UNIVERSIDADE PRESBITERIANA MACKENZIE

PROJETO APLICADO II – CURSO CIÊNCIA DE DADOS

TURMA 201825166.000.03A – GRUPO 19
GUILHERME AUGUSTO LEAL OLIVEIRA
GUILHERME ROCHA DE SOUZA DUARTE
GUILHERME SANTOS OLIVEIRA
RICARDO ZULIAN DE SOUZA AMARAL

ESTUDO DE ATAS DO COPOM COMO FERRAMENTA PREDITORA

São Paulo 2025

TURMA 201825166.000.03A – GRUPO 19

GUILHERME AUGUSTO LEAL OLIVEIRA
GUILHERME ROCHA DE SOUZA DUARTE
GUILHERME SANTOS OLIVEIRA
RICARDO ZULIAN DE SOUZA AMARAL

ESTUDO DE ATAS DO COPOM COMO FERRAMENTA PREDITORA

Projeto aplicado apresentado à Universidade Presbiteriana Mackenzie como requisito parcial para conclusão da disciplina Projeto Aplicado II.

Orientador: Professor Felipe Albino dos Santos

São Paulo 2025

1 – SUMÁRIO

2 – TABELAS, QUADROS E FIGURAS

2.1- QUADROS

2.2– FIGURAS

3 – TERMOS CHAVE

BACEN, COPOM, SELIC, INFLAÇÃO.

4 – GLOSSÁRIO

BACEN – Abreviatura utilizada no mercado financeiro pra Banco Central do Brasil, órgão federal

responsável pela política monetária nacional.

COPOM – Comitê de política monetária. Formado pelos diretores do BACEN, se reúne regularmente para definir os parâmetros da política monetária nacional.

SELIC – A taxa de juros básica da economia nacional.

INFLAÇÃO – Medida estatística da variação de preços de produtos e serviços no mercado nacional.

5- RECURSOS EXTERNOS

Os documentos e o código desenvolvidos para a realização deste estudo podem ser encontrados no Github.

Segue o repositório: https://github.com/guilhermersduarte/Projeto-Aplicado2-Grupo19

6- INTRODUÇÃO

Este projeto que tem por objetivo compilar e processar os dados produzidos pelas reuniões periódicas do COPOM e através de técnicas de processamento de linguagem natural criar indicadores para provisão de comportamento de juros e inflação.

A empresa escolhida como usuária da informação é o Banco do Brasil. Antecipar corretamente o comportamento da inflação e das taxas de juros tem grande valor nas operações no mercado financeiro e no gerenciamento de carteiras de crédito.

Ferramentas em Python e bibliotecas de processamento de linguagem natural serão utilizadas para capturar, processar e analisar os produtos do COPOM e dados históricos de taxa de juros e inflação.

7- A EMPRESA

O Banco do Brasil (BB) foi fundado em 1808 e é o primeiro banco a operar no Brasil e um dos primeiros da américa latina. Criado por Dom João VI, incialmente operava como banco emissor e financiador do governo. Hoje é um dos maiores bancos no país, sendo uma empresa de economia mista, com o governo federal sendo seu maior acionista.

O Banco do Brasil é reconhecido pela excelência na gestão de recursos financeiros e oferta de produtos de investimento. A tesouraria atua na administração de liquidez, estão de riscos financeiros e negociação de ativos no mercado financeiro. Sua carteira de crédito é uma das maiores no país, com destaque ao financiamento do agronegócio. O BB também to uma gestora de recursos de terceiros, a BB Asset Managemente, com fundos de renda fixa, ações e multimercado.

8 – O COPOM

O COPOM se reúne a cada 45 dias para definir a taxa básica de juros vigente no Brasil pelos próximos 45 dias.

Na reunião se define a taxa SELIC, que a taxa de juros média praticada nos títulos da dívida federal de um dia útil. Influencia todas as outras taxas de juros praticadas no Brasil, de Títulos Federais de longo prazo ao rotativo do cartão de crédito.

A taxa de juros é a ferramenta utilizada no controle da inflação. Uma inflação elevada indica uma economia aquecida e excesso de dinheiro em circulação. Nesse cenário o COPOM eleva a taxa de juros, tornando o crédito mais caro e a poupança mais interessante, reduzindo assim a quantidade de dinheiro em circulação e consequentemente a evolução dos preços. Uma inflação em queda acentuada ou até uma deflação (queda de preços) indicam economia estagnada e falta de dinheiro. Uma redução da taxa de juros busca aumentar a quantidade de dinheiro em circulação e assim aquecer a economia.

Cada reunião tem três produtos:

-A **taxa** propriamente dita, expressada em percentual (na data da produção deste documento a taxa selic era de 15% ao ano);

-Um **comunicado**, publicado ao final da reunião. O comunicado tem um título informativo – já apresenta a decisão no próprio título – e uma justificativa curta, de uma lauda, justificando essa decisão. Apresenta também o resultado da votação do comitê, destacando os membros favoráveis e contrários à decisão.

-Uma **ata**, publicada uma semana depois da decisão. Documento mais extenso, detalhando os fatores o processo observado na tomada da decisão.

9 – OBJETIVO

As atas e comunicados tem um formato estrito e a linguagem utilizada é recorrente. Pequenas nuances em terminologia indicam intenções e expectativas. Durante o desenvolvimento dos sistemas da Biruta serão exploradas duas possibilidades envolvendo o processamento de linguagem natural:

9.1 Produção de resumos ou sumários

Utilizar técnicas de processamento de linguagem natural para agilmente apresentar versões concisas e com terminologia acessível dos documentos produzidos pelo COPOM.

9.2 Produção de indicadores

Rotulando cada decisão com o comportamento da economia (inflação) e do próprio COPOM (taxa de juros) em períodos subsequentes desenvolver um modelo de classificação e previsão dessas variáveis em prazos a serem estipulados.

10 – A BASE DE DADOS

As bases primárias de dados são obtidas diretamente dos sistemas do BACEN via API. Aqui os 4 pontos de consulta e a informação fornecida:

- Lista de atas do COPOM
https://www.bcb.gov.br/api/servico/sitebcb/copom/**atas**

Tem registros a partir de 28/01/1998

Nome do Campo	Tipo do Campo	Descrição	Exemplo
nroReuniao	integer	255	255
dataReferencia	string<date>	Data do último dia da reunião do Copom, com formato conforme especificação RFC-3339 (ex:	2023-06-21

		2023-06-21)	
dataPublicacao	string<date>	Data do último dia da reunião do Copom, com formato conforme especificação RFC-3339 (ex: 2023-06-21)	2023-06-21
titulo	string	Título da publicação	255ª Reunião – 20-21 junho, 2023

Quadro 1 – Lista de Atas do COPOM

- Detalhes sobre uma ata do COPOM
https://www.bcb.gov.br/api/servico/sitebcb/copom/**atas_detalhes**

Tem registros a partir de 28/01/1998

Nome do Campo	Tipo do Campo	Descrição	Exemplo
nroReuniao	integer	255	255
dataReferencia	string<date>	Data do último dia da reunião do Copom, com formato conforme especificação RFC-3339 (ex: 2023-06-21)	2023-06-21
dataPublicacao	string<date>	Data do último dia da reunião do Copom, com formato conforme especificação RFC-3339 (ex: 2023-06-21)	2023-06-21
titulo	string	Título da publicação	255ª Reunião - 20-21 junho, 2023
urlPdfAta	string	Caminho completo para acesso ao arquivo PDF do relatório. Disponível só a partir da reunião de número 200	https://www.bcb.gov.br/content/copom/atascopom/Copom255-not20230621255.pdf
textoAta	string	Conteúdo da ata, em formato texto com tags html	<div id=\"atacompleta\"><div id=\"ataconteudo\"><h3 class=\"secao\">A) Atualização da conjuntura econômica e do cenário do ...

Quadro 2 –Ata do COPOM

- Lista de comunicados do COPOM
https://www.bcb.gov.br/api/servico/sitebcb/copom/**comunicados**

Tem registros a partir de 19/04/2000.

Nome do Campo	Tipo do Campo	Descrição	Exemplo
nroReuniao	integer	255	255
dataReferencia	string<date>	Data do último dia da reunião do Copom, com formato conforme especificação RFC-3339 (ex: 2023-06-21)	2023-06-21
titulo	string	Título da publicação	255ª Reunião - 20-21 junho, 2023

Quadro 3 –Lista de Comunicados do COPOM

- Detalhes sobre um comunicado do COPOM
https://www.bcb.gov.br/api/servico/sitebcb/copom/**atas_detalhes**

Tem registros a partir de 19/04/2000

Nome do Campo	Tipo do Campo	Descrição	Exemplo
nroReuniao	integer	255	255
dataReferencia	string<date>	Data do último dia da reunião do Copom, com formato conforme especificação RFC-3339 (ex: 2023-06-21)	2023-06-21
titulo	string	Título do comunicado	Copom mantém a taxa Selic em 13,75% a.a.
textoComunicado	string	Conteúdo do comunicado, em formato texto com tags html	<p style=\"text-align:justify;\">Em sua 252ª reunião, o Comitê de Política Monetária (Copom) decidiu manter a taxa Selic ...

