palavra, pois consegue interpretar que algo próximo de rei e seja mulher, mas não seja homem, será mapeado para um vetor próximo ao de rainha.

A primeira arquitetura proposta foi o Continuous Bag-of-Words (CBOW), segundo Mikolov et. al (2013), é uma rede neural que prediz a palavra dado um contexto, sendo o contexto interpretado como uma sentença. Este modelo pode prever tanto palavras prévias como posteriores em um contexto, como exemplificado na Figura 1.

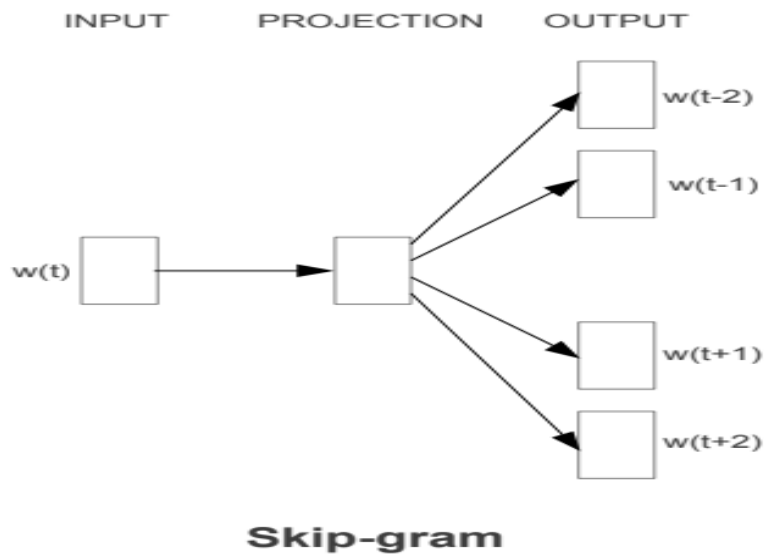
Figura 1 : Representação da arquitetura CBOW



Fonte: [6]

Já O modelo Skip-gram desempenha uma função inversa do apresentado em CBOW, ou seja, dado uma palavra prediz calculando as palavras mais prováveis ao contexto. Como as palavras mais distantes estão geralmente menos correlacionadas com a palavra atual, Mikolov dá um peso menor para essas palavras amostrando uma quantidade menor de exemplos no treinamento do modelo, priorizando as palavras mais próximas, representado na Figura 2.

Figura 2 : Representação da arquitetura Skip-gram



Fonte: [6]

## 3.2. MÉTODOS PARA PREVISÃO

### 3.2.1. ARIMA

Conforme Ediger e Akar (2007), o ARIMA (Autoregressive-integrated-moving-average) é um dos mais populares modelos para análise da previsão em séries temporais. Os modelos ARIMA são modelos que utilizam apenas dados históricos de séries temporais com o intuito de expressar como as séries reagem de acordo com a variação estocástica anterior (BABAI et al, 2013). Os modelos ARIMA podem ajudar a entender a dinâmica dos dados em uma determinada aplicação (BABU; REDDY, 2014). O modelo ARIMA originou-se a partir dos modelos autorregressivos (AR) e dos modelos médias móveis (MA) e da combinação entre AR e MA (modelo ARMA), (GUJARATI, 2000; MORETTIN; TOLLOI, 2004; EDIGER; AKAR, 2007).

Conforme Gujarati (2000) “muitas séries temporais econômicas são não-estacionárias, ou seja, são integrais”. No entanto, dada uma série temporal, se esta for integrada de ordem 1  $I[1]$ , suas primeiras diferenças serão  $I[0]$ , ou seja, demonstram-se estacionárias. Assim, em geral, se uma série temporal não estacionária, representada por  $I[d]$ , se diferenciarmos  $d$  vezes esta série, será obtida uma série estacionária  $I[0]$  (GUJARATI, 2000). Conforme o autor, didaticamente adota-se  $I[0]$  para indicar que a série é estacionaria. Logo, se para análise de uma série temporal for necessária à sua diferenciação  $d$  vezes para torná-la estacionária, diz-se que esta série temporal é ARIMA (Autorregressiva Integrada de Média Móvel). Assim, será representada por  $ARIMA(p,d,q)$ , onde  $p$  indica o número de termos autorregressivos,  $d$  o número de vezes que a série deve ser diferenciada para se tornar estacionária e,  $q$  indica o número de termos de média móvel (GUJARATI, 2000). A estrutura do modelo ARIMA é expressa por:

$$\Phi(B) \cdot (1 - B)^d \cdot \hat{X}_t = \theta(B)\varepsilon_t$$

Como o foco deste trabalho é a modelagem do retorno do índice Ibovespa é esperado que a série seja estacionária ou muito próximo disso.

### 3.2.2. SARIMA

Visando englobar séries sazonais, o modelo ARIMA é generalizado, e obtém-se, então, o modelo SARIMA (Seasonal-autoregressive-integrated-moving-averages), que adiciona três parâmetros,  $P$ ,  $D$  e  $Q$ , referentes à, respectivamente, autoregressão,

diferenciação e média móvel da componente sazonal. Logo, as equações referentes ao modelo SARIMA com sazonalidade multiplicativa são:

$$\begin{aligned}\theta(B^s) &= 1 - \theta_1 B - \theta_2 B^{2s} - \dots - \theta_Q B^{Qs} \\ \phi(B^s) &= 1 - \phi_1 B - \phi_2 B^{2s} - \dots - \phi_p B^{Ps} \\ \Phi(B^s) \cdot \phi(B) \cdot (1 - B)^d \cdot (1 - B^s)^D \cdot \hat{X}_t &= \theta(B^s) \cdot \theta(B) \cdot \varepsilon_t\end{aligned}$$

Onde  $s$  é o período sazonal e  $\Phi$  e  $\theta$  são os coeficientes de autorregressão e média móvel, respectivamente (Ehlers, 2007). Um modelo SARIMA é, analogamente a um modelo ARIMA, denotado como SARIMA(p,d,q)(P,D,Q,s).

### 3.2.3. SARIMAX

Há diversos métodos desenvolvidos na literatura para inserção de variáveis exógenas na análise e modelagem de séries temporais, sendo os mais consagrados o método de regressão, redes neurais artificiais, SARIMAX e ARIMAX, respectivamente (Maçaira et al., 2018). Como este trabalho tem como objetivo verificar se a inclusão de features textuais das notícias melhora a previsão dos ativos, os modelos que permitem features exógenas serão os modelos de interesse e os que utilizam apenas dados da série temporal serão os modelos de referência ou benchmarks.

Modelos SARIMAX (ou ARIMAX) é um modelo SARIMA (ou ARIMA) com adição de uma ou mais variáveis exógenas. Evidenciando  $X_t$  na equação do modelo SARIMA, e adicionando os termos relacionados à variável exógena  $y$ , obtém-se a seguinte equação que descreve um modelo SARIMAX.

$$\hat{X}_t = f(y) + \frac{\theta(B^s)\theta(B)\varepsilon_t}{\phi(B)\phi(B)(1 - B)^d(1 - B^s)^D}$$

### 3.2.4. DTSF

Conforme Costa et al. (2021), o modelo Dynamic Time Scan Forecasting (DTSF) também é um método para previsão em séries temporais no qual um procedimento de varredura é aplicado para identificar padrões, nomeados como melhores correspondências, ao longo da série temporal.

