

## Análise de Sentimentos em textos sobre o BNDES

Agosto de 2022

Gustavo Dutra Martins

Pontifícia Universidade Católica do Rio de Janeiro

Projeto Final do Curso de Pós Graduação "Business Intelligence Master"



# Análise de Sentimentos em Textos sobre o BNDES

Aluno: [Gustavo Dutra Martins](#)

Orientador: [Leonardo Mendoza](#).

Trabalho apresentado ao curso [BI MASTER](#) como pré-requisito para conclusão de curso e obtenção de crédito na disciplina "Projetos de Sistemas Inteligentes de Apoio à Decisão".

- [Link para o código](#).

## Resumo

O BNDES (Banco Nacional de Desenvolvimento Econômico e Social) possui indicadores estratégicos para medir o sentimento dos artigos e postagens publicadas nas redes sociais (Facebook, Twitter, Instagram etc.) ou na imprensa. Atualmente existe uma empresa contratada para realizar o clipping e medição do sentimento (positivo, neutro e negativo).

A proposta do trabalho é utilizar os textos obtidos pelas ferramentas de clipping e suas respectivas análises de sentimento para treinar uma rede neural capaz de avaliar o sentimento de textos postados sobre o BNDES na internet.

## Abstract

BNDES (National Bank for Economic and Social Development) has strategic indicators to measure the sentiment of articles and publications published on social networks (Facebook, Twitter, Instagram, etc.) or in the press. Currently there is a company carried out to carry out the clipping and negative of the sentiment (positive, neutral and negative).

The proposal of the work is to use the texts by the clipping tools and their validly evaluated sentiment for a neural network capable of evaluating the sentiment of texts posted on the internet about BNDES.

## 1. Introdução

O BNDES como instituição pública responsável por investimentos em setores relacionados ao desenvolvimento econômico e social do país tem alta exposição na imprensa e nas mídias sociais. A polarização dos sentimentos nas publicações é considerável:

- As reportagens sobre a atuação do banco são em sua maioria neutras porque relatam os objetivos de programas e linhas de crédito criadas.
- Os responsáveis pelo investimento e beneficiados diretos ou indiretos pelas linhas de financiamento normalmente tem opiniões positivas.
- A oposição ao governo geralmente tem opiniões negativas ou entende que o investimento deveria priorizar outro tipo de atividade.
- Existem os grupos que criticam a existência do banco, a remuneração dos funcionários e condenam qualquer tipo de ação realizada, gerando mensagens negativas que buscam prejudicar a imagem do BNDES.

Priorizada pela alta hierarquia do BNDES, a melhora da imagem externa é um indicador estratégico, sendo a análise de sentimentos sobre artigos e posts publicados uma das variáveis utilizadas para medição.

Atualmente o BNDES possui contrato com uma empresa que realiza o clipping e análise manual de sentimentos. Nosso projeto utilizou os textos extraídos e os rótulos atribuídos (positivo, neutro, negativo) para treinar os diversos modelos.

## 2. Modelagem

### 2.1 - Carregamento dos Textos

A equipe de comunicação do BNDES encaminhou planilhas com artigos da imprensa, mas estes não continham o conteúdo necessário para o processamento, somente os links. Para consultar o conteúdo foi necessário navegar pelas referências, sendo criado um programa com essa finalidade.