Quadro 4 – Comunicado do COPOM

- Histórico de taxas de juros SELIC
https://api.bcb.gov.br/dados/serie/bcdata.sgs.432/dados?formato=json&dataInicial=01/01/2016&dataFinal=31/12/2025

Tem registros diários a partir de 05/03/1999.

Nome do Campo	Tipo do Campo	Descrição	Exemplo
data	string<date>	Apresentação da taxa anual diariamente.	21/06/203
valor	float	Taxa anual no dia	15.00

Quadro 5 –Histórico da Taxa SELIC

- Histórico de taxas de inflação IPCA
https://api.bcb.gov.br/dados/serie/bcdata.sgs.433/dados?formato=json

Tem registros mensais a partir de

Nome do Campo	Tipo do Campo	Descrição	Exemplo
data	string<date>	Apresentação da taxa	21/06/203

		mensal mensalmente	
valor	float	Taxa no mês anterior	15.00

Quadro 6 –Histórico do IPCA

11 - ANÁLISE EXPLORATÓRIA

A base de dados selecionada para o projeto é composta pelas Atas das Reuniões do Comitê de Política Monetária (COPOM), obtidas diretamente através da API pública do Banco Central do Brasil (BACEN).

11.1 - AQUISIÇÃO DE DADOS E ESTRUTURA INICIAL

A análise exploratória (AED) iniciou-se com a fase de aquisição de dados, utilizando a biblioteca requests para realizar uma requisição HTTP GET à *URL* da API do COPOM. A resposta em formato JSON foi então processada e estruturada em um **Pandas DataFrame**.

```
import requests
import pandas as pd
from pandas import json_normalize
import matplotlib.pyplot as plt
from datetime import datetime


# URL da API do COPOM para pesquisa de ATAS disponíveis
url = "https://www.bcb.gov.br/api/servico/sitebcb/copom/atas?quantidade=300"

response = requests.get(url)


# Verificando se a requisição foi bem-sucedida
if response.status_code == 200:
    # Expande o resultado recorrente da API
    json_data = response.json()
    data_list = json_data.get('conteudo', [])

    # Converte para DF
    df = pd.DataFrame(data_list)
    numero_atas = len(df)
    df.describe()
```

Figura 1 – Captura de Dados

import requests e import pandas: Importação das bibliotecas para requisições *web* e manipulação de dados.

- **url:** Define o *endpoint* da API para as atas, solicitando até 300 registros.
- **response.status_code == 200:** Garante que a requisição foi bem-sucedida.
- **pd.DataFrame(data_list):** Converte a lista de dicionários (data_list) extraída do JSON da API em uma estrutura de dados tabular (df).

O *dataset* inicial contém **253 registros**, hoje a lista tem 253 registros, com a primeira reunião em 28 de janeiro de 1998 e a última em 17 de setembro de 2025, a primeira publicação ocorreu em 22 de maio de 1998 e a última em 23 de setembro de 2025.
As colunas chave incluem nroReuniao, dataReferencia, dataPublicacao e título
(que implicitamente remete ao conteúdo da Ata).

11.2 - ENGENHARIA DE VARIÁVEIS E ESTATÍSTICAS DESCRITIVAS

Para aprofundar a AED, as colunas de datas foram convertidas para o tipo datetime e foi calculada uma *feature* temporal: o prazo de publicação.

```
df['dataPublicacao'] = pd.to_datetime(df['dataPublicacao'])
df['dataReferencia'] = pd.to_datetime(df['dataReferencia'])
df['prazoPublicacao'] = df['dataPublicacao'] - df['dataReferencia']
df['prazoPublicacao'] = df['prazoPublicacao'].dt.days
df.describe()
```

	nroReuniao	dataReferencia	dataPublicacao	prazoPublicacao
count	253.000000	253	253	253.000000
mean	147.000000	2010-09-12 02:39:22.055335936	2010-09-25 16:30:21.343873536	13.577075
min	21.000000	1998-01-28 00:00:00	1998-05-22 00:00:00	2.000000
25%	84.000000	2003-05-21 00:00:00	2003-05-28 00:00:00	6.000000
50%	147.000000	2009-12-09 00:00:00	2009-12-17 00:00:00	8.000000
75%	210.000000	2017-10-25 00:00:00	2017-10-31 00:00:00	8.000000
max	273.000000	2025-09-17 00:00:00	2025-09-23 00:00:00	155.000000
std	73.179004	NaN	NaN	24.986484

Figura 2 - Variáveis

11.3 - ANÁLISE GRÁFICA E DEFINIÇÃO DO PERÍODO DE ESTUDO

A análise da *feature* prazoPublicacao foi aprofundada por meio de visualizações e estatísticas específicas para identificar a distribuição e o comportamento temporal do prazo de divulgação.

Figura 3 – Período do Estudo

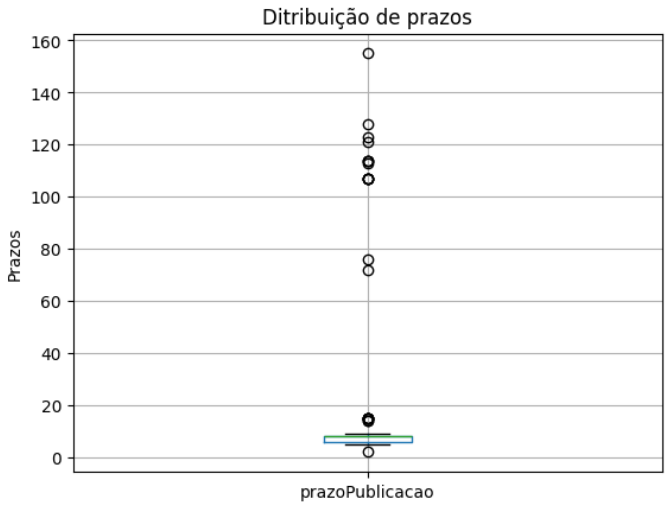

Figura 4 – Prazos de Publicação

Prazo mínimo: 2
Prazo médio: 13.577075098814229
Prazo mediano: 8.0
Prazo máximo: 155

11.4 - DELIMITAÇÃO E FILTRAGEM DO CONJUNTO DE DADOS

Após a constatação de uma quebra estrutural na variável prazoPublicacao coincidente com a adoção do **Regime de Metas de Inflação** em junho de 1999, o *dataset* foi filtrado para garantir que a análise se concentre apenas no período sob o regime monetário atual.

Código Explicado (Filtragem do DataFrame):
O código a seguir cria o novo *DataFrame* (df1999), retendo apenas os registros cuja dataReferencia é posterior a **1º de junho de 1999**. Em seguida, são geradas novas estatísticas descritivas para o subconjunto filtrado.

```
df1999 = df[df['dataReferencia'] > pd.to_datetime('1999-06-01')]
df1999.describe()
```

	nroReuniao	dataReferencia	dataPublicacao	prazoPublicacao
count	238.000000	238	238	238.000000
mean	154.500000	2011-06-14 20:10:05.042016768	2011-06-22 06:39:19.663865600	7.436975
min	36.000000	1999-06-23 00:00:00	1999-07-08 00:00:00	2.000000
25%	95.250000	2004-04-22 18:00:00	2004-04-30 18:00:00	6.000000
50%	154.500000	2010-11-13 12:00:00	2010-11-21 12:00:00	8.000000
75%	213.750000	2018-05-02 00:00:00	2018-05-08 00:00:00	8.000000
max	273.000000	2025-09-17 00:00:00	2025-09-23 00:00:00	15.000000
std	68.848868	NaN	NaN	1.672002

Figura 5 - Filtragem

Validade do Estudo: Ao focar nos **238 registros** posteriores a junho de 1999, o estudo garante que o

conteúdo textual das Atas (a ser usado como variável preditora) está contextualizado sob o **regime de política monetária moderna** (Metas de Inflação), aumentando a validade e a relevância das futuras análises.

11.5 - ANÁLISE GRÁFICA DO SUBCONJUNTO DELIMITADO

Para confirmar a homogeneidade do período selecionado e validar a eliminação da quebra estrutural, foram geradas novas visualizações e estatísticas focadas apenas no *DataFrame* pós-1999 (df1999).

Gráfico de Linha: Diferentemente da série original, o novo gráfico demonstra que a variável prazoPublicacao opera dentro de uma faixa estreita de valores (entre 2 e 15 dias). A alta volatilidade e os picos de longa duração (acima de 100 dias) foram eliminados, indicando um **comportamento estável e previsível**.

Figura 6 – Dados Filtrados

Boxplot: O *boxplot* confirma a **baixa dispersão** dos dados. O *range* interquartil (IQR) é pequeno, e os *outliers* remanescentes (até 15 dias) são significativamente menos extremos em comparação ao *dataset* completo.