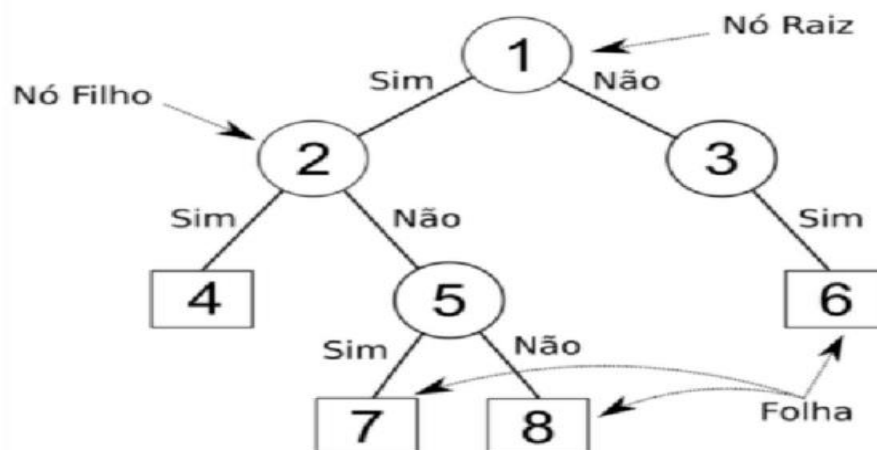
O método de previsão por varredura baseia-se na premissa de que o padrão mais importante em uma série temporal precede a janela de previsão, ou seja, os últimos valores observados. Assim, um procedimento de varredura é aplicado para identificar padrões semelhantes, ou melhores correspondências, ao longo da série temporal. Ao contrário da distância euclidiana, ou de qualquer função de distância, uma função de similaridade é estimada dinamicamente para combinar os valores anteriores com os últimos valores observados. Estatísticas de adequação são usadas para encontrar as melhores correspondências. Utilizando as respectivas funções de similaridade, os valores observados procedentes das melhores correspondências são utilizados para criar um padrão de previsão, bem como intervalos de previsão. O método proposto em Costa et al (2021) superou as abordagens estatísticas e de aprendizado de máquina em um problema real de previsão da velocidade do vento e já foi explorado em outros problemas de previsão de energia, tal qual Hu e Man (2023). Mas sua aplicação na previsão do retorno de investimentos ainda parece pouco explorada.

### 3.2.5. CATBOOST

A maioria dos algoritmos de boosting utilizam árvores de decisão como aprendizes básicos. Uma árvore é uma coleção de elementos chamados de nós, dentre os quais um é distinguido como uma raiz, juntamente com uma relação de “paternidade” que impõe uma estrutura hierárquica sobre os nós.

A ideia do algoritmo é escolher as divisões internas da árvore que melhor explicam os dados e é, em si, uma pequena árvore. Na Figura 3 é mostrado um exemplo de uma árvore de decisão treinada Sato et al. (2013).

Figura 3 : Exemplo de árvore de decisão



Fonte: SATO, et al. (2013)

Boosting é uma das ideias de aprendizado mais poderosas introduzidas nas últimas décadas. Ele foi originalmente projetado para problemas de classificação, mas pode ser utilizado para produzir um “comitê” poderoso, ou seja, criar um aprendiz "poderoso" a partir de uma combinação de aprendizes "fracos"(HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Em resumo, o CatBoost se encaixa nessa categoria de algoritmos de boosting, pois ele constrói modelos de árvores de decisão de forma sequencial, onde cada árvore tenta corrigir os erros das árvores anteriores. O resultado é um modelo de ensemble, que geralmente tem um desempenho superior em relação a modelos individuais.

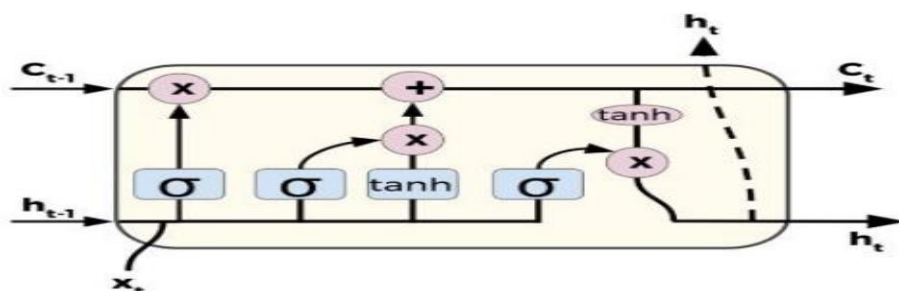
### 3.2.6. LSTM

Redes neurais LSTM (Long Short-Term Memory) são um tipo de rede neural recorrente (RNN) usada principalmente para processar e prever dados sequenciais, como séries temporais, texto ou áudio. As LSTM surgiram para aumentar a capacidade de memória de uma rede neural. Chen (2017) divide em três as etapas de confecção da arquitetura de uma rede neural LSTM:

- 1) Utilização de um mecanismo de esquecimento: mecanismo que interpreta se um dado de entrada deve ser de fato utilizado ou descartado.
- 2) Utilização de um mecanismo de salvamento: mecanismo responsável por salvar as partes úteis dos dados de entrada na memória a longo prazo.
- 3) Transformar memória de longo prazo em memória útil: o modelo deve aprender quais partes da memória a longo prazo devem ser utilizados a cada dado instante.

A Figura 4, ilustra esses mecanismos em um neurônio de uma rede neural LSTM:

Figura 4 : Estruturas de um neurônio de uma rede neural recorrente LSTM.



Fonte: Portilla (2019).

Onde  $C_t$  representa a memória de longo prazo (ou estado) do modelo,  $x_t$  são os dados de entrada e  $h_t$  os dados de saída no instante  $t$ . Ainda, no modelo LSTM, a função sigmoide ( $\sigma$ ) retorna valores de 0 a 1, e atua como um filtro que define a passagem ou não do sinal. A função tangente hiperbólica ( $\tanh$ ) atua na criação de novos candidatos para a memória de longo prazo, variando-os de -1 a 1. Ambas as definições para este caso seguem descritas pelas equações a seguir (Castelão, 2018).

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Castelão (2018) e Portilla (2019) descrevem as funções de cada um dos portões em uma rede neural LSTM.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

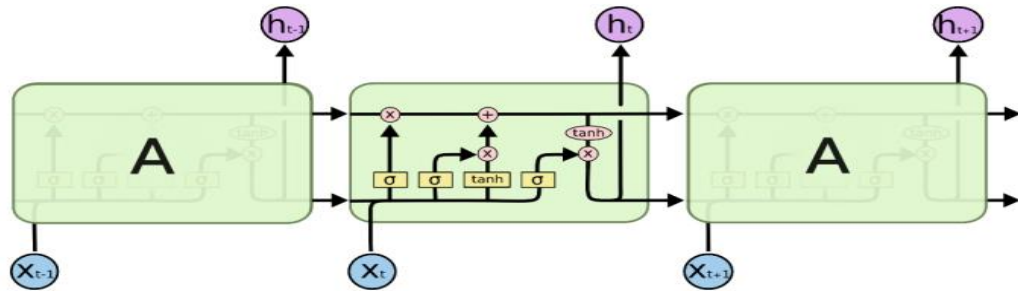
$$C_t = f_t C_{t-1} + i_t (\tanh(W_C[h_{t-1}, x_t] + b_C))$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \tanh(C_t)$$

Onde  $f_t$  é o portão de esquecimento (forget gate),  $i_t$  é o portão de entrada (input gate) e  $o_t$  é o portão de saída (output gate).  $C_t$  é a memória a longo prazo, ou estado (cell state). Os parâmetros de ajuste  $W$  e  $b$  são, respectivamente os pesos e os ajustes por viés que ocorrem dentro da célula da rede neural. Replicando essa estrutura para todas as unidades da rede neural, chega-se na arquitetura de uma rede neural recorrente LSTM, ilustrada pela Figura 5.

Figura 5 : Arquitetura de uma rede neural recorrente LSTM.



Fonte: Olah (2015).

Em resumo, devido essa complexa estrutura, na qual a informação flui entre diversas camadas dentro da célula, a rede neural LSTM é uma ferramenta poderosa para trabalhar com dados sequenciais, permitindo que o modelo "lembra" informações importantes por longos períodos.

### 3.3. ANÁLISE E MÉTRICAS DE AVALIAÇÃO

O desempenho dos modelos é avaliado inicialmente com métricas de regressão visto que, a variável de interesse é uma variável contínua que representa o delta do retorno entre Ibovespa e CDI semana. Segundamente os modelos são avaliados utilizando métricas de classificação pois em um cenário prático o interesse real é que o modelo acerte a alocação entre o Ibovespa ou o CDI na semana seguinte.

