UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS EXATAS, NATURAIS E DA SAÚDE
DEPARTAMENTO DE COMPUTAÇÃO

ANDREI DA SILVA ESPÍRITO SANTO

PERCEPÇÃO DOS USUÁRIOS DE REDES SOCIAIS QUANTO AO FUTURO ECONÔMICO DO PAÍS PÓS-PANDEMIA

ALEGRE - ES
2024

PERCEPÇÃO DOS USUÁRIOS DE REDES SOCIAIS QUANTO AO FUTURO ECONÔMICO DO PAÍS PÓS-PANDEMIA

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Antonio Almeida de Barros Junior.

ALEGRE – ES

2024

ANDREI DA SILVA ESPÍRITO SANTO

PERCEPÇÃO DOS USUÁRIOS DE REDES SOCIAIS QUANTO AO FUTURO ECONÔMICO DO PAÍS PÓS-PANDEMIA

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 30 de setembro de 2024.

BANCADA EXAMINADORA

Documento assinado digitalmente
ANTONIO ALMEIDA DE BARROS JUNIOR
Data: 08/10/2024 17:48:33-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Antonio Almeida de Barros Junior
Universidade Federal do Espírito Santo

Documento assinado digitalmente
EDMAR HELL KAMPKE
Data: 08/10/2024 21:01:06-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Edmar Hell Kampke
Universidade Federal do Espírito Santo

Documento assinado digitalmente
MARCELO OTONE AGUIAR
Data: 09/10/2024 08:38:19-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Marcelo Otone Aguiar
Universidade Federal do Espírito Santo

RESUMO

As redes sociais têm se mostrado uma fonte relevante de opiniões e informações, especialmente no contexto da pandemia, que forçou os governos a adotarem medidas restritivas para conter o vírus, resultando em uma crise financeira que afetou a economia brasileira. Este trabalho busca analisar o sentimento expresso pelos usuários do *Twitter* em relação ao futuro econômico do Brasil no período pós-pandemia. Para isso, foi utilizada a técnica de Análise de Sentimentos, com uma base de dados rotulada manualmente como positiva ou negativa. O algoritmo de *Naive Bayes* foi escolhido para construir o classificador, utilizando essa base rotulada para o treinamento. O desempenho do algoritmo se mostrou significativo, com uma acurácia de 88,3% na fase de treinamento e 84,6% nos dados de teste, demonstrando boa capacidade de generalização. Quando aplicado à base de dados original, a acurácia foi de 77,2%. Esses resultados indicam que o modelo é capaz de identificar e classificar as opiniões dos usuários de maneira consistente, oferecendo uma compreensão das percepções da população sobre o futuro econômico do país. Esta abordagem é útil para captar o sentimento coletivo e as expectativas em um cenário econômico desafiador.

Palavras-chaves: Análise de Sentimentos, *Naive Bayes*, *Twitter*.

ABSTRACT

Social media has proven to be a valuable source of opinions and information, especially in the context of the pandemic, which forced governments to adopt restrictive measures to contain the virus, resulting in a financial crisis that impacted Brazil's economy. This study aims to analyze the sentiment expressed by Twitter users regarding the country's economic future in the post-pandemic period. Sentiment Analysis was applied using a dataset manually labeled as positive or negative. The Naive Bayes algorithm was chosen to build the classifier, using this labeled dataset for training. The algorithm's performance was significant, achieving an accuracy of 88.3% during training and 84.6% on test data, demonstrating a good generalization capability. When applied to the original dataset, the accuracy was 77.2%. These results indicate that the model can consistently identify and classify users' opinions, providing insights into the population's perceptions of the country's economic future. This approach is valuable for capturing collective sentiment and expectations in a challenging economic scenario.

Keywords: Sentiment Analysis, Naive Bayes, Twitter.

LISTA DE QUADROS

Quadro 1 - Palavras-chaves usadas na coleta de tweets. 12
Quadro 2 - Exemplo de estrutura de tweet em formato JSON. 13
Quadro 3 - Resultado do Pré-Processamento. 14

LISTA DE FIGURAS

Figura 1 - Modelo de treino do Algoritmo Supervisionado. 6
Figura 2 - Processo de Filtragem de tweets relevantes.....12
Figura 3 - Processo de desenvolvimento.....18

LISTA DE TABELAS

Tabela 1 - Acurácia calculada para cada um dos classificadores. 9
Tabela 2 - Sentimentos dos usuários do Twitter em relação aos Filmes.....10
Tabela 3 - Categorias de emoções.....11
Tabela 4 - Informações sobre etapa de coleta de tweets. 13
Tabela 5 - Desempenho em diferentes conjuntos de dados.....19
Tabela 6 - Resultados de acurácia obtidos em cada trabalho. 22

SUMÁRIO

1 INTRODUÇÃO 1
1.1 OBJETIVOS 3
1.1.1 Objetivo Geral 3
1.1.2 Objetivos Específicos 3
2 REVISÃO DE LITERATURA 4
2.1 Conceitos tecnológicos 4
2.1.1 Twitter 4
2.1.2 Linguagem de Programação 4
2.1.3 Python 4
2.1.4 Aprendizado de máquina 5
2.1.5 Funcionamento do Algoritmo *Naive Bayes* 6
2.1.6 Aplicações do *Naive Bayes* 7
2.2 Trabalhos Relacionados 8
2.2.1 Carvalho Filho (2014) 8
2.2.2 Tannús (2017) 9
2.2.3 Chaves (2017) 10
3 METODOLOGIA 12
3.1 Coletas de Dados 12
3.2 Pré-Processamento dos Dados 13
3.3 Ferramentas Utilizadas 15
3.4 Classificação dos Dados 15
3.4.1 Acurácia e Avaliação do Modelo 16
3.5 Análise de Sentimentos 17
3.6 Criando Modelo de Classificação 18
4 RESULTADOS 19

4.1 Discussão dos Resultados 20
4.2 Comparação com Outros Trabalhos 21
4.2.1 Carvalho Filho (2014) 21
4.2.2 Chaves (2017) 21
4.2.3 Tannús (2017) 21
4.2.4 Síntese Comparativa 22
5 CONCLUSÃO 23
6 REFERÊNCIAS 25
APÊNDICE A - CÓDIGO DE PRÉ-PROCESSAMENTO DOS *TWEETS* 27
APÊNDICE B: CÓDIGO DE TREINAMENTO E AVALIAÇÃO DO MODELO DE NAIVE BAYES 28

1 INTRODUÇÃO

O novo agente do coronavírus (nCoV-2019) foi descoberto em 31 de dezembro de 2019 após casos registrados na cidade chinesa de Wuhan. Com uma alta taxa de transmissão, o vírus se espalhou rapidamente pelo mundo. No dia 11 de março de 2020 a Organização Mundial da Saúde (OMS) declarou a Covid-19 como pandemia. Com a falta de um medicamento eficaz contra o vírus foram adotadas medidas de segurança para evitar a propagação do vírus.

De acordo com o Ministério da Saúde (2020), a Covid-19 chegou no Brasil no dia 26 de fevereiro 2020. Dando início a uma das maiores crises de saúde do país. Para controlar o avanço do vírus, muitos estados e municípios adotaram restrições mais rígidas, como isolamento social, redução dos horários de funcionamento do comércio e redução de atendimento ao público, provocando um impacto econômico na vida de muitos brasileiros. De acordo com informações do UOL (2021), em 2020, o PIB do Brasil registrou uma retração de 4,1%, a maior diminuição em 24 anos.

Com as restrições impostas pelos governantes de diversos países, observou-se um aumento no uso da internet em todo mundo, principalmente uso das redes sociais. As redes sociais têm sido o lugar onde as pessoas expressam as suas opiniões, sobre um determinado assunto ou acontecimento, gerando uma quantidade muito grande de informações.