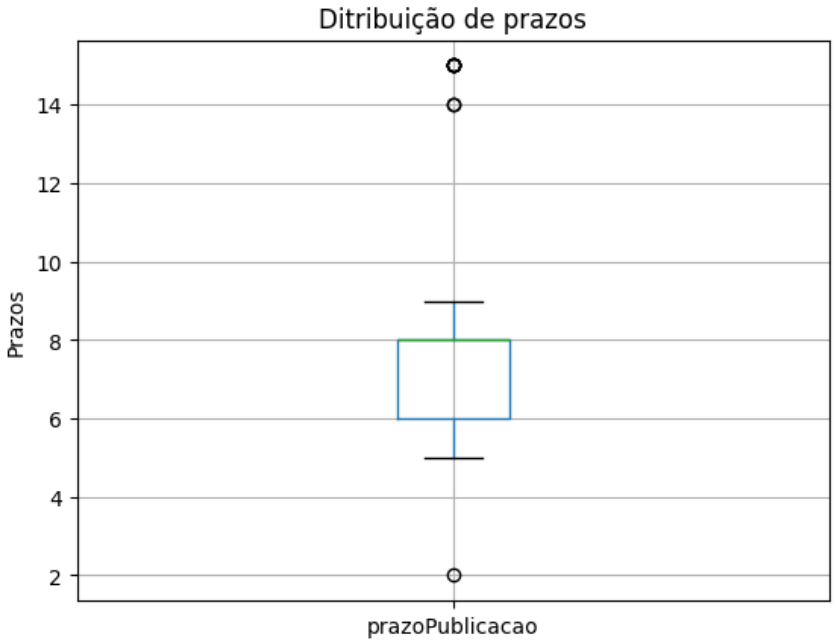

Figura 7 – Nova Distribuição de Pazos

Prazo mínimo: 2
Prazo médio: 7.436974789915967
Prazo mediano: 8.0
Prazo máximo: 15
Estatísticas: A **Média (7.44 dias)** e a **Mediana (8.0 dias)** são muito próximas, sugerindo uma distribuição mais simétrica e menos distorcida pelos valores extremos (já removidos).

Conclusão da AED: O *dataset* df1999 está agora **homogêneo** e **pronto** para a próxima fase do **Tratamento de Dados**, que é a extração de *features* textuais e a preparação do conjunto de treinamento e teste.

11.6 - FORMAÇÃO DA BASE DE DADOS CONSOLIDADA E AQUISIÇÃO DO TEXTO INTEGRAL

A base filtrada (df1999) contém 238 registros válidos, iniciados em junho de 1999, cobrindo todas as reuniões sujeitas à regulamentação atual do Regime de Metas de Inflação. Nesta etapa, foi realizada a **aquisição programática** dos textos completos das Atas e dos Comunicados para cada reunião, utilizando os *endpoints* específicos da API do Banco Central.

O processo iterativo abaixo utiliza o número de cada reunião (nroReuniao) para fazer duas requisições HTTP distintas: uma para obter o texto integral da Ata e outra para o Comunicado. Os dados textuais (textoAta e textoComunicado) são então consolidados em um novo *DataFrame* (df_copom), que será a base primária para o Processamento de Linguagem Natural (PLN).

Out[7]:

›Reuniao	dataReferencia	dataPublicacao	tituloAta	textoAta	tituloComunicado	textoComunicado
269	2025-03-19	2025-03-25	269ª Reunião - 18-19 março, 2025	<h...	Copom eleva a taxa Selic para 14,25% a.a.	<div class="ExternalClassAE4A4113CD704413AF90A...
270	2025-05-07	2025-05-13	270ª Reunião - 6-7 maio, 2025	<h...	Copom eleva a taxa Selic para 14,75% a.a.	<div class="ExternalClassB27C33A0668D4D38BE755...
271	2025-06-18	2025-06-24	271ª Reunião - 17-18 junho, 2025	<h...	Copom eleva a taxa Selic para 15,00% a.a.	<div class="ExternalClass276B5992DAD145FD92ED7...
272	2025-07-30	2025-08-05	272ª Reunião - 29-30 julho, 2025	<h...	Copom mantém a taxa Selic em 15,00% a.a.	<div class="ExternalClass1BE7096A76144A1BB7BCE...
273	2025-09-17	2025-09-23	273ª Reunião - 16-17 setembro, 2025	<h...	Copom mantém a taxa Selic em 15,00% a.a.	<div class="ExternalClass89BEC3FC87B24E52A58CC...

Figura 8 - Texto Integral

Análise: O sucesso na aquisição dos dados é confirmado pelo preenchimento das colunas textoAta e textoComunicado. É notável, contudo, que o texto bruto contém tags HTML (..., ...). Este é um problema de ruído que será tratado na próxima fase: o **Processamento de Linguagem Natural (PLN)**.

11.7 - ANÁLISE ESTATÍSTICA DA EXTENSÃO DOS TEXTO

Com o *DataFrame* df_copom consolidado e contendo o texto integral das Atas e Comunicados, a próxima fase da Análise Exploratória focou na inspeção das *features* textuais. Esta análise preliminar da extensão (número de caracteres) é vital para planejar as etapas subsequentes de limpeza, tokenização

e vetorização no Processamento de Linguagem Natural (PLN).

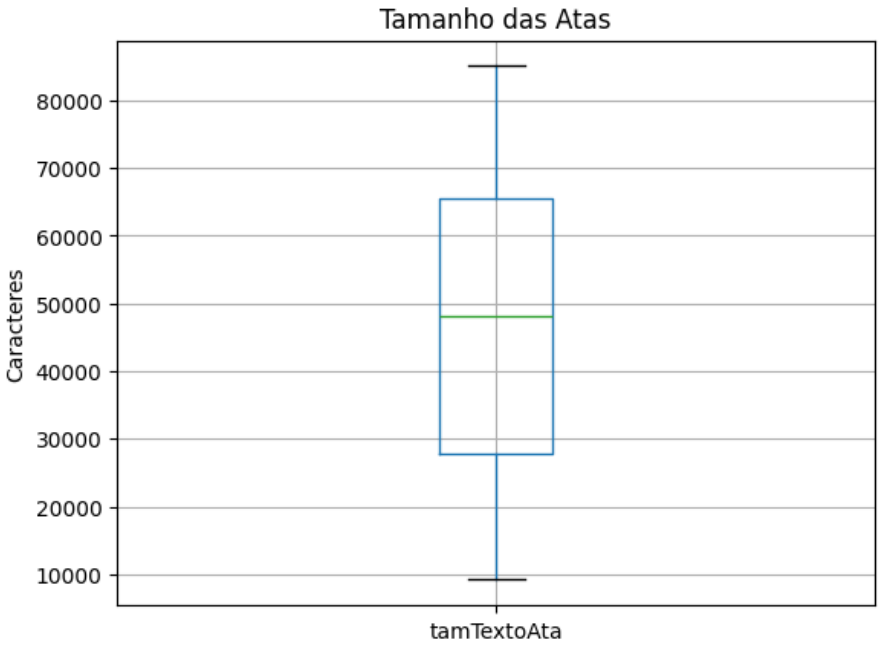

Figura 9 - Extensão dos Textos

Tamanho mínimo das atas: 9177.0
Tamanho máximo das atas: 85056.0
Tamanho médio das atas: 46994.7572815534
Tamanho mediano das atas: 48215.0

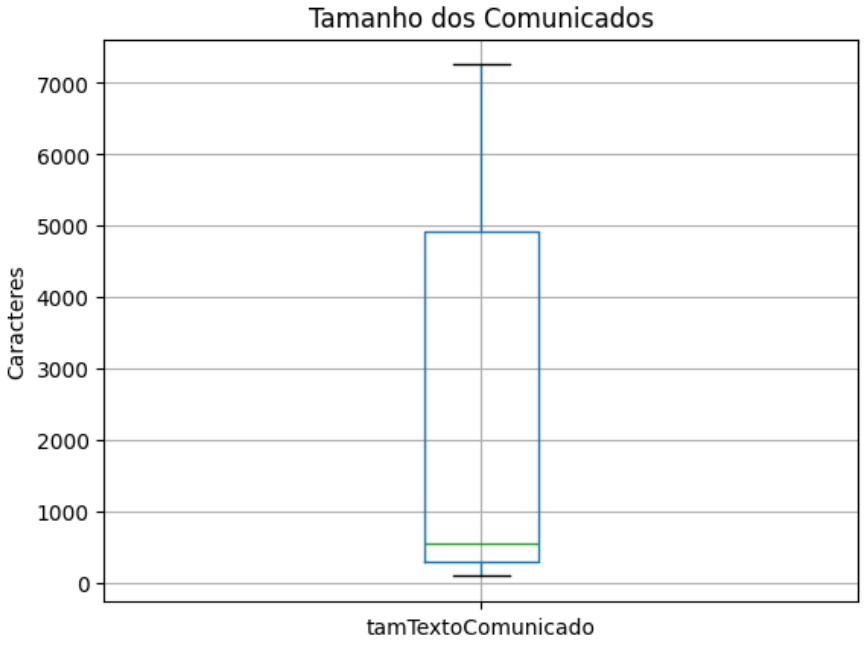

Figura 10 - Tamanho dos Comunicados

Tamanho mínimo dos comunicados: 99.0
Tamanho máximo dos comunicados: 7254.0
Tamanho médio dos comunicados: 2098.5526315789475

Tamanho mediano dos comunicados: 560.5

11.8 - AQUISIÇÃO E ANÁLISE DA SÉRIE HISTÓRICA DO IPCA

A fase de tratamento de dados exige a integração de variáveis macroeconômicas cruciais ao lado das *features* textuais. Esta subseção detalha a aquisição e a análise da série histórica do **IPCA (Índice Nacional de Preços ao Consumidor Amplo)**, que será utilizada como uma **variável preditora numérica**.