**MAE (Mean Absolute Error):** MAE é a média dos erros absolutos entre as previsões do modelo e os valores reais. Em termos simples, é a média das diferenças absolutas entre o valor predito e o valor observado.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i|$$

**RMSE (Root Mean Square Error):** RMSE é a raiz quadrada da média dos erros quadrados entre as previsões do modelo e os valores reais, é frequentemente utilizado em problemas de regressão quando se quer penalizar grandes erros mais do que pequenos. Como ele calcula a média dos erros quadrados antes de tirar a raiz quadrada, erros maiores têm um impacto maior na métrica. Isso torna o RMSE apropriado quando erros maiores são mais críticos no contexto da aplicação.

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

**Acurácia:** Acurácia é a proporção de previsões corretas (tanto positivas quanto negativas) feitas pelo modelo em relação ao total de previsões, é muito usada em problemas de classificação, especialmente quando as classes estão balanceadas.

$$Acurácia = \frac{VP + VN}{total\ de\ predições}$$

**Precisão:** Precisão é a proporção de verdadeiros positivos sobre todas as previsões positivas feitas pelo modelo. O uso dessa métrica é bem difundido quando o custo de falsos positivos é alto, ou seja, quando é importante que, ao prever uma instância positiva, o modelo esteja, de fato, correto. Portanto, é a métrica de maior interesse nesse trabalho.

$$Precisão = \frac{VP}{VP + FP}$$

No contexto desse artigo o verdadeiro positivo seria equivalente a alocar em IBOV dado que performou melhor CDI e o falso positivo seria alocar em IBOV em uma semana que o CDI performou melhor.

## 4. APLICAÇÃO

### 4.1. PRÉ-PROCESSAMENTO

Antes de iniciar a aplicação dos métodos de processamento textual nas notícias das foram necessárias algumas alterações no banco de dados a fim de facilitar ou evitar vieses nos resultados.

#### 4.1.1. Remoção Caracteres Especiais e Normalização

Remoção de caracteres especiais tais como pontuações, símbolos matemáticos e numéricos, pois não existindo um tratamento adequado para estes caracteres, são criados valores incompatíveis na representação vetorial para o modelo.

Esse processo diminui o número de palavras distintas do conjunto. Usando as funções de misspelling e lowercase na normalização, a lista “pouco”, “Pouco” e “poUco” é convergida para a mesma palavra “pouco”.

#### 4.1.2. Tokenizer

O Processo visa transformar textos em tokens. Na análise sentimento normalmente é considerado um token como sendo uma palavra, porém existem variações para definir o seu tamanho. Sendo uma estratégia simples separar palavras por espaços em brancos ou pontuações sendo os delimitadores, nesse projeto a separação utilizou os espaços em brancos dos textos.

#### 4.1.3. Remoção de Stop words

Gary Miner (2012), as stop words são palavras muito comuns na linguagem como exemplo para a língua inglesa o artigo “an” e a preposição “on” são necessárias nas montagens de frases. Assim como as palavras flexionadas a lista de stop words varia pela estrutura morfológica do idioma. Foram removidos os tokens com uma lista contendo as stop words a serem removidas, pois não possuem um forte valor explicativo.

### 4.2. ANÁLISE DESCRITIVA

#### 4.2.1. Aplicação dos Métodos de Processamento Textual

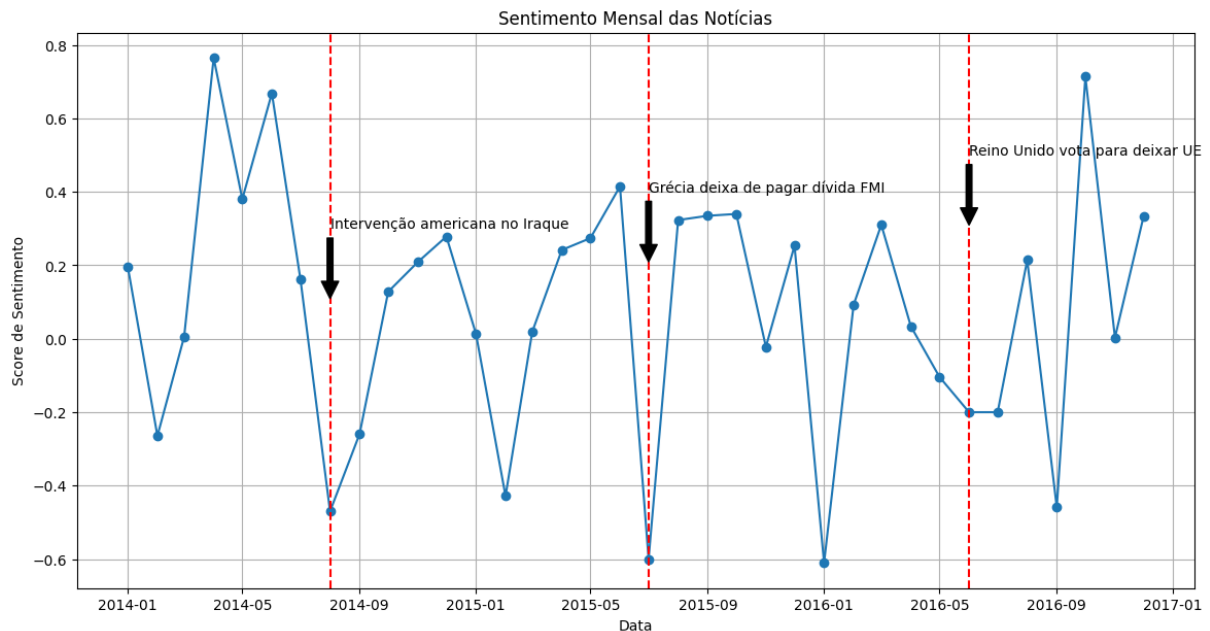
Após o tratamento na base de notícias sobre macroeconomia e geopolítica extraídas do jornal The Diplomat, foi aplicado a metodologia Bag-of-Words com o pacote NLTK Natural Language Toolkit (Edward Loper et al, 2009) em Python para a

criação de um índice de sentimento das notícias, sendo essa a primeira variável explicativa criada para a previsão dos retornos dos investimentos.

A metodologia desse pacote proposta em Gilbert, E.E. (2014) utiliza um mapeamento no qual cada palavra do vocabulário é avaliada quanto a ser positiva ou negativa e uma métrica de sentimento normalizada na faixa de -1 a 1 é calculada com base em todas as palavras representadas no Bag-of-Words.

Aplicando essa metodologia no histórico de notícias é possível identificar como o índice de sentimento se comportou em momentos importantes do cenário geopolítico, ou seja, quando ocorreram grandes quedas nos mercados.

Figura 6 : Índice de sentimento de 2014 até 2017



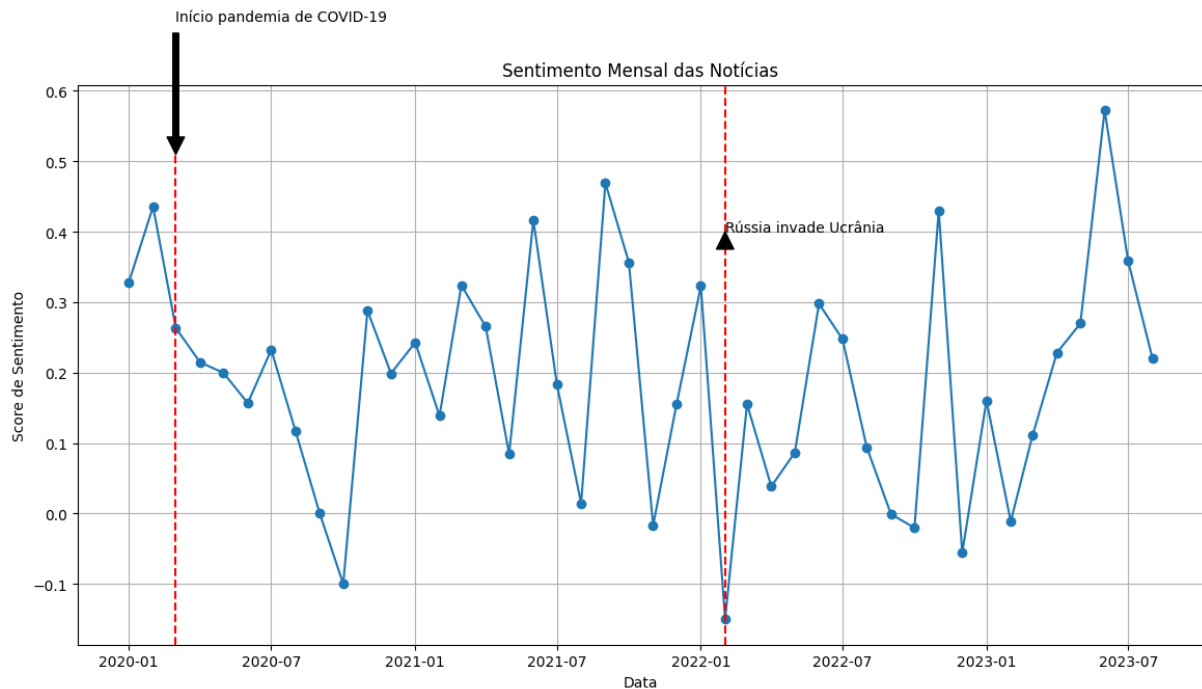
Fonte: Elaborado pelo Autor (2024)

Na figura 6 é possível ver como a feature criada se comportou em momentos de tensão, por exemplo em 2014 com a intervenção americana no Iraque que gerou grande instabilidade no Oriente Médio, resultando em volatilidade nos preços das ações e commodities principalmente do petróleo.