Com o grande volume de dados gerados pelas redes sociais, é possível extrair informações sobre vários assuntos. Neste trabalho utilizaremos a rede social *Twitter*, como fonte para definir uma base de dados para pesquisa. O *Twitter* é uma rede social, aonde são publicados cerca de 656 milhões de *tweets* por dia (OSMAN, 2021). Os *tweets* são textos publicados pelos usuários no *Twitter* que possuem até 140 caracteres.

Com uma imensidão de dados trafegados diariamente na Internet, pessoas e empresas tem cada vez mais utilizando estes dados com o objetivo de tomada de decisão. A Internet tornou-se um ambiente com uma grande riqueza de informação, exemplo disso são as empresas que utilizam esses dados para identificar a opinião pública em relação a um produto, a uma marca ou qualquer ação que envolve a organização. Algumas utilizam técnicas de Análise de Sentimentos (AS) para saber a

opinião dos usuários em relação aos seus serviços ou produtos, com isso as empresas conseguem melhorar seus produtos ou serviços, tomando como base as opiniões do seu público.

A Análise de Sentimentos é uma técnica robusta na área de processamento de linguagem natural, que visa extrair e identificar a polaridade das emoções expressas em textos. Os algoritmos de aprendizagem de máquina e processamento de linguagem natural, são técnicas capazes de determinar se o conteúdo de um texto é positivo, negativo ou neutro. Ela é amplamente aplicada em diversas áreas, desde a análise de opiniões de consumidores em redes sociais até a avaliação de discursos políticos e análises de mercado. Através da análise de padrões linguísticos e contexto semântico, a Análise de Sentimentos proporciona uma compreensão mais profunda das emoções e opiniões expressas em grandes volumes de texto.

Diante desse cenário, torna-se fundamental compreender as percepções dos usuários de redes sociais, especialmente do *Twitter*, sobre o futuro econômico do Brasil no período pós-pandemia. A análise dessas percepções é crucial, pois, em tempos de incerteza econômica, as opiniões e sentimentos expressos nas redes sociais oferecem um reflexo valioso das expectativas da população em relação à economia.

Neste trabalho, a Análise de Sentimentos foi utilizada para processar e examinar dados coletados do *Twitter*, a fim de identificar se as opiniões dos usuários são predominantemente positivas ou negativas em relação ao futuro econômico do país. Isso permite uma análise mais precisa e abrangente das expectativas da população, fornecendo *insights* valiosos sobre o sentimento coletivo em um cenário econômico desafiador.

Assim, o objetivo deste trabalho foi verificar a percepção dos usuários de redes sociais quanto ao futuro econômico do Brasil no período pós-pandemia, utilizando uma base de dados criada a partir de dados extraídos do *Twitter*. A Análise de Sentimentos foi aplicada para realizar essa verificação e compreender as perspectivas expressas nas redes sociais.

1.1 OBJETIVOS

A seguir são apresentados o objetivo geral e os objetivos específicos relacionados a este trabalho.

1.1.1 Objetivo Geral

Este trabalho tem o objetivo de verificar a percepção dos usuários de redes sociais, mais precisamente da rede social *Twitter,* quanto ao futuro econômico do país no período pós-pandemia, utilizando a técnica de Análise de Sentimentos com um algoritmo de aprendizado de máquina para classificar e interpretar as emoções expressas nos *tweets* coletados.

1.1.2 Objetivos Específicos

- Construir uma base de dados com informações coletadas do *Twitter*.
    - Definir palavras-chave relacionadas ao futuro econômico do Brasil no período pós-pandemia. Utilizar a *API (Application Programming Interface)* do *Twitter* para coletar os *tweets*, filtrando-os de acordo com as palavras-chave. A base de dados resultante incluirá informações essenciais, como o texto dos *tweets*.
- Aplicar técnica de Análise de Sentimentos na base de dados criada.
    - Neste passo, será realizada a Análise de Sentimentos utilizando algoritmo de aprendizado de máquina, como o *Naive Bayes* (NB) para classificar os *tweets* coletados em sentimentos positivos, negativos, permitindo identificar as emoções expressas pelos usuários em relação ao futuro econômico.
- Analisar os resultados obtidos quanto ao futuro econômico do país no período pós pandemia.
    - Avaliar as opiniões dos usuários com base na Análise de Sentimentos, considerando a acurácia do modelo utilizado, como Naive *Bayes*. Identificar a proporção de *tweets* classificados como positivos ou negativos e a precisão dessa classificação. Essa análise permitirá compreender melhor as expectativas e preocupações da população em relação à economia brasileira após a pandemia

2 REVISÃO DE LITERATURA

A revisão da literatura é essencial para contextualizar o trabalho e compreender os conceitos que fundamentam a pesquisa realizada.

2.1 Conceitos tecnológicos

Nessa seção são apresentados brevemente os conceitos tecnológicos que apoiaram o desenvolvimento deste trabalho.

2.1.1 Twitter

A rede social *Twitter* foi criada em 2006 e hoje soma mais de 1,3 bilhão de contas ativas, através dela, você pode compartilhar opiniões sobre um determinado assunto ou até divulgar produto, ou serviço (*TWITTER*, 2020).

As publicações feitas no *Twitter*, são chamadas de *tweets*, que representam as mensagens que os usuários postam, podendo ter até 140 caracteres. Um diferencial do *Twitter* são as *hashtags*, que consiste em uma palavra-chave antecipado pelo símbolo #. As *hashtags* indicam que o *tweet* faz referência a um certo tópico.

O *Twitter* disponibiliza, *APIs (Application Programming Interfaces)* que permitem o acesso aos *tweetes* que os usuários compartilham através de palavras chaves, o que torna possível o acesso a publicações recentes dos últimos dias ou em publicações em tempo real, que temos acesso na hora da publicação (*TWITTER*, 2020).

2.1.2 Linguagem de Programação

É uma linguagem formal que especifica um conjunto de instruções que permite que o programador escreva um conjunto de ordem, para criar um programa (VALINOR, 2023).

2.1.3 Python

Dentre as diversas linguagens de programação, o *Python* emerge como uma linguagem de destaque, oferecendo uma abordagem única e acessível para desenvolvedores. Criada por Guido Van Rossum no início da década de 90, *Python*

foi moldada com a missão de tornar o código elegante, legível e eficiente (Portal Desenvolvedor, 2023).

O poder do *Python* reside não apenas em sua simplicidade, mas também na robusta comunidade que a sustenta, colaborando e compartilhando conhecimento para impulsionar constantemente a evolução da linguagem. Em projetos de análise de sentimentos, como a investigação de *tweets*, *Python* se destaca como uma ferramenta capaz de proporcionar não apenas eficiência no processamento, mas também uma abordagem amigável para os desenvolvedores, incentivando a exploração e a inovação.

2.1.4 Aprendizado de máquina

Aprendizado de máquina também conhecido como *Machine Learning* (ML) é uma área de Inteligência Artificial (IA) que estuda o desenvolvimento de sistemas capazes de aprender com os dados e tomar decisões sem intervenção humana.

No mundo da aprendizagem de máquina, há diversas abordagens que ajudam a resolver problemas de maneiras diferentes. Uma delas é a Aprendizagem de Máquina Não Supervisionada, onde os modelos são treinados em conjuntos de dados sem rótulos, explorando padrões e estruturas sem a orientação de saídas conhecidas (NATALI, 2021). Adicionalmente, a Aprendizagem de Máquina por Reforço é outra abordagem intrigante, em que os algoritmos aprendem através de interações com um ambiente, recebendo *feedbacks* em forma de recompensas ou penalidades (ENGINEERING, 2022).