Contexto e Necessidade de Limpeza Textual
Conforme a análise de extensão textual (Seção 11.7), o conteúdo das atas e comunicados possui **formatação HTML** (...), necessitando de um pré-processamento para **remoção de *tags*** e outros caracteres especiais antes da vetorização. A média de caracteres de aproximadamente **47.000 (Atas)** e **2.098 (Comunicados)** confirma a riqueza da informação textual a ser explorada.

Nota Institucional: Observa-se que os comunicados começaram a ser publicados a partir da reunião 46, em 26 de abril de 2000. Para reuniões anteriores (dentro da base filtrada pós-1999), somente o texto das Atas estará disponível, o que será tratado na modelagem como ausência de *feature* textual do comunicado.

Aquisição e Tratamento do IPCA
A série histórica do IPCA foi obtida através da **API de Séries Temporais (SGS)** do Banco Central do Brasil, utilizando o código **13522** (IPCA Acumulado 12 Meses) para garantir a relevância da informação de inflação de longo prazo para as decisões do COPOM.

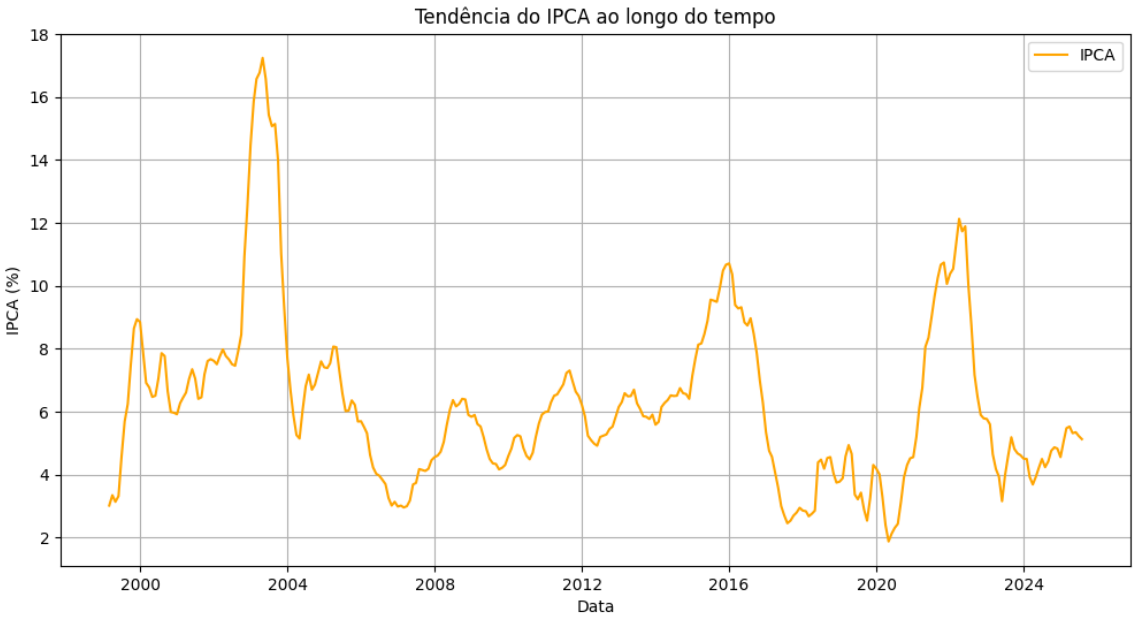

Figura 11 - Histórico do IPCA

Estatísticas descritivas do IPCA:
count 318.000000
mean 6.228553
std 2.697288
min 1.880000
25% 4.500000
50% 5.875000
75% 7.262500
max 17.240000

Name: valor, dtype: float64

Código e Gráfico Explicados (Tendência):
O gráfico de linha confirma a natureza de série temporal volátil do IPCA, mostrando picos (como o máximo de 17.24%) e períodos de estabilidade.

11.9 - AQUISIÇÃO E ANÁLISE DA SÉRIE HISTÓRICA DA TAXA SELIC

A segunda variável macroeconômica fundamental para a previsão é a **Taxa Selic Diária (SGS 432)**, que representa o *target* das decisões do COPOM. Devido à sua natureza de indicador diário e às limitações impostas pela API do Banco Central para o intervalo de datas, foi necessária uma abordagem iterativa para garantir a integridade da série histórica completa.
Estratégia de Aquisição Segmentada
Para contornar as restrições de *range* da API, a busca pela série diária da Selic foi segmentada em blocos de tempo menores. O código implementa uma função (buscar_selic_periodo) que é executada sequencialmente por cinco períodos, desde fevereiro de 1999 até a data mais recente disponível na base de Atas.

```
=== BUSCANDO PERÍODOS HISTÓRICOS ===

Buscando período: 01/02/1999 a 31/12/2004
  URL: https://api.bcb.gov.br/dados/serie/bcdata.sgs.432/dados?formato=json&dataInicial=01/02/1999&dataFinal=31/12/2004
Encontrados 2129 registros
Período 01/02/1999 a 31/12/2004: 2129 registros adicionados

Buscando período: 01/01/2005 a 31/12/2009
  URL: https://api.bcb.gov.br/dados/serie/bcdata.sgs.432/dados?formato=json&dataInicial=01/01/2005&dataFinal=31/12/2009
Encontrados 1826 registros
Período 01/01/2005 a 31/12/2009: 1826 registros adicionados

Buscando período: 01/01/2010 a 31/12/2014
  URL: https://api.bcb.gov.br/dados/serie/bcdata.sgs.432/dados?formato=json&dataInicial=01/01/2010&dataFinal=31/12/2014
Encontrados 1826 registros
Período 01/01/2010 a 31/12/2014: 1826 registros adicionados

Buscando período: 01/01/2015 a 31/12/2019
  URL: https://api.bcb.gov.br/dados/serie/bcdata.sgs.432/dados?formato=json&dataInicial=01/01/2015&dataFinal=31/12/2019
Encontrados 1826 registros
Período 01/01/2015 a 31/12/2019: 1826 registros adicionados

Buscando período: 01/01/2020 a 17/09/2025
  URL: https://api.bcb.gov.br/dados/serie/bcdata.sgs.432/dados?formato=json&dataInicial=01/01/2020&dataFinal=17/09/2025
Encontrados 2087 registros
Período 01/01/2020 a 17/09/2025: 2087 registros adicionados

RESULTADO FINAL:
Total de registros: 9694
Período coberto: 1999-03-05 00:00:00 até 2025-09-17 00:00:00
Número de períodos consolidados: 5

Primeiros 5 registros:
        data  valor
0 1999-03-05  45.00
1 1999-03-06  45.00
```

Figura 12 - Períodos da Taxa SELIC

A execução desta estratégia foi bem-sucedida, resultando na consolidação de **9.694 registros** diários da Taxa Selic, cobrindo o período contínuo de **1999-03-05 a 2025-09-17**.

Análise Estatística e Gráfica
Para fins de visualização e *feature engineering*, a série diária foi agrupada por mês.

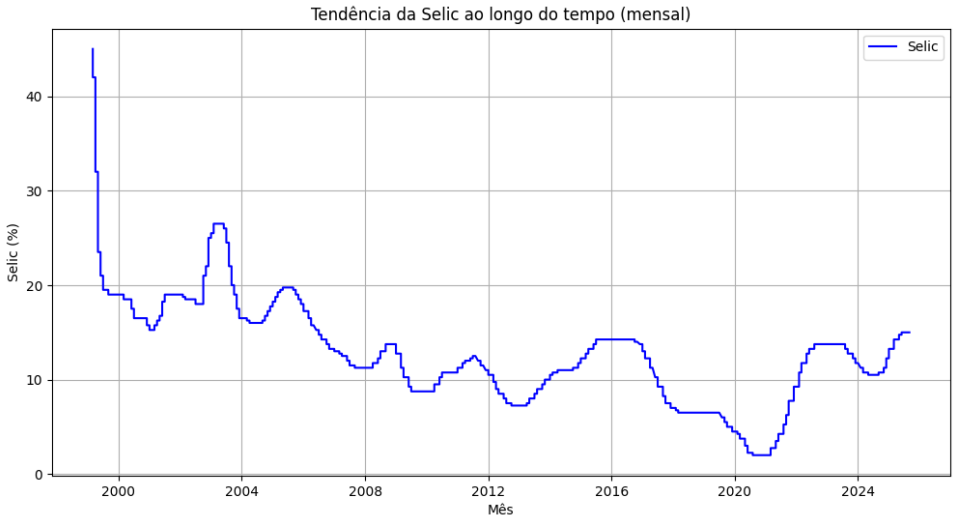

Figura 13 - Histórico da Taxa SELIC

Estatísticas descritivas da Selic:
count 9694.000000
mean 12.727151
std 5.480852
min 2.000000
25% 9.250000
50% 12.250000
75% 16.000000
max 45.000000
Name: valor, dtype: float64

Análise: O gráfico de tendência revela a **alta volatilidade** histórica da taxa Selic, uma característica esperada para a política monetária brasileira, que utiliza a taxa de juros como principal ferramenta de controle inflacionário. O **Desvio Padrão (≈6.27)** é elevado, demonstrando grandes oscilações ao longo do período de estudo. A série exibe picos elevados no início (máximo de 45.00%), seguidos de uma tendência decrescente de longo prazo, intercalada por ciclos de alta e baixa.