Além disso, o índice atingiu seu ponto mais negativo em 2015 no mês em que a Grécia deixou de pagar uma parcela de sua dívida ao Fundo Monetário Internacional (FMI), tornando-se o primeiro país desenvolvido a dar um calote no FMI. Esse evento exacerbou a crise financeira na zona do euro, causando uma queda significativa nas bolsas de valores globais.

Em contrapartida em 2016 com a votação do Reino Unido para deixar a União Europeia não houve uma queda tão acentuada no índice de sentimento, embora esse evento tenha resultado em forte choques cambiais como a libra esterlina caindo drasticamente em relação ao dólar e com investidores buscando títulos mais seguros em detrimento ao investimento em bolsas.

Figura 7 : Índice de sentimento de 2020 até 2023



Fonte: Elaborado pelo Autor (2024)

Já na Figura 7, em um período mais recente, observa-se a queda do índice de sentimento em 2020 após o descobrimento do Covid-19 e uma queda nos meses consecutivos com o início da pandemia e o agravamento do isolamento com lockdowns. A disseminação da Covid-19 em 2020 teve um impacto profundo nos mercados financeiros globais.

Por fim, o último evento marcante foi o início da guerra entre Rússia e Ucrânia que também demonstrou ter tido um impacto no índice construído.

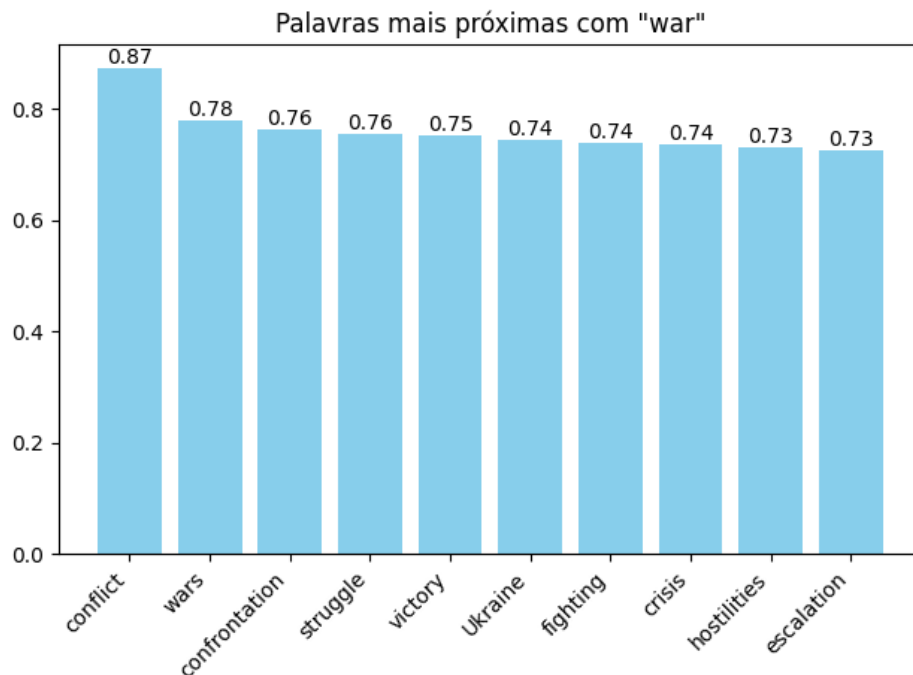
Esses eventos ilustram como crises políticas e econômicas podem gerar incertezas nos mercados financeiros, afetando desde ações até moedas e commodities, e como os investidores tendem a reagir, muitas vezes buscando proteção em ativos menos arriscados, dessa forma, espera-se que esse monitoramento das notícias globais por meio do PLN possa ser útil na previsão de mercado.

#### 4.2.2. Aplicação Word2Vec

A fim de construir mais features que poderiam ser úteis para os modelos utilizou-se o pacote Gensim em Python com implementações de Word2Vec. Nessa etapa a tentativa de treinar um modelo com os dados extraídos encontrou limitações pois seria necessário um grande volume de dados com muita variedade e representatividade, além do tempo e recursos computacionais necessários. Por esse motivo optou-se pela utilização de uma rede neural já treinada do Google visto que, já contaria com toda a parte de configuração do modelo Word2Vec, como ajuste de hiperparâmetros, tamanho da janela de contexto, a dimensionalidade dos vetores, etc.

Dessa forma, o passo necessário foi baixar e carregar esse modelo de Word2Vec treinado e aplicá-lo na base de notícias com o pacote Gensim. Feita a representação vetorial das palavras é possível calcular medidas de similaridade entre elas como mostrado na Figura 8.

Figura 8 : Aplicação do modelo De Word2Vec



Fonte: Elaborado pelo Autor (2024)

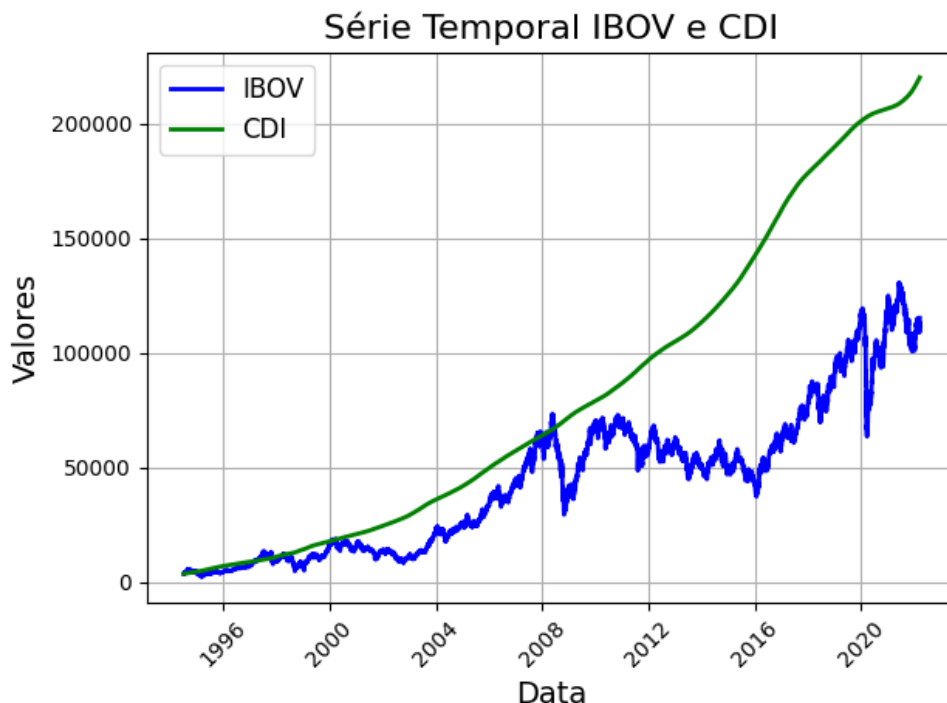
O modelo representa cada palavra em um espaço vetorial de 300 dimensões e a Figura 8 é a representação das palavras mais próximas em relação a palavra war, como o interesse é na representação vetorial da notícia como um todo e não em suas palavras individualmente foi feita a média de todas as palavras presentes em

cada artigo para se obter uma representação também em 300 dimensões de cada notícia, portanto é possível obter mais 300 features textuais para auxiliar na previsão dos retornos.