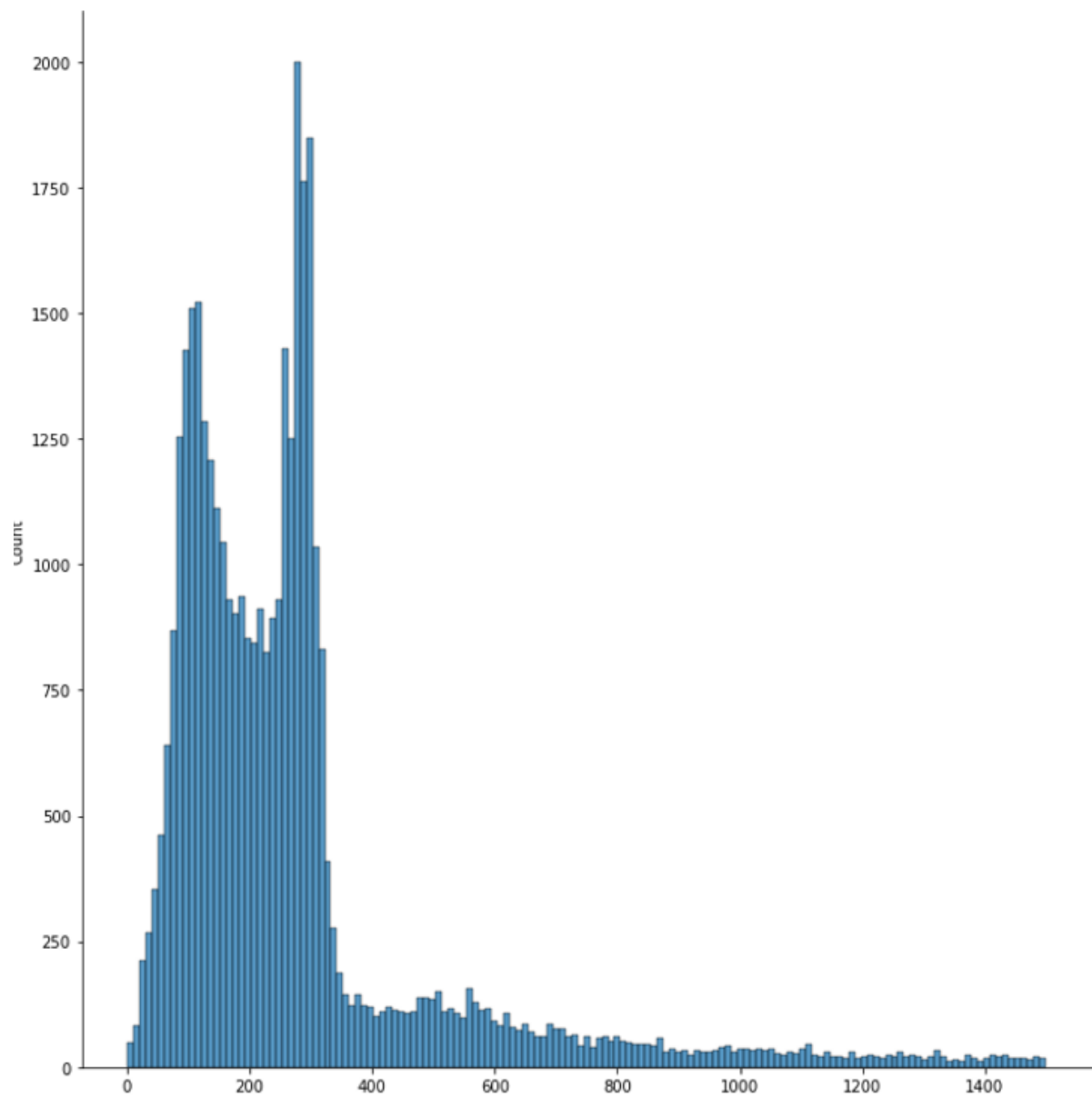
Os textos das mídias sociais estavam disponíveis em planilhas, sem necessidade de consultar links, no entanto continham colunas desnecessárias para nossa análise, linhas vazias ou repetidas (geralmente tweets repostados).

O carregamento extraiu 45.440 textos utilizando dados de 2020 (imprensa e mídias), 2021 (imprensa e mídias) e maio de 2022 (somente mídias sociais).

### 2.2 Análise dos Textos

Os artigos de imprensa são bem grandes, a maior parte continha aproximadamente 1.000 tokens, contudo existiam artigos com mais de 14.000 tokens. Já as publicações nas redes sociais possuem, normalmente, tamanho bem menor, variando entre 100 e 350 tokens.