Por fim, a Aprendizagem Semi-Supervisionada combina elementos das abordagens supervisionada e não supervisionada, usando conjuntos de dados parcialmente rotulados para treinamento. A Aprendizagem Profunda (*Deep Learning*) é uma subcategoria que se destaca por sua capacidade de processar dados complexos, utilizando redes neurais profundas para realizar tarefas complexas, como reconhecimento de imagem e processamento de linguagem natural (ENGINEERING, 2022).

É importante ressaltar que, para o desenvolvimento deste trabalho, utilizou-se a Aprendizagem de Máquina Supervisionada. Os algoritmos de aprendizado supervisionado precisam de uma base de dados rotulada para relacionar a entrada com a saída. Isso significa que o algoritmo recebe exemplos de dados corretos e

incorretos durante o treinamento, para que ele aprenda a prever a saída correta com base nas entradas fornecidas. A Figura 1 ilustra o processo dos algoritmos supervisionados.

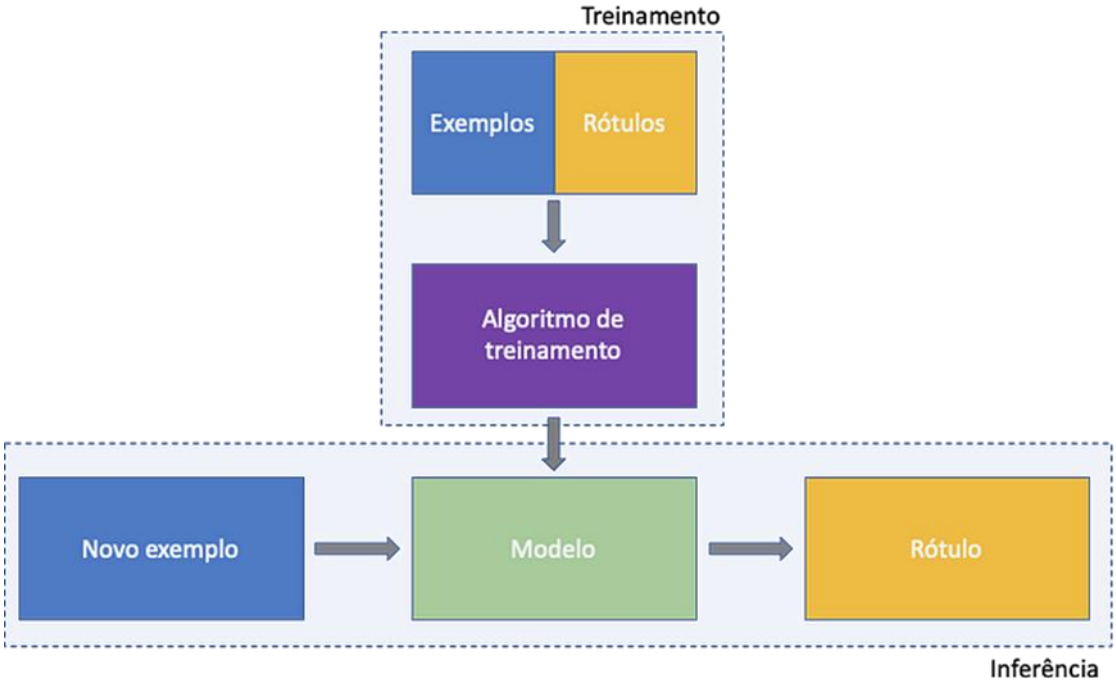

Figura 1 - Modelo de treino do Algoritmo Supervisionado.

Fonte: Araujo (2020).

A escolha dessa técnica foi motivada por diversas razões estratégicas e técnicas, alinhadas aos objetivos de classificação de sentimentos dos usuários em redes sociais.

2.1.5 Funcionamento do Algoritmo *Naive Bayes*

O *Naive Bayes* é um dos algoritmos mais amplamente utilizados em aprendizagem supervisionada, especialmente em tarefas de classificação de textos. Ele se baseia no Teorema de *Bayes*, que calcula a probabilidade de uma determinada classe (como positivo ou negativo) dada uma série de características ou atributos (neste caso, as palavras dos *tweets*).

O Teorema de *Bayes* é representado pela seguinte expressão:

$$P(A|B) = (P(B|A) \cdot P(A)) / P(B)$$

Sendo que:

- *P(A/B)* é a probabilidade da classe *A* (por exemplo, "positivo") dado o conjunto de características *B* (por exemplo, as palavras do *tweet*), chamada de probabilidade posterior.
- *P(B/A)* é a probabilidade de encontrar as características *B* dada a classe *A*, chamada de probabilidade de verossimilhança.
- *P(A)* é a probabilidade a priori da classe *A*, ou seja, a probabilidade de um *tweet* ser positivo independentemente das palavras.
- *P(B)* é a probabilidade a priori das características *B*, ou seja, a probabilidade de observar o conjunto de palavras independentemente da classe.

Na aplicação do *Naive Bayes*, o algoritmo usa essa fórmula para calcular a probabilidade condicional de um *tweet* pertencer a uma determinada classe com base nas palavras que ele contém.

2.1.6 Aplicações do *Naive Bayes*

O *Naive Bayes* tem uma ampla gama de aplicações, especialmente em problemas de classificação de texto e processamento de linguagem natural (PLN). Um dos exemplos mais clássicos para ilustrar a importância do *Naive Bayes* é o classificador de *spam*. Nesse contexto, o algoritmo analisa e-mails e tenta determinar se eles são spam ou não, com base nas informações e na estrutura do e-mail. Isso é feito através da análise de padrões de palavras que frequentemente aparecem em e-mails de *spam*, como termos promocionais ou links suspeitos (Sacramento, 2021).

Além disso, o *Naive Bayes* é amplamente utilizado em classificadores de sentimento, como o aplicado neste trabalho. Nesses casos, o algoritmo analisa textos e tenta identificar a emoção expressa, classificando-os como neutro, positivo ou negativo. Essa abordagem é particularmente útil em análise de redes sociais, onde há uma grande quantidade de opiniões e sentimentos a serem analisados e categorizados (Sacramento, 2021).

Contudo, as aplicações do *Naive Bayes* não se limitam à classificação de texto. Ele também é utilizado em aplicações de *Data Science* voltadas para a saúde, como sistemas que determinam se um paciente tem uma doença ou não, com base em características como sintomas e histórico médico. Nesse caso, o *Naive Bayes* é capaz de calcular a probabilidade de uma doença com base nas características

observadas no paciente, oferecendo suporte à tomada de decisões clínicas (Sacramento, 2021).

2.2 Trabalhos Relacionados

Trabalhos anteriores que também aplicaram Aprendizagem Supervisionada para análise de sentimentos, como os de Carvalho Filho (2014) e Tannús (2017), comprovaram que essa técnica é altamente eficaz para classificar sentimentos em redes sociais. Esses estudos demonstraram que, ao utilizar dados rotulados para o treinamento, algoritmos de aprendizado supervisionado, como o *Naive Bayes*, conseguiram atingir altos níveis de acurácia, o que incentivou o uso dessa abordagem no presente trabalho.

Para nortear o presente estudo, foram realizados alguns estudos envolvendo o tema proposto. Dentre os diversos trabalhos relacionados à Análise de Sentimentos disponíveis na literatura, três foram escolhidos por terem similaridades com o que foi desenvolvido nesta pesquisa.