#### 4.2.3. Análise Retornos Dos Investimentos

Uma vez aplicado os métodos de PLN na base de notícias, procura-se, através da análise descritiva na base com os ativos, constatar características gerais das variáveis que serão importantes para a construção de modelos. Mais especificamente, características como o retorno acumulado dos dois ativos escolhidos Ibovespa e CDI.

Figura 9 : Retorno Acumulado IBOV x CDI

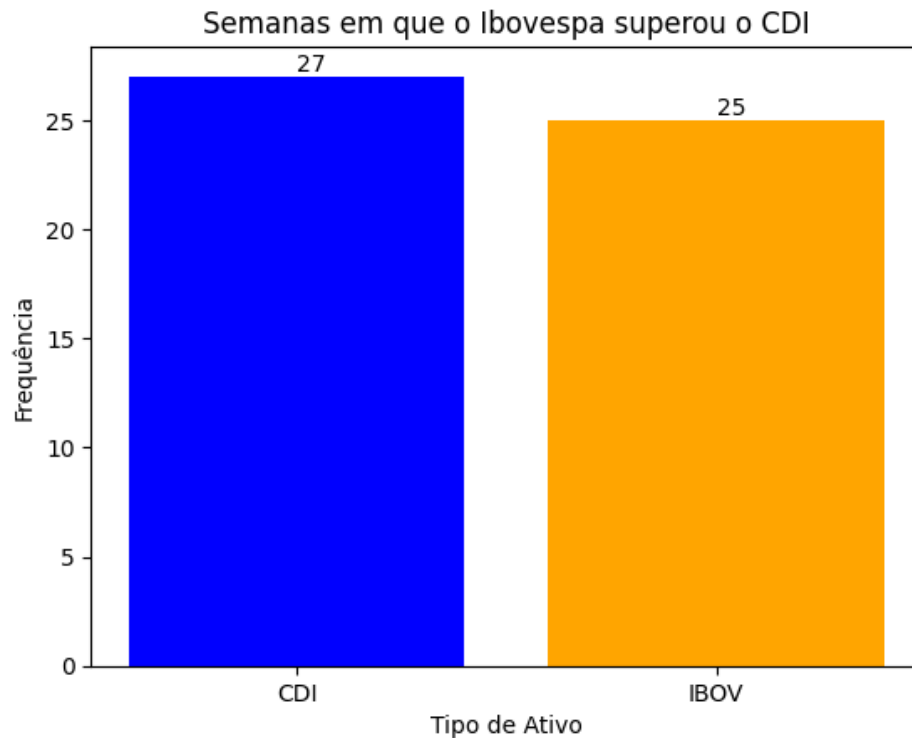


Fonte: Elaborado pelo Autor (2024)

Observa-se, na Figura 9, que embora o CDI seja um ativo com o risco menor ele apresenta um retorno acumulado muito acima do Ibovespa. Contudo, é conhecido na teoria financeira clássica que deveria existir uma relação positiva entre risco e retorno. Apesar disso, observa-se recorrentemente que na prática tal relação é, no mínimo, neutra, e, em muitas vezes, até negativa: ativos menos arriscados colocam retorno realizado médio maior. Um exemplo bastante familiar disso é o prêmio de risco do mercado acionário brasileiro. Tal anomalia é abordada por Yu e Yuan (2011) onde apresentam uma das explicações mais aceitas: a relação risco/retorno é positiva exceto quando há presença de sentimento excessivo modificando o comportamento

racional do investidor. O trabalho mais recente de Ung, Gebka e Anderson (2023) também estendeu essas as análises empíricas, empregando diversos sentimentos diferentes, para concluir que de fato o comportamento irracional dos investidores pode explicar esse dilema do risco/retorno.

Figura 10 : Semanas com retorno acumulado IBOV > CDI



Fonte: Elaborado pelo Autor (2024)

Analisando os dados de 2023, ilustrados na Figura 10, último ano com dados completos no momento de realização do trabalho, observa-se que das 52 semanas do ano, o CDI performou melhor que o Ibovespa em 27 ocasiões. A partir disso e da Figura 9 dos retornos acumulados, conclui-se que, o pior tipo de erro para o algoritmo seria alocar em bolsa de forma errada, por isso, na sessão seguinte de resultados os modelos serão discriminados com base nas métricas de precisão, por penalizar mais a ocorrência de falsos positivos, ou seja, ficar alocado em bolsa indevidamente. Vale ressaltar que nesse caso em estudo o falso negativo não é tão grave pois representa uma alocação segura em CDI, não tendo riscos de perda pois os retornos semanais são sempre positivos nesse caso. Portanto, deseja-se obter um modelo que aloque pouco em Ibovespa e que o faça apenas quando tiver um alto nível de confiança.

## 5. RESULTADOS

Os resultados aqui apresentados foram obtidos após a transformação dos dados diários em semanais. Dessa forma, o retorno dos ativos é o retorno acumulado da semana e o sentimento em cada semana é a média do sentimento obtido de todas as notícias daquela semana.

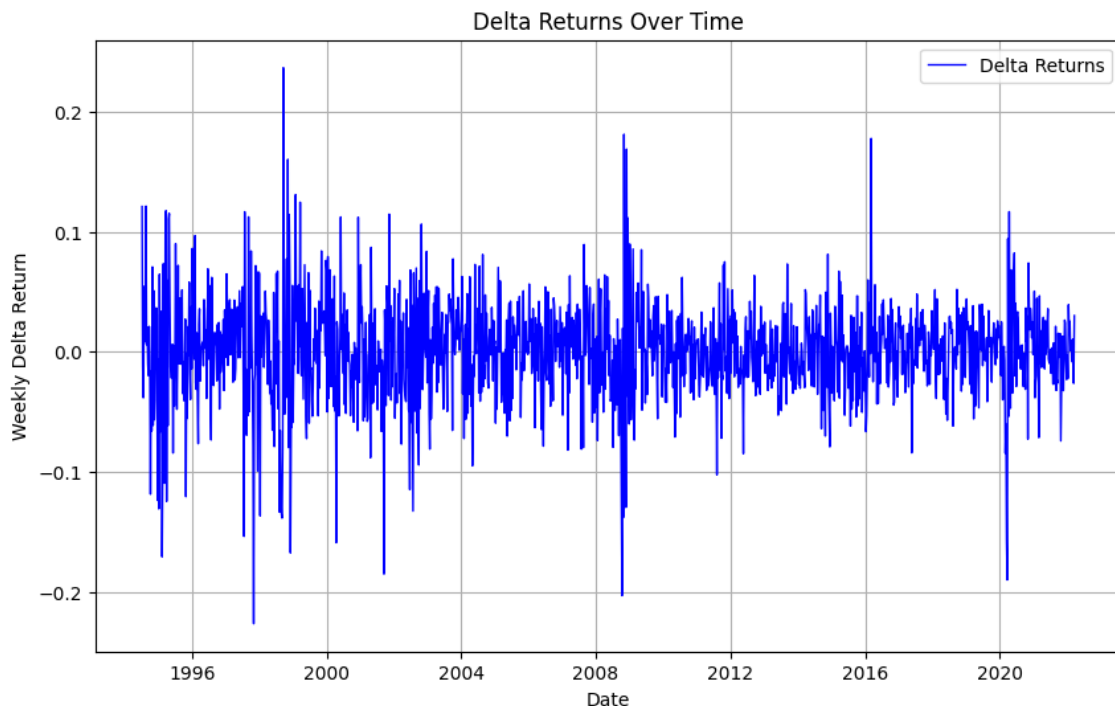
A justificativa para se utilizar dados semanais se deu pela dificuldade de explorar um modelo com dados diários pois está mais sujeito a ruídos (geopolíticos ou não), e em um cenário real, acarretaria em maiores custos de operação.

Além disso, observa-se certa indisponibilidade de dados intradiários, e uma dificuldade para operar no mercado em alta frequência devido à impossibilidade de controlar o impacto que uma notícia do meio do dia causa nos mercados. Portanto, o modelo busca operar mudanças no sentimento geopolítico de médio prazo.