11.10 - LIMPEZA E NORMALIZAÇÃO DO CONTEÚDO TEXTUAL (PLN)

A Análise Exploratória (Seções 11.6 e 11.7) confirmou que as colunas textoAta e textoComunicado contêm **formatação HTML**, que deve ser removida antes de qualquer técnica de Processamento de Linguagem Natural (PLN). A remoção desse ruído é um passo fundamental para garantir que a vetorização textual capture apenas o significado semântico das palavras.

Esta subseção detalha a aplicação da função de limpeza de HTML e a normalização do texto da Ata.
Função de Limpeza e Aplicação em df_atas

A função limpar_html foi implementada para remover *tags* e normalizar o texto:
BeautifulSoup: Utilizada para fazer a raspagem do conteúdo e extrair eficientemente o texto puro, descartando todas as *tags* HTML.

re (Expressões Regulares): Aplicada para normalizar o texto, substituindo múltiplos espaços, quebras de linha (\n) e tabs (\t) por um único espaço, e removendo caracteres especiais desnecessários.

```
Criando DataFrame de atas...
Limpando HTML das atas...
Estatísticas do DataFrame de atas:
Total de atas: 238
Período: 1999-06-23 00:00:00 até 2025-09-17 00:00:00

Tamanho médio original: 46995 caracteres
Tamanho médio limpo: 31747 caracteres
Redução média: 22.0%
Tamanho mínimo das atas limpas: 9177 caracteres
Tamanho máximo das atas limpas: 85056 caracteres
Tamanho mediano das atas limpas: 48215 caracteres

=== EXEMPLO DE LIMPEZA ===
Texto original (primeiros 200 chars):


    <p><strong>Sum&aacute;rio</strong></p>

    <div>
      <a href="#_Toc455996804" title="Pre&ccedil;os e N&iacute;vel de Atividade">Pre&ccedil;os e N&iacute;vel de Atividade
</a>

      <div>


Texto limpo (primeiros 200 chars):
Sumário Preços e Nível de Atividade Agregados Monetários e Crédito Finanças Públicas Balanço de Pagamentos Ambiente Externo
Evolução do Mercado de Câmbio Doméstico e Posição das Reservas Internacionai

DataFrame df_atas criado com 238 registros

Primeiros registros:
  dataReferencia                                          textoAta
0     1999-06-23  Sumário Preços e Nível de Atividade Agregados ...
1     1999-07-28  Sumário Preços e nível de atividade Agregados ...
2     1999-09-01  Sumário Demanda e oferta agregadas Preços Agre...
3     1999-09-22  Sumário Demanda e oferta agregadas Ambiente ex...
4     1999-10-06  Sumário Demanda e oferta agregadas Ambiente Ex...
```

Figura 14 - Limpeza do Texto

A redução média de **22.0%** no tamanho das Atas após a limpeza (de 46.995 para 31.747 caracteres) é significativa e confirma que uma grande quantidade de *tags* HTML, metadados e outros elementos não textuais foram removidos com sucesso.

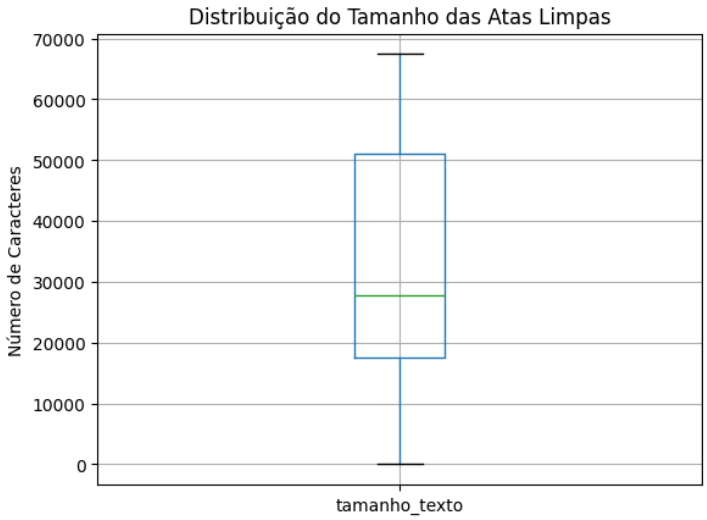

Figura 15 - Tamanho dos Textos Limpos

O processo resultou na criação do **DataFrame df_atas** com **238 registros** válidos e a coluna textoAta perfeitamente limpa e normalizada. O período coberto é de **1999-06-23 a 2025-09-17**, consistente com o filtro temporal estabelecido. A etapa de pré-processamento textual está, portanto, concluída para as Atas.

11.11 - CONSOLIDAÇÃO FINAL DOS DADOS E ENGENHARIA DA VARIÁVEL ALVO

A etapa final do pré-processamento consiste na consolidação dos *datasets* filtrados e tratados: o texto limpo das Atas (df_atas), a série diária da Selic (df_selic) e a série mensal do IPCA (df_ipca). Além disso, é criada a **Variável Alvo (y)**, que define o objetivo de regressão do modelo.
Estratégia de Junção e Mapeamento de Variáveis
Para garantir a coerência temporal, a junção dos *datasets* foi realizada através de funções customizadas

que mapeiam os valores mais relevantes para a data de referência de cada Ata:

- **Selic (Selic):** Utiliza-se a função buscar_selic para encontrar o valor da Selic em vigor na Data exata da Ata. Caso a data não exista (feriado ou fim de semana), busca-se o valor imediatamente anterior, garantindo 100% de preenchimento.
- **IPCA (IPCA):** A função buscar_ipca_anterior mapeia o valor do IPCA (acumulado 12 meses) **referente ao mês anterior** à data da Ata, uma vez que a inflação é divulgada com *lag*.

Análise e Integridade do Dataset Final

O processo de *feature engineering* e *merge* resultou em um *DataFrame* final (df_final) com **238 registros**. A estratégia de mapeamento foi bem-sucedida, eliminando valores nulos (NA) nas variáveis preditoras numéricas.

```
# Verificar registros com dados faltantes
registros_incompletos = df_final[(df_final['Selic'].isna()) | (df_final['IPCA'].isna())]
if not registros_incompletos.empty:
    print(f"\n{len(registros_incompletos)} registros com dados faltantes:")
    print(registros_incompletos[['Data', 'Selic', 'IPCA']])

print(f"\nDataFrame final criado: df_final com {len(df_final)} registros")
```

```
=== CRIANDO DATAFRAME CONSOLIDADO ===
Base de atas: 238 registros
Dados SELIC: 9694 registros
Dados IPCA: 318 registros

Buscando valores SELIC e IPCA para cada ata...

=== ESTATÍSTICAS DO DATAFRAME FINAL ===
Total de registros: 238
Período: 1999-06-23 00:00:00 até 2025-09-17 00:00:00

SELIC:
  Valores válidos: 238/238
  Valores faltantes: 0
  Min: 2.00%
  Max: 26.50%
  Média: 13.28%

IPCA:
  Valores válidos: 238/238
  Valores faltantes: 0
  Min: 1.88%
  Max: 17.24%
  Média: 6.49%
```

Figura 16 - Análise de Integridade

Abaixo estão os primeiros e últimos registros do *dataset* final, confirmando a correta integração dos dados

```
=== PRIMEIROS REGISTROS DO DATAFRAME FINAL ===
        Data                                              Texto  Selic  IPCA  \
0 1999-06-23  Sumário Preços e Nível de Atividade Agregados ...   22.0  3.32
1 1999-07-28  Sumário Preços e nível de atividade Agregados ...   21.0  4.57
2 1999-09-01  Sumário Demanda e oferta agregadas Preços Agre...   19.5  5.69
3 1999-09-22  Sumário Demanda e oferta agregadas Ambiente ex...   19.5  6.25
4 1999-10-06  Sumário Demanda e oferta agregadas Ambiente Ex...   19.0  7.50

   Selic (4m)
```

Figura 17- Registros do Dataset

Criação do rótulo

Foi criada uma nova coluna, ‘Selic (6m)’, com o valor da taxa seis meses após a data da reunião. Comparando o valor da taxa no futuro com a atual, foi criada uma nova coluna, de sentimento, categórica com 3 valores:

‘hawkish’: se a taxa subiu após a reunião;

‘neutral’: se se manteve;
‘dovish’: se a taxa caiu.