2.2.1 Carvalho Filho (2014)

O trabalho de Carvalho Filho (2014), intitulado de “Mineração de textos: Análise de Sentimento utilizando *tweets* referentes à copa do mundo 2014” identificou o sentimento da população sobre a Copa do Mundo, através de postagem no *Twitter*. Para coletar os dados sobre as postagens relacionadas ao tema Copa do Mundo, o autor criou um *script* na linguagem *Python* utilizando *hashtags* relacionadas a Copa do Mundo. Na etapa de pré-processamento o autor removeu os caracteres especiais e *stopwords*, por meio da *NLTK* (*Natural Language ToolKit*).

Na técnica de Análise de Sentimento, o autor utilizou o algoritmo de *Naive Bayes*. Separou-se uma pequena parte dos dados originais e classificou manualmente. Após esta etapa, os dados foram separados em dois subconjuntos chamados de treino e teste. A partir do conjunto de treino, criou-se um modelo de classificação para ser aplicado na base de dados total. Com isso, o autor analisou a precisão dos resultados obtidos nas categorias definidas como negativo, positivo e neutro. O modelo gerado pelo autor, classificou 85,57% das instâncias corretamente.

O trabalho de Carvalho Filho (2014) concluiu que o uso da Análise de Sentimento por meio do algoritmo *Naive Bayes* foi eficaz na identificação do

sentimento da população em relação à Copa do Mundo de 2014, a partir de postagens no Twitter. Com uma acurácia de 85,57%**,** o modelo desenvolvido demonstrou capacidade de classificar corretamente a maioria das instâncias como positivas, negativas ou neutras. A metodologia empregada, que incluiu o pré-processamento dos dados e a divisão em conjuntos de treino e teste, evidenciou que o uso de técnicas de aprendizado de máquina pode ser uma ferramenta poderosa para analisar grandes volumes de dados textuais de redes sociais. Além disso, a pesquisa destacou a importância de um pré-processamento rigoroso, como a remoção de *stopwords* e caracteres especiais, para melhorar a precisão do modelo. Em suma, o estudo validou a viabilidade de algoritmos supervisionados como o *Naive Bayes* para a classificação de sentimentos em eventos de grande repercussão pública, como a Copa do Mundo.

2.2.2 Tannús (2017)

O trabalho desenvolvido por Tannús (2017) verificou a existência de correlação entre os vencedores do Oscar de 2017 e os sentimentos expressos pelas pessoas que utilizam *Twitter*, em relação aos filmes indicados. Foram coletados *tweets* da ferramenta *GetOldTweets*, que contém diversos recursos não oferecidos pela API do *Twitter*.

Após a etapa de coleta dos dados, o autor selecionou uma amostra aleatória da base de dados original e a separou para criar uma base rotulada, sendo utilizada no modelo de aprendizagem supervisionado. Para cada *tweet* foi atribuído um rótulo de sentimento, positivo, negativo ou neutro. Foram testados três classificadores diferentes, *TextBlob*, *Naive Bayes* e *Sentiment140*. Cada um foi avaliado para saber qual seria usado como classificador no trabalho. A Tabela 1 apresenta a acurácia dos classificadores utilizados. Para tanto, o autor optou pelo classificador *Naive Bayes*, por ser o que apresentou um melhor desempenho.

Tabela 1 - Acurácia calculada para cada um dos classificadores.

Classificador	***Naive Bayes***	***TextBlob***	***Sentiment140***
Acurácia	74,1%	63,9%	26,2%

Fonte: Produzida pelo autor com base nos dados de TANNÚS (2017).

Os resultados da aplicação do algoritmo *Naive Bayes* em toda a base de dados podem ser observados na Tabela 2. Observa-se que o filme "*La La Land*" teve o melhor desempenho em relação aos sentimentos expressos no *Twitter*. O filme também foi o mais premiado no Oscar de 2017, levando 6 estatuetas (G1, 2017).

Tabela 2 - Sentimentos dos usuários do Twitter em relação aos Filmes.

Filme	**Sentimentos**					
	Positivos		**Negativos**		**Neutros**	
Arrival	68.485	51%	27.002	20%	39.727	29%
Fences	20.199	48%	8.687	21%	12.796	31%
HacksawRidge	31.998	66%	6.380	13%	10.362	21%
Hellor High Water	7.077	53%	2.064	15%	4.179	31%
Hidden Figures	77.314	56%	16.212	12%	43.625	32%
La La Land	104.507	43%	57.237	23%	82.469	34%
Lion	75.906	50%	36.978	25%	37.757	25%
Manchester bytheSea	14.879	52%	5.410	19%	8.312	29%
Moonlight	49.668	55%	17.130	19%	23.480	26%
Total	**450.033**	**50,6%**	**177.100**	**19,9%**	**262.707**	**29,5%**

Fonte: TANNÚS (2017).

2.2.3 Chaves (2017)

A pesquisa realizada por Chaves (2017), intitulada “Análise de Sentimentos em redes sociais: uma abordagem baseada em emoções”, propôs analisar o sentimento expresso nas postagens em grupos de redes sociais. O autor construiu um *software* para auxiliar as pesquisas por conteúdo em postagens do *Facebook*. Ainda foi incluída a funcionalidade de analisar o sentimento contidos nas postagens.

Para realizar a análise de sentimento no software, o autor utilizou *WordNet-Affect*, que é uma extensão do léxico, definido como o conjunto de palavras pertencentes a uma determinada língua, incluindo suas definições e significados, conforme explicado por Souza (2021). O *WordNet-Affect* serve como uma base de palavras para identificar emoções em três páginas do *Facebook*, contando com cerca de 5 mil termos divididos em seis grupos e categorias de emoções, conforme apresentado na Tabela 3.

Tabela 3 - Categorias de emoções.

Categoria	**Palavras**
Alegria	animação, ânimo, comédia
Desgosto	maldoso, chatear, abominável
Medo	desespero, escuridão, friamente
Raiva	assassinar, bravejar, detestar
Surpresa	chocante,espanto,impressionado
Tristeza	cabisbaixo, chorão, culpa

Fonte: Chaves (2017).

3 METODOLOGIA

Para o desenvolvimento deste trabalho, modelos foram treinados com conjuntos de dados rotulados, capacitando-os para prever saídas de novos dados. Essa abordagem foi escolhida de forma criteriosa e implementada para a Análise de Sentimentos em *tweets*, mostrando eficácia em contextos específicos de classificação e interpretação de dados.

3.1 Coletas de Dados

Para a coleta de dados deste trabalho, o *Twitter* foi selecionado como fonte primária, e um *script* em *Python* foi empregado nessa etapa. O script foi desenvolvido para interagir com a *API* do *Twitter*, permitindo a busca específica de *tweets* por meio da utilização de palavras-chave. As palavras-chave utilizadas durante a coleta estão listadas no Quadro 1.

Quadro 1 - Palavras-chaves usadas na coleta de tweets.

Palavras-chave
futuro econômico, recuperação econômica, crise econômica, pandemia e economia, impacto econômico da pandemia, recessão econômica, desemprego, inflação, PIB

Fonte: O autor.

Adicionalmente, foram aplicados critérios de filtragem, como geolocalização, idioma e outros, para refinar a busca dos *tweets* relevantes para a Análise de Sentimentos sobre o futuro econômico pós-pandemia como mostrado na Figura 2.

.

```
# Buscar tweets
query = " OR ".join(keywords) + " OR " + " OR ".join(hashtags)
tweets = tweepy.Cursor(api.search_tweets, q=query, lang="pt")
```

Figura 2 - Processo de Filtragem de tweets relevantes

Fonte: O autor.