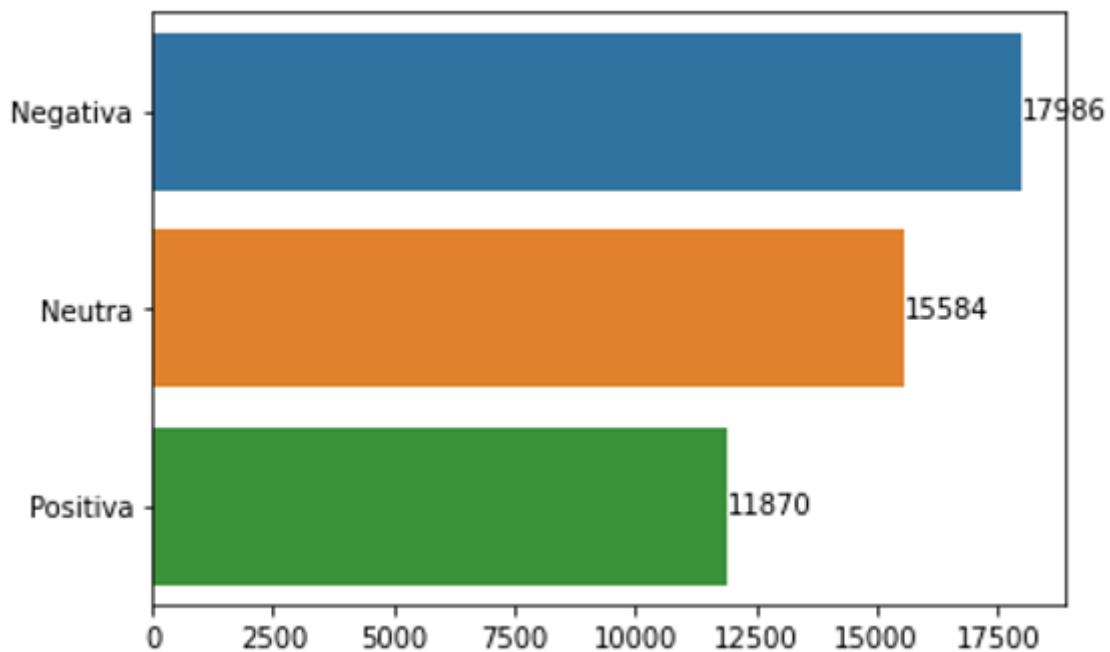
No gráfico abaixo limitamos o número máximo de tokens para facilitar a visualização.



### **Balanceamento**

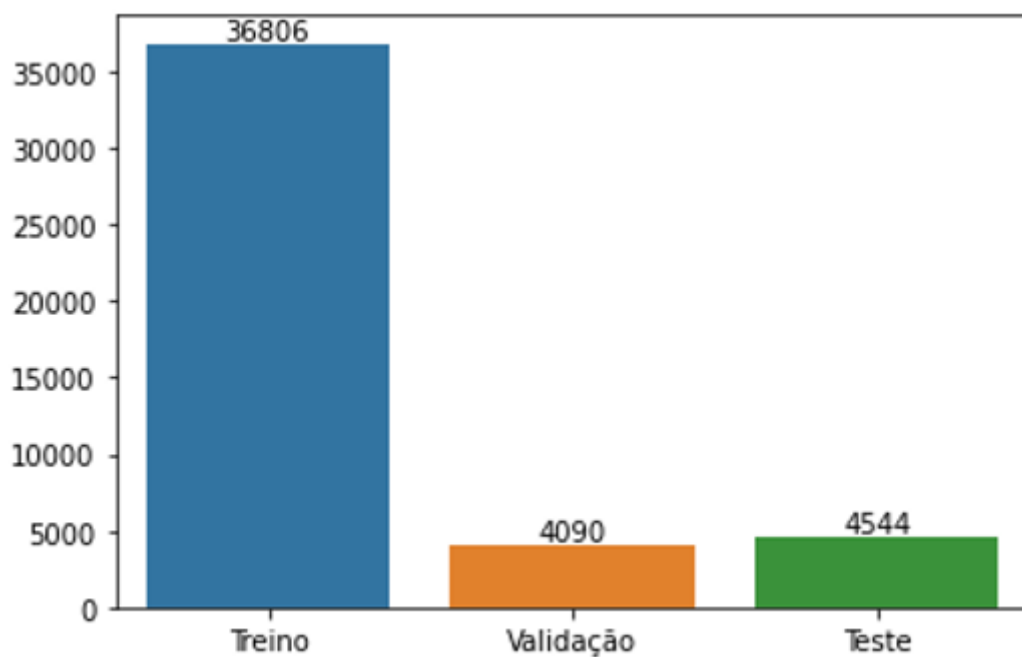
A base não estava completamente balanceada, mas a quantidade de textos adquiridos permitiu um treinamento adequado, no gráfico abaixo pode-se visualizar as proporções entre os sentimentos.

Importante ressaltar que artigos da imprensa são na sua maioria neutra, contudo elas compõem aproximadamente 15% do conjunto de textos.



### Separação das Bases

As bases de treino, validação e teste foram separadas conforme o gráfico abaixo.



## 2.3 Modelos

Foram utilizados dois tipos de modelos:

- LSTM – Long Short Term Memory
- BERT - Bidirectional Encoder Representations from Transformer

## Modelo 1 - LSTM

Para auxiliar na contextualização dos textos a rede neural utilizou uma camada de Embedding pré-treinado disponibilizados em português pelo NILC - Núcleo Interinstitucional de Linguística Computacional. Diferentes dimensões e tipos de embedding foram testados e apresentaremos posteriormente um comparativo entre os resultados.