Definição do Dataset de Modelagem
O *DataFrame* **df_final** está totalmente preparado, sem dados faltantes nas colunas essenciais, e define claramente as variáveis para o modelo:
Variáveis Preditora (X): Texto (limpo)
Variável Alvo (y): Sentimento

12 - EMBASAMENTO

12.1 – DEFINIÇÃO DA LINGUAGEM DE PROGRAMAÇÃO

Foi escolhida PYTHON como a linguagem de desenvolvimento para o projeto. A ampla disponibilidade de literatura e bibliotecas nas disciplinas de Ciências de Dados é o fator principal. A familiaridades de todos os integrantes do grupo com a mesma também é decisiva.

12.2 – BASE TEÓRICA E MÉTODOS

Foi tomada a decisão de abordar o problema inicialmente como um problema de classificação. Os objetos do estudo, inflação e taxa de juros, são numéricos e poderiam ser alvos de regressão. Contudo, para uma primeira abordagem foi decidido que um estudo de sentimento, avaliando se estas variáveis devem subir, cair ou se manter em prazos pré-determinados é mais prático e atende aos objetivos do estudo, que é indicar o posicionamento ideal da instituição considerando a direção de variação dos índices, e não a intensidade.

As atas são textos extensos, com repetições de terminologia e uso ambíguo de palavras. Isto torna o uso de regressão logística ou ‘bag of words’ menos indicados. O estudo aqui proposto é uma análise de sentimento, uma das várias possibilidades dentro do que comummente se chama de ‘Estudo de Linguagem Natural, ou natural linguage processing – NLP - em inglês. O processamento de NLP pode envolver transformação ou interpretação de texto, capacidade que exploraremos aqui. Redes neurais recorrente ou convolucionais (CNN/RNN) capturam contextos sequenciais de palavras ou frases. Modelos avançados com uso de transformadores (BERT) utilizam mecanismos de atenção para priorizar elementos relevantes do texto.

Para a implementação do modelo de NLP selecionamos o TensorFlow, desenvolvido e mantido pela Google. A biblioteca permite a implementação de redes neurais com grande flexibilidade e inclui ferramentas de suporte na preparação dos dados, definição de pipelines e avaliação dos modelos.

Um fator muito decisivo na escolha da ferramenta é a existência do TensorFlow HUB e a possibilidade de se integrar camadas personalizadas pré-treinadas. As atas a serem estudadas são extensas, mas não são muitas. Desta forma a qualidade do modelo que será obtido a partir das mesmas é incerto no início do processo e a posterior adição de camadas pré-trinadas virá suprindo a deficiência na massa de dados do projeto.

12.3 – CÁLCULO DE ACURÁCIA

Para a avaliação da qualidade do modelo foi definido a utilização das seguintes ferramentas:

- **Matriz de confusão:** Permite a visualização da qualidade dos resultados gerados pelo modelo
- **F1-Score:** Oferece um valor mensurável para a qualidade dos resultados obtidos, considerando tanto a *Precisão* quanto o *Recall*.

13 - MODELAGEM E RESULTADOS

13.1 - DEFINIÇÃO DO MODELO

Foram testadas inicialmente duas abordagens para o desenvolvimento do modelo. Ambas utilizam RNNs (redes neurais recorrentes). Nos dois casos as entradas devem ser convertidas em tokens, convertendo os textos em vetores numéricos. Em ambos os casos a unidade de texto escolhida foram palavras. Também em ambos os casos os rótulos são convertidos para One Hot, onde um campo com três valores possíveis é convertido para três campos binários.

A primeira abordagem foi baseada em um exemplo sugerido pelo Dr. Ernesto Lee (Ref. 6) em um artigo no site de internet Medium. Utilizando uma arquitetura de múltiplas camadas com um a camada de incorporação, que transforma palavras em vetores numéricos significativos, seguidas de duas camadas LTSM (LTSM se traduz para o português como Memória de Curto e Longo Prazo), que é um tipo de camada recorrente de RNN projetada para lidar com sequências de dados e manter informações de sequência.
A segunda abordagem baseada no modelo desenvolvido por Géron (Ref.5, Cap. 16) emprega uma camada GRU (Unidade Recorrente de Portas em português), também uma camada recorrente para RNN. Como as camadas, LTSM é indicada para processar sequência de dados, mas é mais leve em consume de memória e processamento.

As duas abordagens também diferem no manuseio dos dados. A primeira processa vetores tokenizados diretamente, sendo a preparação feita de forma independente. A segunda emprega o modelo de datasets da biblioteca TensorFlow, que encapsula e automatiza as tarefas de preparação de dados.

13.2 - TREINAMENTO DO MODELO E MEDIDAS DE ACURÁCIA

Para o modelo LTSM foram empregados 64% para treinamento, 16% para validação e 20% reservados para testes posteriores à modelagem. Para o modelo GRU foram reservados 80% para treinamento, 10% para validação. Outros 10% reservados para. Em ambos os casos a divisão foi feita de maneira aleatória e o tamanho das partições é o sugerido nas referências.

Os modelos foram processados em um mesmo computador, sem GPU. O modelo LTSM rodou por 72 horas, 150 épocas, com os seguintes resultados:

Acuracidade de treinamento	1.0000
Perda de Treinamento	0.0012
Acuracidade de Validação	0.7576
Perda de Validação	1.5932

Tabela 1 – Treinamento LTSM

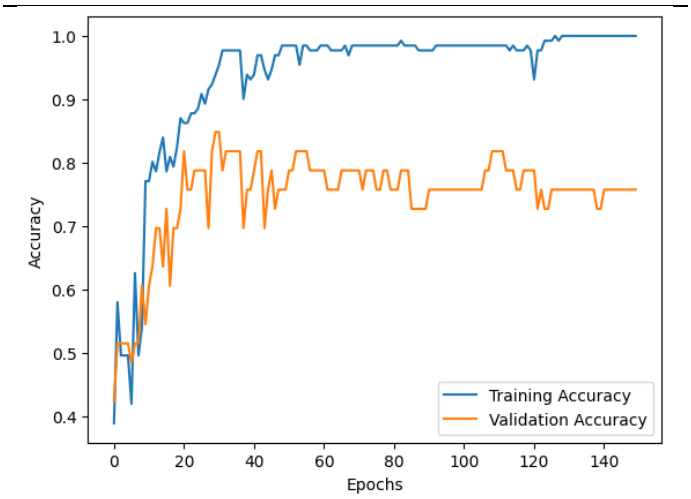

Figura 18 – Evolução do Treinamento LTSM

Os testes com a base não vista no treinamento:

Label	**Precision**	**Recall**	**F1-Score**	**Support**
dovish	0.62	0.67	0.64	12
hawkisk	0.83	0.76	0.79	25
neutral	0.33	0.40	0.36	5
Accuracy			0.69	42
macro avg	0.59	0.61	0.60	42
weighted avg	0.71	0.69	0.70	42

Tabela 2 – Relatório de Classificação – Modelo LTSM

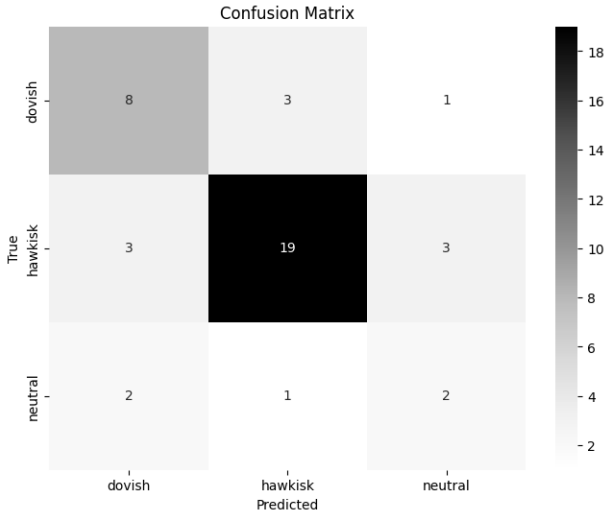

Figura 19 - Matriz de Confusão LTSM

O modelo GRU rodou por 2h30', com 180 épocas e os seguintes resultados:

Acuracidade de treinamento	0.9939
Perda de Treinamento	0.0192
Acuracidade de Validação	1.0000
Perda de Validação	0.0014

Tabela 3 – Treinamento GRU

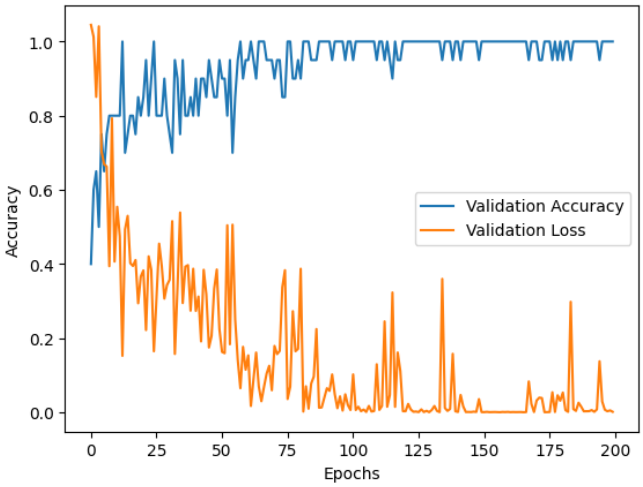

Figura 20 – Evolução do Treinamento GRU

Os testes com a base não vista no treinamento:

Label	Precision	Recall	F1-Score	Support
dovish	1.00	1.00	1.00	9
hawkisk	1.00	1.00	1.00	9
neutral	1.00	1.00	1.00	4
Accuracy				22
macro avg	1.00	1.00	1.00	22
weighted avg	1.00	1.00	1.00	22

Tabela 4 – Relatório de Classificação – Modelo GRU

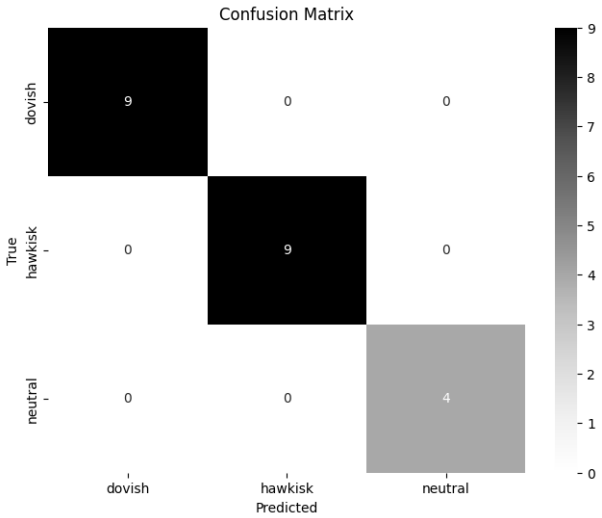

Figura 21 – Matriz de Confusão GRU

Analisando os resultados dos dois treinamentos é visível que os dois modelos convergem. Nos dois casos a presença de spikes nos valores de acurácia e perda indica o uso de estabilização do treinamento, reduzindo a volatilidade de função de perda.

O modelo LTSM contudo tem um custo muito maior e é inviável o processamento até uma solução de

qualidade na infraestrutura disponível.

O modelo GRU, sendo muito mais leve, convergiu rapidamente. Os resultados com 100% de acerto abrem margem para que se suspeite de overfitting. Para reavaliar o modelo foi realizado um novo treinamento do modelo GRU com 60% para treinamento, 10% para validação e 30% para teste posterior. O modelo rodou por 1h45'e 180 épocas. Seguem os resultados:

Acuracidade de treinamento	0.9919
Perda de Treinamento	0.0118
Acuracidade de Validação	1.0000
Perda de Validação	9.7292e-04

Tabela 5 – Retreinamento GRU

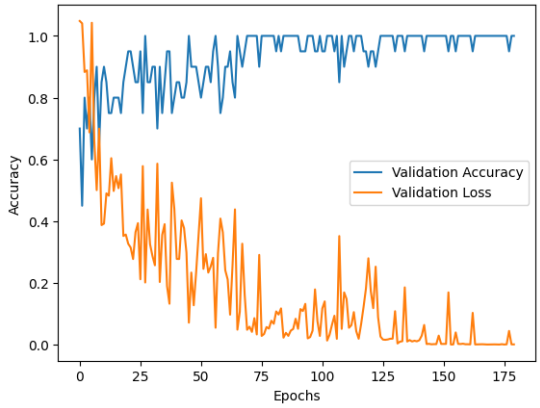

Figura 22 – Evolução do Retreinamento GRU

Os testes com a base não vista no treinamento:

Label	Precision	Recall	F1-Score	Support
dovish	1.00	1.00	1.00	28
hawkisk	1.00	1.00	1.00	32
neutral	1.00	1.00	1.00	3
Accuracy				63
macro avg	1.00	1.00	1.00	63
weighted avg	1.00	1.00	1.00	63

Tabela 6 – Relatório de Classificação – Modelo GRU 60%

O F1-Score de 1.00 e a acuracidade de 100% sugerem um eventual sobre-ajuste. Foi rodado mais um modelo utilizando 40% da base para treinamento, 10% para validação e 50% para teste. O resultado se repetiu. Ainda existem outras abordagens a serem tentadas, como por exemplo a segmentação temporal da base antes da divisão, garantindo distribuição uniforme ao longo do tempo entre os grupos de treino, validação e teste.

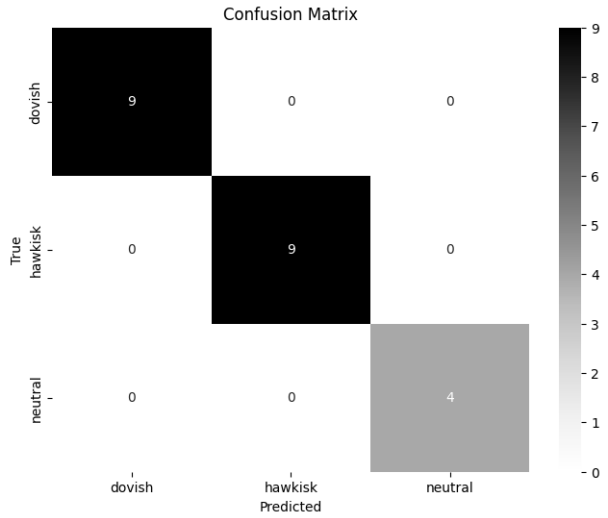

Figura 23 – Matriz de Confusão GRU 60%

Conclui-se que o modelo é realmente preciso, a redução na base de treinamento e aumento na base de testes não alterou a qualidade do resultado obtido.

13. 3 – DISCUSSÃO DOS RESULTADOS

O estudo produzir um modelo de rede neural com unidades GRU, implementado em TensorFlow com o objetivo de prever a variação das taxas de juros básica do Banco Central do Brasil num horizonte de 6 mesas após cada reunião. O modelo classifica as atas em ‘Neutral’ (taxa deve se manter), ‘Hawkish’ (a taxa deve subir) e ‘Dovish’ (a taxa deve cair).

Os dados consistem em atas históricas das reuniões do COPOM para extração de características textuais, e a taxa básica SELIC nos períodos estudados.

Os resultados apresentaram acurácia de validação de 99% depois de 180 épocas. O F1-Score na base de testes é 1.00, com previsões exatas, mostrando que o modelo tem capacidade de capturar nuances no texto que indicam as intenções a médio prazo na condução da política monetária.

O modelo como está já apresenta capacidade preditiva e emprego na definição de políticas institucionais de crédito e no direcionamento de posicionamento da tesouraria do banco.

Ficam abetas outras venidas de desenvolvimento, buscando correlações em outros prazos para a taxa SELIC e a busca de previsibilidade e inflação e até taxa de câmbio.

Na data da redação deste relatório as seguintes previsões se encontram em aberto, indicando uma reversão de tendência e início de período de redução de taxas de juros:

text	date	rate	next_date	sentiment
a atualização da conjuntura econômica e do cen...	2025-05-07	14.25	2025-11-05	dovish
a atualização da conjuntura econômica e do cen...	2025-06-18	14.75	2025-12-10	dovish
a atualização da conjuntura econômica e do cen...	2025-07-30	15.00	2026-01-28	dovish

text	date	rate	next_date	sentiment
a atualização da conjuntura econômica e do cen...	2025-09-17	15.00	2026-03-18	dovish

Tabela 7 – Previsões

14 – STORYTELLING

14.1 - CONTEXTO

O Comitê de Política Monetária, conhecido como COPOM, é responsável por definir a taxa básica de juros do Brasil, a Selic, que é o principal instrumento de controle da inflação. Cada decisão tomada pelo comitê tem efeito direto sobre o crédito, os investimentos e o consumo no país, influenciando desde os grandes bancos até o cotidiano das pessoas.

As atas e comunicados divulgados após as reuniões trazem explicações sobre as escolhas feitas e analisam o cenário econômico. Apesar de serem textos técnicos, eles carregam mudanças sutis no tom e na escolha das palavras que revelam a visão do COPOM em relação à economia. Essas nuances ajudam a entender se a política monetária tende a ser mais rígida, mais flexível ou se deve permanecer estável.

Ter a capacidade de interpretar essas informações de forma rápida e confiável é muito importante para instituições financeiras, gestores de investimento e órgãos de política pública. Para o Banco do Brasil, que foi a instituição escolhida como referência no estudo, antecipar os movimentos da Selic e da inflação pode apoiar na gestão de crédito, na definição de preços de ativos e em estratégias de alocação de recursos.

Nesse sentido, usar ciência de dados e técnicas de processamento de linguagem natural permite transformar esses documentos do COPOM em indicadores práticos que ajudam a prever tendências e embasam decisões de mercado com mais eficiência.

14.2 - PROBLEMA

As atas do COPOM são documentos extensos, com dezenas de milhares de caracteres, e os comunicados, embora mais curtos, também mantêm uma linguagem formal e técnica. Esse formato dificulta a leitura rápida e a extração de informações úteis para quem precisa agir com agilidade no mercado. Além disso, as atas costumam ser publicadas com um intervalo de alguns dias após a reunião, o que reduz a velocidade de acesso a informações que podem impactar decisões financeiras imediatas.

Outro ponto é que, apesar de seguirem uma estrutura padronizada, as atas apresentam variações sutis no vocabulário e na forma de expor os argumentos. Essas pequenas mudanças podem sinalizar a direção futura da política monetária, mas exigem experiência e tempo de análise para serem percebidas.