Devido à limitação de coleta imposta pela *API* do *Twitter*, o volume máximo de dados coletados foi de 500.000 *tweets* ao longo de 30 dias. Para armazenar essa quantidade de informações, foi utilizado o *MongoDB*, um banco de dados *NoSQL*, conhecido por sua flexibilidade e capacidade de lidar com grandes volumes de dados. Os *tweets* foram armazenados no formato *JSON* (*JavaScript Object Notation*), conforme fornecido pela *API* do *Twitter*, contendo atributos como o texto associado a cada *tweet*. Conforme apresentado na Tabela 4, os dados foram coletados entre 1 de maio e 30 de maio de 2021, totalizando 500.000 *tweets*.

Tabela 4 - Informações sobre etapa de coleta de tweets.

Descrição	**Detalhes**
Tweets coletados	500.000
Número de dias	30
Período de Coleta	1 de maio a 30 de maio de 2021

Fonte: O autor.

No Quadro 2, apresentamos um exemplo de como os *tweets* são estruturados em formato JSON, com informações sobre o conteúdo textual, dados do usuário e outras metainformações fornecidas pela *API*.

Quadro 2 - Exemplo de estrutura de tweet em formato JSON.

Atributo	**Valor**
id	987654321012345679
text	Não aguento mais essa pandemia e a economia so afunda 😩... governo n faz nd, só piora.
user.screen_name	anasouza_90
lang	pt

Fonte: O autor.

3.2 Pré-Processamento dos Dados

O pré-processamento dos dados desempenha um papel crucial na Análise de Sentimentos de conteúdos que vêm de plataformas sociais, como o *Twitter*. Durante esta etapa, são realizadas diversas ações para preparar os dados e torná-los adequados à análise. Conforme apresentado no Quadro 3, realizamos a remoção de

usuários e *links* contidos nos tweets. Esse processo contribui para simplificar o texto, eliminando elementos que não são relevantes para a análise de sentimentos.

Quadro 3 - Resultado do Pré-Processamento.

Antes	Depois
@randolfeap Poderia falar sobre como está a economia Europa pós COVID em comparação ao Brasil, aqui está horrível!!! https://t.co/v8pAxfbfqh	Poderia falar sobre como está a economia Europa pós COVID em comparação ao Brasil, aqui está horrível!!!

Fonte: O autor.

Além dessas ações, outros procedimentos comuns de pré-processamento incluem a eliminação de pontuações, caracteres especiais e emojis que não contribuem diretamente para a análise de sentimentos.

O Quadro 3 ilustra o processo de pré-processamento aplicado aos dados extraídos do *Twitter*. No exemplo apresentado, podemos observar a transformação de um *tweet* original, que continha menções a outros usuários e *links*, em uma versão mais simplificada e adequada para a Análise de Sentimentos. Antes do pré-processamento, o *tweet* incluía a menção ao usuário "@randolfeap" e um *link* externo, ambos irrelevantes para a classificação do sentimento. Após o pré-processamento, esses itens foram removidos, deixando apenas o conteúdo textual relevante para a análise.

Esse processo é essencial para garantir que o algoritmo de Análise de Sentimentos foque no conteúdo emocional da mensagem, sem ser influenciado por informações desnecessárias, como nomes de usuários e *URLs*, que não contribuem diretamente para a identificação do sentimento. Ao eliminar essas distrações, o modelo pode classificar com maior precisão as emoções expressas nas postagens.

3.3 Ferramentas Utilizadas

- ***Python***: A linguagem de programação *Python* foi escolhida por sua simplicidade e pela vasta quantidade de bibliotecas disponíveis para análise de dados e aprendizado de máquina. Foi utilizada na criação dos *scripts* que realizaram a coleta, pré-processamento e análise dos dados.
- ***Twitter API***: Para coletar os dados, utilizamos a *API* do *Twitter*, que permite o acesso programático aos *tweets* publicados na rede social. Por meio de palavras-chave específicas, foram coletados 500.000 *tweets* ao longo de 30 dias.
- ***MongoDB***: Os dados coletados foram armazenados em um banco de dados *MongoDB*, um sistema *NoSQL* que facilita o armazenamento de grandes volumes de dados não estruturados, como os *tweets*.
- ***Naive Bayes***: Para a análise de sentimentos, utilizamos o algoritmo de aprendizado supervisionado *Naive Bayes*, que foi treinado com uma base de dados rotulada manualmente e posteriormente aplicado aos *tweets* coletados.

3.4 Classificação dos Dados

Neste trabalho, o algoritmo *Naive Bayes* foi escolhido para o desenvolvimento da Análise de Sentimentos, pois apresentou os melhores resultados no estudo realizado por Carvalho Filho (2014**).**

A criação de uma base rotulada exige tempo e cuidado. Para este estudo, uma parte dos *tweets* coletados foi selecionada para serem rotulados manualmente. A classificação de cada *tweet* foi feita de forma manual, categorizando os textos como positivos ou negativos. Esses dados rotulados foram utilizados para treinar e validar o modelo *Naive Bayes,* utilizado na técnica de Análise de Sentimentos.

Ao utilizar essa base rotulada para treinar e testar o modelo *Naive Bayes*, foi possível validar e ajustar o algoritmo, garantindo sua capacidade de generalizar a Análise de Sentimentos para *tweets* não rotulados. Esse processo permitiu a identificação e classificação dos sentimentos expressos pelos usuários do *Twitter*,

aplicando as regras aprendidas durante o treinamento para classificar os *tweets* restantes.

Além disso, uma validação adicional foi realizada utilizando uma amostra de *tweets* extraída da base original já filtrada. O objetivo foi comparar os rótulos manuais com as previsões geradas pelo modelo, proporcionando uma análise prática de seu desempenho em um cenário real, complementando os dados já obtidos anteriormente.

3.4.1 Acurácia e Avaliação do Modelo

A acurácia é uma das métricas mais utilizadas para avaliar o desempenho de modelos de classificação, como o *Naive Bayes* utilizado neste trabalho. Ela indica a proporção de total de acertos feitos pelo modelo em relação ao número total de total de itens. Em outras palavras, a acurácia mede o quão bem o modelo consegue classificar corretamente os *tweets* como positivos ou negativos. A fórmula utilizada para calcular a acurácia é a seguinte:

```
Acurácia = Total de acertos / Total de itens
```

Onde:

- Total de acertos*:* Refere-se ao número de vezes que o modelo classificou corretamente um *tweet* (por exemplo, classificando corretamente um *tweet* positivo como positivo).
- Total de itens*:* Refere-se ao total de exemplos (*tweets*) que o modelo classificou, incluindo as previsões corretas e incorretas.

Durante o desenvolvimento deste modelo, a acurácia foi a única métrica de avaliação utilizada. Essencialmente, ela fornece uma visão geral da porcentagem de previsões corretas feitas pelo modelo em comparação com o número total de amostras avaliadas.

A análise da acurácia do modelo de classificação *Naive Bayes* nos permite avaliar quão bem o algoritmo está desempenhando a tarefa de classificar os sentimentos expressos nos *tweets*. Entender a precisão das previsões é crucial para determinar a confiabilidade do modelo na identificação correta das opiniões expressas.

3.5 Análise de Sentimentos

A análise de sentimento é uma área que nos permite transforma informações não-estruturadas, em dados estruturados com opiniões públicas sobre um determinado assunto (CONTROLF5, 2018). Para tanto, com o grande número de dados na internet, que se encontram desestruturados e desorganizados, a maior parte deles em formato de texto, a Análise de Sentimentos nos permite melhor compreender as informações não-estruturadas.

Ao processar e analisar grandes conjuntos de texto, essa abordagem identifica e categoriza as opiniões expressas, seja em redes sociais, fóruns, comentários de produtos ou artigos. Dessa forma, possibilita a extração de *insights* valiosos, revelando padrões, tendências e o sentimento predominante em relação a um determinado tópico ou produto.