O target que está sendo modelado é o delta do retorno semanal entre os ativos, ou seja:

$$\text{TARGET} = (\text{Retorno Acumulado IBOV}) - (\text{Retorno Acumulado CDI})$$

Figura 11 : Comportamento histórico do Target



Fonte: Elaborado pelo Autor (2024)

A Figura 11 mostra o comportamento histórico da variável resposta no qual é possível ver um comportamento estacionário com momentos de maior e menor volatilidade. Como variáveis explicativas foram utilizadas as 300 features provenientes do Word2Vec mais a variável de polaridade de sentimento das notícias.

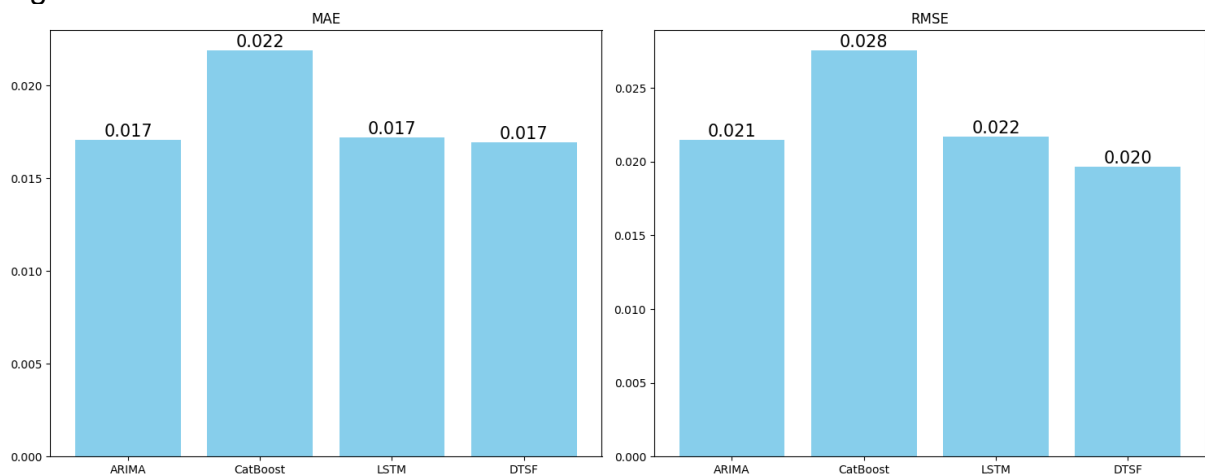
O processo modelagem utilizou um treinamento contínuo, no qual novos dados são incorporados semanalmente para retreinar o modelo e prever a semana seguinte.

Ou seja, o modelo é completamente retreinado usando todos os dados disponíveis, incluindo os novos dados. Isso garante que o modelo capture padrões que possam mudar ao longo do tempo.

A vantagens desse método é que o modelo se mantém atualizado com as últimas tendências dos dados, o que pode levar a previsões mais precisas e relevantes, principalmente em cenários onde os dados são dinâmicos e voláteis, como no mercado financeiro. O desafio é o custo computacional que pode ser caro, especialmente se os dados forem volumosos e os algoritmos lentos para serem treinados.

Inicialmente foram treinados os modelos que não utilizam variáveis explicativas que foram utilizados como referências (benchmarks).

Figura 12 : Métricas modelos sem features



Fonte: Elaborado pelo Autor (2024)

Dessa forma, com a previsão dos modelos para a semana seguinte, é definida a alocação do portfólio da seguinte maneira, se a previsão do modelo para o delta do retorno for maior do que zero a alocação é feita em Ibovespa, e se a previsão do delta do retorno for negativo a alocação é feita em CDI, portanto seguindo esses passos é possível calcular as métricas para um problema de classificação.

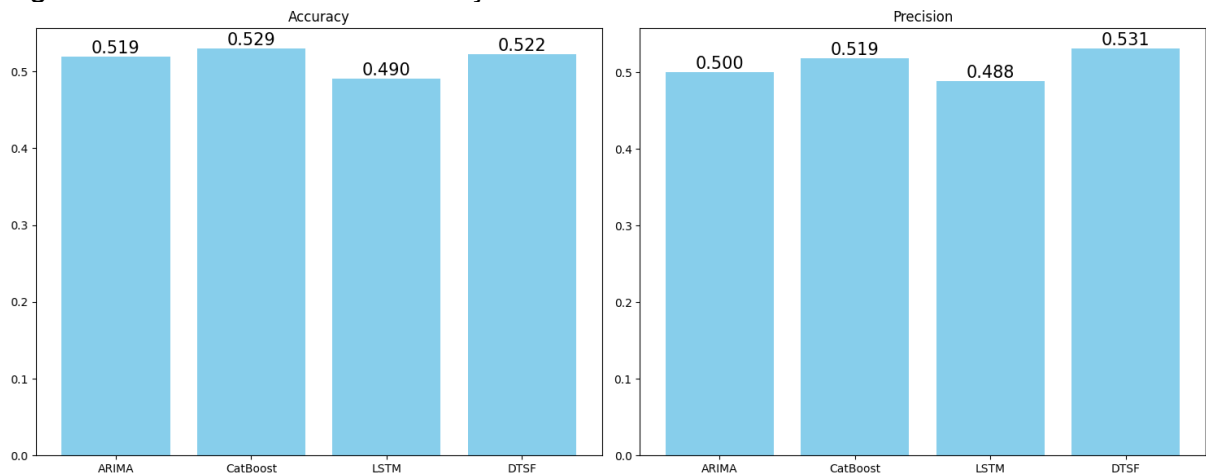
VP = Alocar em IBOV em uma semana que o IBOV superou o CDI

FP = Alocar em IBOV em uma semana que o CDI superou o IBOV

VN = Alocar em CDI em uma semana que o CDI superou o IBOV

FN = Alocar em CDI em uma semana que o IBOV superou o CDI

Figura 13 : Métricas de classificação modelos sem features

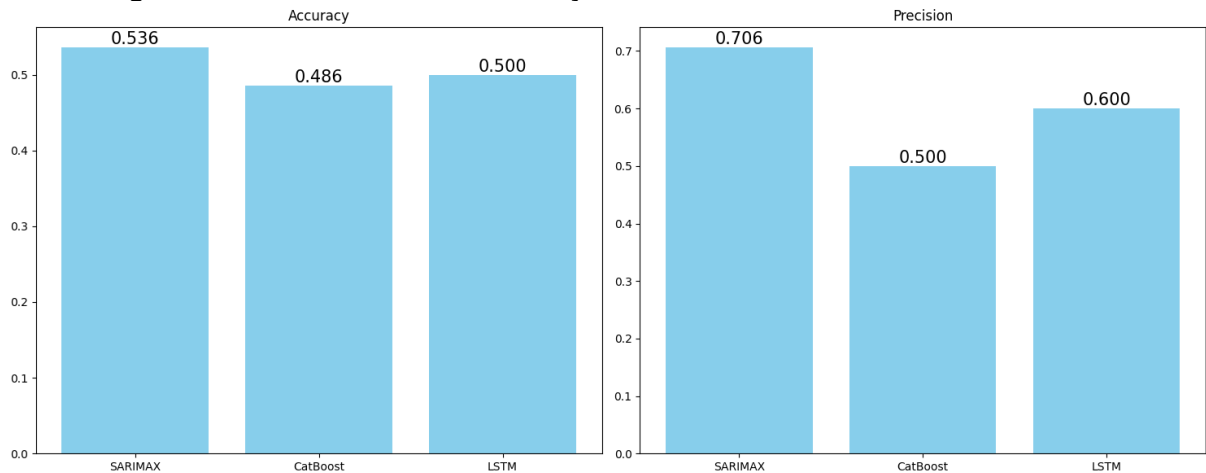


Fonte: Elaborado pelo Autor (2024)

Dentre esses Modelos benchmarks destaca-se a performance do modelo Dynamic Time Scan Forecasting (DTSF) que obteve as melhores métricas com um MAE de 0.017, RMSE de 0.020 e uma precisão de 53.1%, contudo, esse modelo não aceita a inclusão de features externas e não pode ser testado na próxima etapa, sendo um possível avanço futuro que deve ser testado. Vale ressaltar também que alguns dos modelos sem features obtiveram uma acurácia muito próxima a escolha aleatória que estaria por volta de 50% de acurácia.