As tarefas de pré-processamento aplicadas no texto foram as seguintes:

- Transformação em minúscula
- Remoção de stop words
- Remoção de palavras que não existem no Embedding
- Correção de erros comuns de escrita com alta incidência, exemplo: negociacao, inflacao, etc.
- Limitar o tamanho do texto a 1.000 tokens, truncando aqueles com tamanho superior ao limite estipulado.
- Lemetização e Stemming não foram aplicados.

A tabela abaixo contém o comparativo dos modelos testados, importante ressaltar os seguintes pontos:

- Nossos testes iniciais foram realizados com a base de imprensa porque os textos de mídias ainda não haviam sido disponibilizados.
- A métrica utilizada para comparação foi a acurácia, sendo que a nota final do modelo deveria ser baseada na base de teste, que não foi vista ainda pelo modelo.
- As LSTM's de dois níveis, bem como as LSTM's bidirecionais, em nossos testes, não conseguiram superar o modelo com uma camada de LSTM.

Textos	MAX_LENGTH	Embedding	Camada 1	Camada 2	Camada 3	Camada 4	Camada 5	Epochs	Acc Treino	Acc Validação	Acc Teste
Imprensa	1000	Glove/50	LSTM(150, RD=0.3)					1	78,57	78,66	78,5
Imprensa	1000	W2VEC 300	LSTM(150, RD=0.3)					1	79,16	78,66	78,37
Imprensa	1000	W2VEC 300	LSTM(150, RD=0.3)	GlobalMaxPooling1D				3	85,36	84,55	81,09
Imprensa	1000	W2VEC 300	LSTM(150, RD=0.3)	LSTM(150, RD=0.3)	GlobalMaxPooling1D			1	77,55	81,69	80,66
Imprensa	1000	W2VEC 300	LSTM(150, RD=0.3)	LSTM(150, RD=0.3)	Dense(150, ReLu)	GlobalMaxPooling1D		2	80,38	81,53	80,52
Imprensa	1000	W2VEC 300	LSTM(300, RD=0.3)	GlobalMaxPooling1D				5	82,19	82,17	81,23
Imprensa	1000	W2VEC 300	LSTM(300, RD=0.3)	GlobalMaxPooling1D	Dense(300, ReLu)			3	81,35	81,21	79,51
Imprensa	1000	W2VEC 300	LSTM(300, RD=0.3)	Dropout(0.3)	GlobalMaxPooling1D	Dense(300, ReLu)		5	90,54	84,24	83,24
Imprensa	1000	W2VEC 300	BILSTM(300, RD=0.3)	Dropout(0.3)	GlobalMaxPooling1D			5	85,34	83,12	80,8
Imprensa	1000	GLOVE 1000	LSTM(300, RD=0.3)	Dropout(0.3)	GlobalMaxPooling1D	Dense(300, ReLu)		7	88,17	80,41	78,8
Imprensa	1000	W2VEC 300	BILSTM(300, RD=0.3)	BILSTM(200, RD=0.3)	Dropout(0.3)	GlobalMaxPooling1D	Dense(300, ReLu)	1	78,25	81,53	80,52
Imprensa	1000	W2VEC 600	LSTM(300, RD=0.3)	Dropout(0.3)	GlobalMaxPooling1D	Dense(300, ReLu)		3	80,86	82,01	81,52
Imprensa Mencões (21) Menc (05/22) 45.440	1000	W2VEC 300	LSTM(300, RD=0.3)	Dropout(0.3)	GlobalMaxPooling1D	Dense(300, ReLu)		6	92,61	82,52	83,67
Imprensa Mencões (21) Menc (05/22) 45.440	400	W2VEC 300	LSTM(300, RD=0.3)	Dropout(0.3)	GlobalMaxPooling1D	Dense(300, ReLu)		1	70,83	69,73	71,46

Nosso melhor resultado com os textos de imprensa foi 83,24%, já com imprensa e mídias sociais conseguimos 83,67%, resultado bastante satisfatório. Segue a descrição do modelo com o melhor resultado:

1. Input de textos com tamanho de 1.000 posições
2. Embedding Word2Vec (300)
3. LSTM (300, RecurrentDropout=0,3)
4. Dropout (0,3)
5. GlobalMaxPooling1D
6. Dense(300, ReLu)
7. Dense (3)

Observações sobre os modelos testados:

- No teste que diminuiu o tamanho máximo tivemos perdas significativas de performance, alcançando somente 71,46 %.