Essa técnica se torna essencial para empresas, pesquisadores e até mesmo governos, pois oferece uma maneira de compreender o contexto e a percepção pública em meio à vastidão de informações disponíveis na internet, fornecendo uma base sólida para tomadas de decisão mais informadas e estratégicas.

3.6 Criando Modelo de Classificação

Com as etapas descritas nas seções 3.1, 3.2 e 3.4 concluídas, foi possível iniciar o desenvolvimento do modelo de classificação que utiliza o algoritmo *Naive Bayes*. Este modelo foi treinado para analisar e classificar as opiniões expressas nos *tweets* em categorias de sentimento, tais como positivo e negativo

A base rotulada foi particionada em dois conjuntos distintos: treino e teste. O conjunto de treino foi empregado na construção do modelo de classificação dos *tweets*, permitindo que o algoritmo aprendesse a discernir e classificar os sentimentos expressos nos mesmos. Posteriormente, o modelo de classificação gerado foi validado usando o conjunto de teste. Esta etapa consistiu na aplicação do modelo de classificação ao conjunto de teste, para avaliar sua eficácia na classificação dos sentimentos dos *tweets* como mostrado na Figura 3.

Adicionalmente, o modelo passou por uma validação prática em um cenário controlado. Durante essa etapa, uma amostra previamente rotulada foi utilizada para verificar a consistência das previsões geradas pelo algoritmo em comparação aos rótulos manuais. Essa abordagem complementa a avaliação do desempenho do modelo e também está representada no processo ilustrado pela Figura 3.

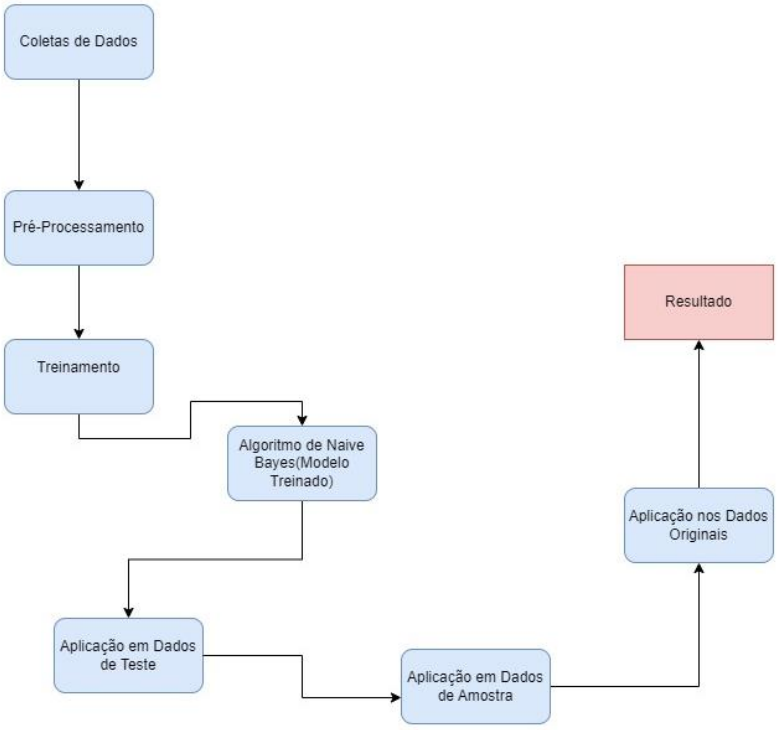

Figura 3 - Processo de desenvolvimento.

Fonte: O autor.

4 RESULTADOS

A avaliação do modelo de classificação utilizando *Naive Bayes* revelou métricas importantes ao longo de diferentes conjuntos de dados, conforme demonstrado na Tabela 5. Durante o treinamento, o modelo atingiu uma acurácia de 88,3% nos 40.000 *tweets* do conjunto de treino, indicando uma alta capacidade de generalização. Esse valor representa a porcentagem de *tweets* corretamente classificados como positivos ou negativos, com base nos dados rotulados manualmente. Nos 10.000 *tweets* do conjunto de teste, a acurácia foi ligeiramente reduzida, alcançando 84,6%, o que sugere que o modelo pode lidar adequadamente com novos conjuntos de dados. Na base de dados original, composta por 449.900 *tweets*, a acurácia caiu para 77,2%, refletindo os desafios de classificar novos dados em um cenário de aplicação real.

Tabela 5 - Desempenho em diferentes conjuntos de dados.

Métricas	**Dados de treino (40.000)**	**Dados de teste (10.000)**	**Base de dados Original (449.900)**	**Amostra (100)**
Acurácia	88,3%	84,6%	77,2%	79%
(Categoria Positiva)	78%	72%	66,7%	82%
(Categoria Negativa)	84%	78%	71,4%	76%

Fonte: O autor.

Ao analisar cada categoria separadamente, observou-se uma variação. Na categoria positiva, o modelo apresentou 78% de acurácia durante o treinamento, diminuindo para 72% nos dados de teste e 66,7% na base de dados original. Na categoria negativa, o desempenho foi mais elevado, com o modelo alcançando 84% de acurácia no treinamento, reduzindo para 78% nos dados de teste e 71,4% na base de dados original, conforme detalhado na Tabela 5. Esses números sugerem uma tendência de queda ao lidar com novos dados, especialmente em categorias onde o sentimento pode ser mais ambíguo ou difícil de definir com precisão.

Por outro lado, para a categoria negativa, o modelo obteve 84% de acurácia durante o treinamento, reduzindo para 78% nos dados de teste e 71,4% na base de

dados original, conforme evidenciado na Tabela 6. Esses valores indicam uma certa consistência, mas também sugerem uma redução ao lidar com novos conjuntos de dados.

Como resultado da validação prática com a amostra de 100 *tweets*, o modelo demonstrou um desempenho eficiente. Dos 50 *tweets* positivos, 41 foram corretamente classificados, correspondendo a uma taxa de acerto de 82%. Já na categoria negativa, dos 50 tweets, 38 foram classificados corretamente, resultando em uma precisão de 76%. No geral, o modelo alcançou 79% de acurácia na amostra, reforçando sua eficácia em um cenário de aplicação prática.

Essas métricas destacam a importância de avaliar o desempenho do modelo em diferentes conjuntos de dados e sua capacidade de generalização. No entanto, mesmo com uma boa acurácia geral, a análise mostra a necessidade contínua de ajustes e refinamentos para aprimorar o modelo, especialmente ao lidar com dados inéditos, garantindo uma melhor performance em diferentes contextos.

4.1 Discussão dos Resultados

Os resultados obtidos neste estudo reforçam a eficácia do *Naive Bayes* na classificação de sentimentos em grandes volumes de dados, como os *tweets*. No entanto, eles também indicam que há espaço para melhorias. A adição de mais dados de treinamento, o uso de técnicas avançadas de *Machine Learning*, como *Deep Learning*, e a inclusão de mais categorias de sentimento, como as usadas por Chaves (2017), poderiam proporcionar uma análise mais detalhada e refinada das emoções expressas nas redes sociais.

Além disso, o trabalho de Tannús (2017), que utilizou classificadores alternativos como *TextBlob* e *Sentiment140*, mostra que explorar outras técnicas de classificação, ou mesmo algoritmos mais sofisticados, como redes neurais, pode proporcionar uma Análise de Sentimentos ainda mais precisa. No entanto, o *Naive Bayes* se demonstrou uma abordagem robusta e eficiente, particularmente por sua simplicidade e velocidade de treinamento em grandes bases de dados.

4.2 Comparação com Outros Trabalhos

Para contextualizar os resultados obtidos neste trabalho, é importante realizar uma comparação com estudos anteriores que também utilizaram o *Naive Bayes* (NB) ou técnicas similares para Análise de Sentimentos ou classificação de texto.