Adentrando nos modelos com as features obteve-se alguns resultados mais promissores. O modelo SARIMAX, foi o que apresentou o melhor resultado dentre os modelos testados, conforme abordado no tópico 4.2.3 deste artigo esse modelo é a versão do modelo SARIMA capaz de utilizar variáveis exógenas.

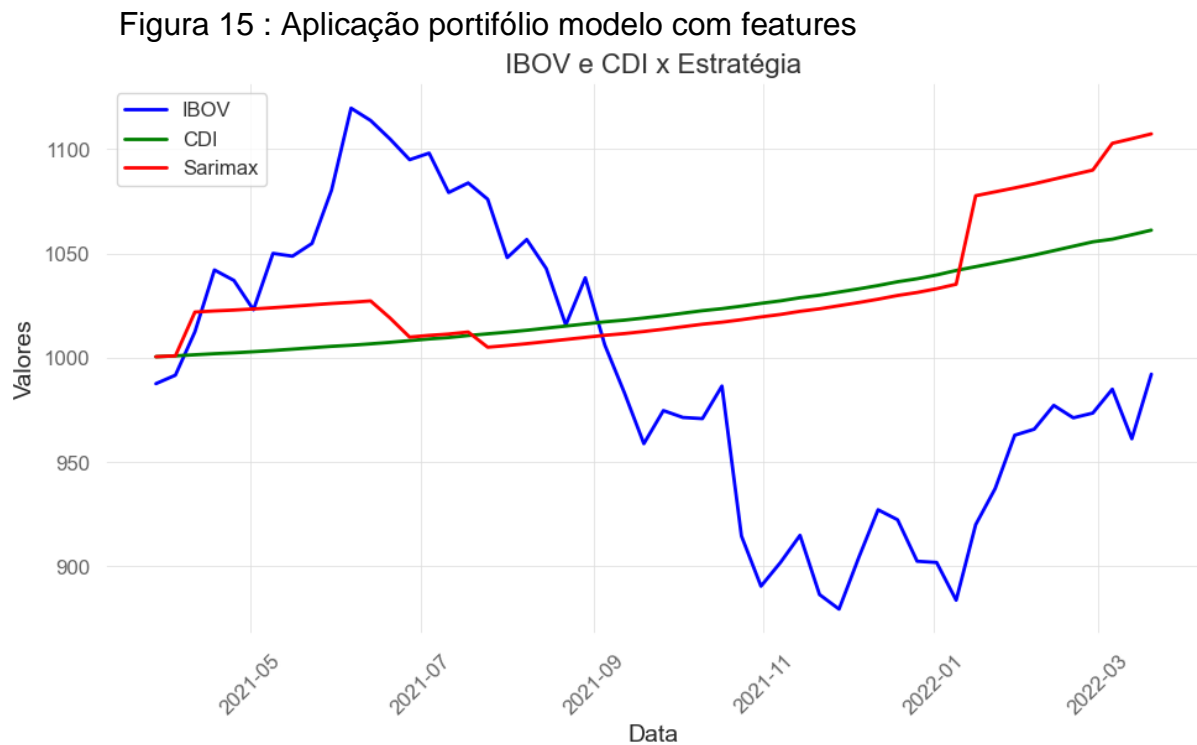
Figura 14 : Métricas de classificação modelos com features



Fonte: Elaborado pelo Autor (2024)

Observa-se uma precisão de 70,6% embora o modelo ainda tenha uma acurácia próxima dos 50%, isso se deve pela ocorrência de muitos falsos negativos, alocação em CDI em semanas que o Ibovespa performou melhor, contudo, isso não é um problema tão grave desde que o algoritmo esteja certo quando faz a alocação em bolsa, precisão alta.

Com o objetivo de avaliar como essa estratégia teria performado caso tivesse sido empregada utilizou-se a metodologia de backtest abordada por Marcos López de Prado em seu livro *Advances in Financial Machine Learning* (2018), que consiste em realizar uma simulação com os dados de teste expondo o portfólio aos retornos obtidos de acordo com as entradas e saídas feitas pelo algoritmo em cada ativo.



Fonte: Elaborado pelo Autor (2024)

Observa-se, na Figura 15, que a estratégia do SARIMAX obteve um retorno acumulado melhor que o Ibovespa e o CDI, contudo, na maior parte do tempo a estratégia ficou alocada em CDI e por esse motivo o retorno acumulado foi uma linha constante por boa parte do período.

Entretanto, o backtest deve evitar “*look-ahead bias*”, onde o modelo acidentalmente usa informações futuras que não estariam disponíveis no momento da tomada de decisão. Para isso, as janelas de previsão devem ser claramente definidas e garantidas que não haja acesso a dados do futuro durante a simulação. Em um cenário de aplicação real as notícias da semana que o algoritmo está prevendo o retorno ainda não estariam disponíveis, portanto, seria necessário treinar os modelos usando apenas as notícias da semana anterior conforme exemplificado na Tabela 3.

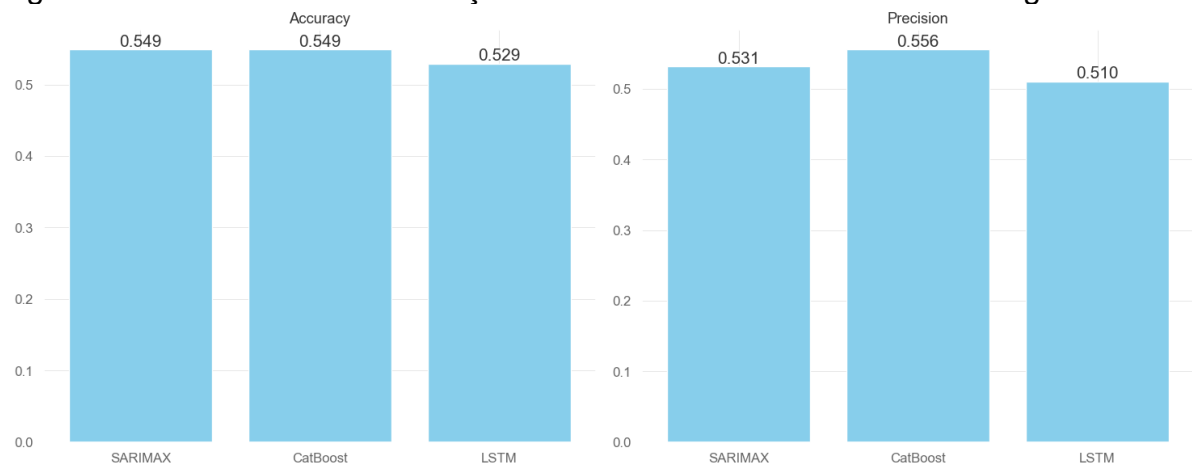
Tabela 3 : Exemplificação Data Leakage

Semana	Target	Delta Target	Features Textuais
Semana 1	0.01	-0.07	w1 .... W300
Semana 2	-0.07	0.02	w1 .... W300
Semana 3	0.02	0.017	w1 .... W300
Semana 4	0.017	-0.03	w1 .... W300
Semana 5	-0.03	-0.01	w1 .... W300
Semana 6	-0.01	Null	w1 .... W300

Fonte: Elaborado pelo Autor (2024)

Portanto, visando evitar o viés de look-ahead ou Data Leakage foi aplicado um lag na variável do target, como demonstrado na Tabela 3, para que fosse previsto o retorno da semana 6 com as informações das notícias até a semana 5 por exemplo, desse modo, obteve-se novos resultados.