- Embeddings com dimensão superior a 300 pioraram o resultado do modelo, com 1.000 posições tivemos o resultado de 78,8% e com 600 81,52%
- O resultado obtido pela rede construído é satisfatório, mas o tempo de treinamento é muito longo, em média 16 horas.

## Modelo 2 - BERT

Utilizamos o modelo BERT da Neuralmind (BERTimbau) e treinamos a última camada para que o modelo aprendesse a classificação de textos com sentimento positivo, negativo e neutro. Nosso primeiro teste não tratou os textos e conseguimos o resultado de aproximadamente 86% de acurácia. Posteriormente achamos que o resultado melhoraria com outros tratamentos de texto, mas foi inferior.

Abaixo mostramos a tabela comparando alguns testes que realizamos com o BERT.

Textos	MAX_LENGTH	Embedding	Camada 1	Camada 2	Camada 3	Camada 4	Camada 5	Epochs	Acc Treino	Acc Validação	Acc Teste	F1
Imprensa Menções (21) Menc (05/22) 45.440	512	BERT-BASE (Bertimbau)	Nenhum tratamento no texto					3	-	85,28	86,00	0,8567
Imprensa Menções (21) Menc (05/22) 45.440	512	BERT-BASE (Bertimbau)	Limando palavras não encontradas, removendo stop_words, pontuação e fazendo substituição de palavras comumente erradas.					3	-	83,32	84,00	0,8374
Imprensa Menções (21) Menc (05/22) 45.440	512	BERT-BASE (Bertimbau)	Limando URL's e substituindo BNDES por banco, já que o token BNDES não existe como token único					3	-	-	84,20	-
Imprensa Menções (21) Menc (05/22) 45.440	512	BERT-BASE (Bertimbau)	Recuperando palavras que incorretas, ex (negociacao, inflacao)					3	-	85,28	84,88	0,8457
Imprensa Menções (21) Menc (05/22) 45.440	512	BERT-BASE (Bertimbau)	Transformando tudo para caixa baixa (lowercase)					3	-	85,45	85,21	0,8484

Para os modelos BERT também medimos o F1-score.

Os modelos BERT eram treinados em 2 horas, 8 vezes menor que o tempo de treinamento das redes LSTM.

## 3. Resultados

Melhor acurácia conseguida com o modelo LSTM foi de 83,67%.

Melhor acurácia conseguida com o modelo BERT foi de 85,67%.

## 4. Conclusões

A jornada até os melhores modelos é singular, no papel de estudantes imaginamos diversas formas de melhorar o resultado e quase sempre nos enganamos, muitas vezes um pequeno detalhe encontramos um caminho diferente do planejado, mas que leva ao melhor resultado, é uma experiência ímpar!

Finalizo o curso ao qual dediquei muitas horas de estudo com a sensação de dever cumprido, aprendi muito e espero poder aprender ainda mais com novos desafios, levando o conhecimento adquirido para outras frentes de trabalho.

Os responsáveis na área de comunicação do BNDES ficaram satisfeitos com a performance do modelo, embora as perguntas finais fossem sobre o que poderia ser melhorado para alcançar um resultado superior. Alguns caminhos foram imaginados nesse sentido:

- Não foram realizados testes com embeddings próprios, mas existe um corpus com tamanho suficiente para treinar um embedding próprio. Nesse caminho é provável que o resultado das redes LSTM's melhorasse, contudo seria difícil superar o BERT.
- A construção de um BERT próprio seria um teste interessante porque acredito na possibilidade de melhorar consideravelmente o resultado. O BNDES por exemplo não é um token reconhecido pelo Tokenizador do BERTimbau, sendo gerado quatro tokens no caso de maiúscula [B, ##N, ##DE, ##S] e dois no caso de minúscula [b, ##ndes]. O BERTimbau também é treinado com textos sensíveis a caixa alta e baixa, já que sua principal função foi reconhecimento de entidades (NER), mas para a classificação de textos é provável que um modelo BERT treinado em caixa baixa tenha melhor acurácia.

Para instruções sobre como rodar os programas utilizados para treinamento e teste [clique aqui](#).

Por fim meus sinceros agradecimentos ao orientador Leonardo Mendoza que pacientemente me ajudou nos vários projetos e caminhos diferentes no curso até finalizarmos com a análise de sentimentos em textos publicados sobre o BNDES.

---

Matrícula: 202.190.058

Pontifícia Universidade Católica do Rio de Janeiro

Curso de Pós Graduação *Business Intelligence Master*