4.2.1 Carvalho Filho (2014)

O estudo de Carvalho Filho (2014) aplicou o *Naive Bayes* para analisar sentimentos em *tweets* relacionados à Copa do Mundo de 2014. O modelo conseguiu uma acurácia de 85,57%, enquanto o presente estudo, ao aplicar o modelo à sua base de dados original de 449.900 *tweets*, obteve 77,2% de acurácia. Esse resultado pode ser explicado pela diferença de escopo entre os temas analisados. Enquanto o trabalho de Carvalho focava em um evento específico (Copa do Mundo), o presente estudo aborda sentimentos sobre um tema mais complexo e amplo, como o futuro econômico do Brasil no período pós-pandemia.

4.2.2 Chaves (2017)

O estudo de Chaves (2017) focou na Análise de Sentimentos em postagens do *Facebook*, categorizando emoções em grupos como alegria, desgosto, medo, raiva, surpresa e tristeza. Ao contrário do presente estudo, Chaves não relatou métricas de acurácia quantitativa para essas categorias. Em vez disso, utilizou o léxico *WordNet-Affect* para identificar emoções nas postagens. Comparado ao presente estudo, que utilizou apenas as classificações de polaridade (positivo e negativo), a inclusão de emoções mais específicas, como no trabalho de Chaves, poderia fornecer uma análise mais rica. No entanto, os desafios de precisão tendem a ser maiores, devido à subjetividade envolvida na classificação de emoções específicas.

4.2.3 Tannús (2017)

No trabalho de Tannús (2017), o *Naive Bayes* atingiu uma acurácia de 74,1% ao classificar sentimentos em filmes indicados ao Oscar de 2017. No presente estudo, o *Naive Bayes* obteve 77,2% ao classificar os *tweets* da base original. A ligeira superioridade do presente estudo pode ser atribuída ao pré-processamento aplicado, ao volume dos dados e à adequação dos dados ao contexto de Análise de Sentimentos sobre o futuro econômico do Brasil.

4.2.4 Síntese Comparativa

A Tabela 6 resume os resultados de acurácia obtidos em cada um desses trabalhos, facilitando a comparação entre eles.

Tabela 6 - Resultados de acurácia obtidos em cada trabalho.

Estudo	**Modelo**	**Acurácia**
Presente estudo	NB-pós pandemia	77,2% (dados originais)
Carvalho Filho (2014)	*Naive Bayes*	85,5%
Chaves (2017)	*WordNet-Affect*	Não aplicável
Tannús (2017)	*Naive Bayes*	74,1%
Tannús (2017)	*TextBlob*	63,9%

Fonte: O autor.

Os resultados obtidos no presente trabalho demonstram uma performance competitiva em relação aos estudos anteriores, com uma acurácia superior ao trabalho de Tannús (2017). Isso indica que, embora o *Naive Bayes* seja uma técnica simples, ele pode alcançar resultados eficazes dependendo da qualidade dos dados, do pré-processamento e da natureza do problema.

5 CONCLUSÃO

Neste trabalho, foi apresentado o processo para realizar a Análise de Sentimentos sobre o tema "Percepções dos usuários do Twitter em relação ao futuro econômico do Brasil no período pós-pandemia". O processo iniciou com a extração de dados do *Twitter*, seguido da criação de um modelo de classificação que permitiu identificar e analisar as opiniões dos usuários sobre o tema, categorizando os *tweets* como positivos ou negativos.

O modelo gerado demonstrou uma capacidade eficaz de analisar os *tweets* coletados sobre o futuro econômico do país, alcançando uma acurácia de 77,2%. Nos novos dados, o modelo classificou 66,7% dos tweets como positivos e 71,4% como negativos. Esses resultados indicam a eficácia do modelo ao lidar com novos conjuntos de dados, destacando sua capacidade de generalização e adaptação às sutilezas dos sentimentos expressos nas redes sociais.

Além dos resultados obtidos, é importante ressaltar os desafios enfrentados ao realizar a Análise de Sentimentos em dados extraídos da internet. A natureza dinâmica e não estruturada desses dados apresenta obstáculos, como o tratamento de ambiguidades linguísticas, sarcasmo e ironia, além da necessidade de lidar com grandes volumes de informações em constante evolução. Estes desafios ressaltam a complexidade da tarefa e a necessidade contínua de aprimoramento de algoritmos e técnicas para garantir resultados mais precisos e robustos na Análise de Sentimentos em contextos online.

Para futuras iterações deste trabalho, seria interessante considerar o uso de técnicas de *Deep Learning*, como redes neurais, que têm se mostrado promissoras na Análise de Sentimentos por sua capacidade de identificar padrões mais complexos e lidar melhor com nuances linguísticas. Outra possibilidade seria a incorporação de análises multiníveis de emoções, como as propostas por Chaves (2017), classificando os tweets não apenas como positivos ou negativos, mas também em categorias como alegria, medo, raiva e tristeza. Essa abordagem poderia proporcionar uma visão mais detalhada sobre as percepções do público.

Além disso, a aplicação do modelo a outras plataformas de redes sociais, como *Facebook* ou Instagram, poderia fornecer uma análise mais ampla e comparativa do sentimento coletivo em diferentes contextos. Essas expansões não apenas

ampliariam o escopo da análise, mas também permitiriam verificar a robustez do modelo em diferentes cenários e fontes de dados.

Ao oferecer uma visão clara sobre as opiniões expressas pelos usuários do *Twitter*, o modelo construído neste trabalho evidencia sua capacidade de discernir e categorizar os *tweets* de maneira consistente, fornecendo insights valiosos sobre as percepções em relação ao cenário econômico pós-pandemia. Esses resultados destacam a utilidade da Análise de Sentimentos como uma ferramenta poderosa para a compreensão do sentimento coletivo manifestado nas redes sociais, auxiliando na tomada de decisões e na formulação de estratégias por parte de governos e empresas que buscam entender as reações da população em tempos de incerteza.

6 REFERÊNCIAS

ARAUJO, Ricardo. **Aprendizado Supervisionado**. Ricardo Matsumura, 2020. Disponível em: <https://ricardomatsumura.medium.com/aprendizado-supervisionado-5bfacca7566e>. Acesso em: 09 jul. 2024.

BECKER, L. **Algoritmo de Classificação Naive Bayes**. Organi Cadigital, 2019. Disponível em: <https://www.organicadigital.com/blog/algoritmo-de-classificacao-naive-bayes/#:~:text=O%20algoritmo%20de%20Naive%20Bayes,dado%20que%20tem%20a%20doen%C3%A7a%E2%80%9D>. Acesso em: 04 mar. 2021.

BRASIL confirma primeiro caso da doença. **GOV**, 2020. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/noticias/brasil-confirma-primeiro-caso-de-novo-coronavirus>. Acesso em: 01 abr. 2021.

CARVALHO FILHO, J. **Mineração de textos: análise de sentimento utilizando tweets referentes à copa do mundo 2014**. TCC (Graduação) - Engenharia de Software, Universidade Federal do Ceará, Quixadá, 2014.

CHAVES, F. **Análise de sentimentos em redes sociais: uma abordagem baseada em emoções**. TCC (Graduação) - Engenharia de Controle e Automação, Universidade de Caxias do Sul, Caxias do Sul, 2017.

CONTROLF5. **O que é análise de sentimentos?** 2018. Disponível em: <https://controlf5.com.br/o-que-e-analise-de-sentimentos>. Acesso em: 16 mar. 2021.

ENGINEERING. **Tipos de aprendizado de máquina**. 2022. Disponível em: <https://blog.engdb.com.br/tipos-de-aprendizado-de-maquina/>. Acesso em: 01 abr. 2022.