Na prática, isso cria um desafio: transformar um volume grande de texto técnico em dados objetivos e indicadores confiáveis, capazes de apoiar decisões em tempo hábil. A ausência de um processo automatizado para leitura e interpretação das atas e comunicados limita o aproveitamento do potencial dessas informações e reduz a competitividade de instituições financeiras que dependem delas para orientar suas estratégias.

14.3 – OBJETIVO

O projeto tem como meta transformar as atas e comunicados do COPOM em informações mais acessíveis e úteis para análise econômica. A ideia é usar técnicas de ciência de dados para resumir automaticamente os documentos, reduzindo o tempo necessário para interpretar cada reunião, e também para criar indicadores preditivos que apontem a tendência da política monetária.

Esses indicadores serão construídos a partir da relação entre o texto das atas e os movimentos posteriores da Selic e da inflação. O objetivo não é prever valores exatos, mas identificar a direção provável se a taxa de juros tende a subir, cair ou se manter estável. Esse tipo de classificação pode

servir como apoio para bancos, gestores de ativos e empresas que precisam reagir rapidamente às mudanças do cenário econômico.
Com isso, o projeto busca mostrar que é possível converter linguagem técnica em sinais práticos, usando processamento de linguagem natural como ferramenta para antecipar tendências de mercado e dar suporte a decisões estratégicas.

14.4 – METODOLOGIA

A primeira etapa foi a coleta das informações por meio das APIs públicas do Banco Central. A partir delas foram obtidas as atas e comunicados do COPOM, além das séries históricas da Selic e do IPCA. Esses dados foram organizados em uma base única, relacionando cada reunião do comitê às variáveis econômicas correspondentes.
Depois da coleta, os textos passaram por um processo de limpeza. Como eram disponibilizados em formato HTML, foi necessário remover tags e caracteres que não traziam valor para a análise. Em seguida, os documentos foram normalizados e unificados em um campo chamado “texto consolidado”, que reúne a ata e o comunicado de cada reunião. Esse campo é a base para as técnicas de Processamento de Linguagem Natural.
Na parte numérica, os valores da Selic foram associados à data exata de cada reunião, e o IPCA acumulado de 12 meses foi vinculado considerando o mês anterior à publicação. Essa integração garantiu que não houvesse falhas de alinhamento temporal entre os textos e os indicadores.

14.5 – RESULTADOS ESPERADOS

Com a base pronta, a metodologia seguiu para a análise exploratória, onde foram estudados os prazos de publicação, a extensão dos documentos e a distribuição dos valores macroeconômicos. A partir daí, foi possível preparar o terreno para a criação da variável alvo, que classifica cada reunião em três categorias: hawkish quando a Selic sobe, neutral quando se mantém e dovish quando cai. Essa transformação permitiu estruturar o problema como uma tarefa de classificação em aprendizado de máquina.
O modelo foi então desenvolvido em Python, utilizando bibliotecas de PLN e redes neurais, com suporte do TensorFlow. Essa escolha foi feita pela flexibilidade da ferramenta e pela disponibilidade de modelos pré-treinados que podem ser adaptados para lidar com textos longos como os das atas do COPOM.
A coleta trouxe 253 reuniões registradas entre 1998 e 2025, mas após o recorte temporal a base final ficou com 238 reuniões dentro do regime de metas de inflação. Esse ajuste foi importante porque estabilizou o prazo de publicação, que passou a variar de 2 a 15 dias, com média em torno de 7,4 dias. O recorte eliminou distorções de períodos anteriores e deixou os dados mais consistentes para análise.
Na etapa de limpeza textual, o tamanho médio das atas caiu de quase 47 mil para cerca de 32 mil caracteres, uma redução de 22%. Isso confirma que grande parte do conteúdo inicial era ruído de formatação e HTML. Os comunicados, por sua vez, mantiveram tamanho bem menor, em torno de 2 mil caracteres, o que os torna complementares às atas. A criação do “texto consolidado” uniu esses dois tipos de documento em uma única fonte de informação para cada reunião, simplificando a análise.
As séries macroeconômicas também mostraram boa qualidade. O IPCA acumulado em 12 meses oscilou em média 6,2%, com forte volatilidade ao longo do período, o que garante variabilidade suficiente para treinar modelos. A Selic, por sua vez, apresentou ciclos bem definidos, com quedas e altas ao longo dos anos, reforçando sua importância como variável de contexto e de validação do modelo.
A integração final dos dados gerou um conjunto completo, sem valores ausentes nas variáveis-chave. Além disso, foi criada uma coluna de rótulo que classifica cada reunião em três categorias: hawkish quando a Selic subiu nos seis meses seguintes, neutral quando se manteve e dovish quando caiu. Essa estrutura transforma o problema em uma tarefa clara de classificação, adequada para aplicar modelos de aprendizado de máquina.
Em resumo, os resultados da preparação confirmam que a base está consistente, equilibrada e com boa representatividade para capturar padrões de linguagem nas atas e comunicados. Esse é o ponto de partida para a modelagem, que poderá mostrar se a comunicação do COPOM antecipa de forma confiável a direção da política monetária.

14.6 – IMPACTO

Os resultados esperados com este projeto vão além do aspecto acadêmico, pois demonstram como a análise de linguagem natural pode ser aplicada em situações reais do mercado financeiro. Ao transformar as atas e comunicados do COPOM em informações mais acessíveis e em indicadores preditivos, abre-se espaço para que instituições financeiras tomem decisões mais rápidas e embasadas.
No caso do Banco do Brasil, usado como referência no estudo, a aplicação prática seria direta. O banco poderia antecipar tendências de inflação e de política monetária e, com isso, ajustar sua gestão de crédito, as taxas oferecidas aos clientes e as estratégias de tesouraria. Esse tipo de inteligência também fortalece a atuação da BB Asset Management, ao fornecer sinais adicionais para orientar fundos de investimento e operações no mercado.
Outro impacto importante é mostrar que ferramentas de ciência de dados podem reduzir a dependência da leitura manual de documentos longos e técnicos. Isso não apenas economiza tempo dos analistas, mas também garante consistência na interpretação, evitando que nuances de linguagem passem despercebidas.
Por fim, o projeto evidencia o potencial de aproximar a análise econômica da tecnologia. Ao aplicar modelos de aprendizado de máquina sobre textos oficiais, demonstra-se que é possível criar soluções práticas que aumentam a competitividade das instituições e oferecem valor tanto para o setor financeiro quanto para a sociedade, já que decisões monetárias mais bem interpretadas podem refletir em políticas públicas e no planejamento econômico do país.

14.7 – CONCLUSÃO

O projeto mostrou que é possível transformar documentos extensos e técnicos, como as atas e comunicados do COPOM, em informações acessíveis e úteis para apoiar decisões econômicas. A partir da coleta sistemática de dados, da limpeza textual e da integração com séries históricas da Selic e do IPCA, foi construída uma base consistente e adequada para análise com técnicas de processamento de linguagem natural.
Os resultados da preparação demonstraram que as atas, mesmo sendo longas e complexas, carregam padrões de linguagem que podem antecipar movimentos de política monetária. A criação de um rótulo simples, dividindo as decisões em hawkish, neutral e dovish, possibilitou enquadrar o problema como uma tarefa clara de classificação. Isso abriu caminho para aplicar modelos de aprendizado de máquina capazes de extrair dessas nuances sinais úteis para o mercado.
Mais do que comprovar a viabilidade técnica, o estudo evidenciou o potencial de uso prático dessa abordagem. Instituições financeiras podem se beneficiar ao automatizar a leitura dos documentos do COPOM, economizando tempo, reduzindo subjetividade e ganhando velocidade para reagir a mudanças no cenário econômico. Para o Banco do Brasil, em especial, a aplicação direta seria no ajuste de crédito, gestão de tesouraria e definição de estratégias de investimento.
Assim, a principal conclusão é que a combinação entre ciência de dados e economia pode gerar valor real, aproximando a análise monetária da tecnologia e criando ferramentas que tornam a interpretação das decisões do COPOM mais ágil, precisa e útil para diferentes agentes do mercado.

15 - REFERÊNCIAS

1. COPOM - https://www.bcb.gov.br/conteudo/dadosabertos/BCBDeinf/elements_copom.html#/

2. BACEN TIME SERIES - https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries

3. GitHub - Projeto Aplicado 2: https://github.com/guilhermersduarte/Projeto-Aplicado2-Grupo19

4. Taxa SELIC - https://www.bcb.gov.br/controleinflacao/taxaselic%20?modalAberto=administracao-da-taxa-selic

5. GÉRON, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. 2. ed. Sebastopol, CA: O'Reilly Media, 2019.

6. Dr. Ernesto Lee. Build an NLP Model for Sentiment Analysis Using TensorFlow in 10 Minutes. Medium. Retrieved October 20, 2025, from https://drlee.io/build-an-nlp-model-for-sentiment-analysis-using-tensorflow-in-10-minutes-a6d3de84b17f

16 - APRESENTAÇÃO

O Vídeo de apresentação do pode ser encontrado aqui:

https://www.youtube.com/watch?v=KarV3ZgmUiU