Figura 16 : Métricas de classificação modelos com features e sem leakage

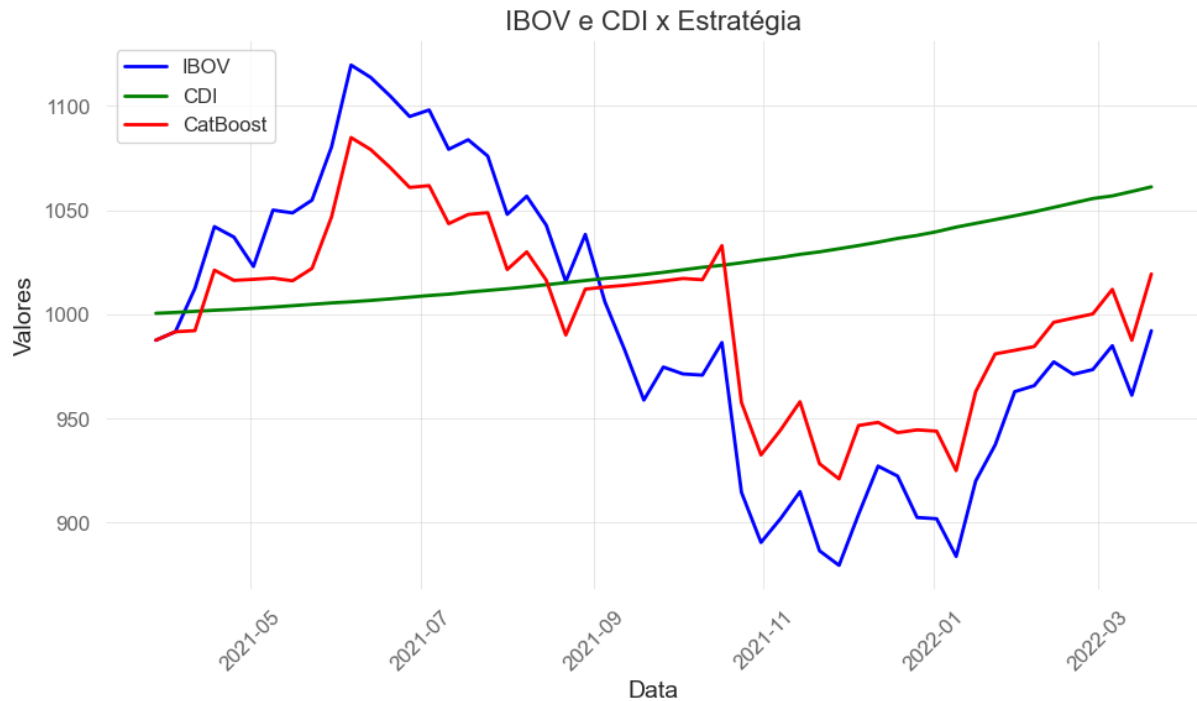


Fonte: Elaborado pelo Autor (2024)

Diferentemente do modelo oráculo o algoritmo que performou melhor com o cenário real foi o CATBOOST com uma precisão de 55.6% e uma acurácia de 54.9%, igualando em acurácia com o SARIMAX, mas com uma precisão um pouco superior. Vale destacar que o modelo LSTM não teve uma performance tão boa possivelmente

devido a falta de uma melhor otimização dos hiperparâmetros, sendo esse outro possível avanço futuro desse artigo.

Figura 17 : Aplicação portfólio modelo com features e sem Leakage



Fonte: Elaborado pelo Autor (2024)

Com a aplicação do backtest utilizando o algoritmo CATBOOST em um cenário de aplicação real o retorno acumulado do portfólio teve um retorno intermediário entre o CDI e o Ibovespa.

## 6. CONCLUSÃO E AVANÇOS FUTUROS

Este trabalho apresentou uma abordagem para a construção de uma estratégia quantitativa de investimento, integrando técnicas de ML e PNL com embasamentos na teoria financeira clássica. O portfólio proposto buscou integrar o cenário macroeconômico e risco geopolítico trazido pelas notícias na construção de uma carteira de investimentos, o que se mostrou promissor para otimizar a alocação de ativos.

Os resultados mostraram que o modelo baseado em notícias da semana foi capaz de superar o CDI, enquanto o modelo real apresentou um desempenho intermediário entre o CDI e o Ibovespa. Esses achados reforçam paralelamente o potencial e os desafios das técnicas de ML para aprimorar as estratégias de investimento, visto a complexidade e o dinamismo do mercado financeiro.

Visando aprimorar os resultados do portfólio diversos avanços futuros, podem ser explorados como a melhoria do Web Scraping, visando a integração de outras fontes de notícias visto que apenas artigos do jornal The Diplomat foram utilizados neste artigo. A inclusão de outras fontes de notícias pode fornecer uma visão mais abrangente do cenário geopolítico global.

Além disso, não foram testados a utilização de outros ativos além do CDI e Ibovespa, contudo espera-se que a operação em outros tipos de investimento, como moedas e commodities, pode ajudar os modelos a encontrarem alguma relação mais significativa além de diversificar mais a estratégia de investimento.

## 7. REFERÊNCIAS

TVERSKY, Amos; KAHNEMAN, Daniel. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* 1974.

CALDARA, Dario; IACOVIELLO, Matteo. Measuring geopolitical risk. *American Economic Review* 2022.

AHMED, Shaker; HASAN, Mostafa M.; KAMAL, Md Rajib. Russia–Ukraine crisis: The effects on the European stock market. **European Financial Management**, 2023.

FANG, Yi; SHAO, Zhiqun. The Russia-Ukraine conflict and volatility risk of commodity markets. *Finance Research Letters*, , 2022.

GÉRON, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.", 2022.

DE PRADO, Marcos Lopez. *Advances in financial machine learning*. John Wiley & Sons, 2018.

PYTHON SOFTWARE FOUNDATION. Python Language Site. Página de documentação. Disponível em: <<https://www.python.org/doc/>>. Acesso em: abr. de 2024.

RICHARDSON, L. Beautiful Soup Documentation. Version 4.9.3. Available at: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Accessed: abr. de 2024.

JURAFSKY, Daniel. *Speech and Language Processing*. 2000.

MIKOLOV, Tomas. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

EDIGER, Volkan Ş.; AKAR, Sertac. ARIMA forecasting of primary energy demand by fuel in Turkey. *Energy policy*, 2007.

COSTA, Marcelo Azevedo et al. Dynamic time scan forecasting for multi-step wind speed prediction. *Renewable Energy* 2021.

PROKHORENKOVA, Liudmila et al. CatBoost: unbiased boosting with categorical features. **Advances in neural information processing systems**, 2018.

GOLDBERG, Yoav. Neural network methods for natural language processing. Springer Nature, 2022.

SILGE, J. Text Mining with R: A Tidy Approach. O'Reilly Media/Sebastopol, 2017.

BABAI, Mohamed Zied et al. Forecasting and inventory performance in a two-stage supply chain with ARIMA (0, 1, 1) demand: Theory and empirical analysis. *International Journal of Production Economics* 2013.

GUJARATI, D. N. *Econometria Básica* Tradução: Ernesto Yoshida. São Paulo: Pearson Makron Books, 2000.

EHLERS, Ricardo S. Análise de séries temporais. Laboratório de Estatística e Geoinformação. Universidade Federal do Paraná 2007.

MAÇAIRA, Paula Medina et al. Time series analysis with explanatory variables: A systematic literature review. *Environmental modelling & software* 2018.

HU, Yusha; MAN, Yi. Energy consumption and carbon emissions forecasting for industrial processes: Status, challenges and perspectives. *Renewable and Sustainable Energy Reviews* 2023.

SATO, Luciane Yumie et al. Análise comparativa de algoritmos de árvore de decisão do sistema WEKA para classificação do uso e cobertura da terra. XVI Simpósio Brasileiro de Sensoriamento Remoto 2013.

HASTIE, Trevor et al. The elements of statistical learning: data mining, inference, and prediction. New York: springer, 2009.

CHEN, Edwin. Exploring Istms. 2017.

OLAH, Christopher et al. Understanding lstm networks. 2015.

MINER, Gary. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Academic Press, 2012.

BIRD, Steven; LOPER, Edward and KLEIN, Ewan (2009), Natural Language Processing with Python. O'Reilly Media Inc.

HUTTO, Clayton; GILBERT, Eric. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI conference on web and social media. 2014.

YU, Jianfeng; YUAN, Yu. Investor sentiment and the mean–variance relation. Journal of Financial Economics, 2011.

UNG, Sze Nie; GEBKA, Bartosz; ANDERSON, Robert DJ. Is sentiment the solution to the risk–return puzzle? A (cautionary) note. Journal of Behavioral and Experimental Finance, 2023.