MACHINE Learning. **SAS**, 2021. Disponível em: <https://www.sas.com/pt_br/insights/analytics/machine-learning.html>. Acesso em: 16 mar. 2021.

NATALI, Lucas. **Tipos de aprendizado de máquina e algumas aplicações**, 2021. Disponível em: <https://www2.decom.ufop.br/terralab/tipos-de-aprendizado-de-maquina-e-algumas-aplicacoes/>. Acesso em: 01 abr. 2021.

OSMAN, M. **Estatísticas e Fatos do Twitter Sobre a Nossa Rede Favorita (2021)**. Kinsta, 2021. Disponível em: <https://kinsta.com/pt/blog/estatisticas-e-fatos-do-twitter/#:~:text=Apesar%20de%20um%20grande%20n%C3%BAmero,milh%C3%B5es%20de%20Tweets%20por%20dia>. Acesso em: 06 mar. 2021.

PORTAL DESENVOLVEDOR. **Quando Python foi criado? Conheça a história da linguagem.** Disponível em: <https://portaldesenvolvedor.com/blog/quando-python-foi-criado-conheca-a-historia-da-linguagem/>. Acesso em: 21 set. 2024.

SACRAMENTO, Gabriel. **Naive Bayes: como funciona esse algoritmo de classificação**. Somos Tera, 2021. Disponível em: <https://blog.somostera.com/data-science/naive-bayes>. Acesso em: 12 set. 2023.

SOUZA, Raynara Aparecida Lopes. **Você sabe o que é léxico?.** Mastigando Letras, 2021. Disponível em: <https://www.mastigandoletras.com.br/in%C3%ADcio/l%C3%ADngua-portuguesa/voc%C3%AA-sabe-o-que-%C3%A9-l%C3%A9xico>. Acesso em: 02 out. 2024.

TANNÚS, I. **Análise dos sentimentos expressos na rede social Twitter em relação aos Filmes indicados ao Oscar 2017**. TCC (Graduação) - Sistemas de Informação, Universidade Federal de Uberlândia, Uberlândia, 2017.

TWITTER. **Help.Twitter**, 2020. Sobre as APIs do Twitter. Disponível em: <https://help.twitter.com/pt/rules-and-policies/twitter-api#:~:text=Nossa%20plataforma%20de%20API%20permite,escolheram%20compartilhar%20com%20o%20mundo>. Acesso em: 01 mar. 2021.

UOL. **Com pandemia, PIB do Brasil cai 4,1% em 2020, pior queda em 24 anos**. São Paulo, 03 de mar. de 2021. Disponível em: <https://economia.uol.com.br/noticias/redacao/2021/03/03/pib-brasil-2020-ibge.htm?cmpid=copiaecola>. Acesso em: 01 abr. 2021.

VALINOR, Rodrigo. **O que é linguagem de programação**. 2023. Disponível em: <https://www.remessaonline.com.br/blog/linguagem-de-programacao/>. Acesso em: 26 out. 2023.

APÊNDICE A - CÓDIGO DE PRÉ-PROCESSAMENTO DOS *TWEETS*

```
import re
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer

class PreProcessador():
    def __init__(self):
        self.lemmatizer = WordNetLemmatizer()
        self.stopwords = set(stopwords.words('portuguese'))
    # Função para processar texto
    def processar(self, texto):
        # Convertendo para minúsculas
        texto = texto.lower()
        # Remover URLs e caracteres especiais
        texto = re.sub(r'http\S+', '', texto)  # Remove URLs
        texto = re.sub(r'\W', ' ', texto)  # Remove caracteres especiais
        texto = re.sub(r'\s+[a-zA-Z]\s+', ' ', texto)  # Remove letras
isoladas
        # Tokenizar o texto
        tokens = word_tokenize(texto)
        # Remover stopwords
        tokens = [token for token in tokens if token not in self.stopwords]
        # Lematizar
        tokens = [self.lemmatizer.lemmatize(token) for token in tokens]
        # Retornar o texto processado como string
        return ' '.join(tokens)
```

Etapas do pré-processamento:

- **Converte para minúsculas:** Padroniza o texto, garantindo que "Economia" e "economia" sejam tratados igualmente.
- **Remove *URLs* e caracteres especiais**: Elimina *links* e pontuações que podem interferir na análise, deixando apenas palavras relevantes.
- **Tokenização**: Divide o texto em palavras (*tokens*), para que cada palavra possa ser analisada separadamente.
- **Remove *stopwords***: Exclui palavras comuns e irrelevantes, focando nas palavras mais importantes.
- **Lematização:** Converte as palavras para suas formas base, como transformar "correndo" em "correr", o que ajuda a manter a consistência no texto.
- No final, o método retorna uma string com o texto limpo e normalizado, pronto para ser utilizado em modelos de machine learning ou análise de sentimentos.

APÊNDICE B: CÓDIGO DE TREINAMENTO E AVALIAÇÃO DO MODELO DE NAIVE BAYES

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

class GerenciadorDados():
    def __init__(self):
        self.pre_processador = PreProcessador()

    # Função para carregar e processar dados de arquivos CSV
    def carregar_dados(self, caminho_positivo, caminho_negativo):
        # Carregar dados positivos
        dados_positivos = self.extrair(caminho_positivo)
        dados_negativos = self.extrair(caminho_negativo)

        # Processar os dados (remover stopwords, lematização, etc.)
        dados_positivos_processados =
[self.pre_processador.processar(dado[1]) for dado in dados_positivos]
        dados_negativos_processados =
[self.pre_processador.processar(dado[1]) for dado in dados_negativos]

        # Criar DataFrames
        df_positivos = pd.DataFrame(dados_positivos_processados,
columns=['tweet'])
        df_positivos['sentimento'] = 'positivo'

        df_negativos = pd.DataFrame(dados_negativos_processados,
columns=['tweet'])
        df_negativos['sentimento'] = 'negativo'

        # Combinar os dados
        dados_combinados = pd.concat([df_positivos, df_negativos],
ignore_index=True)
        return dados_combinados

    # Função para extrair dados de um CSV
    def extrair(self, caminho):
        with open(caminho, 'r', encoding='utf-8') as arquivo:
            dados_carregados = reader(arquivo, delimiter=";")
            return list(dados_carregados)

# Carregando e pré-processando os dados
gerenciador = GerenciadorDados()
dados = gerenciador.carregar_dados("tweets_positivos.csv",
"tweets_negativos.csv")

# Dividindo os dados em treino e teste
X = dados['tweet']
y = dados['sentimento']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Vetorizando os tweets
vectorizer = CountVectorizer()
X_train_vect = vectorizer.fit_transform(X_train)
X_test_vect = vectorizer.transform(X_test)
```

```
# Treinando o modelo Naive Bayes
modelo = MultinomialNB()
modelo.fit(X_train_vect, y_train)

# Fazendo previsões e calculando a acurácia
y_pred = modelo.predict(X_test_vect)
acuracia = accuracy_score(y_test, y_pred)
print(f'Acurácia: {acuracia * 100:.2f}%')
```

Etapas do processamento:

- **Carregamento de Dados:** Os *tweets* rotulados (positivos e negativos) são carregados a partir dos arquivos CSV e processados pelo PreProcessador.
- **Divisão dos Dados:** Os dados são divididos em treino (80%) e teste (20%) para avaliar o desempenho do modelo.
- **Vetorização:** O texto é transformado em uma matriz numérica para que o Naive Bayes possa processar as informações.
- **Treinamento do Modelo**: O modelo Naive Bayes é treinado usando o conjunto de treino.
- **Avaliação do Modelo**: A acurácia é calculada comparando as previsões com os rótulos reais dos dados de teste, mostrando o desempenho do modelo.